English Proficiency Testing and the Prediction of Academic Achievement

Raphael Gamaroff

Supervised by

Sinfree Makoni

This thesis is presented for the degree of

DOCTOR OF PHILOSOPHY

In the Department of ENGLISH

UNIVERSITY OF CAPE TOWN

November 1998

Abstract

The study investigates the ability of English proficiency tests (1) to measure levels of English proficiency among learners who have English as the medium of teaching and learning, and (2) to predict long-term academic achievement (Grade 7 to Grade 12). The tests are "discrete-point" tests, namely, error recognition and grammar tests (both multiple-choice tests), and "integrative" tests, namely, cloze tests, essay tests and dictation tests.

The sample of subjects consists of two groups: (1) those taking English as a First Language subject and those taking English as a Second Language subject. These groups are given the familiar labels of L1 and L2. The main interest lies in the L2 group. The main educational context is a high school in the North West Province of South Africa.

The empirical investigation is divided into four parts:

(1) A description of the battery of English proficiency tests. *(Chapter 3).* These tests were given to Grade 7 school entrants.

(2) An examination of the validity and reliability of the battery of the English proficiency tests . *(Chapter 4).* High correlations were found between all of the tests and a substantial difference in English proficiency was found between the L1 and L2 groups.

(3) A longitudinal investigation of predictive validity, where the English proficiency tests were used as the predictors, and academic achievement (Grades 7 to 12) as the criterion. *(Chapter 5).* The main interest of the longitudinal investigation lies in long-term prediction. It is generally believed that low English proficiency is a major cause of academic failure. The longitudinal study corroborates this belief empirically and also shows that very high English proficiency is a good predictor of success. The matriculation exemptions of the L1 group, scored substantially higher on

the English proficiency tests than the L2 group, were three times higher than those of the L2 group.

(4) A longitudinal investigation of the predictive validity of the Grade 6 reports. (*Chapter 5).* These Grade 6 reports served as the main criterion for admission to Grade 7 at the high school. Almost all of the Grade 6 reports of the L2 group emanated from former Department of Education and Training (DET) schools. Most of the Grade 6 reports of the L1 group emanated from a "feeder" school in close proximity to the high school. The L1 Grade 6 reports were found to be good predictors, while the L2 Grade 6 reports were found to be poor predictors. A probable reason for the poor predictions of the L2 Grade 6 reports was that these reports were inflated, and therefore unreliable.

The outline of the chapters is as follows:

Chapter 1 describes the scope of the study.

Chapter 2 deals with theoretical issues in the testing of language proficiency and academic achievement. The chapter comprises a review of the literature on language testing and a discussion of germane concepts such as *ability, competence, proficiency, authenticity, norm-referenced tests, discrete-point tests, integrative tests, assessment, validity* and *reliability.*

Chapter 3 describes the sample of subjects and sampling procedures, and the structure and administration of the tests.

Chapter 4 presents the results of the English proficiency tests and discussion. Included in the chapter is an investigation of rater reliability among a group of educators of teachers of English.

Chapter 5 deals with the prediction of academic achievement, investigates the reliability of the Grade 6 reports from previous schools, summarises the findings and examines the generalisability of the findings.

Chapter 6 discusses the implications of the study for English testing and presents the conclusions. The four main implications dealt with are: (1) the viability of

the distinction between English *first language* and English *second language*, (2) the kind of English proficiency tests or tasks that should be used, (3) the problem of rater reliability, and (4) the necessity of psychometric measurement. Woven into the discussion of the implications is a description of a few contemporary initiatives to improve language testing in South Africa and elsewhere.

Acknowledgements

I have received feedback - positive and negative - from many applied linguists over the last ten years regarding this research. Both kinds of criticism have been invaluable to me, the former because of the encouragement, the latter because without it, firstly, complacency could have set in, and secondly, but more importantly, this study would have had far less to pitch my wits against. There was a time when I was advised by and "expert" in the field to give up the task and "try something else". On the bright side, there have been others who have considered this research worthwhile.

One of these people is my supervisor, Sinfree Makoni. This study would have been, for what it is worth, the poorer without his incisive yet self-effacing contribution. Besides providing guidance on core issues, he has also increased my awareness of the fact that language is not only power, but a sickness; unto death. And *that* has much ado about the "reality" of language, where applied linguistics can play a valuable role.

I started this research in 1987 and finished it in November 1998, 35 years after obtaining my BA degree at UCT in 1963. The years at UCT where I did philosophy under Martin Versfeld and Andrew Murray were the most formative of my academic and spiritual life and have permeated much of my thinking over the years. One of Andrew Murray's pearls during his lectures on Plato's "simile of the cave", which I am sure would be remembered by many of his students, and which sums up the ethos of this study, was this (unfortunately, *how* he said it cannot be captured in print):

There's the real,
the really real,
the really really real

and so forth....

I dedicate this study to these two great UCT philosophers. I thank the school principals and teachers who made it possible to conduct the tests in their schools. And to my wife, Cathy, thank you for your support and also waiting on many occasions so patiently in the wings. A special appreciation for modern information technology, without which it would have been much harder to reorganise, reformat, reprint, reprocess and distil the procession of ideas in the pages that follow. Perhaps the product might add a touch of spice to the applied linguistics pot.

Table of Contents

CHAPTER 3: Sampling, and Structure and Administration of the English Proficiency Tests

CHAPTER 4: Results of the proficiency tests

Tables

Figures

CHAPTER 1

Scope of the Study

1.1 Introduction: The problem and purpose of the study

Language testing draws on three areas: (1) the nature of language, (2) assessment and (3) language ability.[1] Language ability is closely related to language proficiency, which is a key term in this study. Central concepts in the testing of language ability are: (1) validity (what one is testing), (2) reliability (how one is testing), (3) practicability (economics of time and expense) and (4) accountability (why one is testing). If a test is not practicable, even if judged to be valid and reliable, it would be uneconomical and, accordingly, of little use. Overarching all these concepts is the problem of test authenticity. This problem is dealt with extensively in the study.

The educational context of this study is a High School in the North West Province of South Africa, which will be referred to as MHS, where I taught and did language research for over seven years (January 1980 to April 1987).

The study consists of four interrelated topics:

I. The importance of psychometric, i.e. statistical, or quantitative, measurement in assessment and how this clarifies (a) the construct of language proficiency, and (b) the use of proficiency tests as predictors of academic achievement. Statistical measurement in scoring procedures is also closely related to the structure and administration of tests. This study re-examines and defends this interdependence in terms of the three key notions in testing: validity, reliability and practicability.

---

1 Davies, A. *Principles of language testing*. 1990, p.4.

II. The examination of a battery of traditional English proficiency tests that will be used to predict academic achievement. The tests consist of "integrative" tests such as cloze, dictation and essay tests, and "discrete-point" tests such as a grammar test and an error recognition test.

III. The prediction of academic achievement where English proficiency tests are used as the predictors. A longitudinal study is undertaken of the prediction of academic achievement from Grade 7 to Grade 12. Part of the predictive investigation involves a comparison between the predictive validity of previous school reports (Grade 6) of entrants to MHS and the predictive validity of the English proficiency tests. These reports were the main criterion for admission to the School. Few entrants with aggregates under 60% were admitted to MHS.

IV. Implications of the study for language testing. The three main implications are: (1) the viability of the distinction between *first language* and *second language*, (2) the kind of tests or tasks that should be used, and (3) the problem of rater reliability.

V. This study is limited in that it does not examine what "really matters in first or second language proficiency and academic achievement"[2] or "develop an adequate theoretical framework for relating language proficiency to academic achievement"[3]. Nor does it deal with individual differences in cognitive styles of learning[4] and the many causes of academic failure.

The central educational context is English proficiency and academic achievement in *minority* education. There are two urgent needs in minority education:

---

[2] Saville-Troike, M. 'What really matters in second language learning for academic achievement.' *TESOL Quarterly*, 18(2):199-219.

[3] Cummins, J. *Wanted: A theoretical framework for relating language proficiency to academic achievement among bilingual students'*, in Rivera, C. (ed.). *Language proficiency and academic achievement*, 1984.

[4] (1) Diller, K.C. *Individual differences and universals in language learning aptitude*, 1981.
(2) Skehan, P. *Individual differences in second language learning*, 1989.
(3) Skehan, P. *A cognitive approach to language learning*, 1998.

> *(1) to pursue fundamental research on the nature of language proficiency and how it can be measured, and (2) to provide teachers with up-to-date knowledge of language proficiency assessment so they can improve their classroom assessment practices.*[5]

The term *minority* has much more than a numerical meaning. In South Africa the majority of learners use English as an *additional/second* language, but tradition refers to such learners as originating from *minority* language backgrounds. The term has an obvious discriminatory ring that implies that some acceptable level has not been reached. Yet, tests have to distinguish between levels of proficiency for them to have construct validity. Chapters 3, 4 and 5 deal with the statistical issues of assigning people to the same or different groups, and Chapter 6 deals with the educational and political implications of doing so.

Two main levels of proficiency are examined that are given the well-known labels of L1 and L2. The L1 and L2 labels are central to the study but are used differently to the normal connotation of *first language* and *second language*. These well-known terms together with the terms *mother tongue* and *native language* have been the occasion of much controversy. *In the empirical investigation of the study the labels L1 and L2 will refer to the sample of subjects (i.e. test takers) who take the subject English as a First Language and English as a Second Language, respectively,* at MHS. This definition of L1 and L2 needs constant reminding throughout the study. In Chapter 6 (section 6.2), various other definitions of "L1" and "L2" are examined. The sample of subjects is described in detail in Chapter 3.

Psychometric questions, and discrete-point and integrative testing have been discussed at length in the literature on language testing and assessment (the distinction between *testing* and *assessment* is explained shortly). So one may ask why the need to, and where is the merit and originality of, devoting a PhD to such old and outdated issues, which have not been research issues in language assessment for over 15 years? There is indeed a pressing need

---

[5] Rivera, C. *The ethnographical/sociolinguistic approach to language proficiency assessment*, 1983, p.xii.

because although communicative methods have become the prevalent source of testing with a "richer conceptual base for characterizing the language abilities to be measured, it has presented language testers with a major challenge in defining these abilities and the interactions among them with sufficient precision to permit their measurement."[6] It is this problem of reconciling authentic subjectivity and objective precision that is the major problem in testing, indeed the major problem of cognition and language.[7] The *authenticity* issue has wider ramifications. It is not only central to testing but also to syllabus design and materials development. This study focuses on testing.

In spite of decades of attempts to define it, the how[8] and the why[9] of language proficiency remain a conundrum. Although we may no longer stand before an "abyss of ignorance"[10] and may be able to agree with Alderson (in Douglas[11]) that language testing has "come of age", there are still many problems in language testing, the greatest one being, I suggest, the problem of reliability[12] and specifically rater reliability (see Alderson and Clapham's case studies of this problem[13]). There are two kinds of rater reliability: interrater reliability and intrarater reliability. (These are dealt with in section 2.9.1).

Owing to our ignorance of the processes of language learning and learning processes in general, much of what we know about language testing, and therefore also about teaching, remain tentative. ("Language testing is rightly central to language teaching"[14]). A major

---

6 Bachman, L.F. 'Assessment and evaluation. *Annual Review of Applied Linguistics (1989)*, 10:210-226 (1990a), p.210

7 Lakoff, G. *Women, fire and dangerous things*, 1987.

8 Bachman, L.F. *Fundamental considerations in language testing*, 1990b.

9 Davies, A. *Principles of language testing*, 1990.

10 Alderson, J.C. 'Who needs jam?', in Hughes, A. and Porter, D. *Current developments in language testing*, 1983, p.90.

11 Douglas, D. 'Developments in language testing.' *Annual Review of Applied Linguistics*, 15, 167-187 (1995), p.176.

12 Moss, P. 'Can there be validity without reliability?' *Educational Researcher*, 23 (2), 5-12 (1994).

13 Alderson, J.C. and Clapham, C. 'Applied linguistics and language testing: A case study of the ELTS test.' *Applied Linguistics*, 13 (2), 149-167 (1992).

14 Ibid., p.2.

obstacle in test development has been the lack of agreement on what it means to know a language, on what aspects of language knowledge should be tested - and *taught* - and how they should be tested and assessed.

This problem is not a surprising one because language is closely connected to human rationalities, imaginations, motivations and desires, which comprise an extremely complex network of biological, cognitive, cultural and educational factors. As a result, all language testing theories are inadequate owing to the difficulties involved in devising tests that test authentic language reception and production. This does not mean that we should stop measuring until we've decided what we are measuring. We do the best we can by taking account of generally accepted views of the nature of language proficiency, of modern views and dated ones. In the modern literature on testing there seems to be an overemphasis on up-to-date theories, which gives the impression that "what is dated is outdated".[15] Widdowson's up-to-date admonition that we should take dated views more seriously is taken to heart in this study.

What is a test? It is "the most explicit form of description, on the basis of which the tester comes clean about his/her ideas".[16] What all testers are looking for are systematic elicitation techniques on which they can base useful decisions. The three underlying issues in testing are: to infer abilities, to predict performance and to generalise from context to context.[17] This means that tests should be valid, reliable and practicable. *Communicative* testers would add the notions of "impact" (i.e. face validity) and "interactionist".[18] Opponents of discrete-point tests (such as grammar tests) and integrative tests (such as cloze tests and dictation tests) would probably concede that such tests are reliable and practicable, but they would argue that they are not valid, i.e. they tell us little or nothing about the learner's knowledge of authentic language. I shall argue that, on the contrary, they tell us a great deal about authentic language

---

15 Widdowson, H.G. 'Skills, abilities, and contexts of reality.' *Annual Review of Applied Linguistics*, 18, 323-333 (1998), p.323.

16 Davies, A. *Principles of language testing*, 1990, p.2.

17 Skehan, P. *A cognitive approach to language learning*, 1998, p.153.

18 Bachman, L.F. and Palmer, A.S. *Language testing in practice*, 1996, p.17.

and that these old issues are not outdated and are still worthy of attention within the curriculum. A curriculum framework consists of the following components[19]:

- Needs analysis
- Objectives
- Materials
- Teaching
- Evaluation/Assessment/Testing.

Accordingly, the curriculum is concerned with the syllabus as well as everything to do with pedagogical matters, i.e. teaching what to whom, when and how.[20] *Syllabus* is defined as the content and sequence of content of the programme selected in order to make learning and teaching effective.[21] Although testing is the last component in the curriculum framework, this is only so chronologically, and not logically, because testing permeates the whole of the curriculum. This is the reason why there is the possibility - and the temptation; perhaps justifiably so - of teaching to the test.

A major part of testing is concerned with *assessment*. In this study I use the term *tests* to refer to "elicitation techniques"[22] and the term *assessment* to refer to the procedures used to control raters' judgements and scoring techniques. (Assessment is discussed in detail in section 1.3).

---

19 Brown, J.D. 'Language programme evaluation: A synthesis of existing possibilities', 1989, p.235.

20 Stern, H. H. *Fundamental concepts of language teaching*, 1983.

21 Wilkins, D.A. 'Notional syllabuses revisited.' *Applied Linguistics*, 2 (1), p.83-89 (1981), p.83.

(2) Brumfit, C.J. 'Notional syllabuses revisited: A response.' *Applied Linguistics*, 2 (1), 90-92 (1981), p.90.

22 Ur, P. *A course in language teaching: practice and theory*, 1996, p.37.

There are a variety of language test uses (Pollitt in Yeld[23]). The basic four uses are mentioned[24]:

- Proficiency tests, which evaluate present knowledge in order to predict future achievement, usually at the beginning of a course of study. Proficiency tests are based on knowledge that has been gained independent of any specific syllabus but not independent of typical syllabuses because the knowledge to be tested must have been gained from some syllabus or other.

- Achievement tests, which evaluate how much has been learnt of a particular syllabus, where the focus is on success, usually at the end of a teaching programme.

- Diagnostic tests, which evaluate points not yet mastered, where the focus is on failure and consequent therapy. Diagnostic tests, therefore, may be considered to be the reverse of achievement tests.[25] Proficiency tests often involve diagnosing items that have not been mastered, and therefore diagnostic testing may be part of proficiency testing.

- Aptitude tests, which evaluate abilities for language mastery, and are thus, like proficiency tests, of predictive value. Unlike the three other kinds of test uses, aptitude tests have no specific or general content, and are thus difficult tests to compile. They require, arguably, the most knowledge and care in their construction and application, for it is far worse to be told that one has no aptitude than to be told that one has low proficiency or has failed an achievement test. No aptitude means no hope at all, unless it is possible to have potential without aptitude.

Both proficiency and achievement are concerned with present knowledge. It may occur that a proficiency test contains material previously contained in an achievement test, but this difference is irrelevant to the validity of the proficiency test because, unlike an achievement test, a proficiency test is not concerned with whether the content of a test was previously

---

23 Yeld, N. *Communicative language testing.* Report on British Council Course 559 offered at Lancaster University from 8 September to 20 September 1985, (published in) 1986, p.36.

24 (1) Corder, S.P. *Error analysis and interlanguage*, 1981, p.20.
(2) Davies, A. *Principles of language testing*, 1990, pp.20-21.

25 Davies, ibid., p.21.

taught. The American Council of the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines are a case in point:

> *Because these guidelines identify stages of proficiency, as opposed to achievement, they are not intended to measure what an individual has achieved through specific classroom instruction but rather to allow assessment of what an individual can and cannot do, regardless of where, when or how the language has been learned or acquired: thus the words learned and acquired are used in the broadest sense.*[26]

Proficiency is concerned with what somebody knows and can do here and now. Achievement should be ultimately concerned with proficiency as well. That is why Spolsky omits the term achievement in the following definition of language tests: "Language tests involve measuring a subject's knowledge of, and proficiency in, the use of language."[27]

There are four important considerations in language testing[28]:

1. How valid is the test?
2. How easy is it to compose?
3. How easy is it to administer?
4. How easy is it to mark?

The first, which is concerned with the purpose of a test, is the most important theoretical issue in testing. Ur feels so strongly about practicability that her next three considerations for choosing a test have to do with practicability. (The fourth is related to rater reliability). A test may be everything communicative testers require, but it would not be useful if it took too long to do or was too difficult to administer and assess. The more *objective* the test, the less the danger of rater unreliability. An essay test is a supreme example of a *subjective* test, because it is vulnerable to fluctuations in judgements between raters. The problem is finding the

---

[26] Byrnes, H. and Canale, M. *Defining and developing proficiency: Guidelines, implementations and concepts*, 1987, p.15.

[27] Spolsky, B. *Conditions for second language learning*, 1989, p.138.

[28] Ur, P. *A course in language teaching: practice and theory*, 1996, p.37

appropriate balance between the different testing considerations, where it is difficult, indeed, impossible, to give all of them equal prominence.

The reason why the value of discrete-point tests and many integrative tests such as cloze tests and dictation tests must be reassessed is mainly because of their practicability; if only one could solve the problem of *authenticity*. The basic problem is whether indirect tests such as grammar tests, cloze tests and dictation tests can predict real-life performance, which many authors (these authors are discussed at length in the study) equate with *authentic* language, and thus reject the notion that an indirect elicitation procedure of real-life language can be authentic. Of course, this problem of authenticity of indirect tests is not limited to language tests but to all kinds of indirect tests, e.g. intelligence tests. A major part of this study is concerned with the meaning of *authenticity* in testing.

Every test is an operationalisation about certain beliefs and values about language, whether the test is called authentic or not . These beliefs and values determine to a certain extent our mental and emotional reactions to language and to knowledge in general.

The modest aims of this research are to investigate the measurement characteristics of five types of tests with a specific school cohort and draw conclusions about the predictive validity of these tests. Included is the less modest aim of predictions beyond the specific school cohort to a population of which the specific cohort is a sample. Whether this research contributes to language learning or to educational improvement are not the focus, though such matters are of great interest to proponents of *authentic* tasks

1.2 Psychometrics and norm-referenced testing

In language testing the opposition to psychometrics is closely connected to the "suspicion of

quantitative methods"[29] and the opposition to "reductionist approaches to communicative competence".[30]

The history of the quantitative/qualitative controversy can be viewed from two diametrically opposite angles: (1) qualitative research has been dominated by quantitative research for many decades and is only in recent years becoming accepted as a legitimate scientific approach[31] or (2) qualitative research has been for more than two decades challenging quantitative methods and also setting itself up as the only legitimate form of research.[32]

Galton's view is that scientists should "devise tests by which the value of beliefs may be ascertained, and to feel sufficiently masters of themselves to discard contemptuously whatever may be found untrue"[33] (Rushton 1995; his frontispiece). For Galton tests must be statistically validated. There is no doubt that statistical measurement in language testing has been given a undeserved bad press, e.g. Spolsky[34], Lantolf and Frawley[35] and Macdonald[36].

The increasing number of studies in purely ethnographical/sociolinguistic approaches to language proficiency assessment[37] is witness to the opposition to the *objectivist, or*

---

29 Davies, A. *Principles of language testing*, 1990, p.1.

30 Lantolf, J.P. and Frawley, W. 'Proficiency: Understanding the construct.' *Studies in Second Language Acquisition (SLLA)*, 10 (2), 181-195 (1988), p.182.

31 Lazaraton, A. 'Qualitative research in applied linguistics: A progress report.' *TESOL Quarterly*, 29 (3), 455-471 (1995, p.455.

32 Magnan, S.S. *Review of Creswell, J.W. 1994. Research design: qualitative and quantitative approaches*, 1997.

33 Rushton, J.P. *Race, evolution and behaviour*, 1995.

34 Spolsky, B. 'The limits of authenticity in language testing.' *Language Testing*, 2, 31-40 (1985).

35 Lantolf, J.P. and Frawley, W. 'Proficiency: Understanding the construct.' *Studies in Second Language Acquisition (SLLA)*, 10 (2), 181-195 (1988).

36 (1) Macdonald, C.A. *English language skills evaluation (A final report of the Threshold Project)*, 1990a.

(2) Macdonald, C.A. *Crossing the threshold into standard three in black education: The consolidated main report of the Threshold Project*, 1990b.

37 (1) Bennett, A. and Slaughter, H. 'A sociolinguistic/discourse approach to the description of the communicative competence of linguistic minority children', in *Rivera, C. The ethnographical/ sociolinguistic approach to language proficiency assessment*, 1983.

(2) Jacob, E. 'Studying Puerto Rican children's informal education at home', in

(post-)positivistic, or reductionist methods of psychometric research. For Harrison, psychometric measurement is inappropriate due to its subjective nature:

> *Testing is traditionally associated with exactitude, but it is not an exact science...The quantities resulting from test-taking look like exact figures - 69 per cent looks different from 68 pert cent but cannot be so for practical purposes, though test writers may imply that they are distinguishable by working out tables of precise equivalences of test and level, and teachers may believe them. These interpretations of scores are inappropriate even for traditional testing but for communicative testing they are completely irrelevant. The outcome of a communicative test is a series of achievements, not a score denoting an abstract 'level'.*[38]

Thus, "the quantities resulting from test-taking [which] look like exact figures" (in the quotation above) appear to measure objectively, but in fact they measure subjectively. (See Morrow[39] for a similar view). Lantolf and Frawley[40] maintain that

> *[w]hat must be done is to set aside the test-based approach to proficiency and to begin to develop a theory of proficiency that is independent of the psychometrics. Only after such a theory has been developed and is proven to be consistent and exhaustive by empirical research should we reintroduce the psychometric factor into the picture, with the full realization that such a reintroduction may not be possible, given our earlier remarks on the scalability of human behavior.*

The earlier Spolsky was contemptuous of "psychometrists":

> *In the approach of scientific modern tests, the criterion of authenticity of task is generally submerged by the greater attention given to psychometric criteria of*

Rivera, C. (ed.). *The ethnographical/sociolinguistic approach to language proficiency*, 1983.

(3) Phillips, S. 'An ethnographic approach to bilingual language proficiency assessment', in Rivera, C. (ed.). *The ethnographical/sociolinguistic approach to language proficiency assessment*, 1983.

[38] Harrison, A. 'Communicative testing: Jam tomorrow?', in Hughes, A. and Porter, D. (eds.). *Current developments in language testing*, 1983. p.84.

[39] Morrow, K. 'Communicative language testing: Revolution or evolution', in Alderson, J (ed.). *Issues in language testing*, 1981, p.12.

[40] Lantolf, J.P. and Frawley, W. 'Proficiency: Understanding the construct.' *Studies in Second Language Acquisition (SLLA)*, 10 (2), 181-195 (1988), p.185.

> *validity and reliability. The psychometrists[41] are 'hocus-pocus' scientists in the fullest sense; in their arguments, they sometimes even claim not to care what they measure provided that their measurement predicts the criterion variable: face validity receives no more than lip service.*[42]

Spolsky's recent "postmodern" approach to psychometrics is that it should be used in conjunction with "humanist" approaches.[43] The view in this study, which, for some researchers is taken for granted but for others is highly contested, is that "language testing cannot be done without adequate statistics".[44]

Psychometric assessment in language assessment traditionally means norm-referenced assessment, which is not concerned with individual scores but with the dispersion of scores within a group, where the concern is with maximising individual differences between test takers on the variable that is being measured.[45] For many, *psychometrics* has become synonymous with *quantitative measurement* and *statistical measurement* [46], and this is how I use *psychometrics* in this study.

The term *psychometric* has another meaning. For example, at a conference on academic development, where I presented a paper on this topic[47], a member of the audience took offence (she said she was "boiling") because *psychometrics,* she insisted, was far more than norm-referenced measurement". Her view was that psychometric tests *measured* the *psyche*

---

41 *Psychometrist* has two meanings: 1. a statistician, and 2. somebody with the paranormal power to find lost objects. I guess the double meaning is not lost on Spolsky.

42 Spolsky, B. 'The limits of authenticity in language testing.' *Language Testing*, 2, 31-40 (1985), pp.33-34.

43 Spolsky, B. *Measured words*, 1995, p.357.

44 Davies, A. *Principles of language testing*, 1990, p.16.

45 Cziko, G.A. 'Improving the psychometric, criterion-referenced, and practical qualities of integrative testing.' *TESOL Quarterly*, 16 (3), 367-379 (1982), pp.27-28.

46 (1) Spolsky, B. 'The limits of authenticity in language testing.' *Language Testing*, 2, 31-40 (1985).

(2) Lantolf, J.P. and Frawley, W. 'Proficiency: Understanding the construct.' *Studies in Second Language Acquisition (SLLA)*, 10 (2), 181-195 (1988), p.185.

47 Gamaroff, R. *Psychometrics and reductionism in language assessment*. Paper presented at the SAAAD/SAARDHE conference 'Capacity-building for quality teaching and learning in further and higher education', University of Bloemfontein, 22-24 September, 1998e.

(the literal meaning of *psychometrics)* and was embedded, she insisted (correctly), in the flesh-and-blood context of individuals, and psychometric tests, therefore, involve far more than a comparison between individuals and groups. I was taken aback by her outburst, not because she did not have a valid definition of *psychometrics*, but because this definition of *psychometrics*, owing to the different context, had not entered my mind, owing, no doubt, to my maximum attention to what interested me.

The method used in this study is mainly quantitatively based, where the emphasis is on norm-referenced testing. In language testing, as in second language acquisition research in general, quantitative measurement has been challenged for more than two decades by qualitative methods of research: indeed, qualitative measurement has been setting itself up as the only legitimate form of research.[48] Terre Blanche distinguishes between "two different constituencies" of qualitative researchers:

> *those who would use qualitative methods as a humanist, emancipatory tool to access authentic subjective experiences so easily censored out by more hard-nosed quantitative methods, and those who want to use qualitative methods such as discourse analysis to critique the semantic practices of both 'scientific' and 'humanist' psychologies.*[49]

Both constituencies reject the domination of the norm over the individual. "Norm" has at least two meanings, which are sometimes not distinguished. For example, at the conference of the National Educators of Teachers of English (NAETE) at Potchefstroom (September 17-18, 1998) I was discussing the concept of "norm" with Johan van der Walt. We were in one accord that the individual without the norm is an abstraction. It was only during Van der Walt's presentation that I realised that we did not mean the same thing by the term - yet our different meanings were related, as I shall explain shortly. In this regard, consider the following extract from a repartee between the two English professors, Johan van der Walt

---

[48] Magnan, S.S. *Review of Creswell, J.W. 1994. Research design: qualitative and quantitative approaches*, 1997.

[49] Terre Blanche, M. 'Crash.' *South African Journal of Psychology*, 27 (2), 59-63 (1997), p.61.

and Colyn Davey at the 1998 National Association of Educators of Teachers of English (NAETE) conference[50] that was concerned with the topic of establishing norms of English.

> *Van der Walt.* Then you agree that there should be a norm.
> *Davey.* Yes but learners should be able to choose the norm they prefer.

The question arises of whether it is possible to use the term "norm" in the sense of (1) Van der Walt's imperative of conforming to a standard, by which he means "Standard English", and (2) Davey's imperative of freedom to choose the norm that one refers, which could be "Standard" English, or, say, institutionalised black South African English (IBSAE). The latter comprises ubiquitous constructions such as "I am having a problem", "He write English perfectly" and "When I was in Town I see my English teacher".[51] It is indeed possible to use the term "norm" in these two senses, but neither of these senses explicitly evokes the comparison between individuals within a group, which is what norm-referenced tests are concerned with. To clarify the distinction between the "norms" of Van der Walt and Davey, on the one hand, and norm-referenced tests, on the other, I introduce the notion of criterion-referenced tests.

Criterion-referenced tests are concerned with how well an individual performs relative to a fixed *criterion*, e.g. how to ask questions. Norm-referenced tests are concerned with how well an individual performs compared to a group. This is traditional psychometric testing. Both Van der Walt and Davey believe in norms; the former a standardised norm, the latter an unstandardised norm. Both kinds of norm are concerned with how well an individual performs relative to a fixed criterion, which is the concern of criterion-referenced tests. The

---

50 Van der Walt, J. *The implications for language testing of IBSAE (Institutionalised Black South African English)*. National Association of Educators of Teachers of English (NAETE) conference " Training teachers for the South African context, Potchefstroom College of Education, September 17-18, 1998.

51 Van der Walt's paper followed immediately after the presentation of Makalela's (1998) paper "Institutionalised Black South African English" (IBSAE) in which Makalela advocates that IBSAE be adopted as the norm among blacks in South Africa. The examples of IBSAE cited above are those given by Makalela.

difference between Van der Walt's and Davey's norm is that the former is imposed from above (the standardised norm), whereas Davey's is bottom-up, where the group chooses the norm it wishes to aspire to. To reiterate, this meaning of *norm* is not the same meaning as *norm-referenced* tests, but - an important point - they are related:

> *Consider a test whose results we are to interpret by comparison with criteria. To do so we must already have decided on a standard of performance and we will regard students who attain it as being significantly different from those who do not...The question is: How do we establish the criterion level? What is to count as the standard? Naturally, we can't wait to see how students actually do and base our criterion on the average performance of the present group: this would be to go over into blatant norm-referencing. So suppose we base our criterion on what seems reasonable in the light of past experience? Naturally, if the criterion is to be reasonable, this experience must be of similar groups of students in the past. Knowing what has been achieved in the past will help us avoid setting the criteria inordinately high or low. But isn't this very close to norm-referencing? It would even be closer if we were to base the criterion not just on that of previous students but on students in general.*[52]

Norm-referenced tests can be distinguished from *criterion-referenced* and *individual-referenced* tests:

1. Norm-referenced tests are concerned with how well an individual performs compared to a group which he or she is a member of. This is traditional psychometric testing.

2. Criterion-referenced tests are concerned with how well an individual performs relative to a fixed criterion, e.g. how to ask questions. This is what Cziko calls "edumetric" testing.[53]

3. Individual-referenced are concerned with how individuals perform relative to their previous performance or to an estimate of their ability.

Strictly speaking it is not the test that is norm-referenced or criterion-referenced or individual-referenced but the purpose for which it is used. Similarly, tests in themselves are

---

52 Rowntree, D. *Assessing students: How shall we know them*, 1977, p.185.

53 Cziko, G.A. 'Improving the psychometric, criterion-referenced, and practical qualities of integrative testing.' *TESOL Quarterly*, 16 (3), 367-379 (1982).

not valid, but rather it is the purpose that they are used for that makes them valid. (Validity is discussed in section 2.8).

The idea that norm-referenced tests, on the one hand, and criterion-referenced tests and individual-referenced tests, on the other, are mutually exclusive is based on two contrasting philosophical positions: the former "positivistic", the latter "humanistic". The former is interested in what makes people different. An extreme view of the "humanistic" position is that because it is morally reprehensible to compare people, one should focus instead on common goals. This view is represented most vocally in South Africa by the protagonists of "outcomes-based education" where a "learner's progress will be measured against criteria that indicate attainment of learning outcomes, rather than against other learners' performances".[54]

The point that I shall be emphasising is that norm-referenced tests are important because without data on the variance between individuals within a *group*, it is not possible to distinguish *what* (which is the concern of criterion-referenced tests) an individual knows from what other people know (which is the concern of norm-referenced tests). Individual-referenced tests also cannot be separated from what other people know. The differences between individuals actually clarify the matter under test. In other words, the construct validity of a test is dependent on some people doing well and others doing less well, for if everybody did equally well, we would have little idea of what we were testing. What we think is going on in each individual's invisible mind can be scientifically inferred and described only when one has some idea of what is going on in a many individual minds, i.e. what is going on in a group. Emphasising the individual over the group or vice versa is "somewhat metaphysical [because both] types of test sampling (for that is what norm and criterion referencing do: they sample) need one another".[55] In sum, ***"norm" can refer to a criterion (!) such as Standard English or to the comparison between individuals within a group.*** These

---

[54] Gultig, J., Lubisi, C., Parker, B. and Wedekind, V. *Understanding outcomes-based education: Teaching and assessment in South Africa*, 1998, p.12.

[55] Davies, A. *Principles of language testing*, 1990, p.19.

two meanings of *norm* are distinguishable, but nevertheless closely related, as Rowntree pointed out in his quotation above.

Ranking individuals and generating scores are two purposes of norm-referenced tests. Another purpose is to gain understanding of the nature of the constructs under examination, which cannot be achieved if an individual is not compared with what other individuals do.

1.3 Assessment, evaluation and summative assessment

Evaluation and assessment are - as John Locke said about words - knotty bundles to unravel. The main focus in this study is on assessment, specifically, summative assessment. There are different kinds of assessment and a diversity of definitions. The descriptions of Rea[56] and Rowntree are discussed.[57]

TABLE 1.1

Rea's Schema of Assessment

	Formative Assessment	Summative Assessment
Quantitative Methods	Assessment	Evaluation
Qualitative Methods	Appraisal	

Rea[58] uses the term "evaluation" to refer to formal testing activities, which are external to the teaching situation, and which involve "test scores". She uses the terms "assessment" and "appraisal" to refer to activities which are internal to the teaching program. Grades are given for assessment but not for appraisal. In Rea's schema, assessment and evaluation both use measurement, i.e. quantitative methods, while appraisal does not.

---

56 Rea, P. 'Language testing and the communicative language teaching curriculum', in Lee, Y.P. et al. *New directions in language testing*, 1985.

57 Rowntree, D. *Assessing students: How shall we know them*, 1977, p.185.

58 Rea, P. 'Language testing and the communicative language teaching curriculum', in Lee, Y.P. et al. *New directions in language testing*, 1985, p.29

Before I comment further on Rea, it is appropriate to say something about "evaluation", which is a term used by Rea and Rowntree. Before these authors are discussed, consider some other definitions of the term. There are many different definitions of evaluation. Bachman defines evaluation as the "systematic gathering of information for the purpose of making decisions"[59], while Brown defines evaluation as the

> *systematic collection and analysis of all relevant information necessary to promote the improvement of a curriculum, and assess its effectiveness and efficiency, as well as the participants' attitudes within the context of the particular institutions involved.*[60]

"Evaluation" has been contrasted with "grading".[61] Dreyer argues that grading, i.e. summative testing, causes people to fail. If, however, one doesn't grade, most learners will not take learning seriously, because *they are not so much interested in the love of knowledge as in passing a grade.* (See the conclusion to the study, section 6.7, for further comment).

To return to Rea: what may be confusing in her schema is that "assessment" is used generically to cover everything to do with testing as well as specifically to refer to "formative quantitative" assessment. Consider now Rowntree's schema below (Table 1.2):

---

59 Bachman, L.F. *Fundamental considerations in language testing,* 1990b, p.20.

60 Brown, J.D. 'Language programme evaluation: A synthesis of existing possibilities', in Johnson, R.K. (ed.). *The second language curriculum,* 1989, p.223.

61 Dreyer, C. *Testing: The reason why pupils fail.* National Association of Educators of Teachers of English conference (NAETE) " Training teachers for the South African context, Potchefstroom College of Education, September 17-18, 1998.

TABLE 1.2

Rowntree's Schema of Assessment and Evaluation

	Assessment (focus on learner)	Evaluation (focus on teaching)
Quantitative Methods	Summative	Summative
Qualitative Methods	Formative (diagnostic appraisal)	Formative

Rowntree's "assessment" is "put[ting] a value on something" , which translates into everything concerned with "obtaining and interpreting information" of any kind about another person in order to "[try] and discover what the student is becoming or has accomplished".[62] Rowntree's "formative (pedagogic) assessment" emphasises "potential", while his "summative (classificatory) assessment" emphasises "actual achievement".[63]

For Rowntree, "evaluation" is "an attempt to identify and explain the effects (and effectiveness) of the teaching."[64] Rowntree's "formative evaluation is intended to develop and improve a piece of teaching until it is as effective as it possibly can be...[s]ummative evaluation on the other hand, is intended to establish the effectiveness of the teaching once it is fully developed."[65] Rowntree's "formative evaluation" is concerned with the washback effect of a syllabus and/or teaching programme, while "summative evaluation" is concerned with *how the teacher* compiles, administers and scores "terminal tests and examinations coming at the end of the student's course, or indeed [with] any attempt to reach an overall description or judgement of the student, e.g. in an end-of-term report or a grade or class-rank".[66]

I now focus on *summative assessment.* For Rea summative assessment is "terminal", "formal" and "external", and is only concerned with the beginning and end of a course, i.e. with

62 Rowntree, D. *Assessing students: How shall we know them,* 1997, p.4.
63 Ibid., p.8.
64 Ibid., p.7.
65 Ibid.
66 Ibid.

classifying individuals in terms of numerical products, or scores.[67] (In Rea, quantitative methods are also used in formative assessment). This study uses this meaning of summative assessment.

Although this study is concerned with the summative assessment of the learner, it is also concerned with how a teacher (i.e. a rater) assesses a learner (Rowntree's "summative evaluation"); in other words, rater reliability. The latter consists of two major kinds of judgements: (1) the order of priority of performance criteria for individual raters (criteria such as grammatical accuracy, appropriateness of vocabulary and factual relevance) and (2) the agreement between raters on the scores that should be awarded if or when agreement is reached on how to weight different criteria .[68] Rater reliability is discussed in sections 2.9.1, 4.8ff and 6.5.

Weir believes that there is a more pressing need for research in formative testing as opposed to research in summative testing.[69] On the contrary, there is still a pressing need for research in summative testing, because mainstream language testing, e.g. in South Africa at least, is, wrongly, on my view, taking a radically different turn, as manifested in parts of, for example, "outcomes-based education" (OBE). OBE's attitude to summative testing is discussed in section 6.4.

1.4 The One Best Test

A major empirical problem in language testing is establishing valid and reliable criteria for the assessment of language proficiency, which is basically concerned with fluency and accuracy. Three important issues in language testing are:

---

67 Rea, P. 'Language testing and the communicative language teaching curriculum', in Lee, Y.P. et al. *New directions in language testing*, 1985. p. 29.

68 Gamaroff, R. 'Language, content and skills in the testing of English for academic purposes.' *South African Journal of Higher Education*, 12 (1), 109-116 (1998b).

69 Weir, C.J. *Understanding and developing language tests*, 1993, p.68.

1. The kinds of tests that should be used to assess levels of language proficiency.

2. The relationship between statistical significance (numerical data) and their meaning (information).

3. Whether language proficiency tests can validly predict academic achievement.

Academic achievement in this study is represented by: 1. End-of-year aggregate and 2. Pass rate. An applied linguist, who was at one time involved in my research, maintained that the prediction of academic achievement had no place in applied linguistics and, accordingly, belonged to education. I can't accept such a view, for surely the most important reason for the study of *academic* language proficiency is its twofold role in academic achievement: (1) what really matters in academic achievement and (2) predicting academic achievement. *What is a problem, though - owing to the close relation between academic achievement and academic language proficiency - is where a study such as this one should be undertaken: in a language department or an education department.*

The above three issues in language testing[70] are directly related to the search for the "One Best Test". In the 70s a major issue in language testing was whether it was possible to find the "One Best Test". The "One Best Test" question is closely related to the controversy of whether language proficiency consists of a unitary factor analogous to a *g* factor in intelligence, or of a number of independent factors. This controversy is known as the Unitary Competence Hypothesis (UCH) versus the Divisible Competence Hypothesis (DCH).

Bachman and Palmer relate the concerns they had 25 years ago:

> *[W]e shared a common concern: to develop the "best" test for our situations. We believed that there was a model language test and a set of straightforward procedures - a recipe, if you will - that we could follow to create a test that would be the best one for our purposes and situations.*[71]

---

[70] 1. The kinds of tests that should be used to assess levels of language proficiency.
2. The relationship between statistical significance (numerical data) and their meaning (information).
3. Whether language proficiency tests can validly predict academic achievement.

[71] Bachman, L.F. and Palmer, A.S. *Language testing in practice*, 1996, p.4.

Yet, almost two decades ago, Alderson had already graduated from this kind of thinking and suggested that

> *regardless of the correlations, and quite apart from any consideration of the lack of face validity of the One Best Test, we must give testees a fair chance by giving them a variety of language tests, simply because one might be wrong: there might be no Best Test, or it might not have the one we chose to give, or there might not be one general proficiency factor, there may be several.*[72]

High correlations between different kinds of tests show that the UCH, in its weak form, remains a force to be dealt with.[73] The weak form of the UCH adopts an interactionist approach between global and discrete components of language. Oller describes this approach:

> *[N]ot only is some sort of global factor dependent for its existence on the differentiated components which comprise it, but in their turn, the components are meaningfully differentiated only in relation to the larger purpose(s) to which all of them in some integrated (integrative?) fashion contribute.*
> *(See also Oller and Khan*[74] *and Carroll*[75] *for similar views).*[76]

---

[72] Alderson, J.C. 'Report of the discussion on general language proficiency', in Alderson, J.C. and Hughes, A. *Issues in language testing: ELT Documents III*, 1981a, p.190.

[73] 1) Brown, J.D. *A closer look at cloze: Validity and reliability*, 1983.
(2) Hale, G.A., Stansfield, C.W. and Duran, R.P. *TESOL Research Report 16*, 1984.
(3) Oller, J.W., Jr. 'Cloze tests of second language proficiency and what they measure.' *Language Learning*, 23 (1), 105-118 (1973).
(4) Oller, J.W., Jr. 'A consensus for the 80s', in Oller, J.W., Jr. *(ed.). Issues in language testing research*, 1983.
(5) Oller, J.W., Jr. 'Cloze, discourse, and approximations to English', in Burt, K. and Dulay, H.C. *New directions in second language learning, teaching and bilingual education*, 1976.
(6) Oller, J.W., Jr. '"g", "What is it?', in Hughes, A. and Porter, D. (eds.). *Current developments in language testing*, 1983a.
(7) Oller, J.W., Jr. *(ed). Issues in language testing research*, 1983b.
(8) Stubbs, J. and Tucker, G. 'The cloze test as a measure of English proficiency.' *Modern Language Journal*, 58, 239-241 (1974).

[74] Oller, J.W., Jr. and Kahn, F. Is there a global factor of language proficiency?, in Read, J.A.S. *Directions in language testing*, 1981.

[75] Carroll, J.B. *Psychometric theory and language testing*, 1983, p.82.

[76] Oller, J.W., Jr. 'A consensus for the 80s', in Oller, J.W., Jr. (ed.) *Issues in language testing research*, 1983, p.36.

If one is no longer searching for that one Grand Unified Test (GUT), one should be still looking for good tests, indeed for the best tests available. This implies, I suggest, that what one is looking for has an "objective" reality, which, of course, does not mean that we can completely grasp it.[77]

If we have given up on finding or constructing that elusive (and illusory?) one best test, we are nevertheless looking, indeed are compelled to look, for a plurality of the best tests that we can find. The problem remains what tests to choose to test language proficiency, and ultimately to predict academic achievement. A useful test has been defined as one that "correspond[s] in demonstrable ways to language in non-test situations."[78] These non-test situations are described in the "new" paradigm of language testing as *authentic*, *direct*, *real-life*, *natural(istic)* or *communicative*. An important part of this study consists of a critical analysis of these terms.

1.5 Hypotheses of the study

The following three *null* hypotheses are investigated:

1. Discrete-point tests and/or integrative tests are *not* valid measures of levels of language proficiency.

2. Discrete-point tests and/or integrative tests are *not* valid long-term predictors of academic achievement.

3. Many of the reports (Grade 6) from former schools that were used as criteria for admission to MHS were *not* valid predictors of academic achievement. Many of the entrants with high Grade 6 report scores did not get beyond Grade 9 at MHS. I investigate the

---

[77] Many scientists, in contrast to many applied linguists, have not given up looking for Grand Unified Theory (GUT). Many applied linguists would probably say that physics deals with non-living matter whereas language testing deals with human beings. But this, in my view, is no justification for rejecting the search for unifying linguistic principles in humans; if one is interested in linguistic *science*, that is, and not just in linguistic thought.

[78] Bachman, L.F. and Palmer, A.S. *Language testing in practice*, 1996, p.9.

question of indiscriminate advancement in DET[79] (Department of Education and Training) schools in the light of the consistently poor Grade 12 ("matric") results from most former DET schools over the years. The predictive validity of the tests are examined in an attempt to shed clarity on this question. (I shall henceforth refer to "DET schools" and not "former DET schools", because at the time of this investigation, the DET was still in existence). A major issue in this study is the relationship between the predictive validity of the tests and that of the DET reports (section 5.5).

Although the study is not directly concerned with investigating mother-tongue[80] proficiency it cannot be separated from a discussion of first language and second language proficiency. The notion of mother tongue is given specific attention in the last chapter (section 6.2), where it is related to native language, first language and second language.

Most of the tests in this study belong to the "old paradigm". I did not devise new tests because it was not germane to the objective of this investigation, which was to examine the validity and reliability and practicability of using traditional tests to predict academic achievement. The fact that most of these tests were already established tests meant that I had more time to devote to this objective.

Although it is possible that annual predictions between English proficiency and academic achievement would yield higher correlations than long-term predictions, the aim in this study is to try and find out what chance Grade 7 learners who entered the School in 1987 had of passing Grade 12.

---

[79] The DET was the education department in charge of black education up to 1994. It is now defunct.

[80] At a translation committee meeting at the University of Fort Hare in April 1998, the secretary of the meeting suggested that the term "mother tongue" was sexist.

1.6 Historical and educational context

The school in this study, Mmmabatho High School (MHS), was established in 1980 and has thus had almost two decades of experience dealing with linguistic, cultural and educational problems that other schools have been dealing with only since the 1994 elections. English is used as the single medium of instruction at the School. The School offered the Joint Matriculation Board (JMB) syllabus up to 1992, and the Independent Examinations Board (IEB) syllabus after 1992. MHS was the only state school that offered the JMB syllabus in a wide area containing hundreds of DET secondary schools. This study shows how DET learners coped at such a school.

One problem that the School has been dealing with since its inception is how to reconcile affirmative action with academic merit. By *affirmative action* I mean the endeavour to put right the imbalances of the past, where the majority of South Africans was discriminated against on the basis of race. The School's policy was to provide education for advantaged as well as disadvantaged learners, where the latter are given the opportunity to learn in an advantaged school situation. Disadvantaged learners are those who have suffered educational, social and economic deprivation - often caused by political injustice - and this was what the School also meant by the term. It is also, paradoxically, the School's policy to accept learners only on merit, which was indicated by high scores on former school reports. The problem with affirmative action is that it is often difficult to marry the idea of redress and the idea of academic merit (high achievement, in this case), potential or aptitude.

This difficulty was evidenced by the School's Prospectus of 1986, which informed parents that their children "are admitted solely on the basis of merit"; by "merit" the School meant high scores on reports from previous schools. Candidates are considered on the basis of the results of an entrance examination and their previous school achievement." Thus, the School's intention was to select only those candidates who could cope with a JMB equivalent

syllabus and simultaneously to uplift the disadvantaged. Unfortunately, many learners dropped out or were pushed out along the way by the system.

MHS's policy was to use admission criteria. The tests in this study, although conducted *after* admission - during the first three days of the first school term - were partly concerned with the admission question because I wanted to find out whether those who were admitted on the basis of their former school reports should have been admitted. Owing to the recent abolition of admission tests in South African state schools there would no longer be any point, it seems, in trying to find the best admission tests. But it would certainly still be useful to find out whether those learners who had been admitted to the School (1) had an adequate level of English proficiency to perform in a school where English was the medium of instruction, (2) whether their former school reports were authentic, i.e. accurate, reflections of this adequate level, and (3) whether they could cope with a JMB syllabus or its equivalent.These three points of research have an important bearing on South African education, as will be shown in the study, especially in Chapter 6.

As far as I am aware, former school reports (point 2 above), as is the general practice in all schools in South Africa, are still considered by the School as an important indication of an entrant's ability - if not a criterion for admission.

It would seem that the School's criteria for admission would generally have pinpointed those candidates who could not cope at the School, but this did not happen. Of concern at the School was the large number of failures in Grades 7, 8 and 9 among the DET learners. At the School there were no automatic internal promotions through the system as is claimed to occur in many DET schools.[81] (This issue is dealt with in section 5.5). When low achievers at the School failed they often left without repeating a year. Many who failed at the School, whether they repeated a year or eventually were asked to leave owing to failure, did not

---

[81] Educamus. *Editorial: Internal promotions*, 36 (9), 3 (1990).

manage to get beyond Grade 9. Table 1.3 shows the Grade 9 pass rate for three consecutive years.

TABLE 1.3

Grade 9 Pass Rate

	Number of learners in Grade 7	Passed Grade 9	% Passes
1	36 (1982)	13 (1984)	**36.1**
2	67 (1983)	25 (1985)	**37.3**
3	81 (1987)	49 (1989)	**60.5**
	184	**87**	**47.3**

Row 3 is the sample used in the prediction of academic achievement in this study. It (1) excludes learners who passed a Grade and then left the School before reaching Grade 9 (N=5), and (2) includes learners who failed between Grades 7 and 9 but passed Grade 9 at a later stage (N=9). Samples 1 and 2 in Table 1.3 (rows 1 and 2) do not take the second fact into account, which means that the pass rate would have been higher.

Learners who got as far as Grade 12 at MHS usually passed Grade 12 and most of these obtained a matriculation exemption, but with disappointing symbols, e.g. D and E symbols. Mcintyre remarks: "Matric students pass successfully but often with symbols that are disappointing."[82]

Mcintyre's statement requires qualification. Although it is correct that there was a high Grade 12 pass rate at the School, this does not take into account the high failure rate between Grade 7 and Grade 9 (see Table 1.3), which means that even though most Grade 12 learners passed, this does not imply that many others that started in Grade 7 didn't drop out along the way. This high failure rate is what had been occurring at the School since its inception in 1980. (I am concerned with the period 1980 to 1993). Table 1.4 shows the number of

[82] Mcintyre, S.P. 'Language learning across the curriculum: A possible solution to poor results.' *Popagano*, 9 and 10, June (1992), p.10.

Grade 12 passes (1992/1993) that originated from the group of Grade 7 learners (1987) used in this study.

TABLE 1.4

Grade 12 Pass Rate

Original number of learners in Grade 7 (1987)	Total Grade 12 passes from original Grade 7
81	41 (50.6%)

Two learners passed Grade 11 with high aggregates and then left the School. These are included in the 41 Grade 12 passes because they would have, without doubt, have obtained a matriculation exemption. (See (1) Note 1 in Table 5.3 and (2) Table 5.4).

Table 1.4 takes into account those who failed and passed Grade 12 in the subsequent year (12 learners) and those who left the school during their schooling for reasons other than failure, for example, relocation. If we compare the Grade 9 pass rate of this sample (row three in Table 1.3) with the Grade 12 pass rate of Table 1.4, we see that 49 learners passed Grade 9, but 41 passed Grade 12. Thus, eight learners dropped out in Grades 10 or 11. The vast majority of dropouts, therefore, were between Grade 7 and Grade 9. A detailed analysis is provided in Chapter 5.

The following criteria of admission to MHS provide important background information: Admission to the School was based on (1) the results of entrance tests administered in October of the previous year and (2) former school achievement. The School's criteria for admission to Grade 7 consisted of:

- Grade 6 reports from former schools *(the aggregate)*.
- A Culture Fair Intelligence Test.[83]
- An English proficiency test, which consisted of a short essay of about half a page. I was not involved in the administration or marking of this test and was not able to obtain the

[83] Cattell, R.B. *Measuring intelligence with culture-fair tests*, 1973.

scores of this test. In any case, the admission essay test was marked by only one rater and so there would have been no way of establishing the interrater reliability of this test.

- A mathematics proficiency test. As in the case of MHS's English proficiency test, I had no information on this test.

The admission tests for the sample in this study were written in October 1986. The Grade 6 reports were considered by the School to be the most important criterion for admission. However, a few learners were admitted with Grade 6 aggregates below 60%. Of the Schools admission criteria only the Grade 6 reports are used in this study.

With regard to the culture-fair entrance test at MHS, my original intention was to include these in the predictive investigation, but owing to the problematic (scientific and political) nature of intelligence tests and the fact that the use of these tests as predictors would not be directly pertinent to the topic, I decided to exclude these tests from this investigation. Suffice it to say that learners who score above average on intelligence tests tend to be better at *formal* first or second language learning.[84] The degree of culture-fairness of such tests is irrelevant to this fact. I say no more on this highly controversial matter.

The School's policy was that at least half of all admissions should consist of disadvantaged learners. *Disadvantaged* does not mean *low scoring*, because the School selected on the basis of good performance as indicated by former (Grade 6) school reports. These disadvantaged entrants came from DET Schools. The investigation will show that the scores of the Grade 6 reports of these DET schools were *radically* higher than the scores on the English proficiency tests.

The full sample of subjects (N=86 [L1=49; L2=37]) is discussed in detail in section 3.2.1, but for the moment I deal briefly with 67 subjects (Table 1.5) in order to show the following

---

[84] Mitchell, R. and Myles, F. *Second language acquisition*, 1998.

comparison. Compare (Table 1.5) the aggregate and English scores of the Grade 6 reports of the L1 and L2 groups of entrants to Grade 7 at the School:

(1) The L1 group (N=33). Most of the L1 group were from CM Primary School, the "feeder" school, which provided (at the time this research was conducted) most of the entrants who took English First Language as a subject at MHS. English was the official medium of instruction from Grade 1 at CM Primary School. The learners from this school were generally advantaged. As mentioned, disadvantaged learners are those who have suffered educational, social and economic deprivation.

(2) The L2 group (N=34). Most of the L2 group originated from DET schools . These entrants took English Second Language as a subject at MHS. English was the medium of instruction from Grade 5 at DET schools. Entrants from DET schools were generally disadvantaged.

TABLE 1.5

Comparison of Grade 6 reports between CM Primary School and 28 DET Schools (N=67)

	Aggregate Grade 6		English Grade 6	
	Mean	STD	Mean	STD
CM Primary (N=33): mostly advantaged and English used as a First Language. (**L1**).	68.9	8.8	72.5	8.4
28 DET Schools (N=34): mostly disadvantaged and English used as a Second Language. (**L2**).	68.6	10.8	71.1	12.6
t Stat	- 0.106		- 0.550	
t Critical two-tail	1.995		1.995	

The T-Test in Table 1.5 shows that there was no significant difference between the two groups because the t Stat is less than the critical value. This equivalence in Grade 6 report scores between these groups plays an important role in the arguments and predictions of the study.

1.7 Measures used in the study

The measures used in the study are now briefly described. The study only commenced after the intake of learners to the School, and thus these measures differed in purpose from the School's criteria, which were admission criteria. A detailed description of the measures, e.g. format, instructions, layout, is given in Chapter 3. For the moment I provide only a brief description of the measures:

I. English proficiency tests. Eight English proficiency tests were administered in January 1987 (Grade 7). I devised the essay tests myself, while all the other tests were obtained from various published sources. The English proficiency test battery consists of:

(i) Two cloze tests from Pienaar's[85] "Reading for Meaning".

(ii) Two dictation tests. These were two restored cloze tests from Pienaar.[86] The passages from Pienaar used for the cloze tests are different to the passages used for the dictation tests, but they both belong to the same level. (I explain later what I mean by "level" in section 3.3.1.2).

(iii) Two essay tests (devised by myself).

(iv) An "error recognition" test.[87]

(v) A "mixed grammar" test.[88]

The tests from Bloor et al. consist of multiple-choice items. The "mixed grammar" test consists of items that test a variety of structures, hence the term "mixed".

---

85 Pienaar, P. *Reading for meaning: A pilot survey of (silent) reading standards in Bophuthatswana*, 1984, pp.59 and 61.

86 Ibid., pp.58 and 62.

87 Bloor, M., Bloor, T., Forrest, R., Laird, E. and Relton, H. *Objective tests in English as a foreign language*, 1970, pp.70-77.

88 Ibid., pp.35-40.

I shall argue that although the tests used in the study, except for the essay test, may be out of fashion with many testers, they are nevertheless still very useful for assessing language proficiency and predicting academic achievement.

II. Grade 6 end-of-year school reports from former schools. The Grade 6 aggregate scores are used.

III. Grade 7 to Grade 11 end-of-year aggregates. These scores were obtained from the School's mark schedules.

IV. Grade 12 results (of 1992 and 1993). These results are those of the JMB (1992) and the Independent Examinations Board (IEB; 1993). The IEB results are also taken into account because also included in the study are those subjects who failed once between Grade 7 and Grade 12, repeated a year and sat for the IEB Grade 12 examination in 1993. (The JMB matriculation examination ceased to exist after 1992 and was replaced by the IEB in 1993).

1.8 Method overview

Some researchers separate statistical research from empirical research. For Lantolf and Frawley empirical research and statistical measurement are distinct.[89] In contrast, when Tremblay and Gardner state that in their opinion "empirical investigation is essential to demonstrate the theoretical and pragmatic value"[90] of research, their "empirical" research is firmly based on statistics, without which they would have very little of what they consider to be "empirical" research (see also Cziko's "empirically-based models of communicative competence"[91]). Some empirical research is statistically based, while other empirical research

---

89 Lantolf, J.P. and Frawley, W. 'Proficiency: Understanding the construct.' *Studies in Second Language Acquisition (SLLA)*, 10 (2), 181-195 (1988), p.181.

90 Tremblay, R.F. and Gardner, R.C. 'Expanding the motivation construct of language learning.' *The Modern Language Journal*, 79 (4), 505-518 (1995), p.505.

91 Cziko, G.A. 'Some problems with empirically-based models of communicative competence.'

is not, e.g. much of ethnographical research. The empirical research in this study is mainly based on statistics.

The empirical investigation consists of:

1. An examination of the structure and administration of the English proficiency tests.

2. A statistical analysis of the results of the English proficiency tests.

3. A predictive investigation, where the English proficiency tests are used to predict academic achievement from Grade 7 to Grade 12. The reliability of the Grade 6 reports of entrants from former schools are also examined.

Under method I subsume, as is usual in most studies, the following:

- Subjects (sampling).
- Structure of the measures.
- Procedures of administration and scoring.

A common design in empirical studies is that data analysis, results and discussion are each reported in separate sections. I depart from this traditional structure and follow Sternberg,[92] who recommends that these be treated together. This, I believe, is a sensible arrangement because the data analysis, discussion and results are closely connected.

A clarification of the following terms are in order: *type, method, procedure*. Sometimes *type* refers to such things as multiple-choice type questions versus gap-filling type tests; sometimes *method* refers to such things as cloze methods versus dictation methods, in other

---

Applied Linguistics, 5 (1), 23-37 (1984).

[92] Sternberg, R.J. *The psychologist's companion: A guide to scientific writing for students and researchers*, 1993, p.53.

words, *methods* is used to mean *tests*. Then there is *procedure*, e.g. the cloze procedure, the dictation procedure, etc., which also can mean *methods* or *tests*. I shall use *tests* to refer to elicitation techniques, and *procedure* to the way in which tests are presented and scored.

1.9 Preview of Chapters 2 to 6

Chapter 2 deals with theoretical issues in the testing of language proficiency and academic achievement, where the main focus falls on assessment. The chapter comprises a review of the literature on the testing of language proficiency and an overview of key concepts such as *assessment, validity* and *reliability*.

Chapter 3 describes the sample of subjects and sampling procedures, and the structure and administration of the tests.

Chapter 4 presents the results of the tests and discussion.

Chapter 5 deals with the prediction of academic achievement, examines the reliability of the Grade 6 reports from previous schools and summarises the findings.

Chapter 6 discusses the implications of the study for language testing and presents the conclusions. The three main implications are: (1) the viability of the distinction between *first language* and *second language*, (2) the kind of tests or tasks that should be used, (3) the problem of rater reliability.

Woven into the arguments is a description of a few contemporary initiatives to improve language assessment.

1.10 Summary of Chapter 1

The purpose, problem, main topics, method, hypotheses and educational context of the study were specified. The study deals with the measurement of differences between learners in English proficiency and with assessing the reliability, validity and practicability of discrete-point and/or integrative tests as predictors of academic achievement. Central to the study is the argument that the "old paradigm" of discrete-point and integrative tests and the statistical methods required to measure them are very useful in language acquisition research and educational measurement. The next chapter deals with the theory of language testing and its relationship to academic learning and academic achievement, where also examined is what it means to call language behaviour and language tests "authentic".

CHAPTER 2

Theoretical Issues in the Testing of Language Proficiency and Academic Achievement

2.1 Introduction

Whatever we talk about originates from a theory, i.e. a combination of knowledge, beliefs, wants and needs. In this chapter I deal with language proficiency and academic achievement. For some authors[1] the testing of language proficiency has been a circular enterprise. Vollmer maintains that "language proficiency is what language proficiency tests measure"[2] and this circular statement is all that can be firmly said when asked for a definition of language proficiency. Although this may be all that can firmly be established about language proficiency, we swim on in the hope of hitting *terra firma.*

In the next section I discuss the general notion of ability and its relationship to cognitive skills followed by a discussion of language ability. Subsequent sections move on to competence and performance, proficiency and the discrete-point/integrative controversy, academic achievement, and validity and reliability.

2.2 Ability, cognitive skills and language ability

As mentioned in the first paragraph of the study, language testing draws on three areas: the nature of language, assessment, and language ability .

---

1 (1) Ingram, E. 'Assessing proficiency: An overview on some aspects of testing', in Hyltenstam, K. and Pienemann, M. *Modelling and Assessing second language acquisition*, 1985, p.218. (2) Vollmer, H.J. 'Why are we interested in general language proficiency?', in Alderson, J.C. and Hughes, A. *Issues in language testing: ELT Documents III*, 1981, p.152.

2 Ibid.

The precise definition of ability is not only seldom explicated, but, unlike competence, is often not even considered; this in spite of the fact that the term is used widely in everyday language as well as in scientific circles. Important issues in the study of abilities, of which language is only one ability, are:

(1) The fixety of abilities. If abilities were highly variable over time they would reflect a state rather than a trait or an attribute. The two latter terms imply a fixed structure, or *construct*, rather than a variable process. The constructs that this study is concerned with belong to the domain of language acquisition. (Construct validity is discussed in section 2.8.3).

(2) Consistency. For example, if an athlete in a one-off streak "accident" breaks a world record but is never able to repeat the performance, or even get near the record again, we still say that he or she has the ability to break a world record. We cannot apply the same logic to cognitive abilities, where consistency of output, not records, is the name of the game. Consistency does not only apply to the ability of learners but also of teachers, who are usually also testers. The consistency, or the reliability, of judgements and scoring is a major issue in language testing. This issue is dealt with in various parts of study.

(3) When we say people have the ability to perform academically we mean that they are able to achieve a certain liminal level, i.e. minimum or threshold level. In trying to set a minimum level one is concerned with what the individual can do in terms of established criteria. What the individual can do cannot be separated from what others can do. Hence the importance of norm-referenced tests. This point would be contested by extreme proponents of criterion based testing and developmental testing.

(4) The variability in ability between individuals obeys a "bell-curve" distribution, as in the case of nature as a whole. The "bell-curve" or "normal" distribution is the foundational principle of psychometrics. *Norm* is used in the sense of an idealisation against which comparisons are made of what scientists call the "real" world. Although this "normal" curve is a mathematical abstraction it is based on the reasoning that if there were an infinitely large population then human abilities (and the milk yield of cows) would be represented by a

perfect bell curve. Taking the four points above into account, Carroll suggests the following definition of ability:

> *As used to describe an attribute of individuals, ability refers to the possible variations over individuals in the liminal levels of task difficulty (or in derived measurements based on such liminal levels) at which, on any given occasion in which all conditions appear favorable, individuals perform successfully on a defined class of tasks.* [3]

Several modern theories of education and psychology reject the notion of *traits*, i.e. the fixety of psychological constructs.[4] In traditional trait theories, e.g. Carroll (above), psychological constructs are like any other human trait, animal trait, or plant trait, where biological differences between living things are distributed according to a bell curve. Differences in human abilities are also distributed according to the bell curve. This does not mean that people cannot improve, but only that the degree of improvement depends on fixed psychobiological constraints.

A few more comments on Carroll's idea that ability is a fixed psychological trait are in order. The notions of "transferable" and "transferring skills" are used to explain the idea of "fixed" ability.

A major problem with learners with limited academic ability is the underdevelopment of "transfer skills".[5] There are two kinds of transfer skills: (1) lower order "transferable skills" and (2) higher order "transferring skills".[6] Transferable skills are skills that are learnt in one situation or one kind of subject-matter that are transferable to another. Examples are: (i) a reading skill such as scanning that is learnt in the English class can be transferred to the

---

3 Carroll, J.B. *Human cognitive abilities: A survey of factor analytic studies*, 1993, p.10.

4 Minick, N.J. L.S. *Vygotsky and Soviet activity theory: New perspectives on the relationship between mind and society*, 1985, pp.13-14.

5 Botha, H.L. and Cilliers, C.D. 'Programme for educationally disadvantaged pupils in South Africa: A multi-disciplinary approach.' *South African Journal of education*, 13 (2), 55-60 (1993).

6 Bridges, P. 'Transferable skills: A philosophical perspective.' *Studies in Higher Education*, 18 (1), 43-52 (1993), p.50.

geography class, (ii) using a dictionary, (iii) making charts and diagrams, (iv) completing assignments, (v) reviewing course material, (vi) learning formulas and dates, and (vii) memorising material. "Transferring skills" are "metacompetences" of a far higher order. These metacompetences are : (i) A sensitive and intelligent discernment of similarities and differences, (ii) Cognitive equipment that one uses to modify, adapt and extend, and (iii) attitudes and dispositions that support both of the above.[7]

These three "metacompetences" are interrelated. For example, without the "cognitive equipment" that enables one to modify, adapt and extend, it would not be possible to sensitively and intelligently discern similarities and differences. With regard to Bridges' third "metacompetence" of "attitudes and dispositions", which have to deal with intention, motivation and resulting approach to a task, I suggest that its successful development is to a large degree dependent on the successful development of the other two "metacompetences". If one has the right healthy cognitive equipment, in working order, as well as the desire and opportunity to develop it, one will understand more; consequently, one will be more motivated to learn. Of course, socialisation into a community of learners and the correct mediation/intervention procedures between learner and task also play an important role in cognitive development, e.g. the development of critical awareness and learning strategies.

Bridges' distinction between lower order "transferable skills" and higher order "transferring skills"[8] is useful in understanding the nature of the problem of transfer. The problem of transfer refers mostly to the higher order "transferring skills". The question is whether higher order cognitive skills (i.e. Bridges' "transferring skills") can be acquired at all (whether independently or through teaching). Millar maintains that courses in skills development (e.g. the development of executive processes) pursue the "impossible" because processes such as classifying and hypothesising cannot be taught, but can only develop (i.e. they are part of

---

7 Ibid.

8 Bridges, P. 'Transferable skills: A philosophical perspective.' *Studies in Higher Education*, 18 (1), 43-52 (1993), p.50.

inborn potential, or ability).[9] For Millar the challenge is to find ways of "motivating pupils to feel that it is personally valuable and worthwhile to pursue the cognitive skills (or processes) they [children] *already possess* to gain understanding of the scientific concepts which can help them make sense of their world"[10]. (Square brackets and italics added).

According to Millar, these cognitive skills, especially the higher order transferring skills (e.g. a sensitive and intelligent discernment of similarities and differences), can only be developed if they are based on something that learners already possess, namely, academic potential, or ability. I have raised some highly controversial issues, but they needed to be raised to explain what I mean by "fixed" ability. I cannot pursue these issues further in this study.[11]

I now relate ability to language. In section 1.1, four major test uses were mentioned; achievement, proficiency, aptitude and diagnosis. These are all manifestations of what Davies' calls "language ability".[12]

"Fixed" ability in Carroll's sense does not mean that people cannot develop and become better. If people couldn't develop, it would be nonsensical to talk about things such as transitional competence and interlanguage, which feature so prominently in the applied linguistic literature.

In the next section I discuss the notions of competence and its sibling, performance.

---

9 Millar, R. 'The pursuit of the impossible.' *Physics Education*, 23, 156-159 (1988), p.157.

10 Ibid.

11 See (1) Gamaroff, R. 'Solutions to academic failure: The cognitive and cultural realities of English as the medium of instruction among black ESL learners.' *Per Linguam*, 11 (2), 15-33 (1995c).

(2) _____ 'Abilities, access and that bell curve.' Grewar, A. (ed.). *Proceedings of the South African Association of Academic Development "Towards meaningful access to tertiaty education*, 1996b.

(3) _____ 'Language as a deep semiotic system and fluid intelligence in language proficiency.' *South African Journal of Linguistics*, 15 (1), 11-17 (1997b).

12 Davies, A. *Principles of language testing*, 1990, p.6,.

2.3 Competence and performance

For Chomsky competence is the capacity to generate an infinite number of sentences from a limited set of grammatical rules.[13] This view posits that competence is logically prior to performance and is therefore the generative basis for further learning.[14] Competence, on this view, is equivalent to "linguistic" (or "grammatical") competence. Chomsky distinguishes between "performance", which is "the actual use of language in concrete situations", and "competence" or "linguistic competence" or "grammatical competence", which is "the speaker-hearer's knowledge of his language".[15] Chomsky's description of language involves no "explicit reference to the way in which this instrument is put to use...this formal study of language as an instrument may be expected to provide insight into the actual use of language, i.e. into the process of understanding sentences."[16] Chomsky's great contribution was to focus on linguistic introspection, without giving introspection (linguistic intuitions) the final word.[17]

Canale and Swain make a distinction between knowledge of use and a demonstration of this knowledge.[18] Knowledge of use is often referred to in the literature as "communicative competence"[19], and the demonstration of this knowledge as "performance". Communicative

---

13 Chomsky, N. *Aspects of the theory of syntax*, 1965, p.6.

14 (1) Brown, K. *Linguistics today*, 1984, p.144.
(2) Leech, G. *Semantics*, 1981, p.69.
(3) Hutchinson, T. and Waters, A. *English for special purposes: A learner-centred approach*, 1987, p.28.

15 Chomsky, N. *Aspects of the theory of syntax*, 1965, pp. 3-4.

16 Chomsky, N. *Syntactic structures*, 1957, p.103.

17 Atkinson, M., Kilby, D. and Roca, I. *Foundations of general linguistics*, 1982, 369.

18 Canale, M. and Swain, M. 'Theoretical bases of communicative approaches to second language teaching and testing.' *Applied Linguistics*, 1 (1), 1-47 (1980), p.34.

19 Hymes, D. 'On communicative competence', in Pride, J.B. and Holmes, J. (eds.). *Sociolinguistics*, 1972.

competence has come to subsume four sub-competences: grammatical competence, sociolinguistic competence, discourse competence and strategic competence[20]:

(1) Grammatical competence is concerned with components of the language code at the sentence level, e.g. vocabulary and word formation.

(2) Sociolinguistic competence is concerned with contextual components such as topic, status of interlocutors, purposes of communication, and appropriateness of meaning and form.

(3) Discourse competence is concerned with: (i) a knowledge of text forms, semantic relations and an organised knowledge of the world; (ii) cohesion - structural links to create meaning, and (iii) coherence - links between different meanings in a text; literal and social meanings, and communicative functions.

(4) Strategic competence is concerned with (i) improving the effectiveness of communication, and (ii) compensating for breakdowns in communication. Strategic competence means something very different in Bachman and Palmer, namely, metacognitive strategies, which is central to communication. For these authors, "language ability" consists of "language knowledge" and "metacognitive strategies".[21] (See Skehan[22]).

According to Widdowson communicative competence should subsume the notion of performance:

> *[T]he idea of communicative competence arises from a dissatisfaction with the Chomskyan distinction between competence and performance and essentially seeks to establish competence status for aspects of language behaviour which were indiscriminately collected into the performance category.*[23]

---

20 (1) Canale, M. and Swain, M. 'Theoretical bases of communicative approaches to second language teaching and testing.' *Applied Linguistics*, 1 (1), 1-47 (1980), p.34.

(2) Swain, S. 'Large-scale communicative language testing: A case study', in Lee, Y., Fok, A., Lord, R. and Low, G. (eds.). *New directions in language testing*, 1985.

(3) Savignon, S.J. *Communicative competence: Theory and classroom practice*, 1983.

21 Bachman, L.F. and Palmer, A.S. *Language testing in practice*, 1996. (See their Chapter 4).

22 Skehan, P. *A cognitive approach to language learning*, 1998, p.16.

23 Widdowson, H.G. *Aspects of language teaching*, 1990, p.40.

How does ability fit into the competence-performance distinction? Chomsky equates "ability" with "performance" ("actual use"), which he regards as a completely different notion from "competence" or knowledge".

> *Characteristically, two people who share the same knowledge will be inclined to say quite different things on different occasions. Hence it is hard to see how knowledge can be identified with ability...Furthermore, ability can improve with no change in knowledge.*[24]

Thus, as Haussmann points out, "it should be noted that Chomsky's original definition of the term [i.e. competence] always excluded this idea [i.e. ability]."[25] There doesn't seem to be any reason, however, why *ability* cannot refer to (linguistic/grammatical) competence, (which is Chomsky's interest), as well as to the knowledge one has of how to use the language in appropriate situations. We can retain *performance* to mean the actual use of this knowledge. For example, Bachman and Clark define "ability" in the following way: "We will use the term 'ability' to refer both to the knowledge, or competence, involved in language use and to the skill in implementing that knowledge, and the term 'language use' to refer to both productive and receptive performance.[26] Weir also equates "ability" with "competence":

> *There is a potential problem with terminology in some recent communicative approaches to language testing. References are often made in the literature to testing communicative 'performance' [e.g. B.J. Carroll 1980[27]]. It seems reasonable to talk of testing performance if the reference is to an individual's performance in one isolated situation, but as soon as we wish to generalise about ability to handle other situations, 'competence' would seem to be involved.*[28] *(Square brackets added)*

---

24 Chomsky, N. *Language and the problem of knowledge*, 1988, p.9.

25 Haussmann, N.C. *The testing of English mother-tongue competence by means of a multiple-choice test: An applied linguistics perspective*, 1992, p.16.

26 Bachman, L.F. and Clark, J.L.D. 'The measurement of foreign/second language proficiency.' *American Academy of the Political and Social Science Annals*, 490, 20-33 (1987), p.21.

27 Carroll, B.J. *Testing communicative performance*, 1980.

28 Weir, C.J. *Communicative language testing*, 1988, p.10.

This is Skehan's position as well: "it is defensible to speak of competence-orientated abilities."[29] In other words, different performances point back to the underlying competence or ability.

"Competency-based education and training" (CBET) has a different set of concepts for the labels of competence, performance and ability to those discussed above. CBET is discussed in section 6.4 where the future of assessment in South Africa is dealt with.

2.4 Proficiency

Proficiency is closely related to ability, competence and performance discussed above. Proficiency is used in at least two different ways: it can refer to (1) the "construct or competence level"[30], which is at a given point in time independent of a specific textbook or pedagogical method[31] or to (2) the "performance level"[32], which is a reflection of achievement in the test situation. The construct level or competence level is the knowledge of the language, and the performance level is the use of language.

Proficiency, like the notions of competence and performance, is very much of a "chameleon" notion[33], because it can be defined not only in terms of knowledge (the construct or competence level) and in terms of specific tasks or functions (the performance level), but also in terms of degrees of behaviour that are observed at different stages (minimum to native-like[34]), in terms of language development (e.g. interlanguage studies), in terms of

---

29 Skehan, P. *A cognitive approach to language learning*, 1998, p.154

30 Vollmer, H.J. 'The structure of foreign language competence', in Hughes, A. and Porter, D. (eds.). *Current developments in language testing*, 1983, p.5.

31 Brière, E. 'Are we really measuring proficiency with our foreign language tests?' *Foreign Language Annals*, 4, 385-91 (1971), p.322.

32 Vollmer, H.J. 'The structure of foreign language competence', in Hughes, A. and Porter, D. (eds.). *Current developments in language testing*, 1983, p.5.

33 Hyltenstam, K. and Pienemann, M. *Modelling and Assessing second language acquisition*, 1985, p.15.

34 The term *native* is problematic. This problem is discussed in sections 3.2.1 and 6.1.1.

situations that require some skills but not others, or in terms of general proficiency, where no specific skill is specified.

Porter uses the term "communicative proficiency"[35], which seems to subsume the notions of "communicative competence" and "performance" discussed above. According to Child, "proficiency" is a "general `across-the-board' potential", while "performance" is the "actualised skill", the "mission performance" involved in "communicative" tasks, i.e. the output.[36] Child has much in common with Alderson and Clapham, who distinguish between "language proficiency" and "language use", where proficiency, not use, is part of output.[37]

I would like to spend some time on Lantolf and Frawley's views on language proficiency because they epitomise the opposition to the view that I am arguing for in this study.[38] These authors will be referred to as L and F. In their abstract they state that they "argue against a definitional approach to oral proficiency and in favor of a principled approach based on sound theoretical considerations."[39] The authors use oral proficiency as a backdrop to their views on language proficiency in general. L and F, in their criticism of "reductionism" in the assessment of language proficiency, leave few authors unscathed; authors that many would consider to be at the vanguard of the real-life/communicative movement, e.g. Hymes, Omaggio and Widdowson.

To adumbrate: in the second section of their article "The tail wagging the dog", L and F use Omaggio's section of her manual entitled :"Defining language proficiency"[40] to lament that the "construct of proficiency, reified in the form of the [American Council on the Teaching

---

35 Porter, D. 'Assessing communicative proficiency: The search for validity', in Johnson, K. and Porter, D. (eds.). *Perspectives of communicative language teaching*, 1983.

36 Child, J. 'Proficiency and performance in language testing.' *Applied Linguistic Theory*, 4 (1/2), 19-54 (1993).

37 Alderson, J.C. and Clapham, C. 'Applied linguistics and language testing: A case study of the ELTS test.' *Applied Linguistics*, 13 (2), 149-167 (1992), p.149.

38 Lantolf, J.P. and Frawley, W. 'Proficiency: Understanding the construct.' *Studies in Second Language Acquisition (SLLA)*, 10 (2), 181-195 (1988).

39 Ibid.

40 Omaggio, A.C. *Teaching language in context: Proficiency-orientated instruction*, 1986.

of Foreign Languages - ACTFL] *Guidelines*, has begun to determine how the linguistic performance of real people *must* be perceived":

> *In her discussion, she considers various models of communicative competence, including those of Hymes, Munby, Widdowson, and Canale and Swain, all of which are reductionist approaches to communicative competence, because they define communicative competence by reference to a set of constitutional criteria. She then proceeds to a subsection entitled "From Communicative competence to Proficiency." However, nowhere in her analysis is there any in-depth consideration of proficiency that is independent of the proficiency test itself.*[41]

I would think that any "consideration of proficiency" independent of the "test itself" (L and F in their quotation) is reductionist: but I am pre-empting the end of the study.

Strange that L and F consider Widdowson[42] a reductionist. I would think that Widdowson fully appreciates the distinction between language structure and language in use, where grammar plays a vital role. By "grammar", Widdowson does not mean merely morphology, phonology and syntax but lexico-grammar, where semantics is included under "grammar". The inclusion of semantics under "grammar", or "linguistic knowledge", is what modern linguistics understands by these terms. The papers of the Georgetown University Round Table Conference[43] were concerned with the reality and authenticity of communicative language proficiency, where Widdowson argued that grammar is not dead, but the life blood of language, communication and social meaning. Such a view is not reductionist!

---

41 Lantolf, J.P. and Frawley, W. 'Proficiency: Understanding the construct.' *Studies in Second Language Acquisition (SLLA)*, 10 (2), 181-195 (1988), p.182.

42 Widdowson, H.G. *Explorations in applied linguistics*, 1979.
_____ 'Knowledge of language and ability of use.' *Applied Linguistics*, 10 (2), 128-137 (1989).
_____ *Aspects of language teaching*, 1990.
_____ 'Communication, community and the problem of appropriate use', in Alatis, J.E. *Georgetown University Round Table on Languages and Linguistics*, 1992.

43 Alatis, J.E. (ed.). *Georgetown University Round Table on Languages and Linguistics*, 1992.

L and F reject the ACTFL's (American Council of the Teaching of Foreign Languages) adoption of a uniform yardstick for the measurement of foreign language ability based on real-life behaviour.[44] The ACTFL's tail (the series of real life descriptors) that is wagging the real dog is not, according to L and F, a real tail. The unreal tail for L and F is the unreal "construct"; the real dog being wagged is real people. The metaphor is clear: it is *researchers* who have fabricated the "construct", and fabrications have no psychological reality. In other words the construct constricts the reality of "the nontest world of human interaction".[45] The test world, which represents the "construct" for these authors, "has come to determine the world, the reverse of proper scientific methodology".[46]

Recall that L and F are arguing in "favor of a *principled* approach based on sound theoretical considerations" (italics added), which L and F seem to think authors such as Widdowson do not use. Yet Widdowson, who was probably not unaware of L and F's criticism, ends his "Aspects of language teaching" with the following:"There needs to be a continuing process of *principled* pragmatic enquiry. I offer this book as a contribution to this process - and as such, it can have no conclusion"[47] (italics added). (See Gamaroff 1996a[48]).

Widdowson perceives the content of both the structural and the notional syllabus to be, in Nunan's words, "synthetic" and "product-orientated"[49], i.e. the content of both syllabuses is static and lacks the power to consistently generate communicative behaviour. Widdowson's argument against structuralist and notional syllabuses is that "[i]t has been generally assumed...that performance is a projection of competence...that once the rules are specified we automatically account for how people use language."[50] His argument is that structural and

---

44 Byrnes, H. and Canale, M. (eds.). *Defining and developing proficiency: Guidelines, implementations and concepts*, 1987.

45 Lantolf, J.P. and Frawley, W. 'Proficiency: Understanding the construct.' *Studies in Second Language Acquisition (SLLA)*, 10 (2), 181-195 (1988), p.182.

46 Ibid.

47 Widdowson, H.G. *Aspects of language teaching*. (Oxford, Oxford University Press, 1990).

48 Gamaroff, R. 'Is the (unreal) tail wagging the (real) dog?: Understanding the construct of language proficiency.' *Per Linguam*, 12 (1), 48-58 (1996a).

49 Nunan, D. *Syllabus design*, 1988, p.28.

50 Widdowson, H.G. *Explorations in applied linguistics*,1979, p.141.

functional-notional syllabuses do not link in past experiences with new experiences because they lack proper learner involvement.[51] Widdowson also believes that "the most effective means towards this achievement [i.e. "complete native-speaker mastery"] is through an experience of authentic language in the classroom."[52]

L and F, and Widdowson are backing the same communicative horse. The main difference between them appears to lie in the value they place on school learning. All three believe in teaching language as communication, with the difference that much of Widdowson's work is concerned with academic achievement and school learning rather than with real-life "natural" contexts.

I examine more closely the cogency of the distinction between "natural" contexts in "real-life" and "unnatural" contexts in the classroom. According to L and F, "tasks cannot be authentic by definition"[53], which implies that very little in school is authentic, i.e. natural. The nub of L and F's criticism is that the exchange between tester and test taker is not a natural one, therefore any kind of test cannot be a natural kind of communication. Communicative *testing*, it seems, would be for L and F a contradiction in terms. What is more, communicative school *tasks* would also be a contradiction in terms. In that case, school, which may be defined as an institution whose role it is to *guide* learners by defining and dispensing tasks, is another tail wagging the world (of "reality"). The ACTFL Guidelines, according to L and F, draw a line between the world and the individual. L and F regard such a situation as scientifically unprincipled and morally untenable. There is very little in tasks such as instructional activities and nothing in tasks such as tests that L and F find authentic in the Proficiency literature. L and F want language tasks to be contextualised in natural settings such as cooking clubs.

---

51 Ibid, p.246.

52 Widdowson, H.G. 'Communication, community and the problem of appropriate use', in Alatis, J.E. *Georgetown University Round Table on Languages and Linguistics*, 1992, p.306.

53 Lantolf, J.P. and Frawley, W. 'Proficiency: Understanding the construct.' *Studies in Second Language Acquisition (SLLA)*, 10 (2), 181-195 (1988), p.183.

Byrnes and Canale caution that the danger of "the proficiency movement" as espoused by L and F and others as "with any movement is that a rhetoric of fear and enthusiasm will develop which is more likely to misrepresent and confuse than to clarify the crucial issues."[54] One confusing issue is that of "natural environment".

"Just what is a 'natural environment' as far as learning or acquiring a second language under any circumstances is concerned?" asks Morrissey[55]:

> *There is no environment, natural or unnatural, that is comparable with the environment in which one learns one's mother tongue. Furthermore, it seems to me that there is a teaching (i.e. unnatural?) element in any L2-L1 contact situation, not just in cases of formal instruction. This element, even if it only consists in the awareness of the communicants that the [teaching or testing] situation exists, may be a more significant factor in L2 learning and L2 acquisition [and L2 testing] than any other factor that is common to [the natural setting of] L1 acquisition and L2 acquisition.*[56]

L and F are seeking a testing situation analogous to the L2 "acquisition" situation (which in terms of Krashen's[57] definition is a "natural" situation). But, as Morrissey above suggests, much of language and learning, like culture, consists of extrapsychological elements, and in this sense are an "imposition" upon nature. However, although the test situation, i.e. school, may be less "authentic" in the sense that the test is more concerned with learning language than with using it, the dichotomy between "natural" and "unnatural" is a spurious one. It is incorrect to assume that "natural" approaches (e.g. Krashen and Terrell[58]) and immersion programmes mirror natural language acquisition and that the ordinary classroom doesn't.[59]

---

54 Byrnes, H. and Canale, M. (eds.). *Defining and developing proficiency: Guidelines, implementations and concepts*, 1987, p.1.

55 His specific context is the second language "acquisition"/second language "learning" controversy of Krashen (1981). See Note 60 below for the reference.

56 Morrissey, M. D. 'Toward a grammar of learner's errors.' *International Review of Applied Linguistics*, 21 (3), 193, 207 (1983), p.200.

57 Krashen, S. *Second language acquisition and second language learning*, 1981.

58 Krashen, S. and Terrell, T. *The natural approach: Language acquisition in the classroom*, 1983.

59 Butzkamm, W. 'Review of H. Hammerly, "Fluency and accuracy: Toward balance in language teaching and learning"'. *System*, 20 (4), 545-548 (1992).

The swimming club, cooking club, tea party or cocktail party are in a sense not more neither less natural than the traditional classroom. That is why one doesn't have to go outside the classroom in search of "real reality".[60] The learning brain needs stimulation, and it can get it in the classroom or at an informal (or formal) cocktail party. In other words, there is much informal learning in classrooms and much formal learning outside the classroom. But both kinds of learning are completely "natural" to the brain that is doing the learning.

2.5 The discrete-point/integrative controversy

L and F point out that in real life one uses far less words than one would use in school "tasks", and this is one of the reasons why, they maintain, that tests are inauthentic by definition.[61] However, as Politzer and Mcgroarty[62] show, it is possible to say or write few words (as one often does in natural settings) in a "communicative competence" test by using a "discrete-point" format. When one uses far less words in natural settings than one would use in many "artificial" school tasks, one is in fact using a "discrete-point" approach to communication. One doesn't merely look at the *format* of a test to decided whether it is a "discrete-point" test, one looks at *what* the test is testing.

It is now opportune to examine what tests are testing. The way I have chosen to do so is by means of an examination of the discrete-point/integrative controversy. This controversy is better understood within the context of a parallel controversy: the structuralism/ functionalism controversy. The discussion of the latter controversy will serve as background to the discrete-point/integrative controversy.

It is impossible to test the structures and functions of language without understanding how language is learnt. Language learning is language processing. The central issue in testing is

---

60 Taylor, B.P. 'In search of real reality.' *TESOL Quarterly*, 16 (1), 29-43 (1982).

61 Lantolf and Frawley, 1988, p.183.

62 Politzer, R.L. and McGroarty, M. 'A discrete-point test of communicative competence.' *International Review of Applied Linguistics*, 21 (3), 179-191 (1983).

assessing this language processing skill. Language processing, as with all knowledge, exists within a hierarchical organisation: from the lower level atomistic "bits" to the higher level discoursal "bytes". The lower level bits traditionally belong to the "structuralist" levels, while the higher level bytes belong to the "functionalist" levels. It is difficult to know where structure ends and function begins.[63]

The following continuum, adapted from Rea[64], includes the concepts of competence and performance discussed in section 2.3.

TABLE 2.1

Functionalist and Structuralist Levels of language

FUNCTIONALISM Communicative		STRUCTURALISM Non-communicative
Performance	Knowledge of rules	
Communicative performance	Communicative competence	Linguistic competence
Pragmatics (discourse level;"use")		Semantics (sentence level;"usage")

63 Entwhistle, W.J. *Aspects of language*, 1953, p.157.

64 Rea, P. 'Language testing and the communicative language teaching curriculum', in Lee, Y.P. et al. *New directions in language testing*, 1985.

This classification of functionalism and structuralism into two distinct categories is a highly controversial one. The structuralist/functionalist controversy is about whether the semantic meaning of words and sentences (structuralism) can be distinguished from the pragmatic (encyclopaedic, i.e. world knowledge) meaning of discourse (functionalism).

Halliday proposes two meanings of the term *function*, namely, "functions in structure" and "functions of language".[65] "Functions in structure" is concerned with the relationship between different words of a sentence. *Structuralism* is traditionally associated with the study of language at the sentence level and below. "Functions of language", on the other hand, goes beyond individual linguistic elements or words (Saussure's[66] "signs") to discourse. *Functionalism* is traditionally associated with the study of discourse.

I understand the terms *linguistic knowledge*, *lexico-grammar* and Halliday's "functions in structure" to be synonymous. Lexico-grammar only deals with linguistic knowledge at the sentence level and below that level. Halliday's "functions of language", what I call functionalism deal with discourse, i.e. the intersentential domain.

Functionalism rejects the Chomskyan idea that grammar is logically and psychologically the origin of "functions in language" (Halliday above). For functionalists like Halliday, the grammar of a specific language is merely "the linguistic device for hooking up the selections in meaning which are derived from the various functions of language".[67]

In functionalism it is communication that is claimed to be logically and psychologically prior to grammar. Givon, for whom the supreme function of language is communication, criticises Chomsky for trying to describe language without referring to its communicative function.[68]

---

65 Halliday, M.A.K. *Learning how to mean*, 1975, p.5.
66 Saussure, F. de. *Course in general linguistics*. 1916 [1974]).
67 Halliday, M.A.K. *Learning how to mean*, 1975, p.2.
68 Givon, T. *Understanding grammar*, 1979, pp.5 and 22.

Givon argues: "If language is an instrument of communication, then it is bizarre to try and understand its structure without reference to communicative setting and communicative function."[69] Rutherford, whose view is similar to Givon's communicative view, rejects the "mechanistic" view that grammatical structure (Givon's "syntax") is logically or psychologically prior to communication.[70] Rutherford sees language as a dynamic process and not as a static "accumulation of entities".[71]

In this regard Spolsky suggests that the "microlevel" is "in essence" the "working level of language, for items are added one at a time", keeping in mind that "any new item added may lead to a reorganisation of the existing system, and that items learnt contribute in crucial, but difficult to define ways to the development of functional and general proficiency."[72] Thus, according to Spolsky, building up the language from the microlevel to the macrolevel need not be a *static* "accumulation of entities" (Rutherford above), but may lead to a dynamic "reorganisation of the existing system" (Spolsky above). Alderson's view is similar to Spolsky's:

> *Another charge levelled against (unidentified) traditional testing is that it views learning as a 'process of accretion'. Now, if this were true, one would probably wish to condemn such an aberration, but is it? Does it follow from an atomistic approach to language that one views the* ***process*** *of language as an accretion? This does not necessarily follow from the notion that the* ***product*** *of language learning is a series of items (among other things). (Original emphasis).*[73]

The process and product methodologies "are too often perceived as generally separate", i.e. they suffer from an "oppositional fallacy".[74] The product is considered to be discrete, static

---

69 Ibid., p.31.

70 Rutherford, W.E. *Second language grammar: Learning and teaching*, 1987, pp.1-5

71 Ibid., pp.4 and 36-37.

72 Spolsky, B. *Conditions for second language learning*, 1989, p.61.

73 Alderson, J.C. 'Reaction to the Morrow paper, in Alderson, J.C. and Hughes, A. *Issues in language testing: ELT Documents III*, 1981c, p.47.

74 Besner, N. 'Process against product: A real opposition?' *English Quarterly*, 18 (3), 9-16 (1985), p.9.

and, accordingly, is not party to language *processing*, while the process is considered to be integrative and dynamic, and accordingly, the process is seen as belonging to language processing. (More about this in section 6.4). It is this oppositional fallacy that is the battleground of the discrete-point/integrative controversy.

It is widely believed that tests such as essay tests test the "use" of language, i.e. authentic communicative language, while tests such as error recognition tests and grammar accuracy tests test the "usage" of language, i.e. the elements of language.[75] Such a distinction between the two kinds of tests, which Farhady describes as the "disjunctive fallacy"[76], is an oversimplification. Many studies report high correlations between "discrete-point tests" and "integrative tests".[77] It may be asked how the construct is able to account for this: "Shouldn't supposedly similar types of tests relate more to each other than to supposedly different types of tests?" An adequate response presupposes three further questions: (1) "What are similar/different *types* of tests?" (2) Wouldn't it be more correct to speak of *so-called* discrete-point tests and *so-called* integrative tests? (3) Isn't the discrete/integrative dichotomy irrelevant to what any test is measuring?

Let us examine some of the issues in the discrete-point/integrative controversy that are related to the questions posed above. The notion of "real-life" tests is also critically examined.

The terms "integrative" and "discrete-point" have fallen out of favour with some applied linguists, while for others these terms are still in vogue. For example, Fotos equates

---

75 Widdowson, H.G. *Explorations in applied linguistics*, 1979.

76 Farhady, H. 'The disjunctive fallacy between discrete-point tests and integrative tests', in Oller, J.W. (Jr.). *Issues in language testing research*, 1983.

77 (1) Hale, G.A., Stansfield, C.W. and Duran, R.P. *TESOL Research Report 16*, 1984.

(2) Henning, G.A., Ghawaby, S.M., Saadalla, W.Z., El-Rifai, M.A., Hannallah, R.K. and Mattar, M. S. 'Comprehensive assessment of language proficiency and achievement among learners of English as a foreign language.' *TESOL Quarterly*, 15 (4), 457-466 (1981).

(3) Oller, J.W., Jr. *Language tests at school*, 1979.

(4) Oller, J.W. (Jr.) and Perkins, K. (eds.). *Language in education: testing the tests*, 1978.

"integrative" skills with "advanced skills and global proficiency"[78], which he contrasts with Alderson's "basic skills".[79] These "basic skills" are Alderson's "low order" skills.[80] Alderson prefers to distinguish between "low order" and "higher order" tests than between "discrete-point" and "integrative" tests.[81] Alderson in 1979 refrained from talking about discrete-point and integrative tests, but preferred to talk of "low order" and "higher order" tests.[82] Yet, in his later collaborative textbook on testing, one of the book's test specifications is that "tasks" should be "discrete point, integrative, simulated 'authentic', objectively assessable".[83] These tests specifications would dovetail with the notion that although these tests do not mirror life, they are nevertheless "good dirty methods [of testing] overall proficiency".[84] Whatever one's classification, all tests, except for the most atomistic of tests, reside along a continuum of "integrativeness".[85] For example, consider two items from Rea[86]:

1. How ---- milk have you got?
(a) a lot (b) much of (c) much (d) many
2. ---- to Tanzania in April, but I'm not sure.
(a) I'll come (b) I'm coming (c) I'm going to come (d) I may come.

Item 1 is testing a discrete element of grammar. All that is required is an understanding of the "collocational constraints of well-formedness"[87], i.e. to answer the question it is sufficient

---

78 Fotos, S. 'The cloze test as an integrative measure of EFL proficiency: A substitute for essays on college entrance examinations.' *Language Learning*, 41 (3), 313-336 (1991), p.318.

79 Alderson, J.C. 'The cloze procedure and proficiency in English as a foreign language.' *TESOL Quarterly*, 13, 219-227 (1979).

80 Ibid.

81 Ibid.

82 Ibid.

83 Alderson, J.C., Clapham, C. and Wall, D. *Language test construction and evaluation*, 1995.

84 Bonheim, H. *Roundtable on language testing*. European Society of the Study of English (ESSE) conference, Debrecen, Hungary, September, 1997.

85 Oller, J.W., Jr. 'A consensus for the 80s', in *Issues in language testing research* 1983, p.137.

86 Rea, P. 'Language testing and the communicative language teaching curriculum', in Lee, Y.P. et al. *New directions in language testing*, 1985, p.22.

87 Ibid.

to know that "milk" is a mass noun (see also Canale and Swain[88]). Item 2 relates form to global meaning. Therefore, all parts of the sentence must be taken into account, which makes it an integrative task. To use Rea's terminology, her item 1 is testing "non-communicative performance", while her item 2 (above) is testing what she calls "communicative performance".[89] Other discrete-point, or "low order", items could be shown to be more integrative, or "higher order", than the items described above. (The terms in inverted commas are Alderson's [1979[90]]).

Above I described an "objective" type of test. Consider now a test that tends toward the "pragmatic" (Oller's use of the term) extreme of the integrative continuum: the cloze test. ("Pragmatic" for many other researchers would mean for more than the language behaviour demonstrated by a cloze test; "pragmatic" language would be full blown "real-life" language. For such researchers a cloze test would probably be linked with Halliday's "functions in structure" mentioned earlier).

Although cloze answers are short, usually a single word, the cloze test can still be regarded as an "integrative" test. A distinction needs to be made between (1) integrative and discrete *formats* and (2) integrative and discrete *processing strategies*. The salient issue in a cloze test or in any test is not the length of the answer or the length of the question, but whether the test measures what it is supposed to measure, in this case integrative processing strategies. One should distinguish between the structure of the test - long answer, short answer, multiple choice - and what one is measuring. One is measuring the natural ability to process language and one component of this ability is the behaviour of supplying missing linguistic data in a discourse.

---

88 Canale, M. and Swain, M. 'Theoretical bases of communicative approaches to second language teaching and testing.' *Applied Linguistics*, 1 (1), 1-47 (1980), p.35.

89 Ibid.

90 Alderson, J.C. 'The cloze procedure and proficiency in English as a foreign language.' *TESOL Quarterly*, 13, 219-227 (1979).

According to the "pop"[91] view, it is only in language use that natural language processing can take place. Although the "pop" view may not conflict with the idea of a continuum of integrativeness, such a view would nevertheless hold that language tests should only test language "use", i.e. direct language, or authentic language. For language "naturalists" the only authentic tests are those presented in a direct real-life situation. Spolsky maintains that "authenticity of task is generally submerged by the greater attention given to psychometric criteria of validity and reliability, where face validity receives no more than "lip service".[92] For Spolsky and others[93], authenticity is closely related to communicative language, i.e. to direct language. Authentic tests for Spolsky would be "direct" tests in contradistinction to "indirect" tests. Owing to the lack of clarity on the relationship between a global skill like composition writing and the components of composition writing, e.g. vocabulary, punctuation and grammar, Hughes recommends that it is best, in terms of one's present knowledge, to try and be as comprehensive as possible, and the best way to do this would be to use direct tests. "Direct" testers argue that in language use we do not process language in a multiple choice way as in the case of discrete-point tests. Yet, many multiple choice tests do test processing strategies, e.g. the making of predictions. Furthermore, multiple choice tests are neutral in themselves, i.e. they can serve any purpose; communicative or non-communicative. Rea gives the following reasons why indirect tests should be used[94]:

1. There is no such thing as a pure direct test.

2. Direct tests are too expensive and involve too much administration.

3. Direct tests only sample a restricted portion of the language, which makes valid inferences difficult. (Of course, no battery of tests can sample the whole language. Rea's point seems to be that indirect tests are able to be much more representative than direct tests).

---

91 Stevenson, D.K.'Pop validity and performance testing', in Lee, Y., Fok, A., Lord, R. and Low, G. (eds.). *New directions in language testing*, 1985.

92 Spolsky, B. 'The limits of authenticity in language testing.' *Language Testing*, 2, 31-40 (1985), p.33-34.

93 For example, Hughes, A. *Testing for language teachers*, 1989, p.15.

94 Rea, P. 'Language testing and the communicative language teaching curriculum', in Lee, Y.P. et al. *New directions in language testing*, 1985.

If it could be shown that indirect test performance is a valid predictor of direct performance, this would be the best reason for using indirect tests. Even if indirect performance is accepted to be a valid predictor of direct "natural" performance, one may object that indirect tests are unnatural, and consequently lack face validity. But, as previously mentioned, the laws of learning and testing apply to all contexts, "naturalistic"[95] and otherwise. One can have authentic indirect tests, because tests are authentic activity types in their own right.[96] The quality of learning outcomes depends, of course, on the quality of input - and more importantly on the quality of intake.

There is a sense, though, in which "real-life" "authentic" tasks in the *classroom*, if not a contradiction in terms, are not possible: in the sense that learners are aware that life in the classroom is a preparation for, and simulation of, life outside the classroom, which comprises not only life skills but content knowledge in specific disciplines and an understanding of their relationship. But this "preparation for life" view of the classroom, does not justify, I suggest, the radical rupture between "real-life" and the classroom, described by Lantolf and Frawley (see end of section 2.4). Tritely, life is one big classroom; and less tritely, the classroom is one small part of life. This does not mean that the classroom has to be turned into a cooking club or a cocktail lounge to get learners to respond authentically to a recipe or to something "stronger" - for instance, a test.

If by some good fortune we come of age in our understanding of what an "authentic" task is (and, accordingly, isn't), it still doesn't follow that it is necessary to do "authentic" tasks in order to prove that we are proficient to do them, because communicative tasks can be tested successfully through indirect tests.[97] For example, an eye test doesn't directly, or

---

95 Omaggio, A.C. *Teaching language in context: Proficiency-orientated instruction*, 1986, p.312-313.

96 Alderson, J.C. 'Who needs jam?', in Hughes, A. and Porter, D. *Current developments in language testing*, 1983, p.89.

97 Politzer, R.L. and McGroarty, M. 'A discrete-point test of communicative competence.' *International Review of Applied Linguistics*, 21 (3), 179-191 (1983).

"holistically", measure whether someone can see the illuminated road clearly, but its a jolly good predictor of whether one will be able to see, if not avoid, that oncoming road-hog on that same illuminated road.

In sum, both direct tests and indirect tests - as in all direct and indirect classroom activities - have communicative, or real-life, language as their aim. The difference lies in this: direct tests, or outcomes, or activities are based on the view that communicative language should be directly taught and tested, while indirect tests are based on the view that indirect teaching materials and tests are a prerequisite and solid basis for ultimate real-life language. But I would go even further and agree with Widdowson that "semantic meaning is primary" (Chomsky's dated! "linguistic competence") where semantic meaning should (naturally, i.e. obviously) be internalised to provide for "communicative capacity"[98]; which is the same idea as Spolsky's (mentioned above) where the building up of language from the microlevel to the macrolevel may be a dynamic and not necessarily a static "accumulation of entities" (Rutherford[99] above), which in turn leads to a dynamic "reorganisation of the existing system".[100]

This raises the contentious issue of separating "semantics" from "pragmatics".[101] From the point of view of the ideational (or conceptualising) function of language, which is what most of language processing is concerned with, or should be concerned with, much more demands are made on semantic and syntactic encoding than the communicative act itself, which, after all, is only the last stage of language in action - unless one speaks before one thinks.[102]

---

98 Widdowson, H.G. 'Skills, abilities, and contexts of reality.' *Annual Review of Applied Linguistics*, 18, 323-333 (1998), p.329.

99 Rutherford, W.E. *Second language grammar: Learning and teaching*, 1987.

100 Spolsky, B. *Conditions for second language learning*, 1989, p.61.

101 Hudson, R. *Word grammar*, 1984.

102 Widdowson, H.G. 'Skills, abilities, and contexts of reality.' *Annual Review of Applied Linguistics*, 18, 323-333 (1998), p.330.

2.6 Cognitive and Academic Language Proficiency and "test language"

Cognitive and Academic Language Proficiency (CALP) is closely related to the ability to do tests. Its features are better understood when compared and contrasted with Basic Interpersonal and Communicative Skills (BICS).[103] BICS refers to salient basic features such as fluency (speed of delivery) and accent, and not to advanced social and communicative skills, which are cognitively demanding skills. For example, the skills of persuading or negotiating in face-to-face communication require relatively much more cognitive involvement than a BICS task, and are therefore cognitively demanding CALP tasks. Thus, it would be incorrect to equate BICS with all face-to-face communication, because face-to-face communication may involve informal as well as formal speaking. Formal speech acts such as persuading and negotiating belong to advanced communicative skills, and are consequently part of CALP. Spoken language can be just as complex as written language. They differ in that speech is dynamic while writing is synoptic, and writing is lexically denser than speech: "written language does not have to be immediately apprehended in flight and does not need to be designed to counter the limitations of processing capacity".[104]

Cummins' BICS and CALP have affinities with Bernstein's "restricted code" and "elaborated code", respectively.[105] The "elaborated code" has the following features: precise verbalisations, large vocabulary, complex syntax, unpredictability, low redundancy, individual differences between speakers. In contrast, the "restricted code" has the following features: loose verbalisations, limited vocabulary, simple syntax, high redundancy where assumptions are based on shared social experience.

---

103 Cummins, J. 'The cross-lingual dimensions of language proficiency: Implications for bilingual education and the optimal age issue.' *TESOL Quarterly*, 14 (2), 175-87 (1980).
_____ 'Language proficiency and academic achievement', in Oller, J.W. (Jr.), (ed.). *Issues in language testing research*, 1983.
_____ 'Wanted: A theoretical framework for relating language proficiency to academic achievement among bilingual students', in Rivera, C. (ed.). *Language proficiency and academic achievement*, 1984.

104 Widdowson, H.G. 'Skills, abilities, and contexts of reality.' *Annual Review of Applied Linguistics*, 18, 323-333 (1998), p.326.

105 Bernstein, B. *Class, codes and control*, 1971.

Wald makes a distinction between "test language" (spoken and written), which he equates with CALP, and "spontaneous language"/"face-to-face" communication.[106] For Wald, test skills are CALP skills, which can involve all four language modes: listening, speaking, reading and writing. For example, an oral cloze test would be a CALP task. In terms of these distinctions, it would be possible to have tests of *basic* language (grammar tests). Basic language tests involve CALP because, according to Wald, they are tests. All tests are *formal*, no matter how "natural" one tries to make them. In terms of Wald's definition of CALP as "test language", the tests in this study are CALP tasks because they are tests. Accordingly, if Wald is correct, and I think he is, one could not have a BICS test.

Ur uses the term "informal" not in the sense of *natural*, but in the sense that test takers are not told in advance what they need to know for a test.[107] One could, for example, spontaneously test learners on their homework. On such an interpretation of "informal", it follows that there can be "informal" tests (i.e. "informal" CALP tasks).

2.7 Language proficiency and academic achievement

Much of the research in second language acquisition involves finding factors which affect language proficiency. In such a research scheme, factors such as intelligence, motivation, mother-tongue interference and socio-economic standing are defined as the independent variables and language proficiency is defined as the dependent variable. Language proficiency in such a research context does not look beyond itself to its effect on academic achievement.

In the investigation of academic achievement the focus changes from considering language proficiency as the dependent variable (the criterion) to considering academic achievement as

---

106 Wald, B. 'A sociolinguistic perspective on Cummins' current framework for relating language proficiency to academic achievement', in Rivera, C. *Language proficiency and academic achievement*. 1984, p.57

107 Ur, P. *A course in language teaching: practice and theory*, 1996.

the dependent variable, as in Saville-Troike.[108] Consider the following schema. (The schema is highly simplified and is thus not a comprehensive "model"):

Table 2.2

Language Proficiency as a Criterion Variable and as a Predictor Variable

FIRST FOCUS - Language proficiency as the criterion

Predictor variables	**Criterion variable**
Intelligence	
Motivation (active participation)	*Language proficiency*
Mother-tongue interference	
Socio-economic standing	
Personality (e.g. emotional maturity)	

SECOND FOCUS - Academic achievement as the criterion

Predictor variables	**Criterion variable**
Intelligence	
Motivation	Academic achievement
Mother-tongue interference	
Socio-economic standing	
Language proficiency	
Subject learning	

One needs to know how language proficiency, which is embedded in other factors, promotes or hinders academic achievement. These other factors comprise a complex network of variables such as intelligence, learning processes and styles, organisational skills and content knowledge, teaching methods, motivation and cultural factors. Owing to the complexity of

[108] Saville-Troike, M. 'What really matters in second language learning for academic achievement.' *TESOL Quarterly*, 18 (2), 199-219 (1984), p.199.

the interaction between these variables, it is often difficult, perhaps impossible, to isolate them from language proficiency, which means that any one or any combination of these above-mentioned variables might be the cause of academic failure. Therefore, care must be taken not to make spurious causal links between any of these variables and academic failure. Although prediction does not necessarily imply causation, this does not mean that prediction should be ignored; on the contrary, prediction plays a very important role in the selection and placement of candidates. What is important is that these predictions be valid. (I stress that this study focuses mainly on those causes of academic failure that are related to the testing situation, e.g. rater (un)reliability and marks inflation).

Although the distinction between the two kinds of focus (see Table 2.2) can be useful, the change from one focus to the other does not merely involve rejuggling the variables that were previously used to predict language proficiency (the first focus) and then assigning them to the new game of predicting academic achievement (the second focus), where language proficiency, previously a criterion variable, would then become another predictor variable among those that were previously used to predict it. The reason is that when academic achievement is brought into the foreground, the predictive mechanism becomes far more complex. One cannot merely shift variables around, because in the second focus, learning in or through a second language is added to the demands of learning the second language itself.

Upshur distinguishes between two distinct general questions: "Does somebody have proficiency?" and "Is somebody proficient?"[109] The first question considers such issues as grammatical competence and the use of language in discourse. The second question is concerned with the ability of language proficiency tests to predict future performance in tasks that require language skills, i.e. with the "prerequisites for a particular job or course of

---

[109] Upshur, J.A. 'English language tests and predictions of academic success', in Wigglesworth, D.C. (ed.). *Selected conference papers of the Association of Teachers of English as a Second Language*. Los Altos, California, National Association for foreign Student Affairs (NAFSA) Studies and Papers, English Language Series 13, 85-93 (1967), p.85.

study".[110] It is this second question, namely, "Is somebody proficient?" (to do a particular task) that has a direct bearing on academic achievement.

2.8 Validity

Validity is concerned with "the purposes of a test"[111], which are basically concerned with the meaning of scores and the ways they are used to make decisions. A major difficulty in this regard is ensuring that one's descriptions of validity are validly constituted, which involves reconciling "objective" reality with one's own interpretation of "objective" reality - a daunting and probably circular task.

For some researchers, validity comprises face validity, content validity, construct validity and criterion validity (concurrent and predictive validity), whereas for others, especially those belonging to the American Psychological Association[112] (APA), construct validity itself is validity.

2.8.1 Face validity

Face validity is concerned with what people (which includes test analysts and lay people) believe must be done in a test, i.e. what the test looks like it is supposed to be doing.

For Clark, face validity, oddly, covers the "whole business" of tests, i.e. looking "at what it's got in it, at the way it is administered, at the way it's scored." [113] Clark's definition is unusual, because it covers everything to do with testing. Clark's meaning of face validity is not what a

---

110 Valette, R.I.. *Modern language testing: A handbook*, 1969, p.5.

111 Carmines, G. and Zeller, A. *Reliability and validity assessment*, 1979, p.15.

112 American Psychological Association.*Standards of educational and psychological measurement*. 1974.

113 Clark, J.L.D. *Theoretical and technical considerations in oral proficiency testing*, 1975, p.28.

test looks like to the non-tester but what it is to the tester, who should know what it is, i.e. "what it's got in it", and not only what it looks like.

Spolsky's meaning of face validity has affinities with Clarke's. Spolsky equates face validity with "authenticity": "authenticity of task is generally submerged by the greater attention given to psychometric criteria of validity and reliability", where "face validity receives no more than lip service".[114]

For Davies "face validity is desirable but not necessary, cosmetic but useful because it helps convince the public that the test is valid."[115] The reason why face validity is desirable, according to Davies, is that, in spite of its "cosmetic" nature, it can still have a "major and creative influence for change and development"[116]. Yeld maintains that face validity should be capitalised on as a point of entry into testing for those "who have not been trained in the use of techniques of statistical analysis and are suspicious of what they perceive as 'number-crunching'."[117]

Thus, face validity (what Stevenson calls "pop" validity) is so popular today because many language teachers have a poor knowledge of language testing and educational measurement, i.e. they are "metrically naive".[118] Accordingly, they could remain satisfied with superficial impressions.

There are others who reject face validity altogether, because it relies too much on the subjective judgement of the observer[119]:

---

114 Spolsky, B. 'The limits of authenticity in language testing.' *Language Testing*, 2, 31-40 (1985), p.33-34.

115 Davies, A. *Principles of language testing*, 1990), p.44.

116 Ibid., p.7.

117 Yeld, N. 'Communicative language testing and validity.' *Journal of Language Teaching*, 21 (3), 69-82 (1987), p.78.

118 Stevenson, D.K.'Pop validity and performance testing', in Lee, Y., Fok, A., Lord, R. and Low, G. (eds.). *New directions in language testing*, 1985, p. 112.

119 (1) American Psychological Association. 1974. *Standards of educational and psychological measurement*, 1974.

> *Adopting a test just because it appears reasonable is bad practice; many a 'good-looking' test has failed as a predictor... If one must choose between a test with 'face validity' and no technically verified validity and one with technical validity and no appeal to the layman, he had better choose the latter.*[120]

Gardner and Tremblay consider face validity to be the lowest form of validity, and should, accordingly, not be generally recommended as a research strategy.[121] The difficulty with face validity in its usual connotation of what a test appears to be is that the prettier the package the worse may be the inherent quality of the tests. No matter what one's opinions of face validity, it does have the following useful features: it increases a learner's motivation to study for the test; it keeps sponsors happy; and it sustains the parents' resolve to pay the ever-escalating school fees.

2.8.2 Content validity

Face validity and content validity can overlap, because what must be done in a test involves content. The latter subsumes subject matter as well as skills. Content validity "implies a rational strategy whereby a particular behavioural domain of interest is identified, usually by reference to curriculum objectives or task requirements or job characteristics".[122] Content validity is concerned with how test items represent the content of a syllabus or the content of real-life situations. Content validity is not only a match between (the situation, topic and style of) tests and real-life situations but also a match between tests and school life, both of which are part of "real" life.

---

(2) Cronbach, L.J. *Essentials of psychological testing*, 1970.

(3) Gardner, R.C. and Tremblay, P.F. 'On motivation: measurement and conceptual considerations.' *The Modern Language Journal*, 78 (4), 524-527 (1994).

(4) Stevenson, D.K.'Pop validity and performance testing', in Lee, Y., Fok, A., Lord, R. and Low, G. (eds.). *New directions in language testing*, 1985.

120 Cronbach, ibid., p.183.

121 Gardner and Tremblay, ibid, p.525.

122 Messick, S. *Constructs and their vicissitudes in educational and psychological measurement*, 1989a, p.1.

2.8.3 Construct validity

The constructs, or human abilities, that this study is interested in belong to the domain of language acquisition. As I mentioned earlier (section 2.2), abilities are fixed attributes, or constructs (in the sense of consistent, not immutable). If behaviour is inconsistent it would be impossible to find out what lies behind the behaviour, i.e. discover the construct. The problem for scientists, whether physical scientists or linguistic scientists, is figuring out the nature and sequence of the contribution of (abstract) theory and (concrete) experience to construct validity.

Consider how evidence for construct validity is assembled. There are two main stages: (1) hypothesise a construct and (2) construct a method that involves collecting empirical data to test the hypothesis, i.e. develop a test to measure the construct. Hypothesising is concerned with theory, while the construction of a method, although inseparable from theory, is largely an empirical issue. The problem is that it is not clear whether theory should be the cart and experience the horse, or vice versa, or some other permutation. Consider some of the problems in assessing the relative contribution of theory and experience in construct validation:

For Messick construct validity is a unitary concept that subsumes other kinds of (sub-)validities, e.g. content validity and criterion validity.[123] Messick defines validity, which for him is construct validity, as a "unitary concept that describes an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the **adequacy** and **appropriateness** of **inferences** and **actions** based on test scores or other modes of assessment[124] (original emphasis).

---

123 Messick, S. *Validity*, 1989b, p.1-2

124 Messick, S. *Meaning and values in test validation: The science and ethics of measurement*, 1988, p.2.

"Test scores" above refers to "quantitative summaries"[125], which is the commonly understood meaning of the term. Messick's longer and much denser definition of construct validity

> *implies a joint convergent and discriminant strategy entailing both substantive coverage and response consistency in tandem. The boundaries and facets of a behavioural domain are specified, but in this case as delineated by a theory of the construct in question. Items are written to cover that domain in some representative fashion, as with content validity, but in this approach the initial item pool is deliberately expanded to include items relevant to competing theories of the construct, if possible, as well as items theoretically irrelevant to the construct... Item responses are then obtained, and items are selected that exhibit response homogeneity consistent with the focal theory of the construct but are theoretically distinct from theoretically irrelevant items or exemplars of competing theories.* [126]

What Messick's definition loses in brevity and simplicity it gains in scientific rigour. In Messick's definition one should have a theory and only then a method. And the theory must be able to specify the problem without prejudging a solution: something very difficult to do. We try and ensure a "substantive coverage" (Messick above) of entities or qualities that are similar (a convergent strategy) and of those that are different (a discriminant strategy), which is "delineated by a theory of the construct in question" (Messick above). The problem is knowing what items to include or exclude in a test, because owing to the infinity of the corpus and the fact that elements and skills hang together, it is difficult to distinguish between "items relevant to competing theories of the construct, *if possible*" and "items theoretically irrelevant to the construct"? (Messick above; italics added).

The Unitary Competence controversy is concerned with the nature and degree of interdependence between elements and skills, i.e. how, how much and which elements and skills hang together. Difficulties exist in the discrimination between items. One of these difficulties is distinguishing between low order (so called "discrete") items and higher order

---

125 Messick, S. *Validity*, 1987. p.3.

126 Messick, S. *Constructs and their vicissitudes in educational and psychological measurement*, 1989a, p.1.

(so called "integrative") items, as was shown in the discussion of the discrete-point/ integrative controversy (section 2.5).

The group-differences approach to construct validity that is used in this study is now explained: The aim of testing is to discern levels of ability. If one uses academic writing ability as an example of a construct, one would hypothesise that people with a high level of this ability would have a good command of sentence structure, cohesion and coherence, while people with a low level of this ability would have a poor command of these. Tests are then administered and if it is found that there is a significant difference between a group of high achievers and a group of low achievers, this would be valid evidence for the existence of the construct. Second-language learners are often relatively less competent than first-language or mother-tongue users.[127]

Important for the arguments presented and the validation of the sample of subjects is that those who take English First Language as a subject are generally more competent than those who take English Second Language as a subject. If a test fails to discriminate between low-ability and high-ability learners there are three possible reasons for this:

- The construction of the test is faulty, e.g. the test may be too easy or too difficult for all or most of the test takers participating in the test.
- The theory undergirding the construct is faulty.
- The test has been inaccurately administered and/or scored, which would decrease the reliability, and hence also the validity of the test. (Reliability is discussed shortly).

2.8.4 Criterion validity: concurrent and predictive validity

Criterion validity is concerned with correlating one test against an external criterion such as another test or non-test behaviour. Ebel maintains that "unless a measure is related to other

---

127 There are many second-language users who have a far better command of academic discourse than mother-tongue users. This is so because the ability to understand and produce academic discourse depends on much more than "linguistic ability": it also depends on academic ability. This point needs to be continually recalled to mind.

measures, it is scientifically and operationally sterile."[128] Criterion validity should not be confused with criterion-referenced tests. Criterion-referenced tests deal with profiles, i.e. with setting a predetermined cut-off score for an individual.

Criterion validity, which relies mainly on empirical methods, ignores the theoretical contribution of construct and content validity. For this reason, some researchers, particularly those of the American Psychological Association, prefer to dissociate validity from descriptions of the criterion, e.g. Loevinger[129] and Bachman[130]. Bachman[131] prefers the term "concurrent relatedness" to "concurrent validity", and "predictive utility" to "predictive validity".

A term used by Messick is "criterion-related validity", where the latter "implies an empirical strategy whereby items are selected that significantly discriminate between relevant criterion groups or that maximally predict relevant criterion behaviours".[132] Messick's "criterion-related validity" is the same notion as the simpler term "criterion validity", which was defined in the first sentence of this section.

Criterion validity consists of concurrent validity and predictive validity. Concurrent validity is also concerned with prediction because there is only a chronological difference between concurrent and predictive validity.[133] So, we could distinguish between concurrent prediction and prediction proper, which is concerned with the ability of one test to predict

---

128 Ebel, R.L. 'Must all tests be valid?' *American Psychologist*, 16, 640-647 (1961), p.645.

129 Loevinger, J. 'Objective tests as instruments of psychological theory', in Jackson, D.N. and Messick, S. *Problems in human assessment*, 1967, p.93.

130 Bachman, L.F. *Fundamental considerations in language testing*, 1990b, p.253.

131 Ibid.

132 Messick, S. *Constructs and their vicissitudes in educational and psychological measurement*, 1989a.

133 Cronbach, L.J. *Essentials of psychological testing*, 1970, p.122.

another test where the predictor and the criterion are not given concurrently and thus are separated from each other by a reasonable period of time.

The reason why predictive validity is easier to measure than other kinds of validity is that predictive validity does not depend on the nature of the test items, but on the consistency of the predictions of performance.[134] It would be possible to ignore all the other kinds of validity and still have a high degree of predictive validity. The question is whether one should be satisfied with predictive validity alone. No. That is why this study is also concerned with construct validity.

If the *construct* validity of one test always depends on the validity of another test, there cannot exist any one test that stands by itself such as an equivalent of a "Prime Mover". Lado's solution is to compare all tests in terms of "some other criterion whose validity is self-evident, e.g. the actual use of the language."[135] The question is: What is self-evident? Is there a self-evident test that pre-exists all other tests? There isn't, because "the buttressing validity of an external criterion is often neither definable nor, when found, reliable".[136] This does not mean, of course, that any test or battery of tests, direct or indirect, will do.

Having said that, one doesn't need to worry about the difficulty of establishing construct validity if one is merely interested in predictive validity. If Test A is a good predictor of Test B, then it seems one doesn't need Test C as a second predictor, because Test A is doing a good job on its own. Recall, however, the discussion of the "One Best Test" question: one can never be sure; furthermore, it doesn't look (face validity) fair to use only one test as a predictor. To do so would be regarded by some researchers as highly unethical. Spolsky is a case in point:

---

134 Weir, C.J. *Communicative language testing*, 1988, p.30

135 Lado, R. *Language testing*, 1961, p.324.

136 Davies, A. *Principles of language testing*, 1990, p.3.

> *Only the most elaborate test batteries, with multiple administrations of multiple methods of testing the multiple traits or abilities that make up language proficiency, are capable of producing rich and accurate enough profiles to be used for making critical or fateful decisions about individuals.*[137]

Such Herculean conditions, however, would probably "paralyze"[138] most testing endeavours, because they are impracticable in their unrealisability. We shouldn't only start measuring when we are clear about what we are measuring; rather we should do the best we can; always taking into account generally accepted theories, but not necessarily following them slavishly if we have cogent reasons why we shouldn't.

2.9 Reliability

If the validity of a test depends on its close approximation to real life then validity would relate to subjectivity. We try to be as objective as possible in the test compilation, administration and assessment. This search for objectivity is the domain of reliability. Reliability in testing is concerned with the accuracy and consistency of scoring and of administration procedures. The less the accuracy and consistency, the more the measurement error.

A major difficulty in testing is how to make the "leap from scores to profiles "[139], i.e. how to define the cut-off points. In norm-referenced testing, one defines cut-off points by computing the measurement error. In criterion-referenced tests, one makes a value judgement of what is progress enough for a specific individual.

To the extent that one can decrease the measurement error, one increases the reliability of the test. Measurement error has important ethical implications. It would be unjust to fail students because they get 49% - perhaps even 47%, where does one draw the line? - instead

---

137 Spolsky, B. *Measured words*, 1995, p.358.

138 Ibid.

139 Yeld, N. *Communicative language testing.* Report on British Council Course 559 offered at Lancaster University from 8 September to 20 September 1985, (published in) 1986, p.31.

of 50%. In subjective tests such as essay tests the problem is more serious, because even the best essay test, owing to its subjective scoring procedures, is often not more than 80%-90% reliable, and therefore measurement error should be calculated in order to make more equitable judgements. According to Perkins, "raters, guided by...holistic scoring guides..., can achieve a scoring reliability as high as .90 for individual writers."[140] Indirect objective tests such as multiple choice grammar and vocabulary tests, on the other hand, can have reliability coefficients as high as .99, because there is no problem of rater reliability involved, i.e. subjective judgements will not affect the scores.[141]

The following "aspects" are germane to reliability:

- Facets: These refer to such factors as the (1) testing environment, e.g. the time of testing and the test setting, (2) test organisation, e.g. the sequence in which different questions are presented, and (3) the relative importance of different questions and topic content. Facets also include cultural and sexual differences between test takers, the attitude of the test taker, and whether the tester does such things as point out the importance of the test for a test taker's future.[142]

- Features or conditions: These refer to such factors as clear instructions, unambiguous questions and items that do or do not permit guessing.

- The manner in which the test is scored. A central factor in this regard is rater consistency. Rater consistency becomes a problem mainly in the kind of tests that involve subjective judgements such as essay tests. (I discuss interrater and intrarater consistency in the next section). According to Ebel and Frisbie, consistency is not only concerned with the correlations between raters, but also with the actual scores, i.e. the equivalence between raters.[143]

---

140 Perkins, K. 'On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability', *TESOL Quarterly*, 17 (4), 651-671 (1983). p.655.

141 Hughes, A. *Testing for language teachers*, 1989, p.29.

142 Bachman, L.F. *Fundamental considerations in language testing*, 1990b, pp. 116ff, 168-172, 244.

143 Ebel, R.L. and Frisbie, D.A. *Essentials of educational measurement*, 1991, p.76.

I would like to clarify a possible confusion between rater reliability and concurrent validity: rater reliability has to do with the consistency between raters' judgements on one test, e.g. an essay test. Concurrent validity, in contrast, has to do with the correlation between two or more different tests e.g. a dictation test and an essay test. In the next section more details are provided on the approaches to reliability.

2.9.1 Approaches to the measurement of reliability

There are five approaches to measuring reliability. Owing to the structure of this study, only approaches 2, 4 and 5 are used:

1. Stability, i.e. consistency over time. The method used to measure stability is the test-retest method, which involves giving the same test a second time and comparing the scores of the two test trials. If the scores are equivalent, the test is considered to be stable. A disadvantage of the test-retest method is that students may not be motivated to do the test a second time, which might affect performance on the retest.

2. Internal consistency. This approach, also called the "split-half" test, divides the test into two halves. The two halves of the test are regarded as two parallel tests. For each student there is a separate score for each half of the test. It is possible to correlate the two sets of scores as if they were parallel tests. This approach is usually used with discrete-point tests.

3. Rater reliability. Rater reliability is particularly important in non-objective tests such as essay tests, where there are liable to be fluctuations in scores between (1) different raters, which is the concern of interrater reliability, and (2) within the same rater, which is the concern of intrarater reliability. In this study I use essay assessment to examine interrater reliability (section 4.8.1).

4. Equivalence (in the form of the test). There are two meanings of equivalence: firstly, the equivalence between test scores, and secondly, between the facets of the tests. The method used to measure equivalence is the parallel test method, which is used for

integrative tests. In parallel tests it is difficult to ensure equivalent, i.e. parallel, conditions within the many facets of a test, especially with regard to the content. (Parallel reliability is discussed further in section 4.2).

5. A combination of stability and equivalence (in forms). The method used is a parallel test which is administered a period of time after the first test. The difficulties are compounded here, because they include the problems of both equivalence and of stability mentioned above.

The degree of reliability required depends on the relative importance of the decisions to be made. For example, an admission test would require more reliability than a placement test, because decisions based on a placement test can be more easily adjusted than decisions based on an admission test. A final evaluation for promotion purposes would require the most reliability of all.

2.10 Ethics of measurement

Validity should not be separated from what Sammonds refers to as the following "ethical" questions. (Most of these ethical questions are scientific questions as well[144]: (The kinds of validity corresponding to each question are the appellations given by Sammonds).

1. Are the measures that are chosen to represent the underlying concepts appropriate? ("Construct validity").

2. Has measurement error been taken into account, because in all measurement there is always a degree of error? ("Statistical conclusion validity"; in other words, reliability).

3. Are there other variables that need to be taken into account? ("Internal validity").

4. Are the statistical procedures explained in such a way that a non- statistical person can understand them? Or sometimes even a statistical person. One reader may find an explanation superfluous or too detailed, while another may find the same information patchy. Much depends on the background knowledge of readers and/or what they are looking for. Pinker maintains that expository writing requires writers to overcome their

144 Sammonds, P. 'Ethical issues and statistical work', in Burgess, R.G. (ed.). *The ethics of educational research*, 1989, p.53.

"natural egocentricism" where "trying to anticipate the knowledge state of the generic reader at every stage of the exposition is one of the important tasks of writing well."[145] True, but there is much more to the issue, namely, the basic expository problem of negotiating a path between under-information and overkill. Getting experts to read and provide comments before one submits one's work to public scrutiny is one way of reducing the expository problem. But this can also be fraught with problems.

4. Has the description of the sample and the data analysis been properly done so that generalisations can be made from it? ("External validity"). This issue is dealt with in section 5.6.

2.11 Summary of Chapter 2

The first part of the chapter dealt with theoretical issues in language proficiency, language learning, language testing and academic achievement. Key concepts such as authenticity, competence, performance, ability, proficiency, test language, integrative continuum and achievement were explained. The second part of the chapter was concerned with explaining the key concepts in summative assessment. The two principal concepts in summative assessment are validity and reliability. Different kinds of validity were discussed, namely, content validity, face validity, construct validity and criterion validity, where the latter comprises concurrent and predictive validity. Other kinds of validity were also referred to in the context of the ethics of measurement. The group-differences approach to construct validity, to be used in the study, was described. Different approaches to the examination of reliability were discussed and those chosen for the study were specified.

---

145 Pinker, S. *The language instinct*, 1995, p.401.

CHAPTER 3

Sampling, and Structure and Administration of the English Proficiency Tests

3.1 Introduction

This chapter describes the sample of subjects and provides a detailed description of the structure and administration of the battery of English proficiency tests. A literature review is provided for each of the test methods used in which an overview of the relevant theoretical issues is provided. After each literature review follows a detailed description of the structure and administration of the specific tests used.

The sample of subjects consists of two main groups: First Language (L1) and Second Language (L2). A major distinction in this study is that between L1 and L2 levels of language proficiency. This distinction has become a controversial one in South Africa where an increasing number of applied linguists and educationists argue that the L1/L2 distinction should be jettisoned. (Specific authors on this issue are discussed in Chapter 6). I use the labels L1 and L2 slightly differently from what is normally meant by these labels. Details are provided in the next section.

3.2 Sampling procedures for the selection of subjects

A crucial issue is how - and whether! - to classify the subjects into *distinct groups* that represent different levels of proficiency. There were 90 entrants to Grade 7 in January 1987 who also sat the Grade 7 end-of-year examinations. Owing to the fact that the battery of English proficiency tests was administered during the first week of the school year, there was some absenteeism during the three-day test period. Thus, not all of the learners did all of the

tests, and four learners did not do any of the tests. These four learners were not included in the sample. The other 86 learners (44 boys, 42 girls) comprise the sample of subjects.

3.2.1 The two main groups of the sample: First language (L1) and second language (L2) groups

Tables 3.1 and 3.2 below provide a clear picture of the details of the L1 and L2 groups. The reader may want to consult these figures in conjunction with the verbal descriptions of the sample below.

At the school there were mother-tongue speakers from diverse linguistic backgrounds, e.g. Tswana, Sotho, English, Afrikaans, Gujarati, and some expatriates, e.g. Greek and Filipino. (The exact numbers are provided in Table 3.1). About two thirds were Tswana-mother-tongue speakers. All learners had to take English as the medium of instruction at the School.

The Tswana speakers could choose from the following language subject combinations:

- *Tswana* as a *First* Language and *English* as a *Second* Language. (After 1987 Afrikaans was also offered as a first language. The battery of tests for this study was administered in 1987).
- *Tswana* as a *First* Language and *English* as a *First* Language.
- *English* as a *First* Language and *Afrikaans* as a *Second* Language. Tswana speakers never took this option.

The English and Afrikaans speakers and speakers of other languages (expatriates and those using English as a "replacement" language[2]) could choose from the following language subject combination:

---

[2] A replacement language is a language that becomes more dominant than the mother tongue, usually at an early age, but is seldom fully mastered, as in the case of, for example, some of the

- *English* as a *First* Language and *Afrikaans* as a *Second* Language.

- *English* as a *First* Language and *French* as a *Second* Language. This combination was taken by the expatriates, because they had not studied Afrikaans in primary school as South Africans had done. The "replacement" learners (see Note 2 below) took Afrikaans as a second language.

All the L2 learners were Bantu-mother-tongue speakers, most of whom were Tswana-mother-tongue speakers. The L1 learners were a mixture of English-mother-tongue speakers, Tswana- mother-tongue speakers and mother-tongue speakers of other languages. The latter consisted of (i) expatriates from other countries and (ii) South Africans who spoke other languages such as Afrikaans and Gujarati.[3] It was not always certain who among the L1 group (i.e. those who took English as a First language at MHS) were English-mother- tongue speakers because although some of them identified themselves as English-mother-tongue speakers, there was little doubt that many in (ii) were using English as a "replacement" language. (I substantiate this at the end of section 4.8).

There were also a few subjects who said that they had more than one mother tongue, of which one of these mother tongues was English. Of course, one can have more than one "mother", or "father", or "native", tongue,[4] because the mother tongue or native language is not something transmitted through the placenta. The notion of *native-speaker* is not a simple one.[5] According to Paikeday[6], the "the native speaker is dead!". Indeed, it is difficult to

---

Coloured and Indian subjects in the sample.

[3] Barkhuizen, G. 'Proposal for an independent English Second Language Department at Mmabatho High School.' *English Language Teaching Centre (ELTIC) Reporter*, 16 (1), 25-32, 1991, p.25.

[4] Paikeday, T.M. *The native speaker is dead!*, 1985, p.5.

[5] (1) Davidson, F. 'Norms appropriacy of achievement tests: Spanish-speaking children and English children's norms.' *Language Testing*, 11:83-95 (1994).
(2) Davies, A. *The native speaker in applied linguistics*, 1991.
(3) Paikeday, T.M. *The native speaker is dead!*, 1985.

[6] Ibid.

- The School did not wish to force the labels of "English second language" (L2) onto learners.

- Limited English proficiency learners might benefit in a class of high English proficiency learners, because the former might benefit from listening to a higher standard of English than their own.

- The School might not have been sure of the actual level of English proficiency of each individual entrant, in spite of the fact that it was aware that the level of English proficiency of disadvantaged entrants was generally low, but once these entrants had become part of the School, it would have been possible to make more accurate judgements of their English proficiency.

- Finally, the School might have been reluctant to use the results until I had produced solid evidence that these tests were valid predictors of academic achievement.

It was not a simple matter to decide how to classify beyond the fact that some took English as a First Language subject (called the L1 group) and others took English as a Second Language subject (called the L2 group. The following variables had to be taken into account (the descriptions are specific to the sample):

(1) Some of the L1 group were (or said they were) English-mother-tongue speakers, while others were mother-tongue speakers of Tswana and other languages.

(2) Some had English as the medium of instruction from Grade 1 (Bantu speakers and non-Bantu speakers), while some had English as the medium of instruction from Grade 5 (only Bantu speakers).

(3) All had the freedom to choose at the beginning of Grade 7 whether they wanted to take English First Language or English Second Language.

There wasn't a clear separation between the subjects that would indicate clearly a difference in levels of English proficiency. A common division is mother-tongue speaker/non-mother-tongue speaker, where English-mother-tongue proficiency is regarded as the level of

English to aspire to. When the essays for this study were marked they were judged in terms of mother-tongue proficiency, and so non-English-mother-tongue speakers' essays were not marked more leniently than those of English-mother-tongue speakers. There were difficulties in deciding on the norms for the other tests, which were all previously standardised published tests, because it is only after the test has been performed on the test-bench that it is possible to decide whether the test is too easy or too difficult.[9] If there are mother-tongue speakers and non-mother-tongue speakers in the same sample, as in this study, one needs to consider whether the norms of the two kinds of speakers should be separated or interlinked. One can only do this if subjects have been precisely classified into mother-tongue/non-mother-tongue groups. This was not a simple matter in the sample. I pursue this issue further:

In the multicultural setting of MHS, a composite of the following cultural-ethnic groups said that they were English-mother-tongue speakers: Ghanaian, Sri Lankan, Indian (South African and expatriate) and Coloured. There was also a Greek, a South Sotho, and a Filipino who said that they had two mother tongues, one of them being English. Although all the above (N=18) obtained an English proficiency test score (in this case a composite of the cloze, dictation and essay test) of 60% and over, there were also quite a number of Tswana-mother-tongue speakers (N=10), who also obtained a score of 60% and over.

What is more, there were five subjects who said that they were English-mother-tongue speakers but obtained scores between 50% and 55%. Such a score is not a good score for somebody claiming to be an English-mother-tongue speaker, because the tests were pitched at the second language level. (More about this in the description of the tests). In the light of these aforementioned observations, it was difficult to tell from the results of the proficiency tests who were native-speakers of English. Although it was difficult to pinpoint native speakers of English this does not mean that the notion of *native-speaker* is a figment. (More

---

[9] Cziko, G.A. 'Improving the psychometric, criterion-referenced, and practical qualities of integrative testing.' *TESOL Quarterly*, 16 (3), 367-379 (1982), p.368.

groups belong to separate populations and therefore cannot be grouped together in a correlational analysis. (See section 4.6 for further discussion of this issue).

Tables 3.1 and 3.2 provide a detailed analysis of the sample. These tables compliment the verbal description of the sample. The subjects originated from 36 different schools:

1. The L1 subjects (N=49) originated from (i) CM Primary School (N=37), (ii) a "white" school (N=1), (iii) a "coloured" school (N=4), and (iv) several DET schools (N=7). (One Sri Lankan came from a DET school where his mother was a teacher).
2. The 37 L2 subjects originated from 28 DET schools (N=34) and three church schools (N=3).

Of the total sample of 86 subjects, there were 60 South African blacks, of which 52 were Tswanas and eight were non-Tswanas. These eight non-Tswana South African blacks (L1:N=4; L2:N=4), as in the case of all the Tswana-speaking learners at MHS, had to take Tswana as a first language.

In Table 3.1, the L1 group is divided into two sub-groups: (1) Non-Tswana L1 (NTL1: N=26) and (2) Tswana L1 (TL1: N=23). In Table 3.2 , the L2 group is also referred to as the Tswana L2 group (TL2) (N=37), because 33 of the 37 subjects were Tswana-mother-tongue speakers.

TABLE 3.1

Detailed Analysis of the L1 Subjects

1. L1 (NTL1 + TL1)		
NTL1 (Non-Tswana L1)		
CM Primary School	**Mother tongue**	**N**
Ghanaian	English	2
Coloured	English (Replacement?)	3
Greek	Greek and English	1
Filipino	Tagalog and English	1
S.A. Indian	English (Replacement?)	9
S.A. Whites	English	3
Sri Lankan	Tamil	1
Total		**20**
Other Schools	**Mother tongue**	**N**
S.A. Coloured	English (Replacement?)	3
S.A. Indian	English (Replacement?)	1
S.A. White	English	1
Sri Lanka	English	1
Total		**6**
TL1 (Tswana L1)		
CM Primary School	**Mother Tongue**	**N**
South Sotho	Sotho	1
South Sotho	Sotho and English	1
Tswana	Tswana	13
Zulu	Zulu	2
Total		**17**
Other Schools (DET)	**Mother Tongue**	**N**
Tswana	Tswana	6
Total		**6**
TOTAL L1 (NTL1 + TL1)		**49**

There were ten L1 subjects who changed from L1 English as a taught subject to L2 English as a taught subject in Grade 8 (January 1988). I did not have any information on why these changes were made. One plausible reason for this change could have been that the School recommended to the learners concerned that it was in their best interest to change, because the change to L2 English as a taught subject at a later stage might have given them a better chance of passing English. Another plausible reason is that learners decided themselves to make the change, because it wasn't necessary to take English as a First Language, because they already had *Tswana* as a First Language. Some members of the L1 group who had obtained Grade 7 English *achievement* scores in the same range as those who changed to English Second Language did not change from English First Language to English Second Language. For example, the Grade 7 English *achievement* scores of the ten L1 subjects who changed to L2 in Grade 8 were (in ascending order in percentages) 50, 51, 53, 53, 55, 55, 58, 58, 61 and 63. (Most of these also had English proficiency scores in the 55% to 70% range). The Grade 7 English scores of five L1 subjects who did not change to L2 in Grade 8 were 45, 52, 53, 59, 62. As a matter of interest, of the 10 L1 subjects who changed to L2, eight obtained a matriculation exemption, while of the five L1 subjects who didn't change to L2 , two left the school after passing their respective grades, two obtained a matriculation exemption, and one failed before reaching Grade 12 and left the school. (More about this in section 5.3).

To reiterate: the "replacement language" subjects (who all belonged to the L1 group) were required to take English First language because they had no other First Language, while the Tswanas in the L1 group could take Tswana as a First Language. The initial choice of language group (L1 or L2) at the beginning of *Grade 7*, as pointed out earlier, was voluntary. (See Note 2 of this chapter for a definition of a "replacement language" learner).

Most of the L2 subjects were Tswana-mother-tongue speakers who originated from rural or peri-urban schools where English was used on a limited scale, which probably explains the low English proficiency of most of them. A few L2 subjects, however, did obtain high

English proficiency scores. None of L2 subjects came from CM Primary School, because all learners at this school had English as the medium of instruction from Grade 1. L2 subjects even if they did very well in English Second Language did not change to English First Language. This was probably because there was no need to complicate their lives unnecessarily. A detailed analysis of the L2 group follows.

TABLE 3.2

Detailed Analysis of the L2 Subjects

L2 (=TL2)		
Other Schools (mostly DET)	**Mother tongue**	**N**
North Sotho	North Sotho	1
South Sotho	South Sotho	2
Tswana	Tswana	33
Venda	Venda	1
Total L2		**37**

As shown in Table 3.2, four of the black L2 subjects were non-Tswanas. These took Tswana as a first language at the School.

In sum, there are two groups in the sample: the L1 group (a composite of the Tswana L1 and the Non-Tswana L1 groups) and the L2 group. *The L2 group can also be referred to as the TL2 (Tswana L2) group*, because 33 of the 37 L2 subjects were Tswana-mother-tongue speakers and the remaining four took Tswana as a first language..

3.3 Structure and administration of the English proficiency tests

Subjects were divided into four groups and the tests were administered in four classrooms by four Grade 7 teachers, where each classroom contained a combination of L1 and L2 subjects.

The time allotted for each test will be indicated in the description of the administration of the individual tests.

The possibility exists that fatigue resulting from a three-day test period may have affected the results of all the tests, but this seems unlikely because subjects were released from all lessons and from all school activities during this three-day period. Also, the test sessions were interspersed with ample rest periods. The structure and administration of the English proficiency tests follow. A theoretical review precedes the description of each of the tests.

I would like to emphasise that no sample of tests can adequately represent the vast variability of language, nor does it have to "because of the generative nature of language which acts as its own creative source"[12]. The controversy, as far as general language proficiency is concerned, is which sample of tests to use: the "reductionist" kind of tests used in this study or "holistic" (formal and informal) outside-of-the-classroom "cocktail party", "tea-party", "cooking club" type tests. In the academic context, what is important is the relationship between general, or overall, proficiency, communicative competence and academic achievement.

3.3.1 The cloze tests

3.3.1.1 Theoretical overview

> *Cloze tests are deceptively simple devices that have been constructed in so many ways for so many purposes that an overview of the entire scope of the literature on the subject is challenging to the imagination not to mention the memory.*[13]

Since 1973 the literature on cloze has more than doubled, adding even more challenges to the imagination if not - thanks to the printed word - to the memory.

---

12 Davies, A. *Principles of language testing*, 1990, p.3.

13 Oller, J.W., Jr. 'Cloze tests of second language proficiency and what they measure.' *Language Learning*, 23 (1), 105-118 (1973), p.106.

The aim of a cloze test is to evaluate (1) readability and (2) reading comprehension. The origin of the cloze procedure is attributed to Taylor[14] who used it as a tool for testing readability. Of all the formulas of readability that have been devised, cloze tests have been shown to be the best indicators of readability.[15] It is also regarded as a valid test of reading comprehension. Bormuth[16] found a multiple correlation coefficient of .93 between cloze tests and other linguistic variables that Bormuth used to assess the difficulty of several prose passages. Bormuth maintains that cloze tests "measure skills closely related or identical to those measured by conventional multiple choice reading comprehension tests."[17]

Many standardised reading tests use cloze tests, e.g. the Stanford Proficiency Reading Test. Johnson and Kin-Lin[18] believe that cloze is more efficient and reliable than reading comprehension, because it is easier to evaluate and does not, as in many reading comprehension tests, depend on long written answers for evaluation. (It is also possible to use multiple choice reading tests). Johnson and Kin-Lin's implication is that although cloze and reading comprehension are different methods of testing, they both tap reading processes. Anderson[19], however, maintains that as there is no consensus on what reading tests actually measure: all that can be said about a reading test is that it measures reading ability. Far more can be said about reading: Notions associated with reading are "redundancy utilization"[20], "expectancies about syntax and semantics"[21] and "grammar of

---

14 Taylor, W. 'Cloze procedure: A new tool for measuring readability.' *Journalism Quarterly*, 30, 414-438 (1953).

15 (1) Bormuth, J. 'Mean word depth as a predictor of comprehension difficulty.' *Journal of Educational Research*, 15:226-231 (1964).

(2) Oller, J.W., Jr. 'Cloze tests of second language proficiency and what they measure.' *Language Learning*, 23 (1), 105-118 (1973), p.106.

16 Bormuth, J., ibid., p.265.

17 Ibid.

18 Johnson, F.C. and Kin-Lin, C.W.L. 'The interdependence of teaching, testing, and instructional materials', in Read, J.A.S. (ed.). *Directions in language testing*, 1981, p.282.

19 Anderson, J. *Psycholinguistic experiments in foreign language testing*, 1976. p.1.

20 Weaver, W.W. and Kingston, A.J. 'A factor analysis of the Cloze procedure and other measures of reading and language ability.' *Journal of Communication*, 13, 252-261 (1963).

21 Goodman, K.S. 'Analysis of oral reading miscues: Applied psycholinguistics.' *Reading Research Quarterly*, 5, 9-30 (1969), p.82.

expectancy"[22]. All these terms connote a similar process. This process involves the "pragmatic mapping" of linguistic structures "into" extralinguistic context.[23] This mapping ability subsumes global comprehension of a passage, inferential ability, perception of causal relationships and deducing meaning of words from contexts.

Oller's[24] use of the word "into" in the previous paragraph might create the impression that pragmatic mapping for Oller is merely a question of language being mapped into or onto world knowledge, as if extralinguistic contexts were the raw material that had to be processed through language. The processes involved are, of course, much more complicated than that. As Oller points out "it is difficult to separate skill in handling information contained in the [cloze] test from previously acquired knowledge...The question then arises 'How do you separate knowledge of the word from language skill?' The answer is that you don't."[25] The reason why you don't, I suggest, is because it is often very difficult to distinguish between language skill, language knowledge and world knowledge. All three are inextricably implicated in language-processing, except, perhaps, for the *basic* levels of semantics and syntactics.

Although cloze answers are short, usually a single word, the cloze test can still be regarded as an "integrative" test. A distinction needs to be made between integrative and discrete test-formats and integrative and discrete processing strategies. The salient issue in a cloze test or in any test is not the length of the answer or the length of the question, but whether the test measures what it is supposed to measure, in this case integrative processing strategies. One should distinguish between the structure of the test - long answer, short answer, multiple choice - and what one is measuring. One is measuring the natural ability to process language; and one component of this ability is the behaviour of supplying missing

---

22 Oller, J.W., Jr. 'Cloze tests of second language proficiency and what they measure.' *Language Learning*, 23 (1), 105-118 (1973), p.113.

23 Oller, J.W., Jr. *Language tests at school*, 1979, p.61.

24 Ibid.

25 Oller, J.W., Jr. 'Cloze tests of second language proficiency and what they measure.' *Language Learning*, 23 (1), 105-118 (1973), p.112.

linguistic data in a discourse, which is what real-life language behaviour is about. Filling the gaps in a cloze test, therefore, is not the same thing as doing a crossword puzzle, as Spolsky[26] points out.

Consider the following views regarding the relationship between cloze tests and other tests. According to Bachman,

> *[t]here is now a considerable body of research providing sound evidence for the predictive validity of cloze test scores. Cloze tests have been found to be highly correlated with virtually every other type of language test, and with tests of nearly every language skill or component.*[27]

Clarke[28], in support of Bachman, is cautiously optimistic that the cloze procedure has a good future in reading research. Alderson is less optimistic:

> *[I]ndividual cloze tests vary greatly as measures of EFL proficiency. Insofar as it is possible to generalise, however, the results show that cloze in general relates more to tests of grammar and vocabulary... than to tests of reading comprehension.*[29]

Hughes[30] and Porter[31] concur with Alderson's findings that individual cloze tests produce different results, and thus they maintain with Alderson that each cloze test "needs to be validated in its own right and modified accordingly"[32]. Johnson and Kin-Lin[33] and Oller[34],

---

26 Spolsky, B. 'The limits of authenticity in language testing.' *Language Testing*, 2, 31-40 (1985), p.35

27 Bachman, L.F. 'The trait structure of cloze test scores.' *TESOL Quarterly*, 16 (1), 61-70 (1982), p.61.

28 Clark, J.L.D. Language testing: Past and current status - Directions for the future. *Language testing*, 64 (4), 431-443 (1983).

29 Alderson, J.C. 'The cloze procedure and proficiency in English as a foreign language.' *TESOL Quarterly*, 13, 219-227 (1979), p.225.

30 Hughes, A. 'Conversational cloze as a measure of oral ability.' *English Language Teaching Journal*, 35 (2), 161-168 (1981).

31 Porter, D. 'Cloze procedure and equivalence.' *Language Learning*, 28 (2), 333-41 (1978).

32 Alderson. J.C., ibid., p.226.

33 Johnson, F.C. and Kin-Lin, C.W.L. *The interdependence of teaching, testing, and instructional*, 1981.

contrary to Alderson, found that a great variety of cloze tests correlates highly with tests such as dictation tests, essay tests and reading tests as well as with "low order"[35] grammar tests.

The concept of closure is important in cloze theory. Alderson states that

> *one must ask whether the cloze is capable of measuring higher-order skills. The finding in Alderson (1978) that cloze seems to be based on a small amount of context, on average, suggests that the cloze is sentence - or indeed clause - bound, in which case one would expect a cloze test to be capable, of measuring, not higher-order skills, but rather much low-order skills... as a test, the cloze is largely confined to the immediate environment of a blank.*[36]

This means that there is no evidence that increases in context make it easier to complete items successfully. Oller maintains, contrary to Alderson, that subjects "scored higher on cloze items embedded in longer contexts than on the same items embedded in shorter segments of prose".[37] Oller used five different cloze passages and obtained similar results on all of them.

Closure does not merely mean filling in items in a cloze , but filling them in a way that reveals the sensitivity to intersentential context, which measures "higher-order skills" (Alderson above). A cloze test that lacks sufficient closure would not be regarded as a good cloze test.

There are two basic methods of deletion: fixed deletion and rational deletion, or "selective" deletion[38]). In the former, every nth word is deleted; which may range between every fifth word - which is considered to be the smallest gap permissible without making the

---

34 Oller, J.W., Jr. *Language tests at school*, 1979.

35 Alderson, J.C., ibid., 1979.

36 Alderson, J.C., ibid., 1979, p.225.

37 Oller, J.W., Jr. Cloze, discourse, and approximations to English, 1976, p.354.

38 Ingram, E. 'Assessing proficiency: An overview on some aspects of testing', in Hyltenstam, K. and Pienemann, M. *Modelling and Assessing second language acquisition*, 1985, p.241.

recognition of context too difficult - and every ninth word. Pienaar's tests, which are used in this study, are based on a rational deletion procedure.[39]

Alderson proposes that the rational deletion procedure should not be referred to as a "cloze" but as a "gap-filling" procedure[40]. Alderson has made a proposal, which some researchers, e.g. Weir[41], have accepted while others haven't, e.g. Maclean's "Using rational cloze for diagnostic testing in L1 and L2 reading"[42] and Markham's "The rational deletion cloze and global comprehension in German."[43]

Alderson proposes that what he calls cloze tests (namely, every nth word deletion), which he contrasts with "gap-filling" tests such a rational deletion, should be abandoned in favour of "the rational selection of deletions, based upon a theory of the nature of language and language processing"[44] (see also Bachman[45]). The evidence does appear strong that the rational deletion procedure increases closure and thus is able to measure "higher order" skills. Accordingly, I have chosen the rational deletion procedure for the cloze tests in this study.

Having considered the arguments for the validity of the cloze test as a test of reading, it seems that a cloze test can be a good test of reading strategies, i.e. it can test long-range contextual constraints, especially if the rational deletion method is used. One must keep in mind, however, that deletion rates, ways of scoring, e.g. acceptable words or exact words,

---

39 Pienaar, P. *Reading for meaning: A pilot survey of (silent) reading standards in Bophuthatswana*, 1984.

40 Alderson, J.C. 'Native and nonnative speaker performance on cloze tests'. *Language Learning*, 30 (1), 59-77 (1980), p.59-60.

41 Weir, C.J. *Understanding and developing language tests*, 1993, p.81.

42 Maclean, M. 'Using rational cloze for diagnostic testing in L1 and L2 reading.' *TESL Canada Journal*, 2, 53-63 (1984).

43 Markham, P. The rational deletion cloze and global comprehension in German. *Language Learning*, 35, 423-430 (1985).

44 Alderson, J.C. 'The cloze procedure and proficiency in English as a foreign language.' *TESOL Quarterly*, 13, 219-227 (1979), p.226.

45 Bachman, L.F. 'The trait structure of cloze test scores.' *TESOL Quarterly*, 16 (1), 61-70 (1982).

and types of passages chosen in terms of background knowledge and of discourse devices may influence the way reading strategies are manifested.

3.3.1.2 The cloze tests used in the study

In a review of Pienaar's[46] pilot survey called "Reading for meaning", Johanson[47] refers to the "shocking" low reading levels in many schools in the North-West Province revealed by Pienaar's survey.

Pienaar tested a variety of learners from different schools: learners whose (1) first language (i.e. mother tongue or a language the learner knows best) was English, (2) replacement language was English, (3) first language was Afrikaans, (4) first language was a Bantu language; these were mostly Tswana speakers. Categories (1) and (2) and (3) came from upper middle class families, while category (4) was split into two sub-categories: (4a) Bantu speakers who generally came from working class families and (4b) Bantu speakers who lived in sub-economic settlements in the environs of Mmabatho. Many of the parents of (4b) were illiterate or semi-literate and were either unemployed or semi-employed. (The category labels used by Pienaar differ from mine: I have changed them for clarity sake). Pienaar's major finding was that 95% of learners in the North West Province (Grade 3 to Grade 12), most of whom belonged to category 4b, were "at risk", i.e. they couldn't cope with the academic reading demands made on them.

There are four reasons why Pienaar's cloze tests are used in this study:
(1) they have already been used in many schools in the North West Province and have produced a solid body of results, (2) their purpose is "to select lexical and structural items

---

[46] Pienaar, P. *Reading for meaning: A pilot survey of (silent) reading standards in Bophuthatswana*, 1984.

[47] Johanson, L. 'Shocking low reading levels in many Bop schools.' *Matlhasedi*, 7 (1 and 2), 27 (1988), p.27.

relevant to the demands of the appropriate syllabuses"[48], (3) Pienaar's cloze tests are based on a rational deletion method, where it is possible to select gaps in such a way that closure , i.e. long-range constraints, is ensured, and (4) Pienaar's data will be compared with the data in this study.

Pienaar's tests comprise five graded levels - "Steps" 1 to 5, where each level consists of four short cloze passages (Form A to Form D) with 10 blanks in each passage[49]:

Step 1 corresponds to Grades 3 and 4 for English First Language and to Grades 5 to 7 for English Second Language.

Step 2 corresponds to Grades 5 and 6 for English First Language and to Grades 7 to 9 for English Second Language.

Step 3 corresponds to Grades 7 and 8 for English First Language and to Grades 9 to 11 for English Second Language.

Step 4 corresponds to Grades 9 and 10 for English First Language and to Grades 11 and 12 for English Second Language.

Step 5 corresponds to Grades 11 and 12 for English First Language and to Grades 12 + (i.e. higher than Grade 12) for English Second Language.

If one Step proves too easy or too difficult for a specific pupil, a higher or a lower Step could be administered. For example, if Step 2 is too difficult, the pupil can be tested on Step 1. In this way it is possible to establish the level of English proficiency for each individual pupil.

Owing to the fact that (1) many of the L1 group were not English-mother-tongue speakers and (2) I had to give the same test to both the L1 and L2 groups in order to make reliable comparisons, I used Step 2 for the L1 group and the L2 group.

---

[48] Pienaar, P. *Reading for meaning: A pilot survey of (silent) reading standards in Bophuthatswana*, 1984, p.3.

[49] Ibid., p.41.

I did not use the other Steps, because it was irrelevant to the purpose of study, which was not concerned with placing learners in the level they belonged to, i.e. for teaching purposes (which was the purpose of Pienaar's tests), but to test for proficiency at the Grade 7 level, and in the process to test the tests themselves, i.e. examine whether the passages chosen for the Step 2 level were valid for that level. *Although annual predictions between English proficiency and academic achievement might yield higher correlations than long-term predictions, the aim was to investigate what chance Grade 7 learners who entered the School in 1987 would have had of passing Grade 12.*

According to Pienaar, a perfect score on a cloze test indicates that the pupil has fully mastered that particular level. A score below 50% would indicate that the learner is at risk.

Pienaar[50] maintains that English second language learners are generally two to three years behind English first language learners in the acquisition of English proficiency, and there is often also a greater age range in the English second language classes, especially in the rural areas. Pienaar's implication is that to fall behind in English language proficiency is also to fall behind in academic performance.

Pienaar standardised his tests in 1982 on 1068 final year JSTC (Junior Secondary Teacher's Certificate) and PTC (Primary Teacher's Certificate) students from nine colleges affiliated to the University of the Transkei. These standardised results became the table of norms for Pienaar's tests. Below are the weighted mean scores achieved by the students of the nine colleges:

	Step 1	Step 2	Step 3	Step 4	Step 5
Weighted means:	67%	53%	37%	31%	24%

50 Pienaar, P., ibid., p.41.

Most of the colleges performed similarly on all five Steps. These results confirmed the gradient of the difficulty of the various steps.

During 1983 Pienaar administered a major part of the test instrument to a smaller group of college students selected from the original large group. No significant difference between the scores of the two administrations was found, which confirmed the test-retest reliability of the instrument.[51]

The tests underwent ongoing item analysis and refinement. By the time the final version was submitted to school learners in the Mmabatho/Mafikeng area in 1984, 30% of the items had been revised. As a result of continuous item analysis, a further 18% of the items were revised[52].

An important point is that these results claim to represent the reading ability of college students, who are supposed to be more proficient in English than school learners. However, final year student teachers only obtained a score of between 40% and 60% on Step 2 - see Pienaar's mean scores above. (Step 2 has been used in this study for Grade 7 learners). These low scores indicate that the reading level of the student teachers, who were to start teaching the following year, was probably no higher than the level of many of the learners they would eventually teach. This *alarming* state of affairs would probably have had a detrimental effect on the academic performance of these learners.

Pienaar's original tests had four passages for each level and he tried to establish the equivalence in difficulty between the passages for each level. In this study I used two cloze passages for Step 2 instead of four. This was done for two reasons: (1) to see whether only two passages were sufficient, and (2) the other two passages, in their "unmutilated" form, were used for the dictation tests, because they belonged to the same level, namely, Step 2.

---

51 Pienaar, ibid., p.9.

52 Ibid.

The question is whether two passages - in the cloze and dictation tests - were enough to ensure reliability and validity. The results of the tests (Chapter 4) deal with this question.

In the test battery I used Pienaar's Form B and Form D passages of Step 2, as shown below:

Pienaar's Practice exercise

(Pienaar does not provide the answers for this practice exercise. Possible answers are provided in brackets).

The 1 (rain) started falling from the sagging black 2 (clouds) towards evening. Soon it was falling in torrents. People driving home from work had to switch their 3 (headlights) on. Even then the 4 (cars, traffic) had to crawl through the lashing rain, while the lightning flashed and the 5 (thunder) roared.

Cloze passage 1: Form B Step 2[53]*:*

A cat called Tabitha

Tabitha was a well-bred Siamese lady who lived with a good family in a shiny white house on a hill overlooking the rest of the town. There were three children in the family, and they all loved Tabitha as much 1 she loved them. Each night she curled up contentedly on the eldest girl's eiderdown, where she stayed until morning. She had the best food a cat could possibly have: fish, raw red mince, and steak. Then, when she was thirsty, and because she was a proper Siamese and did 2 like milk, she lapped water from a blue china saucer.

Sometimes her mistress put her on a Cat show, and there she would sit in her cage on 3 black padded paws like a queen, her face and tail neat and smooth, her black ears pointed forward and her blue 4 aglow.

It was on one of these cat shows that she showed her mettle. The Judge had taken her 5 of her cage to judge her when a large black puppy ran into the hall. All the cats were furious and snarled 6 spat from their cages. But Tabitha leapt out of the judge's arms and, with arched 7 and fur erect, ran towards the enemy.

The puppy 8 his tail and prepared to play. Tabitha growled, then, with blue eyes flashing, she sprang onto the puppy's nose. Her 9 were razor-sharp, and the puppy yelped, shook her off, and dashed for the door. Tabitha then stalked back down the row of cages to where she had 10 the judge. She sat down in front of him and started to preen her whiskers as if to say, "Wait a minute while I fix myself up again before you judge me." She was quite a cat, was Tabitha!

Answers. (The words in round brackets are Pienaar's suggested alternative answers. The words in square brackets are suggested alternative answers):
1. as; 2. not; 3. her [four, soft]; 4. eyes (eye); 5. out; 6. and; 7. back (body); 8. wagged, twitched (waved, lifted); 9. claws (nails); 10. left (seen, met).

Item 9 could also be "teeth". (There are cats who jump on to faces and bite, rather than scratch). Even in easy cloze passages, "acceptable" answers can be a problem.

---

[53] Pienaar, ibid, p.59.

Cloze passage 2: Form D Step 2[54]*:*

A dog of my own

When I was ten all 1 wanted was a dog of my own. I yearned for a fluffy, fat, brown and white collie puppy. We already had two old dogs, but my best friend's pet collie had 2 had seven fluffy, fat, brown and white puppies, and I longed for one with all my heart. However, my mother said no, so the seven puppies were all sold. I had horses, mice, chickens and guinea-pigs, and as my 3 said, I loved them all, but I wasn't so keen on finding them food. Since she had five children to look after, it made here angry to 4 hungry animals calling, so she said crossly, "No more dogs."

This didn't stop me wanting one though, and I drew pictures of collie dogs, giving 5 all names, and left them lying around where she would find them. As it was 6 Christmas, I was sure that she would relent and give me a puppy for Christmas.

On Christmas morning I woke up very excited, 7 the soft little sleepy bundle that I wanted at the bottom of the bed wasn't there. My mother had given me a book instead. I was so disappointed that I cried to myself, yet I tried not to 8 her how sad I was. But of course she noticed.

Soon after that my father went off to visit his brother and when he came back he brought me a puppy. Although it 9 a collie it was podgy and fluffy, and I loved him once. My mother saw that I looked after him properly and he grew up into a beautiful grey Alsatian. We were good friends for eleven happy 10 before he went to join his friends in the Animals' Happy Hunting Ground.

Answers.
1. I; 2. just, recently; 3. mother (mummy, mum, mom); 4. hear; 5. them; 6. near (nearly, nearer, close to; 7. but, however (though); 8. show (tell); 9. wasn't (was not); 10. years.

Pienaar allotted six minutes for each passage. I allotted 12 minutes. Ability is dependent on speed of processing and so if a test taker does badly with an allotted time of six minutes, perhaps performance would significantly improve with an allotted time of 12 minutes. But it would not be difficult to test this hypothesis empirically: one could compare a control group (allotted a time of six minutes) with an experimental group (allotted a time of 12 minutes).

3.3.3 The essay tests

3.3.3.1 Theoretical overview

Language processing involves various components such as linguistic knowledge, content knowledge, organisation of ideas and cultural background. All these factors mesh together

---

[54] Pienaar, ibid., p.61.

into a proficiency network of vast complexity, which makes objective evaluation of essay performance very difficult. It is this vast complexity that makes essay writing the most pragmatic of writing tasks and the main goal of formal education.

Essay writing, is "probably the most complex constructive act that most human beings are ever expected to perform".[55] If getting the better of words in writing is usually a hard struggle for mother-tongue speakers, the difficulties are multiplied for second-language learners, and very difficult for disadvantaged or second-language learners such as those described in this study. Many of the disadvantaged Grade 7 subjects are similar to young mother-tongue speakers of English first learning to write in that much mental energy is expended on attention to linguistic features rather than to content.

What makes essay writing a pragmatic task is that it involves writing beyond the sentence level (at the intersentential, or suprasentential, level). This does not mean that non-pragmatic tasks are not integrative. As discussed in section 2.5, all language resides along a continuum of integrativeness, where pragmatic tasks are the most integrative.

Owing to the fact that the production of linguistic sequences in essay writing is not highly constrained, problems of reliability arise in essay scoring. (In this respect, essay tests have much in common with oral tests). Inferential judgements have to be converted into a score, so "[h]ow can essays or other writing tasks be converted to numbers that will yield meaningful variance between learners?".[56] Oller argues that these inferential judgements should be based on intended *meaning* and not merely on correct structural forms.[57] That is why, in essay rating, raters should rewrite (in their minds, but preferably on paper) the intended meaning. Perhaps one can only have an absolutely objective scoring system with

---

55 Bereiter, C. and Scardemalia, M. 'Does learning to write have to be so difficult?', in Freedman, A., Pringle, I. and Yalden, J. *Learning to write: First language/second language*, 1983, p.20.

56 Oller, J.W., Jr. *Language tests at school*, 1979, p.385.

57 Ibid.

lower-order skills (Allen in Yeld[58]), but Oller is not claiming that his scoring system is absolutely objective, but only that it is a sensible method for assessing an individual's level within a group.

Whatever one's paradigm - structural ("old") or communicative ("new") - when one marks an essay one can only do so through its structure. The paradox of language is that structure must "die" so that meaning may live; yet, if structure is not preserved, language would not be able to mean.

In the normal teaching situation, marking is done by one rater, namely the teacher who teaches the subject. Sometimes if a test is a crucial one, for example an entrance test or an end-of-year examination, more than one rater, usually two, are used. In a research situation, the number of raters depends on the nature of the research and the availability and proficiency of raters. The raters used for the essay tests in this study were all English-mother-tongue speakers, and recognised as such by their colleagues. (In the dictation tests there were three English-mother-tongue presenters and one non-English-mother-tongue presenter).

With regard to the level of English proficiency of raters, it does not follow that because a rater (or anybody else) is not a mother-tongue speaker (of English in this case) that his or her English proficiency is necessarily lower than an English-mother-tongue speaker. In the academic context, there are many non-English-mother-tongue speakers who have a higher level of academic English proficiency than English-mother-tongue speakers. A major reason for this is not a linguistic reason, but because these non-English-mother-tongue speakers are more academically able, i.e. they have better problem-solving abilities and abilities for learning content.[59] (See the last three pages of section 6.2 on problem-solving in a first and second language).

---

[58] Yeld, N. *Communicative language testing.* Report on British Council Course 559 offered at Lancaster University from 8 September to 20 September 1985 (published in) 1986, p.32.

[59] (1) Bley-Vroman, R. 'The logical problem of foreign language learning.' *Linguistic Analysis,*

Kaczmarek[60] reports high correlations between essay raters, while Hartog and Rhodes[61] and Pilliner[62], contrary to Kaczmarek, found essay tests to have low interrater and low intrarater reliability. With regard to scoring procedures in essay testing. Mullen[63] (1980:161) recommends the use of four "scales" (criteria) of writing proficiency: structure, organisation, quantity, vocabulary. According to Mullen[64], a combination of all four scales is required to validly predict proficiency. A major issue in scoring is whether marks should be separately allocated to each of the criteria, i.e. whether one should use an analytic scoring procedure or whether marks should be allocated globally. Global scoring usually refers to two ways of scoring: (1) "Overall impressions" and (2) "Major constituents of meaning", which takes into account global errors, e.g. cohesion and coherence, but not local errors, e.g. grammar and spelling.

The following terms are used interchangeably in the literature: global rating, overall impressions, holistic scoring and global impressions. The term holistic scoring is used by Perkins[65] to refer to overall impressions, which takes into account global as well as local errors. With regard to the two ways of global scoring mentioned in the previous paragraph, it is possible that a rater's "overall impressions" may include quick, yet thorough attention to

---

20 (1-2), 3-49 (1990).

(2) Collier, V.P. 1987. 'Age and rate of acquisition of second language for academic purposes.' *TESOL Quarterly*, 21 (4), 617-641.

(3) Olivier, A. 'The role of input in language development at tertiary level.' *South African Journal of Education*, 18 (1), 57-60 (1998).

(4) Vollmer, H.J. 'The structure of foreign language competence', in Hughes, A. and Porter, D. (eds.). *Current developments in language testing*, 1983.

60 Kaczmarek, C.M. 'Scoring and rating essay tasks', in Oller, J.W. (Jr.) and Perkins, K. (eds.). *Research in language testing*, 1980, p.156-159.

61 Hartog, P. and Rhodes, E.C. *The marks of examiners*, 1936, p.15.

62 Pilliner, A.E.G. 'Subjective and objective testing', in Davies, A. (ed.). *Language testing symposium*, 1968, p.27.

63 Mullen, K. A. 'Evaluating writing proficiency in ESL', in Oller, J.W. (Jr) and Perkins, K, in *Research in language testing*, 1980, p.161.

64 Ibid.

65 Perkins, K. 'On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability', *TESOL Quarterly*, 17 (4), 651-671 (1983).

major constituents of meaning as well as to local errors. In such a case the distinction within global scoring between "overall impressions" and "major constituents of meaning" would no longer apply.

With regard to the relative reliability of analytic and global scoring procedures, Kaczmarek[66] and Ingram[67] have shown that analytic scoring procedures are not more reliable than global scoring. According to Perkins, "[w]here there is commitment and time to do the work required to achieve reliability of judgement, holistic evaluation of writing remains the most valid and direct means of rank-ordering students by writing ability."[68] Zughoul and Kambal report "no significant difference between the two methods"[69], and Omaggio maintains that "holistic scoring has the highest validity"[70] (reliability, not validity, surely).

According to Oller, "judges always seem to be evaluating communicative effectiveness regardless whether they are trying to gauge 'fluency', 'accentedness', 'nativeness', 'grammar', 'vocabulary', 'content', 'comprehension', or whatever."[71] Even if one does this quickly, say a minute per page, the brain of the rater still has to consider the trees to get an overall idea of the wood. The greater the experience and competence of the rater the more unconscious and quicker, but no less rational, is the judgement.

It is arguable whether judges always seem to be evaluating communicative effectiveness, as Oller maintains. Although it seems reasonable that in essays one *should* be looking at the overall impact of a piece of writing (the whole) and that the only way to do this is to look at the various aspects of the writing such as those mentioned by Oller, I would question

---

66 Kaczmarek, C.M. 'Scoring and rating essay tasks', in Oller, J.W. (Jr.) and Perkins, K. (eds.). *Research in language testing*, 1980, pp.151-159.

67 Ingram, E. 'Item analysis', in Davies, A. *Language testing symposium*, 1968, p.96.

68 Perkins, K. 'On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability', *TESOL Quarterly*, 17 (4), 651-671 (1983), p.655.

69 Zughoul, M.R. and Kambal, O. K. 'Objective evaluation of EFL composition.' *International Review of Applied linguistics*, 21 (2), 87-103 (1983), p.100.

70 Omaggio, A.C. *Teaching language in context: Proficiency-orientated instruction*, 1986, p.263.

71 Oller, J.W., Jr. *Language tests at school*, 1979, p.342.

whether raters (in general) do regard communicative effectiveness as the overarching criterion. Unfortunately, I did not manage to obtain the judgements of the raters of the MHS essay protocols and so could not investigate this important issue. I did, however, at a later stage, use some of the MHS protocols on a different set of raters to verify whether Oller's cautious observation was reasonable. (More about this in section 4.8.1).

If global or holistic scoring is more effective or even as effective as analytic scoring, then for reasons of economy global scoring should be used. Global scoring, however, ideally requires at least three raters[72], who would presumably, and hopefully, balance one another out. The effectiveness of global scoring depends on factors such as availability, willingness and qualifications of raters. The unavailability of raters is often a problem. In special circumstances such as proficiency tests used for purposes of admission or placement at the beginning of an academic year, it may be possible to obtain the help of three or four raters. However, in the normal testing situation during the school year, only one rater may be available, who is usually the teacher involved in teaching the subject. Hughes[73] recommends four raters because this has been shown to be the best number. Four raters were used in this study.

To ensure high interrater reliability there should only be a narrow range of scores and judgements between raters. If three or four raters are considered to be required for reliability a serious problem is what to do in the normal education situation where at most two and usually only one rater is available. (More about this in section 4.8.1.3).

3.3.3.2 The essay tests used in the study

There were two essays:

Essay 1: Everybody in this world has been frightened at one time or another. Describe a time when you were frightened. Write between 80 and 100 words.

---

72 Ingram, E. 'Item analysis', in Davies, A. *Language testing symposium*, 1968, p.96.

73 Hughes, A. *Testing for language teachers*, 1989, p.87.

Essay 2: Do (a) or (b) or (c). Do only one of the following topics. Don't forget to write the letter (a) or (b or (c) next to the topic you choose. The topic you choose must not be shorter than 80 words and not longer than 100 words.

(a) Describe how a cup of tea is made.

(b) Describe how shoes are cleaned.

(c) Describe how a school book is covered.

Both the L1 and L2 group's essays were judged in terms of English-mother-tongue proficiency.

The TOEFL (Testing of English as a Foreign Language) Test of Written English recommends spending a rapid one and a half minutes per page using a holistic scoring method. I would imagine that when working at such a speed, the scoring criteria are assumed to be known to the point of automaticity. Raters in this study were recommended to spend about one and a half minutes on each protocol, where protocols were much shorter than a page in length. The TOEFL scoring method seems to be the same as the "overall impressions" approach of Perkins[74], which takes into account global as well as local errors. As I discussed earlier, clarity and consistency of judgements are difficult to ensure.

Four raters - three Grade 7 teachers at MHS and myself - rated 86 protocols. All four of us were English-mother-tongue speakers and were recognised as such by our colleagues. Each rater, in turn, was given the original 86 protocols and was requested to give an impressionistic score based on such considerations as topic relevance, content and grammatical accuracy.

Raters did not provide any judgements. The reason for this was because these raters, who were also the Grade 7 teachers at the School, were fully involved in the three-day administration of the test battery. Accordingly, I did not want to overload the three other

[74] Perkins, K. 'On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability', *TESOL Quarterly*, 17 (4), 651-671 (1983).

raters with too much extra work after the three days test period, because they had to return to a full teaching load. Thus, they merely gave a score based on a global impression. It would have been useful to compare raters' scores and judgements because this would have provided insights into the knotty problem of interrater reliability. As mentioned earlier, I did manage at a later stage to obtain data on the same essay test (given to the Grade 7 subjects in this study) from a workshop on language testing[75] (see section 4.8.1).

Bridgeman recommends that each rater assign a holistic score on a six point scale, where zero is given if the essay is totally off the topic or unreadable.[76] A nine-point scale was used in this study: from Scale 1; 0 to 1 point = totally incomprehensible, to Scale 9; 9 to 10 points = outstanding. The points were converted to percentages.

Raters did not record their scores on the protocols but were each provided with a copy of the list of the names of subjects on which they had to record their scores. Raters were requested not to consult one another on the procedures they used to evaluate the protocols. The results, as with all the tests in the study, are presented in the next chapter.

3.3.4 Error recognition and mixed grammar tests

3.3.4.1 Theoretical overview

Grammar is an important component in most standardised test batteries, e.g. English as a Foreign Language (EFL), English Language Testing Service (ELTS), Testing English as a Foreign Language (TOEFL)[77]. Error recognition has been used in various studies on

---

[75] Gamaroff, R. *Workshop on quantitative measurement in language testing.* National Association of Educators of Teachers of English (NAETE) Conference, East London Teacher's Centre, September, 1996c.

_____ *Interrater reliability, the bug of all bears: Report on the 1996 NAETE Workshop on quantitative measurement in language testing.* National Association of Educators of Teachers of English (NAETE) conference " Training teachers for the South African context, Potchefstroom College of Education, September 17-18, 1998c.

[76] Bridgeman, B. *Essays and multiple-choice tests as predictors of college freshman GPA.* (ETS Research Report, 1991, p.9.

[77] O'Dell, F. *English as a foreign language: Intermediate examinations,* 1986.

language proficiency testing. Olshtain et al.[78] use it as part of a battery of first language proficiency tests to predict second language proficiency. The emphasis in Olshtain et al. is on appropriacy, i.e. language use, and not on acceptability, i.e. grammatical correctness. Different to Olshtain et al.'s aim of trying to find a connection between error recognition and language use, Irvine, Atai and Oller[79] use the multiple-choice error recognition test from the TOEFL battery of tests to find out whether integrative tests such as cloze tests and dictation tests correlate more highly with each other than they do with the multiple-choice tests of TOEFL.

Henning et al.'s[80] revised GSCE (Egyptian General Secondary Certificate Examination) test battery contains an "Error Identification" test and a "Grammar Accuracy" test. Henning et al. found that the highest correlation with their "composition" subtest was with Error Identification (.76). They accordingly maintain that "Error Identification may serve as an indirect measure of composition writing ability."[81] A grammar component has always featured prominently in all the standardised English first and second language proficiency and achievement tests of the Human Sciences Research Council (HSRC). The recent tests of the HSRC range across various school levels from junior secondary school to senior secondary school.[82]

3.3.4.2 Error recognition and mixed grammar tests used in the study

The error recognition test and the mixed grammar test in this study are both multiple choice tests that have been designed for learners who have completed the "elementary" stage of

---

78 Olshtain, E., Shohamy, E., Kemp, J. and Chatow, R. 'Factors predicting success in EFL among culturally different learners.' *Language Learning*, 40 (1), 23-44 (1990).

79 Irvine, P., Atai, P. and Oller, J.W. (Jr.). 'Cloze, dictation, and the test of English as a foreign language.' *Language Learning*, 24 (2), 245-252 (1974), p.247.

80 Henning, G.A., Ghawaby, S.M., Saadalla, W.Z., El-Rifai, M.A., Hannallah, R.K. and Mattar, M. S. 'Comprehensive assessment of language proficiency and achievement among learners of English as a foreign language.' *TESOL Quarterly*, 15 (4), 457-466 (1981).

81 Henning et al., ibid., p.462.

82 Barry, D.M., Cahill, S., Chamberlain, J.C., Reinecke, S. and Roux. *Past and future approaches to language test development with examples*, undated.

second/foreign language learning.[83] Bloor et al. divided their tests into three levels; First Stage/Elementary Stage, Second Stage/Intermediatory Stage, and Third Stage/Advanced Stage. Although the levels have been designated relative to one another, these are merely guidelines, and therefore the tester must use discretion in fitting the level to the relevant group of students. I have used the First Stage level for the Grade 7 subjects. These two tests each comprise 50 items.

Bloor et al. (Teacher's book, p.ix) state that their tests were analysed by themselves over an extended period and subjected to an item analysis and validated under test conditions. No data on this item analysis or validation under test conditions were provided by the authors, probably because that kind of data would not appear in a Teacher's book. I shall show (Chapter 4) that their tests have high (split-half) reliability and high correlations with other test methods.

These two tests were administered at the same sitting over a one-and-a- half-hour period, the emphasis being on completing the task and not on speed. It is true, however, that ability is dependent on speed of processing. Sample items from the error recognition and mixed grammar tests are now provided.

Error recognition test: Sample items

The test used was Test 1 from Bloor et al[84]: Instructions: In some of the following sentences there are mistakes. (There are no mistakes in spelling and punctuation). Indicate in which section of the sentence the mistake occurs by writing its letter on your answer sheet. If there is no mistake, write E.

---

[83] Bloor, M., Bloor, T., Forrest, R., Laird, E. and Relton, H. *Objective tests in English as a foreign language*, 1970.

[84] Bloor, M., et al., *Objective tests in English as a foreign language*, 1970, pp.70-77; Book 2.

Example:

A B C D
Although he has lived in England/since he was fifteen,/he still speaks English/much badly.

Correct - E.

Answer: D.

A B C D
Item 8. Both Samuel and I/are much more richer/than we/ used to be. Correct - E.

Answer: B.

A B C D
Item 19. Some believe that/a country should be ruled/by men who are/too clever than ordinary people. Correct - E.

Answer: D.

A B C D
Item 25. His uncle is owning/no fewer than ten houses,/and all of them/are let at very high rents. Correct - E.

Answer: A.

A B C
Item 27. As I have now studied/French for over three years/I can be able to/make myself

D
understood when I go to France. Correct - E. Answer: C

Mixed Grammar test: Sample items

The test used is from the "First Stage: Test 2 in Bloor et al., Book 1[85]. (By "mixed" grammar is meant a variety of grammatical structures).

The test consists of choosing the correct alternative that fits into the gap within a sentence. The following are the instructions and an example from the test, followed by five selected items from the test:

---

[85] Ibid., p.35-40.

Instructions: Choose the correct alternative and write its letter on your answer sheet. Example: His sister is....than his wife. A) more prettier B) prettier C) very pretty D) most pretty. Answer: B.

Item 1. They often discuss.... A) with me B) about whether there is a problem C) the problem D) about the problem with me. Answer: C.

Item 28. This dress was made... A) by hands B) by hand C) with hands D) with hand. Answer: B.

Item 30. When the door-bell...., I was having a bath. A) rang B) rings C) rung D) ringed. Answer: A.

Item 38. My friend always goes home....foot. A) by B) with C) a D) on. Answer: D.

Item 50. We....our meat from that shop nowadays. A) were never buying B) do never buy C) never buy D) never bought. Answer: C.

The mixed grammar tests and error recognition tests of this study are commonly used tests. Compare these test items with equivalent test items from the Egyptian study of Henning et al.[86] Consider the following two items from their test battery:

Grammar Accuracy

Ahmed enjoys....us. A. helping B. to help C. help to D. helping to.

The item requires the selection of one of the four options. This test has the same format as Bloor et al.'s mixed grammar test.

---

[86] Henning, G.A., Ghawaby, S.M., Saadalla, W.Z., El-Rifai, M.A., Hannallah, R.K. and Mattar, M. S. 'Comprehensive assessment of language proficiency and achievement among learners of English as a foreign language.' *TESOL Quarterly*, 15 (4), 457-466 (1981).

Error Identification

A B C D

In my way to school,/ I met a man/ who told me that/ the school was on fire.

One has to choose the incorrect segment in a sentence, i.e. item A. This format is almost identical to Bloor et al.'s error recognition test. The difference is that Bloor et al.'s test has five options, which makes their test more difficult. I judged Bloor et al.'s "elementary stage" to be appropriate for the Grade 7 subjects. If this is correct it suggests that the standard of grammatical knowledge required for Egyptian university entrants is very similar to the standard required for Grade 7 non-English-mother-tongue speakers! Further, as mentioned, Henning et al's Error Identification test is easier than Bloor et al.'s, because it only has four options, whereas Bloor et al.'s error recognition test has five options. The greater the number of options the more difficult it is to guess. (More about this in section 4.5)

3.3.5 The dictation tests

3.3.5.1 Theoretical overview

Language tests involve one or various combinations of the four language modes, namely listening, speaking, reading and writing. Although the most important combination in education, except for the early school years, is usually reading-writing, the listening-writing combination, which is what a dictation test measures, also plays an essential role. The dictation test is a variation of a listening comprehension test where subjects write down verbatim what they listen to.

Listening comprehension "poses one of the greatest difficulties for learners of English".[87] This section examines language proficiency through the dictation test, which is the most

---

[87] Suenobu, M., Kanzaki, K., Yamane, S. and Young, R. 'Listening comprehension and the process of information acquisition by non-native speakers of English.' *International Review of Applied Linguistics*, 24 (3), 239-248 (1986).

demanding type of listening comprehension test, because it forces the test taker to focus on structure as well as meaning.

Some authors regard the dictation test merely as a test of spelling or of grammar[88]. For Ur[89], dictation "mainly tests spelling, perhaps punctuation, and perhaps surprisingly, on the face of it, listening comprehension". For Lado dictation was only useful as a test of spelling because dictation, he argued, did not test word order or vocabulary, both of which were already given; neither did it test aural comprehension, owing to the fact that the learner could often guess the context. For protagonists of the audio-lingual method the dictation test was considered to be a "hybrid test measuring too many different language features, mixing

listening and writing, and giving only imprecise information".[90] Savignon[91] maintains that dictation does not test communicative proficiency. The reasons for its popularity, she suggests, is that it has high concurrent validity, is easy to develop and score, and has high reliability.

Contrary to these negative views, other authors regard dictation as a robust test of the ability to reconstruct surface forms to express meaning at the sentence level and beyond and are valid measures of communicative proficiency.[92] Spelling, which for Lado was the dictation test's only justification, is disregarded by others.[93]

---

88 (1) Lado, R. *Language testing*, 1961.
(2) Froome, S. *Why Tommy isn't learning*, 1970, 30-31.

89 Ur, P. *A course in language teaching: practice and theory*, 1996, p.40.

90 Tönnes-Schnier, F. and Scheibner-Herzig, G. 'Measuring communicative effectiveness through dictation.' *International Review of Applied Linguistics*, 26 (1), 35-43 (1988), p.35.

91 Savignon, S.J. *Communicative competence: Theory and classroom practice.* (Reading, Mass. Addison-Wesley Publishing Company, 1983), p.264.

92 (1) Bacheller, F. 'Communicative effectiveness as predicted by judgements of the severity of learner errors in dictations', in Oller, J.W. (Jr.) and Perkins, K. (eds.). *Research in language testing*, 1980, p.67
(2) Bott, D. and Satithyudhakarn, V. 'Dictation: Easy and accurate evaluation of "Co-co".' *English Teaching Forum*, 24 (3), 42-44 (1986), p.40.
(3) Cziko, G.A. 'Improving the psychometric, criterion-referenced, and practical qualities of integrative testing.' *TESOL Quarterly*, 16 (3), 367-379 (1982).

There is substantial evidence in the interdisciplinary co-operation between research in linguistic pragmatics and reading comprehension to show that reading and listening employ the same underlying processing strategies.[94] Vellutino et al.[95] found significant correlations between reading and listening comprehension in young children and in adults. Of the four language skills, reading performance was found to be the best predictor of listening performance, and vice versa.

Cloze and dictation have been found to reveal similar production errors in writing[96], and a combination of cloze tests and dictation tests have been used effectively in determining general language proficiency.[97] The reason why a dictation and a cloze test (which are apparently such different tasks) intercorrelate so strongly is that both are effective devices for assessing the efficiency of the learner's developing grammatical system, or language ability, or pragmatic expectancy grammar. This underlying skill is overall, or general, language proficiency. Spolsky describes the overall proficiency claim in the following

---

(4) Larsen-Freeman, D. 'From unity to diversity: Twenty-five years of language teaching methodology.' *English Teaching Forum*, 25 (4), 2-10 (1987).

(5) Oller, J.W., Jr. 'Cloze, discourse, and approximations to English', in Burt, K. and Dulay, H.C. *New directions in second language learning, teaching and bilingual education.* (TESOL, Washington, D.C., 1976), p.69.

(6) Oller, J.W., Jr. *Language tests at school.* London, Longman, 1979), p.42..

93 (1) Bacheller, F., ibid, p.69

(2) Oller, J.W., Jr. 'Cloze, discourse, and approximations to English', in Burt, K. and Dulay, H.C. *New directions in second language learning, teaching and bilingual education*, 1976, p.278.

94 (1) Hoover, W. and Gough, P. 'The simple view of reading.' *Reading and writing: An interdisciplinary journal*, 2, 127-160, (1990).

(2) Horowitz, R. and Samuels, S. *Comprehending oral and written language: Critical contrasts for literacy and schooling*, 1987.

(3) Vellutino, F., Scanlon, D., Small, S. and Tanzman, M. 'The linguistic bases of reading ability: Converting written to oral language.' *Text*, 11, 99-133 (1991).

95 Vellutino, F., Scanlon, D., Small, S. and Tanzman, M. 'The linguistic bases of reading ability: Converting written to oral language.' *Text*, 11, 99-133 (1991), pp.107 and 114.

96 (1) Oller, J.W., Jr. 'Cloze, discourse, and approximations to English', in Burt, K. and Dulay, H.C. *New directions in second language learning, teaching and bilingual education*, 1976, p.278.

(2) Oller, J.W., Jr. *Language tests at school*, 1979, p.57.

97 (1) Stump, T.A. *Cloze and dictation tasks as predictors of intelligence and achievement*, 1978.

(2) Hinofotis, F.B. 'Cloze as an alternative method of ESL placement and proficiency testing', in Oller, J.W. (Jr.) and Perkins, K. (ed.), in *Research in language testing*, 1980.

"necessary" condition: "As a result of its systematicity, the existence of redundancy, and the overlap in the usefulness of structural items, knowledge of a language may be characterized as a general proficiency and measured."[98]

Tönnes-Schnier and Scheibner-Herzig[99] compared Oller's procedure and Bacheller's "scale of communicative effectiveness"[100] - where both procedures distinguish between spelling errors and "real" errors - with the relatively much simpler traditional procedure, where different errors are not distinguished. Tönnes-Schnier and Scheibner-Herzig maintain that "this simplified way of marking errors in dictations chosen in accord with the learners' level of structures and vocabulary proves an effective way to rank a class of learners according to their communicative capacities."[101] The fact that a superficial and reductionist procedure such as the traditional procedure of counting surface errors could rank learners according to their communicative capacities shows that reductionist procedures of testing can predict "pragmatic" language, i.e. language that straddles sentences. (Analogously, an eye that is involved in an eye test is no less alive looking at letters on an optician's screen than reading a book or looking at the sunset).

Tönnes-Schnier and Scheibner-Herzig's "surface" (discrete?) findings expose our ignorance. Constructs may not be lurking beneath the surface after all, but staring us in the *face*; or more accurately lurking beneath the surface *and* staring us in the face. The German term *aufheben (sublation)* illustrates the paradox. This term means "to clear away" as well as "to preserve": the simultaneous preservation and transcendence of the structure/meaning

---

98 Spolsky, B. *Conditions for second language learning*, 1989, p.72.

99 Tönnes-Schnier, F. and Scheibner-Herzig, G. 'Measuring communicative effectiveness through dictation.' *International Review of Applied Linguistics*, 26 (1), 35-43 (1988).

100 Bacheller, F. 'Communicative effectiveness as predicted by judgements of the severity of learner errors in dictations', in Oller, J.W. (Jr.) and Perkins, K. (eds.). *Research in language testing*, 1980, p.71.

101 Tönnes-Schnier, F. and Scheibner-Herzig, G. 'Measuring communicative effectiveness through dictation.' *International Review of Applied Linguistics*, 26 (1), 35-43 (1988), p.38.

antithesis. Language (i.e. language structure) has to be cleared away *and* preserved in order to convey its meaning.

3.3.5.2 The dictation tests used in the study

Excerpts from two restored cloze tests of Step 2 of Pienaar's (1984) "Reading for meaning" were used. Step 2 corresponds to Grade 5 and Grade 6 for English-mother-tongue speakers, and Grades 7 to 9 for non-English-mother-tongue speakers. These dictation passages were different from the passages that were used for the cloze test (which were Forms B and D). For the dictation tests, I used the restored texts of Forms A and C of Pienaar's Step 2. Thus, all four passages - two for the cloze test and two for the dictation test - belong to the same level.

I judged the conceptual difficulty of the word sequences to be within the range of academic abilities required of Grade 7 learners who have to use English as the medium of instruction. I decided to pitch the dictation test at the L2 level and not at the L1 level, because I suspected that STEP 3, which was meant for the Grade 7 L1 level, would be too difficult for the Grade 7 L2 subjects. Accordingly, I used the passages of STEP 2, which were aimed at the Grade 7 L2 level.

Dictation Test 1

The fire

We were returning from a picnic up the river when the fire-engine raced past us. Of course we followed it. We hadn't gone far when we saw black smoke pouring from an old double-storey house in the high street. When we drew nearer we saw angry tongues of flame leaping from the downstairs windows. There was already a curious crowd watching the fire, and we heard people say that there was a sick child in one of the upstairs bedrooms. A black cat was also mentioned.

(86 words)

Dictation Test 2

A close call

It was early evening and we were driving at a steady ninety when a small buck leapt into the road about a hundred metres ahead of us. At the last moment it swerved and ran directly towards us. I flicked on the headlights and swerved at the same time. The car slithered to a halt in a cloud of dust, and it was only then that we saw why the buck had changed direction. A number of sinister shapes were hard on the Duiker's* heels. Wild dogs!

(87 words)

**Duiker* is a South African species of small buck.

The reason for the choice of these dictation passages was the same as Pienaar's for the choice of his cloze passages (see section 3.3.1.2) which was "to select lexical and structural items relevant to the demands of the appropriate syllabuses" (Pienaar 1984:3), i.e. relevant to English as the medium of instruction. However, the relevant demands of the appropriate syllabuses cannot be separated from general language proficiency, which is often the hardest part of learning English for ESL learners.

3.3.5.3 Presentation of the dictation tests

The following procedures were used:

1. The degree of difficulty of the texts was regulated by controlling factors such as speed of delivery and length of segments between pauses. The text was read at a speed that preserved the integrative nature of the sequences, while catering for subjects who might not have been able to write at the required speed. The length of sequences between pauses was also sufficient, which satisfied the requirements of both mechanical speed and speed of information-processing.

2. The background noise level was kept to a minimum.

3. The text was presented three times. Once straight through, which involved listening only, a second time with pauses, and a third time without pauses, but at a speed that allowed for quick corrections.

4. And very importantly for this study, more than one presenter was used. This procedure is explained in the next section.

It is normal procedure in a dictation test to use one presenter for all subjects - in this case all four groups. It has been argued that a " dictation can only be fair to students if its presented in the same way to them all"[102], i.e. using only one presenter. In this study, I used "old" tests, but the procedure of presentation was new. If one is using indirect test *methods* such as dictation, this does not mean that one has to stick to "old" *procedures*. One can still try to be exploratory.

The normal procedure in a dictation test is to use one presenter even when subjects are split up into different venues/classrooms. Owing to the exploratory nature of the dictation tests, four presenters (including myself) were used. The presenters then repeated the process on a rotational basis so that each of them presented the two dictation tests to all four groups. The dictation scores used in this study were the scores of the first presentation of each presenter. Thus, *I did not use any scores for the statistical analysis from dictations that had been heard more than once by the subjects*. Table 3.3 shows the procedure of the first rotational presentation.

---

[102] Alderson, J.C., Clapham, C. and Wall, D. *Language test construction and evaluation*, 1995, p.57.

TABLE 3.3

Dictation with Four Presenters: First Presentation of Each Presenter

Presenter 1	*Presenter 2*	*Presenter 3*	*Presenter 4*
Group 1 (Venue 1)	Group 2 (Venue 2)	Group 3 (Venue 3)	Group 4 (Venue 4)
Group 1: N=23	Group 2:N=21	Group 3: N=21	Group 4: N=21

The reason why I chose such a design instead of doing the usual and simple thing of using one presenter was because I wanted to investigate whether different presenters, i.e. different procedures of presentation, would have any significant effect on the results. Three of the presenters were English-mother-tongue speakers, and one was a Tswana-mother-tongue speaker. In order to test for any significant difference in the means between the results of the four different presenters, I did an analysis of variance (One-way ANOVA). The ANOVA results are presented and discussed in section 4.1. The reason why an ANOVA was used is because I had to test whether there was any significant difference between the results of the four procedures of administration: each rater's presentation represents a different procedure.[103] An ANOVA deals with the four sets of data simultaneously, whereas a T-test can only deal with two at a time.

3.3.5.4 Procedure of scoring the dictation tests

Various procedures of scoring dictation are examined and reasons are given for the scoring procedure used in this study.

Cziko[104] uses the procedure of "scoring by segment" using an exact-spelling criterion where a point is awarded for a correct segment on condition that there is no mistake in the segment.

---

[103] Some researchers might say "method" instead of "procedure", but recall that I have reserved the term *method* for "elicitation technique"; what I have called a *test* or a *test method.*

[104] Cziko, G.A. 'Improving the psychometric, criterion-referenced, and practical qualities of integrative testing.' *TESOL Quarterly*, 16 (3), 367-379 (1982).

Bacheller[105] created a scale of communicative effectiveness, where spelling, unlike in Cziko above, was disregarded. In Bacheller's procedure each segment is rated on a scale of 0 to 5 according to how much meaning is understood, e.g. a score of zero indicates that none of the intended meaning of the segment has been captured; a score of 3 indicates that the subject apparently understands the meaning of the segment"; a score of 5 indicates that the meaning is understood. Owing to the fact that the emphasis in Bacheller is on the top-down process of coherence/meaning, this procedure tends to be subjective, especially if only one rater is involved.

Tönnes-Schnier and Scheibner-Herzig[106] compared Bacheller's procedure with the "traditional German method" of dictation (henceforth called the traditional *procedure*) where all words are counted, *including spelling* , one point for each word. Thus, the total score is the number of possible correct words minus the number of errors. (Recall that I have called the manner of adminstration and scoring of a test, a *procedure*, and have reserved the term *method* for *test elicitation technique).*

In the procedure used by Oller[107] and Stump[108] each correct word is worth one point. One point is deducted for each deletion, intrusion or phonological or morphological distortion. Spelling errors, punctuation and capitalisation are not counted. As in the case of the traditional procedure, the total score is the number of possible correct words minus the number of errors. In the traditional procedure of counting errors, the "different kinds of errors" are not distinguished. Thus spelling errors - unlike in Oller's procedure - are lumped together with omissions, intrusions, lexical and grammatical errors.

---

105 Bacheller, F. 'Communicative effectiveness as predicted by judgements of the severity of learner errors in dictations', in Oller, J.W. (Jr.) and Perkins, K. (eds.). *Research in language testing*, 1980.

106 Tönnes-Schnier, F. and Scheibner-Herzig, G. 'Measuring communicative effectiveness through dictation.' *International Review of Applied Linguistics*, 26 (1), 35-43 (1988).

107 Oller, J.W., Jr. *Language tests at school*, 1979, pp.276 and 282.

108 Stump, T.A. 'Cloze and dictation tasks as predictors of intelligence and achievement scores', in Oller, J.W. (Jr.) and Perkins, K. (eds.). *Language in education: testing the tests*, 1978, p.48.

Cziko found that his "exact-spelling segment scoring procedure...was three to four times faster than an "appropriate-spelling word-by-word scoring system".[109] The reason is that one only has to look for one mistake in each segment, whereas in Oller's procedure one has to take into account each and every error. Cziko[110] found a correlation of .89 between his procedure and Oller's procedure. In my procedure I did not have to count every word or mistake but decided on a maximum possible score of 20 (before the test was given), which was determined by the difficulty of the test. In this study I used a variation of the traditional procedure, where errors were subtracted from a possible score of 20 points. One point was deducted for any kind of error, including spelling, and the actual score was deducted from a possible score of 20. This was done because in my opinion this procedure yielded a valid indication of the level of proficiency of *individual* subjects.

If one is only interested in norm-referenced tests, it wouldn't matter what the possible score was, because in norm-referenced tests one is only interested in the relative position of individuals in a group, not with their actual scores. One could then measure the correlation between this procedure and Oller's procedure. If the correlation is found to be high, one could use the shorter procedure. I did a correlational analysis on the dictation tests between Oller's procedure and my variation of the traditional procedure (a possible 20 points). High correlations were found. These are reported and discussed after the ANOVA results in section 4.3).

All the dictation protocols (N=86) were marked by myself. The main reason for this was that the teachers/presenters did not have time for much marking. Recall that the dictation tests were only a part of a large battery of tests. The teachers/ presenters of the dictation tests were involved in the administration of the whole battery. At the end of section 4.2, I discuss the advantages of using a single rater.

---

109 Cziko, G.A. 'Improving the psychometric, criterion-referenced, and practical qualities of integrative testing.' *TESOL Quarterly*, 16 (3), 367-379 (1982), p.378.

110 Ibid., p.375.

3.4 Summary of Chapter 3

The sampling procedures, and the structure, administration and scoring procedures of the tests were described. Various procedures of administration and scoring in the different tests were appraised to show the reason for the procedures chosen for this study.

Although subjects are divided into an a L1 and L2 groups, the two groups should be treated as a composite group in the correlational analyses (Chapters 4 and 5). The reason for this is that the following conditions were the same for all the subjects:

(1) The admission criteria to the School.

(2) The English proficiency tests and their administration (this investigation).

(3) The academic demands of the School.

(4) The treatment they were given at the School. What is relevant to the statistical rationale of this investigation is not the fact that the entrants to the School had received different treatment prior to entering the School, where some may have been disadvantaged, but only the fact that the all entrants received the same treatment after admission to the School.

(5) The proportion of L1 and L2 learners (as I have defined these labels) was similar from year to year at the School.

All five conditions show that the 1987 Grade 7 sample represented subjects who came from the same population of Grade 7 learners at the school from year to year, specifically from 1980 to 1993, irrespective of their origin and whether they were divided into "L1" and "L2" groups. (More about this in section 4.6).

The next chapter presents the results of the battery of proficiency tests.

CHAPTER 4

Results of the proficiency tests

4.1 Introduction

There are two basic kinds of statistics: descriptive and inferential. Descriptive statistics provides summary data of a whole array of data. Examples of summary data are means, standard deviations, analysis of variance, and reliability and validity coefficients. Inferential statistics indicates the extent to which a sample (of anything) represents the population from which it is claimed to have been drawn.[5]

The population in this study refers to the Grade 7 entrants at MHS from its inception in 1980 up to the present day. *This study is particularly interested in the L2 learners at MHS and the wider population of Grade 6 Tswana-mother-tongue speakers at DET schools* in the North West Province of South Africa who were admitted to Grade 7 at MHS from 1980 onwards.

This chapter provides the descriptive statistics of the English proficiency tests. In the next chapter, descriptive and inferential statistics are provided of the prediction of academic achievement. This chapter also deals with inferential issues regarding the L1 and L2 groups, which has an important bearing on the notion of "levels" of proficiency, a central notion in this study.

From the outset, I need to point out that there was a significant difference between the means of the L1 and the L2 groups. This has important inferential implications, for, if the L1 and L2 groups belong to separate *populations* (in the statistical sense of the word), one couldn't consider the two groups as a uniform group for correlational purposes. I shall argue that the L1 and L2 groups do not belong to separate *populations*.

---

5 Davies, A. *Principles of language testing*, 1990, p.16.

This chapter contains the following sets of results:

(1) Reliability coefficients of all the tests.

(2) Analysis of variance (one-way ANOVA) of the dictation tests only.

(3) Validity coefficients of all tests.

(4) Means and standard deviations of the L1 and L2 groups on all the tests.

4.2 Reliability coefficients

Two kinds of reliability measurements were used:

- The Pearson r correlation formula measures the parallel reliability between two separate, but equivalent, i.e. parallel, tests. The tests involved are the two cloze tests, the two dictation tests and the two essay tests. The procedure used for calculating the reliability of parallel (forms of) tests is to administer the tests to the same persons at the same time and to correlate the results as indicated in the following formula:

rtt = rA,B (Pearson r formula)

where

rtt = reliability coefficient,

and

rA,B = the correlation of test A with test B when administered to the same people at the same time.

- The Kuder-Richardson 20 (KR-20) formula splits a single test in half, and treats the two halves of the test as if they are parallel tests.[6] The tests involved are the error recognition test and the mixed grammar test. The following parallel and KR-20 reliability coefficients are reported:

---

[6] Bachman, L.F. *Fundamental considerations in language testing*, 1990b, p.172.

TABLE 4.1

Reliability Coefficients of All the Tests

Test Type	N	Type of Reliability	n (items)	Coefficient of Reliability
Cloze (1 and 2)	86	Parallel	10X2	.80
Dictation (1 and 2)	86	Parallel	-	.91
Essay (1 and 2)	86	Parallel	-	.90*
Grammar	80	KR-20	50	.91
Error Recognition	80	KR-20	50	.90

* The average scores of four raters for each of the two essay tests were correlated with each other to produce the parallel reliability coefficient between Essay 1 and Essay 2.

Considering the two different genres of essay task it might be questioned in what sense these two tasks are "parallel". As noted, performance was consistent across the two tasks (as will be evidence further by the similar means between the two tasks).
This raises the knotty question of what it means to say that tests are parallel. How does one ensure that two tests are parallel. For example, it is very difficult to ensure parity of content, not only in "integrative" tests such as cloze, dictation and essay but also in "discrete-point", or "objective", tests. This is so because all tests no matter how "objective" they look are subjective.[7] Accordingly, it is better to speak of integrative and discrete-point *formats* than integrative or discrete-point *tests*. From this position it is not a big step to take to speak of parallel *scoring*, because it is only in the sense that test *scores* are found to be "parallel" that we can talk of *tests* being parallel. Statistics becomes not only sensible but indispensable in this matter: (1) if the "parallel" tests ranked individuals in a group in a similar way, i.e. if there were to be a high correlation between the tests, and (2) if there were to be no significant difference between the means of the two tests, this would be pretty good evidence that the tests of similar formats and scores were parallel tests. Table 4.2 shows that there was no

[7] Yeld, N. *Communicative language testing.* Report on British Council Course 559 offered at Lancaster University from 8 September to 20 September 1985, (published in) 1986, p.32.

significant difference between the means within each of the three pairs of integrative tests, because the t Stat was less than the Critical value.

TABLE 4.2

Means and Standard Deviations of Parallel Tests (N=86)

	CLOZE		DICTATION		ESSAY	
	1	2	1	2	1	2
MEAN	49	48	44	49	45	43
STANDARD DEV.	26	25	33	33	17	18
t Stat	.1790		-1.093		1.345	
t Critical two-tail	1.974		1.974		1.974	

Interrater reliability was only a factor in the essay tests, because the essay tests were marked by more than one rater. With regard to essay tests, Henning[8] points out that "because the final mark given to the examinee is a combination of the ratings of all judges, whether an average or a simple sum of ratings, the actual level of reliability will depend on the number of raters or judges." According to Alderson,

> *[t]here is considerable evidence to show that any four judges, who may disagree with each other, will agree* ***as a group*** *with any other four judges of a performance. (It was pointed out that it is, however, necessary for markers to agree on their terms of reference, on what their bands, or ranges of scores, are meant to signify: this can be achieved by means of a script or tape library).*[9] *(Original emphasis).*

If interrater reliability is measured in this way, this would make complex statistical procedures of calculating interrater reliability unnecessary. One would simply compute the average of the four raters' scores for Essay 1 and Essay 2, respectively, and then

[8] Henning, A. *A guide to language testing*, 1987, p.82.

[9] Alderson, J.C. 'Report of the discussion on Communicative Language Testing', in Alderson, J.C. and Hughes, A. *Issues in language testing*, 1981b, p.61

compute the parallel reliability coefficient between the average of Essay 1 and the average of Essay 2. This was the procedure used to compute the reliability coefficient of the essay tests. It may be argued that the rationale for "One would simply compute..." (in the previous sentence) is unclear because I stated that Pearson correlation would also be used in measuring the reliability of the essay tests. I only "simply compute" the average of each subject's four scores given by the respective four raters, but don't abandon the Pearson correlational measurement for the parallel reliability of the two sets of (averaged individual) scores. As shown in Table 4.2, the parallel reliability is .90 (see also the note in Table 4.2).

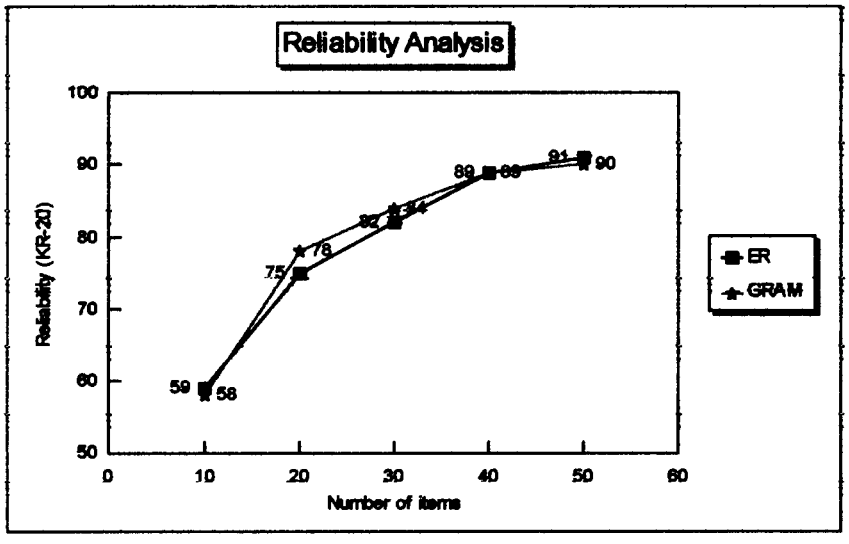
A reliability analysis was also done in a cumulative fashion of the error recognition test (ER) and the mixed grammar test (GRAM). The reason why the reliability coefficients were computed in this cumulative fashion was to find out the minimum items required to ensure high reliability.

As shown in Figure 4.1, I used the first 10 items (items 1-10), then the first 20 items (items 1-20), then 30 items (items 1-30), and so on. The KR-20 formula was used to compute the reliability coefficients.

FIGURE 4.1

Comparison between the reliability coefficients of ER and GRAM

N=80



The reliability coefficients of both ER and GRAM have an almost identical pattern. An important statistical truth is illustrated by these reliability data, namely that less than 40 items are not likely to produce satisfactory reliability coefficients, i.e. of .90 or higher, for discrete, objective, items. In multiple-choice grammar tests a reliability coefficient between .90 and .99 is usually required to be considered a reliable test, whereas in tests such as an essay test, a reliability coefficient of .90 is considered high.[10]. There is also a tapering off of the reliability coefficient after 40 items until it reaches a point of *asymptote*, where any increase in items does not result in a significant increase in reliability.[11]

Some of the reliability calculations may appear odd, because if it is true, as I have shown, that 40 items produce low reliability coefficients, then why (1) use 10 items for CLOZE, and (2) why use the *parallel* method of reliability for CLOZE and the KR-20 (split-half) method of reliability for GRAM and ER. The answer to these questions requires an answer to another question: (3) Why is the parallel reliability of CLOZE with only 10 items as high as .80 while the KR-20 reliability of GRAM and ER with ten items is a low .60. Answers to these questions lie in the relationship between grammatical/linguistic competence (sentence meaning) and discourse competence (pragmatic meaning) and the continuum of "integrativeness" (see section 2.5).

One doesn't merely look at the *format* of a test to decided whether it is a "discrete-point" test. One looks at what the test is testing. As pointed out earlier, it is possible to write few words, i.e. a "discrete-point" format, as in a cloze test (or as in "natural" settings) and still be testing "communicative competence", or "pragmatic" language. In the case of GRAM and ER, each of these tests consists of unrelated "objective" items; that is, there is no "pragmatic" connection between them. The KR-20 formula is used to measure the reliability of objective items. The Pearson r formula is used to measure

---

10 Perkins, K. 'On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability', *TESOL Quarterly*, 17 (4), 651-671 (1983).

11 Henning, A. *A guide to language testing*, 1987, p.78..

parallel reliability of tests at the pragmatic end of the integrative continuum: the cloze, dictation and essay tests. It is true that sometimes the KR-20 formula is used to measure the reliability of cloze tests, because some authors, e.g. Alderson, maintain that many cloze tests test "low order" skills that "in general [relate] more to tests of grammar and vocabulary... than to tests of reading comprehension".[12] This is not a point of view shared by many other authors (see section 3.3.1.1).

A split-half reliability method (of which the KR-20 formula is a sophisticated version) on an "integrative" test such as a cloze test, dictation test, or an essay test may not be a good idea for the very important reason that the two halves of such tests do not consist of clusters of comparable items, owing to their "pragmatic" nature, i.e. items are not completely independent, i.e. they all hang together. If items hang together as in integrative tests one may not have to worry about searching for an "empirical basis for the equal weighting of all types of errors"[13], as Cziko believes it necessary to do for all tests.

If there is only one test form, e.g. as in Cziko's[14] dictation test, one cannot use the parallel reliability method, but one can measure reliability using other methods such as the test-retest method. With regard to the cloze tests, the parallel reliability coefficient of .80 is not only quite acceptable for a "pragmatic" test, but also very good for only ten deletions.

I would like to add a few remarks on rater consistency, or rater reliability. I argued that in the dictation test, presenters and groups were not confounded (i.e. each group had its respective presenter). Therefore, it was legitimate to subsequently do an ANOVA of the four groups/presenters/presentations. One may accept this rationale but still be concerned about the rater reliability of the dictation test (and the cloze test for that

---

12 Alderson, J.C. 'The cloze procedure and proficiency in English as a foreign language.' *TESOL Quarterly*, 13, 219-227 (1979), p.225.

13 Cziko, G.A. 'Improving the psychometric, criterion-referenced, and practical qualities of integrative testing.' *TESOL Quarterly*, 16 (3), 367-379 (1982), p.369.

14 Ibid., 1982.

matter) because only one rater was involved, and not four as in the essay test. The question, therefore, is whether the scoring procedures in these tests lack evidence of consistency of application owing to the fact that there was only one rater (myself). This should not be a problem in the dictation test, because I didn't have to worry about distinguishing between spelling and grammatical errors (which can be a serious problem[15]), owing to the fact that only wrong forms of words, intrusions and omissions were considered in my marking procedure. In the cloze tests special care was taken that all acceptable answers were taken into account. The error recognition test and mixed grammar test had only one possible answer. The answers to the latter two tests were provided by the test compilers.

4.3 Analysis of variance of the dictation tests

Recall that a separate presenter was used for each of four groups of subjects. There were four presentations on a rotational basis (Table 3.3). An analysis of variance (ANOVA) was conducted on the - and I must stress this point - *first* presentation to test for any significant difference between the four presenters' procedure of presentation. As I explained in section 3.3.5.4, no scores for the statistical analysis were used from dictations that had been heard more than once by any group in the rotation of presenters. Accordingly, presenters and groups were not confounded. In other words, Presenter 1 coincided with Group 1, Presenter 2 with Group 2, and so on.

The ANOVA showed (Table 3.3) that there was no significant difference between the four groups, i.e. the null hypothesis was *not* rejected. If the null hypothesis had been rejected this would have demonstrated that there was a significant difference between the four presenters' procedures of presentation. Under these circumstances the use of the dictation in a corrrelational analysis with other tests would be invalid because it would have been illegitimate to combine the four dictation groups into a composite group. The results of the ANOVA are reported below.

---

[15] Alderson, J.C. and Clapham, C. 'Applied linguistics and language testing: A case study of the ELTS test.' *Applied Linguistics*, 13 (2), 149-167 (1992), p.46.

TABLE 4.3

Analysis of Variance of the Dictation Tests with First Presentation

Dictation 1

Group 1: N=23 ; Group 2: N=21 ; Group 3: N=21 ; Group 4: N=21					
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-ratio	Critical[1] Value
Between Groups	4383.292	3	1461.098	1.324	2.72
Within Groups	90509.731	82	1103.777		
Total (corrected)	94893.023	85			

[1] **If the F-ratio is less than the critical value, then the null hypothesis is not rejected. The critical value is determined from the degrees of freedom between groups and within groups (Stoker 1974:10).**

Dictation 2

Group 1: N=23 ; Group 2: N=21 ; Group 3: N=21 ; Group 4: N=21					
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-ratio	Critical Value
Between Groups	2439.24	3	813.08	.714	2.72
Within Groups	93356.108	82	1138.489		
Total (corrected)	95795.349	85			

Dictation Average

Group 1: N=23 ; Group 2: N=21 ; Group 3: N=21 ; Group 4: N=21					
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-ratio	Critical Value
Between Groups	3018.094	3	1006.031	.935	2.72
Within Groups	88233.359	82	1076.017		
Total (corrected)	91251.453	85			

The ANOVA results show that the null hypothesis was not rejected, which means that there was no significant difference between the four presenters/groups. Therefore, the dictation test scores would not have been significantly different if the same presenter was used for all four groups.

Statistical results, in this case the ANOVA, cannot tell us *why* there was no significant difference between the different presenters. To make qualitative comparisons between the results obtained from different presenters, one would have to examine the results of the four different presenters on the *first* presentation given to the four groups of subjects. I stress the first presentation because the possibility exists that the dictation passages would become progressively easier with each subsequent presentation.
I examined a random selection of protocols (from both the L1 and L2 groups) to find out whether there was any difference in the quality of output using different presenters. This analysis of protocols was a lengthy enterprise and would thus take up too much space if reported in this study. It is fully reported elsewhere (Gamaroff, forthcoming). The intention is not at all to treat qualitative data in a cavalier fashion. The point is that this study's main emphasis is on quantitative data. Qualitative data are not at all ignored, however (see section 4.8ff). What I shall do here is summarise the conclusions of the qualitative analysis of the dictation tests:

The dictation passages (Pienaar's[16] restored [or "unmutilated"] cloze passages) for the Grade 7 subjects were intended for the Grades 5 to 7 L2 levels and for the Grades 5 and 6 L1 levels. Consequently, the L1 group would be expected to do well, even if the presenter's prosody were unfamiliar. As the statistics will show (section 4.5, Table 4.5), the L1 group did well and the L2 group did badly.

Recall (section 3.3.5.4) that I used a variation of the traditional procedure, where errors were subtracted from a possible score of 20 points. One point was deducted for any kind of error, including spelling, and the actual score was deducted from a

[16] Pienaar, P. *Reading for meaning: A pilot survey of (silent) reading standards in Bophuthatswana*, 1984.

possible score of 20. This was done because I believed that this procedure would yield a valid indication of the level of proficiency of *individual* subjects. If one was only interested in norm-referenced tests, it wouldn't matter what the possible score was, because in norm-referenced tests one is only interested in the relative position of individuals in a group, not with their actual scores. One could then measure the correlation between this procedure and Oller's procedure. If the correlation is found to be high, one could use the shorter procedure. A correlational analysis was done on the dictation tests between Oller's procedure and my variation of the traditional procedure (a possible 20 points). High correlations were found: .98 for the first dictation passage, and .89 for the second dictation passage. The reason for the high correlations is probably the following:

The word forms of the L2 group were so deviant that I regarded them as grammatical errors. In the L1 group, in contrast, the scores were very high, which meant that no scores were subtracted for spelling or for grammatical errors. As a result, in both groups, spelling had no significant effect, which meant that very few marks were subtracted for spelling. This means that whatever possible score I chose, the correlations between my procedure and Oller's procedure would have been high; hence the high correlations reported in the previous paragraph. (Correlation is not concerned with whether scores are equivalent between two variables, but only with the common variance between two variables, i.e. whether the scores "go together"; see section 5.3). So, if Oller's dictation procedure yielded *relatively* higher scores than my procedure, this doesn't effect the correlation.

One can explain the difference in performance between the L1 and L2 groups in terms of the difference between the information-processing strategies used by low-proficiency and high-proficiency learners. When we process language, we process in two directions: bottom-up from sound input and top-down from the application of the

cognitive faculties[17]. With regard to the dictation test in the study, the words were highly predictable for the L1 group, and therefore this group did not have to rely totally on the sound input. The opposite was the case for the L2 group, where there was almost a total reliance on the bottom-up process of sound recognition. In other words, native listeners or listeners with high proficiency "can predict the main stresses and can use that fact to 'cycle' their attention, saving it as it were, for the more important words.[18] It should be kept in mind, however, that bottom-up processes from sound input plays a major role at all levels of proficiency, not only at the low levels. The difficulties experienced by the L2 group did not only have to do with lexical lacunae: there is much more to knowing a word than knowing the various meanings it may have. To master a word one also needs to know its form, its frequency of use, its context, its relationship to other words.[19] Problems can occur in any of these areas. This applies to all the tests of the test battery.

4.4 Validity coefficients

The singular term *test* will be used to refer to the means of the two cloze tests (CLOZE), of the two essay tests (ESSAY) and of the two dictation tests (DICT). With the single mixed grammar test (GRAM) and the single error recognition test (ER),

---

17 (1) Kelly, P. 'Lexical ignorance: The main obstacle to listening comprehension with advanced foreign language learners.' *International Review of Applied Linguistics*, 29 (2), 134-149 (1991).

(2) Rumelhart, D. E. *Introduction to human information processing*, 1977.

(3) Samuel, A.G. 'Phonemic restoration: Insights from a new methodology.' *Journal of Experimental Psychology*, 110, 474-494 (1981).

18 Suenobu, M., Kanzaki, K., Yamane, S. and Young, R. 'Listening comprehension and the process of information acquisition by non-native speakers of English.' *International Review of Applied Linguistics*, 24 (3), 239-248 (1986), p.244.

19 (1) Kelly, P. 'Lexical ignorance: The main obstacle to listening comprehension with advanced foreign language learners.' *International Review of Applied Linguistics*, 29 (2), 134-149 (1991), p.138.

(2) Laufer, B. 'Why are some words more difficult than others? - Some intralexical factors that affect the learning of words.' *International Review of Applied Linguistics*, 28 (4), 293-307 (1990), pp.294-295.

there are five "tests" altogether. Table 4.4 shows the validity coefficients of the English proficiency tests. The numbers in the top row refer to the tests that appear next to the corresponding numbers in the extreme left hand column.

TABLE 4.4

Validity Coefficients of English Proficiency Tests

$p < .01$

	1.	2.	3.	4.	5.
1. CLOZE (N=86)	1.00				
2. DICT (N=86)	.77	1.00			
3. ESSAY (N=86)	.81	.80	1.00		
4. GRAM (N=80)	.82	.80	.81	1.00	
5. ER (N=80)	.84	.79	.81	.84	1.00
TOTAL (N=80)	.87*	.85*	.84*	.88*	.89*

* *Corrected for part-whole overlap. Part-whole overlap occurs when an individual test score is correlated with the total score of all the tests of which its score is a part. In such a situation, one would not be measuring two variables that are separate from one another; which would result in part-whole overlap between the individual test and the total score. This part-whole overlap would increase the correlation, thus giving an inaccurate picture.*

The high validity coefficients are impressive and perhaps unusually high. For this reason the raw data and computations (using the statistical programme "Statgraphics") were rechecked twice. High validity coefficients, however, are not unusual between these tests. Validity coefficients, unfortunately, do not give a close-up picture and thus often need to be supplemented by other descriptive data such as frequency distributions, means and standard deviations. The next section shows these other descriptive data where a comparison is made between the L1 and the L2 groups.

4.5 Descriptive results of the L1 and L2 groups

The difference between the performance of the L1 and L2 groups are shown. The following data are provided:

1. Means and standard deviations (Table 4.5).
2. A Frequency distribution (Table 4.6).

The following measures appear in the tables:

1. CLOZE - Average of Cloze tests 1 and 2 (N=86).
2. DICT - Average of Dictation tests 1 and 2 (N=-86).
3. ESSAY - Average of Essay tests 1 and 2 (N=86).
4. GRAM - Mixed grammar test (N=80).
5. ER - Error recognition test (N=80)

A statistically significant as well as a substantial difference was found between the means of the two groups as shown in Table 4.5.

TABLE 4.5

Means (M) and Standard Deviations (STD) for the L1 and L2 groups

	L1(N=39)		L2 (N=37)		T-test p<.05	
	M	SD	M	SD	t Stat	t Critical
CLOZE	65	14	26	16	11.29	1.993
DICT	71	17	16	19[1]	13.63	
ESSAY	56	11	29	10	11.74	
	L1(N=43)		L2 (N=37)			1.996
GRAM	77[2]	12	44[2]	12	10.22	
ER	50[2]	18	12[2]	11	9.32	

[1] In the L2 group DICT STD is more than DICT M because of the large number of zero scores.

[2] Adjusted for guessing (to be discussed shortly)

When the t Stat is more than the t Critical value this shows that there is a significant difference between the two groups. (According to Nunan[20], when two sets of scores have substantially different means or standard deviations, it is not necessary to use a T-test to test for a significant difference between means). The frequency distributions are shown in Table 4.6.

TABLE 4.6

Frequency Distribution of all the Tests

	DICT		ESSAY		CLOZE		GRAM		ER	
	L1	L2	L1	L2	L1	L2	L1	L2	L1	L2
10	0	20	0	2	0	9	0	0	1	18
20	0	7	0	3	0	7	0	2	1	11
30	0	4	0	10	1	7	0	9	2	3
40	4	0	4	12	0	8	0	5	11	5
50	5	2	1	8	8	5	2	6	7	0
60	6	3	19	2	15	0	2	10	7	0
70	6	1	14	0	9	1	8	3	8	0
80	13	0	8	0	9	0	10	2	4	0
90	11	0	3	0	7	0	17	0	2	0
100	4	0	0	0	0	0	4	0	0	0
TOT	**49**	**37**	**49**	**37**	**49**	**37**	**43**	**37**	**43**	**37**

The L2 group did very poorly on the dictation and the error recognition tests, less poorly on the cloze and essay tests, and best of all on the grammar test. The L1 group did best on the dictation and the grammar tests, while in the other tests, the order of increasing difficulty are the cloze, the essay and error recognition tests. In Chapter 5 the frequency distributions are analysed in more detail in relation to the prediction of academic achievement

---

20 Nunan, D. *Research methods in language learning*, 1992, p.29

Does the significant difference between the L1 and L2 scores above mean that these two groups come from different *populations* and therefore should not be treated as a composite group in a correlational analysis? I examine this question in section 4.6.

The multiple-choice format is vulnerable to guessing. Sometimes it is recommended that scores be adjusted for guessing, as in Bloor et al.'s[21] GRAM and ER that was used in this study. Guessing was taken into account in the study, which meant that in the mixed grammar (GRAM) test, a score of 88% was reduced to 85%; a score of 64% to 55%; and a score of 40% to 25%. Thus the person who has more, loses proportionately less. The score of 40% in GRAM is used to show how to calculate the adjustment for guessing:

$$\frac{\text{100 minus Actual Score (40\%)}}{\text{Number of options in item (4 options)}} = \frac{60}{4} = 15\%$$

$$\text{40\% (actual score) minus 15\%} = 25\%$$

As shown in the last line of the equation, the result of the first line (15%) is subtracted from the actual score of 40% to give an adjusted score of 25%. The greater the number of options, the less the adjustment, because the test would be more difficult. ER has five options, and so the adjustment is less than for GRAM. Suppose the actual ER score was also 40%, as in the GRAM example above. The adjusted score of ER would be 28%, which is 3% higher than the adjusted score of GRAM:

$$\frac{\text{100 minus Actual Score (40\%)}}{\text{Number of options in item (5 options)}} = \frac{60}{5} = 12\%$$

$$\text{40\% (actual score) minus 12\%} = 28\%$$

One cannot prove that someone is guessing, and without proof, it might be argued that one would be penalising non-guessers as well as guessers: "in multiple-choice formats,

[21] Bloor, M., Bloor, T., Forrest, R., Laird, E. and Relton, H. *Objective tests in English as a foreign language*, 1970.

guessing affects scores and, though statistical procedures are available to correct for it, they necessarily apply indiscriminately whether or not a learner actually has guessed".[22] However, the logic behind correction for guessing is not indiscriminate even though it affects everybody. As shown in the examples above, the less one knows the more the *likelihood* of guessing. Although one cannot be sure who is guessing, the rationale of the adjustment for guessing is based on what we know about learning and test performance. The key point of logic in the adjustment for guessing is that the lower the original score, the greater the possibility that one is guessing. If the scores are not adjusted for guessing, this would of course affect the ranges of scores. But in this study, I am not interested so much in the absolute values of these ranges as in the relative values: the L1 group relative to the L2 group.

I now focus on the cloze results because these cloze tests have been used elsewhere and have produced a solid body of results which one can compare with the results in this study. I shall also introduce cloze data from another school (to be described shortly). Recall that Pienaar[23] tested a variety of learners from different schools. including Bantu speakers living sub-economic settlements in the environs of Mmabatho (category 4b; see section 3.3.1.2). Pienaar used the label "III" for the group that I called 4b). Many of the parents of category 4b were illiterate or semiliterate and were either unemployed or semi-employed. The sample at MHS did not contain learners that belonged to this category, as shown by the occupations of the parents of the L2 group in Table 4.7 below.

---

22 Ingram, E. 'Assessing proficiency: An overview on some aspects of testing', in Hyltenstam, K. and Pienemann, M. *Modelling and Assessing second language acquisition*, 1985, p.237.

23 Pienaar, P. *Reading for meaning: A pilot survey of (silent) reading standards in Bophuthatswana*, 1984, p.13.

TABLE 4.7

Occupation of Parents of the L2 Subjects

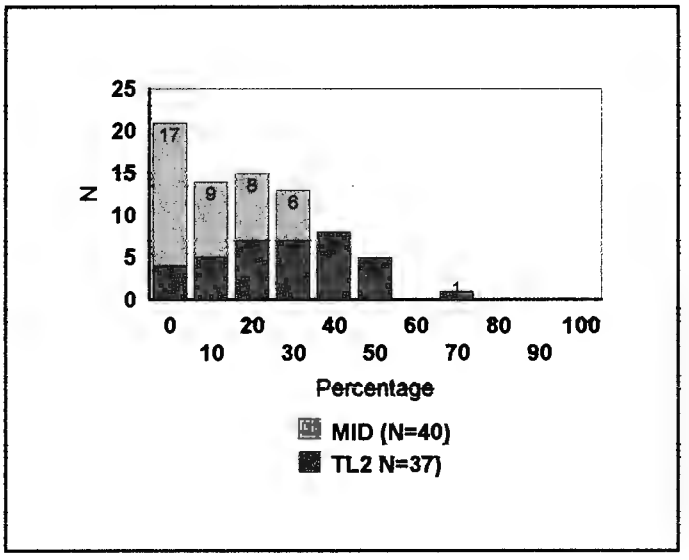
	FATHER'S OCCUPATION	*MOTHER'S OCCUPATION*
1	PRISON SERVICES	PRISON SERVICES
2	FARMER	TEACHER
3	SPORTSMAN	NURSE
4	?	?
5	SUPERVISOR	NURSE
6	SHOP OWNER	SHOP ASSISTANT
7	?	?
8	PRODUCER	TYPIST
9	?	BUYER
10	VET	HOUSEWIFE
11	?	UNEMPLOYED
12	FARMER	OFFICER
13	BROKER	NURSE
14	UNIVERSITY ADMINISTRATOR	LECTURER
15	BUSINESS	TUTOR SISTER
16	UNEMPLOYED	NURSE
17	POLICEMAN	TYPIST
18	INSURANCE	HOUSEWIFE
19	?	?
20	COLLEGE RECTOR	TEACHER
21	ELECTRICIAN	STUDENT
22	BEER HALL OWNER	HOUSEWIFE
23	CLERK	TEACHER
24	CHIEF	HOUSEWIFE
25	TEACHER	TEACHER
26	INSPECTOR OF EDUCATION	TEACHER
27	BUSINESS	TEACHER
28	?	?
29	BOOKSELLER	TEACHER
30	PRIEST	HOUSEWIFE
31	SCHOOL INSPECTOR	TEACHER
32	DOCTOR	HOUSEWIFE
33	FARMER	TEACHER
34	TEACHER	NURSE
35	TEACHER	TEACHER
36	BUILDER	PRINCIPAL
37	TEACHER	TEACHER
38	MINSTER OF RELIGION	TEACHER

The L2 group contains a good number of parents who work in the education field (see highlighted occupations). One cannot infer that the children of educators are usually advantaged, because in South Africa, education was one of the few professions open to blacks.[24]

To make the data more comparable I included in the investigation a Middle school (Grades 7 to 9) situated in the environs of Mmabatho that accommodated learners similar to Pienaar's category 4b[25]. The sample from this school, referred to as MID, consisted of 40 Grade 7 learners. Learners at MID come from many primary schools in the area, owing to the fact that there are far more Primary schools than Middle schools in the area. Figure 4.2 compares the cloze test frequency distributions of the MHS L2 group with the Middle School (MID).

FIGURE 4.2

A Comparison of the MHS L2 Group with the Middle School (MID) on the Cloze Tests



[24] Human, L. and Hofmeyer, K. *Black managers in South African organisations*, 1985, p.5.

[25] Members of Category 4b comprise Bantu speakers living in sub-economic settlements in the environs of Mmabatho (see section 3.3.1.2).

The MID school results are very similar to Pienaar's[26] category of sub-economic learners, namely, his category III (which I have called category 4b). By comparing MID with the MHS sample we see that the L2 group at MHS - poorly as it has done - is better than the MID group. The MID group is comparable with Pienaar's "at risk" group: indeed at high risk. The MHS L2 group is also at high risk, but the MID group is much worse.

4.6 The L1 and L2 groups: Do these two levels represent separate populations?

Section 4.4 showed that there were high correlations between the discrete-point and integrative tests of the test battery. This study is not only concerned with statistical concepts such as correlation but also with the problem of assigning levels of language proficiency to learners. The discussion to follow is relevant to both these issues:

It is only after the test has been performed on the test-bench that it is possible to decide whether the test is too easy or too difficult. Furthermore, if there are L1 and L2 subjects in the same sample, as is the case with the sample in this study, one needs to consider not only whether the norms of the L1 and the L2 groups should be separated or interlinked but also how to ensure the precise classification of the L1 and L2 subjects used for the creation of norms.

As far as the correlational analysis was concerned, I interlinked the L1 and L2 groups and treated them as a composite group. But I also separated the L1 and L2 groups in order to find out whether there was a significant difference between the means of the two groups. *If a significant difference were <u>not</u> to be found between the L1 and L2 groups, this would militate against the construct validity of the tests, because this would mean that the L2 group, who should be weaker than the L1 group, was just as proficient as the L2 group.* Under such conditions, we would have no idea what we were testing.

---

[26] Pienaar, P. *Reading for meaning: A pilot survey of (silent) reading standards in Bophuthatswana*, 1984, p.21.

The question is whether it is legitimate to treat the L1 and L2 subjects as a composite group (for a correlational analysis) as well as two separate groups (for comparing the means between the L1 and L2 groups). One may object that one cannot do both; that one cannot interlink groups and also separate them. I shall argue that one can.

As shown in Table 4.4 the correlations between the different tests were high. The means and standard deviations (Table 4.5), however, show that there was a significant difference between the L1 and L2 groups. Can one, accordingly, maintain that if there was a significant difference between the L1 and L2 groups that these two groups belong to separate *populations*, and thus argue that the correlations were artificially inflated by combining samples that represent two separate populations?[27] A discussion of this question raises the question of the logicality of dividing the subjects into L1 and L2 groups. Is this division arbitrary or does it have a cogent theoretical rationale on which one can make the inference that the L1 and L2 groups represent different populations?

There are two distinct issues, which are also related, namely, levels of proficiency and correlations. The logic of correlation, which is based on a bell curve distribution[28], is that tests that do not have a reasonably wide spread of scores (high achievers and low achievers) could give a false picture because tests that have a large spread of scores around the mean have more likelihood of being replicable[29], owing to the fact that in a representative sample of human beings there is likely to be a wide range of ability.

---

27 This was an objection that was made by a seasoned professional in the field. This is the reason I decided to explain in detail the rationale at issue. (Often when a reader finds in the academic literature phrases such as "one may object", "it may be argued", the reader suspects that the writer is writing about someone - either someone who remains unpublished or someone who should remain anonymous - who actually did object or argue against the writer's position: (someone who, the writer believes, should have known better).

28 As mentioned in section 2.2, the variability in ability between individuals obeys a "bell-curve" distribution, as in the case of nature as a whole. The "bell-curve" or "normal" distribution is the foundational principle of psychometrics.

29 Davies, A. *Principles of language testing*, 1990, p.5.

This does not mean that it is not possible to have a high correlation with a narrow spread, or a low correlation with a wide spread, but it is more likely that a correlation would be higher with a wide spread of scores, say 0% to 80%, than a narrow spread, say 40% to 80%. The sample in this study represented the Grade 7 population at the school throughout the years.

In the assessment of levels of proficiency I separated the high achievers from the low achievers in the sample because they could be distinguished - unsurprisingly - as those who took English as a First Language and English as a Second Language, respectively.

I discuss briefly the theory of investigating the difference between groups. Consider the following example:

In South Africa there are many immigrants from different countries for whom English is a foreign language. If one tested and compared the English proficiency of a group of Polish immigrants and Chinese immigrants and found no significant difference between these two groups, one wouldn't be surprised, because one would probably conclude that there was a wide spread of scores in both groups. If a significant difference were to be found, one may be curious to know why the one national/ethnic group did worse or better than the other.

Replace the Polish and Chinese immigrants with two other groups, the L1 and L2 groups in this study. A significant difference was found between these two groups, but this is not surprising at all, because it is to be expected that the group taking English as a First Language subject (the L1 group) would be better at English than the group taking English as a Second Language subject (the L2 group): assuming that the subjects (test takers) in the sample made a reasonable choice of which group to belong to. (Recall that learners at MHS initially decided themselves whether they were belonged to L1 or L2. In most cases they had a good idea where they belonged).

Accordingly, it is quite logical that there would be a significant difference between the L1 and L2 groups, or levels. To get a clearer grasp of the issue of the respective levels of proficiency of the two groups and the "separate populations" question one has to examine whether:

(1) The reliability *aspects* were the same for the L1 and L2 subjects, e.g. the same tests, same testing *facets* and same testing *conditions*, etc. (see section 2.9). This was so.

(2) The composition of the sample, i.e. the proportion of L1 and L2 learners, was similar from year to year at MHS. This was so. In other words, the 1987 Grade 7 sample represented the *population* of Grade 7 learners at MHS from year to year, specifically from 1980 to 1993.

One would also look at whether there were differences in:

(3) Admission criteria for the L1 and L2 groups.

(4) The background, or former *treatment* of L1 and L2 learners before they entered the school.

(5) What one expected from the L1 and L2 learners.

(6) The treatment they were given in the same education situation.

All the above points except for (4) apply to both the L1 and L2 subjects. I discuss (40):

MHS endeavours to provide disadvantaged learners with the opportunity to learn in an advantaged school situation. In the validation of the sample, the notion of dis-advantage is important. In South Africa the term *disadvantage* often bears the connotation of "consciously manipulated treatment" meted out by apartheid. Treatment can have the following two connotations: (i) consciously manipulated treatment in an empirical investigation and (ii) the long-term treatment - be it educational, social, economic, cultural or political - of human beings in a non-experimental life situation.

What is relevant to the statistical rationale of this investigation is not the fact that the entrants to MHS had received different treatment *prior* to entering MHS, where some may have been victims of apartheid and others not, but only the fact that all entrants received the same treatment after admission to MHS. I am not implying that their background experience is inconsequential as far as the teaching situation - past (at former schools) or future (at MHS) - is concerned, but only that all entrants were expected to fulfil the same academic demands. I discuss later the role of language background, specifically the role of English input.

The vast majority of the 1987 Grade 7 intake had high Grade 6 scores from their former schools . This was the main reason why many of them were admitted to MHS. The disadvantaged group and the advantaged group *both* consisted of *high-scoring* entrants revealed by the Grade 6 school reports. Accordingly, it appeared that all the entrants were extremely able, whether they came from an advantaged or disadvantage background. Now, suppose one found that (i) high Grade 6 scores (from former schools) were obtained by both the L1 and L2 groups but that (ii) while high English proficiency test scores were obtained by the L1 group, low English proficiency test scores were obtained by the L2 group. The findings showed that both these facts were so. This does not mean, however, that the L1 and L2 subjects belong to different populations. What it shows - on condition that the English proficiency tests were valid and reliable, which the findings show was the case - is the true nature of the population, namely, a wide spread of scores.

So, although it seemed, from the good Grade 6 reports of all entrants at MHS (L1 and L2 entrants) from year to year, that MHS only admitted high achievers, the reality was that MHS admitted a mixture of academically weak learners (who were generally disadvantaged) and academically strong learners (who were generally advantaged), as was the case with the 1987 Grade 7 sample. Further, learners at MHS received the same treatment.

The statistical analysis should be kept distinct from educational, social, economic, cultural, political and other deprivations that pre-existed admission to MHS. The principal issue in this study is what learners are expected to do after admission, where all learners are called upon to fulfil the same academic demands, except for the language syllabuses, namely English, Tswana, Afrikaans and French, and where all are required to use English as the medium of instruction.

If it is true that former academic achievement (Grade 6 in this case) is the best predictor of subsequent achievement[30], it would follow that many of these entrants should have had at least a reasonable standard of academic ability. What happened in fact was that although almost all of the 1987 Grade 7 entrants (L1 and L2) obtained high *Grade 6* scores on English achievement and on their aggregates, many of the L2 entrants (who were mostly disadvantaged learners) obtained low scores on the English proficiency tests. For this reason, the sample turned out to be, as far as the English proficiency tests were concerned, a representative mixture of weak and strong learners, i.e. a random sample. This fact is crucial to the validation of any sample, whose essential ingredient is randomness.

I am arguing, therefore, that the L1 and L2 groups do not represent separate populations: they are merely a mixture of weak and strong performers, where it is only logical that weak subjects would prefer to belong to the L2 group than to the L1 group and that the L2 group would also do relatively worse than the L1 group on the English proficiency tests. It turned out that there was a clear distinction between the L1 and L2 groups. Most of the L2 group did poorly and most of the L1 group did relatively much better on the tests, hence the significant difference in the means between the two groups.

---

30 Hale, G.A., Stansfield, C.W. and Duran, R.P. *TESOL Research Report 16*, 1984.

In the traditional distinction between L1 and L2 learners, these two kinds of learners differ only in so far as L2 learners aspire to reach the L1 level. The difference, therefore, between L1 and L2 learners lies in the different levels of mastery. And that is what tests measure within a sample that represents a population: it measures which members are strong, which are weak. If the tests are too difficult for the L2 group or too easy for the L1 group this does not mean that the tests are invalid, i.e. that they have been used for the wrong purpose, if the purpose is to distinguish between weak and strong learners. *One does not look at the actual scores as far as construct validity is concerned but at whether the tests distinguish between weak and strong learners.* Oller elucidates (he is talking about one learner and one task, while I am talking about many learners and several tasks: I make the necessary adjustments in Oller to suit the context):

> *It is probably true that the [tasks were] too difficult and therefore [were] frustrating and to that extent pedagogically inappropriate for [these students] and others like [them], but it does not follow from this that the [tasks were] invalid for [these learners]. Quite the contrary, the [tasks were] valid inasmuch as [they] revealed the difference between the ability of the beginner, the intermediate, and the advanced [learners] to perform the [tasks].*[31]

(Section 4.7 elaborates on the comparison between test scores and the comparison between groups).

If one is or believes one is weak at English, one would sensibly prefer to take English as a Second Language, if one had a choice: one did have a choice at MHS. This is not to say that if one were good at English one would not take English as a Second Language, owing to the fact that somebody good at English could obtain higher marks taking English as a Second Language than taking English as a First Language.

---

31 Oller, J.W., Jr. *Language tests at school*, 1979, p.391.

Accordingly, those Tswana speakers in the L1 group at MHS who later changed to English Second Language could have done so not because they were weak at English but in spite of the fact that they were good at English.

4.7 Comparing groups and comparing tests

It might be argued that measuring the difference in means between groups apportions equivalent scores to each item, and accordingly, does not take into account the relative level of difficulty of items. I suggest that the relative difficulty of items is not important in a language proficiency test, but is indeed in a diagnostic test, which has remediation as its ultimate purpose. With regard to proficiency tests, one is concerned with a specific level (e.g. elementary, intermediate and advanced) for specific people at a specific time and in a specific situation. Within each level there is a wide range of item difficulty. To attain a specific level of proficiency one has to get most of the items right - the difficult and the easy ones. In sum, the different bits of language have to hang together, which is what we mean by general, or overall, language proficiency. As pointed out earlier (section 2.5), the controversy is about which bits do and which bits don't hang together.

We now come to a very important issue. What does a score of 60% on a test for an L2 learner in this study mean? An answer to that question requires a distinction between (1) the comparison between tests and (2) the comparison between groups: the L1 and L2 groups.

As a preliminary I refer to the relationship between norm-referenced tests and criterion-referenced tests. The former is only concerned with ranking individuals in a group and not, as in the case of criterion-referenced tests, with individual scores achieved in different tests. So, in norm-referenced tests one is interested in correlations, which is concerned with how individuals are ranked in a group on the tests involved in the correlation, and not with whether individuals achieved equivalent

scores within a group. The latter is the concern of criterion-referenced tests. But, of course, one needs both kind of information to get a empirically-based idea of language tests.[32] One has to be careful, however, when one compares tests.

In this study I have been comparing different kinds of tests and contrasting different groups of learners. The main focus is on the difference between groups, and thus there was no explicit and sustained attempt to contrast the scores between the different tests. This was deliberate. If one is going to compare the results between different tests, e.g. the dictation test and the cloze test, extreme caution is required, because such comparisons could lead to false conclusions. This is not to say that such comparisons are not useful; they can be very useful, but when one makes such comparisons, one must be aware of the parameters involved. Scores reveal nothing and surface errors reveal little about why a particular score was awarded. One has to look at the construction of the test, i.e. what, why and who is being tested and doing the testing, and how it is being tested. All these parameters are related to the scales of measurement that one uses.[33] Consider the following measurement scales, especially the ratio scale. The ratio scale could be confused with the other scales:

- Nominal scale (also called categorical scale). This is used when data is categorised into groups, e.g. gender (male/female); mother tongue (English/Tswana).

- Ordinal scale. One could arrange proficiency scores from highest to lowest and then rank them, e.g. first, second, etc.

- Interval scales. One retains the rank order but also considers the distances (intervals) between the points, i.e. the relative order between the points on the scale.

- Ratio scales are interval scales with the added property of a true zero score, where the "points on the scale are precise multiples, or ratios, of other points on the

---

[32] Cziko, G.A. 'Some problems with empirically-based models of communicative competence.' *Applied Linguistics*, 5 (1), 23-37 (1984).

[33] Brown, J.D. 'Statistics as a foreign language - Part 2: More things to consider in reading statistical language studies.' *TESOL Quarterly*, 26, (4), 629-664 (1992).

scale".[34] Examples would be the number of pages in a book, or the number of learners in a classroom. If there were 200 pages in a book, 100 pages would be half the book.

Taking these scales into account, consider the proficiency tests of the study:

- *The cloze test.* To be considered proficient enough to cope in a higher grade in the cloze tests, one should obtain a score of at least 60%. (The mean scores of the L1 and L2 groups for the cloze test were 66% and 26%, respectively). (See Table 4.5).

- *The essay test.* A score over 60% in the essay, in contrast to 60% on the cloze test in this study, would be considered a good score. A score of 40% on an essay test or on any test is not half as good as a score of 80%. As far as essay tests are concerned, 80% would be an excellent score, while 30% would be a poor score. But, poor is not half of excellent. (The mean scores of the L1 and L2 groups for the essay test were 56% and 29%, respectively). (See Table 4.5).

- *The dictation test.* I used a score of a possible 20 points; one point for every correct word. But if I had made the score out of 86 or 87 points (the dictation passages consisted of 86 and 87 words, respectively), where every word counted one point, a score of 60% would mean that 40% of the words in the dictation passage would be wrong. It is hardly likely that a dictation protocol with a score of 60% marked in this way would be comprehensible. Accordingly, an *individual's* score of 60% on a dictation test would not mean the same thing at all as 60% on the cloze and essay tests. (The mean scores of the L1 and L2 groups for the dictation test were 71% and 16%, respectively). (See Table 4.5).

- The *error recognition and the mixed grammar tests.* These test scores were adjusted for guessing. If one adjusts for guessing, one must take this into account.

---

[34] Brown, J.D. Ibid., p.633

To sum up, it is the "relative difference in proficiencies"[35] between *learners* of high ability (in this case the L1 group) and low ability (in this case the L2 group) and *not* the equivalence in scores between the *tests* that determines the reliability and construct validity of the tests.

4.8 Error analysis and rater reliability

Although the average of four or even three raters may be a reliable assessment of a "subjective" test such as an essay test, it is usual in the teaching situation to have only one rater available, who is the teacher involved in setting the test. If there are two raters available it is generally only the teacher who sets the test and is overall in charge of the test who has the time or inclination to do a thorough job. The problem of rater consistency is an extremely serious problem in assessment. The nub of the problem is one of interpretation, an issue that fills innumerable tomes in the human sciences, especially during this "postmodern" era. This is what is involved:

> *Logically prior to any question of the reliability and validity of an assessment instrument is the question of the human and social process of assessing...This is a radically interpersonal series of events, in which there is an enormous, unavoidable scope of subjectivity - especially when the competences being assesses are relatively intangible ones to do with social and personal skills, or ones in which the individual's performance is intimately connected with the context.*[36]

It is the interpretation, or judgement, of errors that is the main problem in language testing. Ashworth and Saxton (in their quotation above) are concerned with the lack of equivalence in judgements and scores between raters. The subjectivity question in the battery of tests of this study remains a problem in the essay test. I tried to solve the problem by using four raters. But, in most testing situations only one rater and at most two raters are available. I would like to expand on the issue of rater reliability, because

---

35 Oller, J.W., Jr. *Language tests at school*, 1979, p.394.

36 Ashworth, P. and Saxon, J. On competence. *Journal of Further and Higher Education*, 14 (2), 23 (1990).

this seems to be the major problem in the assessment of "subjective" tests such as essay tests. Error analysis is brought into the picture.

The qualitative analysis of errors and quantitative measurement are closely related in issues of interrater reliability. In this section I discuss more theory, in this case the uses and limitations of error analysis, which serves as a background to the examination of a detailed practical example of the uses and limitations of error analysis and quantitative measurement. I begin the discussion by assessing the value of the quantitative procedures used in this study in relation to the lack of qualitative procedures used so far:

One may feel that the linguistic substance of individual errors obtained in an error analysis has more bite than reductionist "number-crunching"[37] and that consequently this study has overreached itself by limiting itself to something as insubstantial as a statistical investigation. One might want to see additional analyses of a qualitative nature of the proficiency tests, especially of the integrative tests, where writing output is involved. Such a desire is understandable because scores by themselves don't illuminate the linguistic substance behind the numbers owing to the fact that similar scores between raters do not necessarily mean similar judgements, and different scores between raters do not necessarily mean different judgements.[38]

Error analysis can be useful because it provides information on the progress made towards the goal of mastery and provides insights into how languages are learnt and the strategies learners employ. Concerning learning strategies, the making of errors is part of the learning process. (An error analysis need not involve a "linguistic" analysis. For example, in an error analysis of writing one could look for cohesion errors, but if

---

37 Yeld, N. 'Communicative language testing and validity.' *Journal of Language Teaching*, 21 (3), 69-82 (1987), p.78.

38 Gamaroff, R. 'Language, content and skills in the testing of English for academic purposes.' *South African Journal of Higher Education*, 12 (1), 109-116 (1998b).

one were to examine the noun-to-verb ratio in individual protocols, this would not be an error analysis but a linguistic analysis).

This study is mainly concerned with norm-referenced testing. So, to include a linguistic/error analysis of the tests, would, besides being far too long and ambitious a project, go beyond the objectives of this study. The problem would be which tests to use in such an analysis, and how long such an analysis should be. Naturally, qualitative analysis is very important, but in the empirical part of the study I focus on quantitative data. As far as qualitative data are concerned, relevant to this study is examining the problems of error analysis as it relates to rater reliability. As mentioned a few paragraphs earlier, I shall be using a detailed concrete example later on in this section to examine this problem. But first some theory.

Often mother-tongue proficiency is advocated as an absolute yardstick of language proficiency, but, as Bachman and Clark point out "native speakers show considerable variation in proficiency, particularly with regard to abilities such as cohesion, discourse organisation, and sociolinguistic appropriateness."[39] As a result, theoretical differences between testers can affect the reliability of the test. Raters who know the language well and even mother-tongue speakers can differ radically in their assessments of such pragmatic tasks as essay tasks. That is why different raters' scores on a particular protocol are often incommensurate with their judgements. Owing to these problems it is virtually impossible to define criterion levels of language proficiency in terms of actual individuals or actual performance. Bachman and Clark suggest that such levels must be defined abstractly in terms of the relative presence or absence of the abilities that constitute the domain.[40] But again this doesn't solve the problem because the difficulty is how to apply the definition to concrete situations of language behaviour.

---

[39] Bachman, L.F. and Clark, J.L.D. 'The measurement of foreign/second language proficiency.' *American Academy of the Political and Social Science Annals*, 490, 20-33 (1987), p.29.

[40] Ibid., p.30.

Another problem is the representativeness of specific errors. In previous research[41] I did an error analysis of Tswana speakers' English but did not establish statistically whether the errors I was dealing with were common errors, e.g. *cattles (a plural count noun in Tswana "dikgomo") and *advices (a plural count noun in Tswana "dikgakoloko). Under such circumstances one can be duped into believing that errors are common if one comes across them a few times, which may only create the *feeling* that they are common. Error analysis under such circumstances could indeed become merely an idiosyncratic - and mildly interesting - "stamp collection".

Another example: Bonheim[42], coordinator of the Association of Language Testing in Europe, gives an example of a test taker who had done very well on a multiple-choice test, but in one of his/her few incorrect items of the test had circled an option that was an unlikely answer. Bonheim suggested that one should try and find out why this highly proficient learner had circled this option. This idiosyncratic example surely cannot contribute anything to the general principles of error analysis, i.e. tell us whether the error is common enough to warrant a time-consuming investigation. In *proficiency* testing, one is not looking for idiosyncratic errors but for general errors. In diagnostic testing, of course, the situation is quite different because one focuses on both individual and general errors, because the main aim of a diagnostic test is therapy, not finding out the level of a person's present ability, which is what proficiency tests are about.

Obviously, the different types of tests, e.g. proficiency, diagnostic, aptitude and achievement, are related, but it is important to keep their main purposes distinct: otherwise there would be no point in creating these distinctive categories. For

---

41 Gamaroff, R. *Native Language Transfer in Tswana Speaker's English.* Unpublished MA thesis. Potchefstroom: Potchefstroom University for Christian Higher Education, 1986, pp.68-77.

42 Bonheim, H. *Roundtable on language testing.* European Society of the Study of English (ESSE) conference, Debrecen, Hungary, September, 1997.

example, an itemised analysis can reveal the relative strengths and weaknesses of the different groups in different parts of the language. Such an analysis can be used as a diagnostic tool at the beginning or end of a specific course of instruction, or, in this case, as a measurement of specific points of proficiency. However, without quantitative procedures, the data one gathers remain unconvincing. For example, consider the percentage error on individual items of the error recognition test for the Non-Tswana L1 sub-group (NTL1) and the L2 group in Table 4.8:

TABLE 4.8

Error Recognition Test: Percentage Error

	N=20	N=37
Item	**NTL1**	**L2**
8	75	88
19	30	88
25	50	93
27	40	93
34	35	73
40	30	76
44	45	90
42	65	88
46	30	80
49	40	76

Care must be taken in the interpretation of Table 4.8. The ***higher*** the scores, i.e. the ***higher*** the percentage ***error***, the more ***difficult*** the item. The information in Table 4.8 reveals the similarities and differences between groups on each item. For example, in item 8, the difference between the NTL1 and the L2 group is not substantial. Item 8 is given below.

A B C D

Item 8. Both Samuel and I/are much more richer/than we/ used to be. Correct - E.

Answer: B

In item 19 the NTL1 group does substantially better than the L2 group.

A B C D

Item 19. Some believe that/a country should be ruled/by men who are/too clever than ordinary people. Correct - E. Answer: D

ESL learners often confuse intensifier forms such as "*too* clever", "*very* clever" and "*so* clever" and comparative forms such as "*more* beautiful" and "clever*er*." The error in Item 19 is a double confusion between intensifier and comparative forms probably caused by false generalisation, or false analogy, from the English forms.

A quantitative analysis of errors was also found useful in identifying the "replacement language" subjects, which helps in establishing levels of proficiency between learners. Recall (Note 2, Chapter 3) that a "replacement language" is a language that becomes more dominant than the mother tongue, usually at an early age, but is seldom fully mastered, as in the case of some of the Coloured and Indian subjects in the sample, who belong to the Non-Tswana L1 (NTL1) sub-group. (Bantu speakers, of course, can also have replacement languages).

The "replacement" language subjects could be identified, to a certain extent, by the very low scores they obtained in the tests. An examination of particular errors made by those I suspected of being "replacement language" subjects increased the accuracy of the identification of these subjects. (These were Indians and "coloureds" who had been using English as a medium of instruction from the beginning of primary school).

Mother-tongue speakers do make or cannot recognise several grammatical errors. Consider the percentage error of the NTL1 group on items 8 and 19 given above - 75 and 30 respectively. In item 8, it is possible that 12-year old English-mother-tongue

speakers would not be able to recognise that segment B ("are much more richer") is an error. In item 19, it is more likely that 12-year old English-mother-tongue speakers would recognise segment D ("too clever than ordinary people") as an error. But there is still a slight possibility that an English mother-tongue speaker would not recognise the error. There are, however, certain errors that all English-mother-tongue speakers would recognise. Thus if such a mistake were to be made by subjects that one suspected of being "replacement language" subjects, one would be almost certain that they indeed were. For example, consider the percentage error of item 27 of the error recognition test.

A B C

Item 27. As I have now studied/French for over three years/I can be able to/make

D

myself understood when I go to France. Correct - E. Answer: C

Percentage Error of Item 27

NTL1	*L2*
N=20	N=38
40	93

In item 27 segment C ("can be able") is a notorious error among South African black ESL users. The L2 group had a percentage error of 93. What is interesting is that 40% of the NTL1 group got this item wrong. It is highly likely that an English-mother-tongue speaker would recognise this error. Such an example is good, if not absolute, evidence that those in the NTL1 group who didn't recognise this error were "replacement language" subjects.

4.8.1 Rater reliability among educators of teachers of English

Assessment is concerned with the control of (1) rater judgements and (2) scoring techniques. Actually, the definition of assessment is very much like the definition of

rater reliability, which could be defined as the control (1) of rater judgements and (2) of rater scoring techniques. Both assessment and rater reliability - where the latter is not only subsumed under assessment but can almost be identified with it - are concerned with reconciling authentic subjectivity and objective precision. Rater reliability is particularly important in "subjective" tests such as essay tests, where there exist fluctuations in judgements between (1) different raters, which is the concern of interrater reliability, and (2) within the same rater, which is the concern of intrarater reliability. This study focuses on interrater reliability only.

The first step that conscientious raters take to control their judgements and scoring techniques is to try and establish what the test in question is meant to be measuring. Yet, in spite of discussions and workshops on establishing common criteria such as content relevance and grammatical accuracy, there remain large differences in the relative weight that raters attach to different criteria.[43]

In previous research on interrater reliability I examined the assessments of lecturers of English for Academic purposes (EAP) and of Science lecturers on first-year-university student essays.[44] These students were from the University of the North West in Mmabatho (ex-University of Bophuthatswana). The topic was the "Greenhouse Effect". Comparisons were firstly made within the EAP group of raters and within the Science raters, and secondly between the two groups of EAP and Science raters. The findings showed a wide range of scores and judgements within each group as well as between the two groups of raters. In this section I report on some research on interrater reliability that was based on a English workshop on quantitative measurement in

---

[43] (1) Bradbury, J., Damerell, C., Jackson, F. and Searle, R. 'ESL issues arising from the "Teach-test-teach" programme', in Chick, K (ed.). *Searching for relevance: Contextual issues in applied linguistics*, 1990.

(2) Lumley, T. and McNamara, T. 'Rater characteristics and rater bias: implications for training, Language Testing', 12, 55-21 (1995).

(3) Santos, T. 'Professors' reactions to the academic writing of nonnative-speaking students.' *TESOL Quarterly*, 22 (1), 69-90 (1988).

[44] Gamaroff, R. 'Language, content and skills in the testing of English for academic purposes.' *South African Journal of Higher Education*, 12 (1), 109-116 (1998b).

language testing that I conducted at a conference of the National Association of Educators of Teachers of English.[45] The participants consisted of a group of 27 South African educators of teachers of English. These educators taught at a diverse selection of universities, technikons and colleges of education. Three participants were excluded because they were each the fifth member of a group, and in the computations I limited groups to four members for reasons that will be explained shortly. Thus, the results of 24 of the 27 participants of the workshop were taken into account, six groups of raters in all.[46] The protocols that had to be rated belonged to the same MHS subjects as those in the main investigation of this study.

Although I was unable to obtain the judgements of the MHS raters (except my own, of course), I hope to compensate for this by providing the judgements of the NAETE raters in this recent study, and by so doing enlarging the rater base from four raters as in the MHS study to 24 raters as in this study.

Recall that at MHS, three raters, who were also the Grade 7 teachers, and myself were involved in the administration and marking of the essay test. Owing to practical obstacles, such as the limited time that these teachers could devote to the marking of the tests, they did not provide judgements on specific criteria such as content and

---

[45] (1) Gamaroff, R. *Workshop on quantitative measurement in language testing.* National Association of Educators of Teachers of English (NAETE) Conference, East London Teacher's Centre, September, 1996c.

(2) Gamaroff, R. *Interrater reliability, the bug of all bears: Report on the 1996 NAETE Workshop on quantitative measurement in language testing.* National Association of Educators of Teachers of English (NAETE) conference " Training teachers for the South African context, Potchefstroom College of Education, September 17-18, 1998c.

(3) Gamaroff, R. (In Press). 'Rater reliability in language assessment: the bug of all bears.' *System*, 28, 31-53.

[46] Recall that rater reliability and concurrent validity are different notions (see section 2.9). Rater reliability has to do with the consistency between raters' judgements on one test (method), e.g. an essay test. Concurrent validity, in contrast, has to do with the correlation between two or more different tests e.g. a dictation test and an essay test.

grammatical accuracy. They merely gave a score based on a global impressions. These limited data were informative from a norm-referenced point of view, because they distinguished well between weak and strong learners, and had high interrater reliability, but they could not show the relationship between scores and judgements because there were no judgements given.

A test can have high interrater reliability, where raters give equivalent scores but this does not necessarily mean that these scores represent what they are supposed to measure, i.e. that the test is valid. To illustrate, if all raters of an essay believe that spelling should be heavily penalised and, accordingly, give equivalent scores in terms of spelling, the interrater reliability would be high. The question, however, is whether spelling should be the most important criterion. Or, raters may differ in the importance they attach to different criteria. Therefore, similar scores between raters do not necessarily mean similar judgements, and, also, different scores between raters do not necessarily mean different judgements.

The following procedures are followed concerning the data collected from the NAETE workshop:

(1) A comparison between individual rater's scores.

(2) A comparison between the average scores of six groups of raters, four in a group.

(3) An examination of the relationship between judgements and scores of individual raters.

As mentioned, there were originally 27 raters in the NAETE workshop. These were divided into six groups of four or five per group: Groups A to F. Only four raters in each group were used because the average score of any reasonably competent four raters has been found to be reliable, the rationale being that the problems of subjective

judgements will be neutralised using the average of four judges.[47] Consequently, I excluded three of the 27 raters, who were designated by the number 5 in their three respective groups: these raters were C5, D5 and E5.

Raters were asked to assess two essay protocols: one from the MHS L2 group (Protocol 1) and one from the MHS L1 group (Protocol 2). Protocol 2 was chosen at random, while Protocol 1 was chosen because of the interesting spelling errors, where I wanted to see how raters judged these highly visible errors.

The essay question consisted of choice between three topics: describe how to (1) clean a pair of shoes, (2) make a cup of tea or (3) cover a book. The content of these topics should be far easier to assess than the controversial topic of the "Greenhouse Effect"[48], which was the topic in the previous research mentioned above. The protocols are now presented followed by the frequency distribution of the scores on each protocol.

Protocol 1 (Grade 7 L2 learner)

How a school book is covered

> If you cover a book you need several things such as a brown cover, a plastic cover and selotape ect. First you open your couver and put the book on the corver. You folled the cover onto the book and cut it with the sicor and folled it again. You stick the cover with the selotape so that it mast not come out of the book. Same aplies to when you cover with a plastic cover. Then you book is corved well.

---

[47] Alderson, J.C. 'Report of the discussion on Communicative Language Testing', in Alderson, J.C. and Hughes, A. *Issues in language testing: ELT Documents III.*, 1981b, p.61.

[48] Gamaroff, R. *Workshop on quantitative measurement in language testing.* National Association of Educators of Teachers of English (NAETE) Conference, East London Teacher's Centre, September, 1996c.

Protocol 2 (Grade 7 L1 learner)

How a school book is covered

You need a roll of paper cover or plastic cover, A pair of scissors some sellotape. You put the book on the paper or Plastic and cut the length it is better if about 5 cm of cover was left from the book. You cut it into strips You fold the cover over the book. You then put strip of sellotape to keep them down. Then you put plasitic paper over it and stick it down. Then you can put your name and standard.

Participants in the workshop were requested to (1) work individually, (2) spend about one and a half minutes on each protocol[49], and (3) give an impressionistic score based on criteria such as topic relevance, content and grammatical accuracy, and (4) give reasons for their judgements on the criteria they specified.

4.8.1.1 Results of the NAETE workshop

Figures 4.3 and 4.4 show the frequency distribution of the individual scores awarded by the 24 raters for the L2 protocol and the L1 protocol, respectively. A nine-point scale was used; 0 to 1 point = totally incomprehensible, 2 points = hardly readable; 3 points = very poor; 4 points = poor; 5 points = satisfactory; 6 points = good; 7 points = very good; 8 points = excellent; 9 points = outstanding. *Ratings* can refer to points or judgements. To avoid confusion, I shall refer to scores and judgements, and not use the term *rating*.

---

[49] Hughes, A. *Testing for language teachers*, 1989, p.86.

Figure 4.3

Frequency Distribution of the Scores Awarded by the 24 Raters on Protocol 1 (L2)

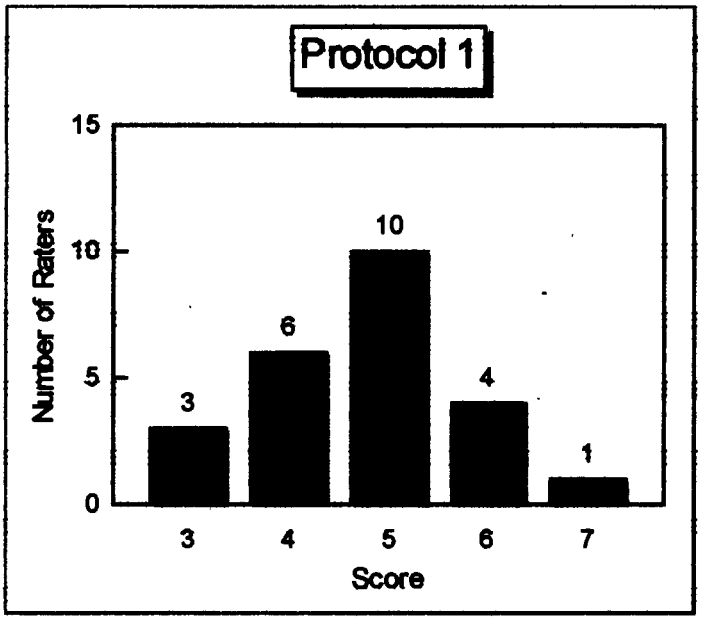
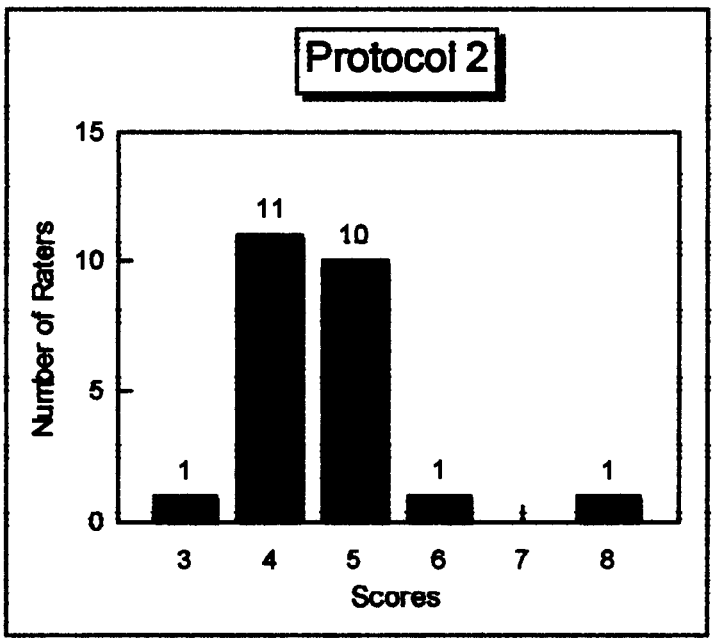


Figure 4.4

Frequency Distribution of the Scores Awarded by the 24 Raters on Protocol 2 (L1)



Although Figure 4.4 (the **L1** protocol) has a wider range of scores (3 to 8) than Figure 4.3 (3 to 7; the **L2** protocol), there is a far more variability in Figure 4.3. Consider Table 4.9 below, which shows the average score for each of the six groups of raters: Groups A to F. Also included in Table 1 is the average score of the four raters at MHS who were involved in the original test battery. These scores appear after Group F.

Table 4.9

NAETE Workshop and MHS: Average Scores on Protocols 1 and 2 of Groups of Raters

Groups of Raters	Protocol 1	Protocol 2
Group A	4.5	4.3
Group B	5.3	4.0
Group C	4.8	4.3
Group D	5.0	4.5
Group E	4.8	5.8
Group F	4.3	5.0
MHS	3.5	5.5

I was surprised that the scores of Groups A, B, C and D for Protocol 1 (L2) were higher than those for Protocol 2 (L1) because I considered Protocol 2 to be of higher quality. In the original research at MHS, I had awarded a score of 4 for Protocol 1 and score of 6 for Protocol 2.

The data in Table 4.9, useful as they are, do not provide enough information. We also need to compare the judgements on different criteria. I did not specify the criterion of "spelling" in the workshop because I suspected that many participants would give prominence to spelling errors, and I wanted to verify this suspicion without having to make spelling a explicit criterion, and thereby influencing the raters to make judgements on spelling. Raters were explicitly asked, however, to take into account the criteria of "topic relevance", "content" and "grammar". Most of the raters didn't distinguish between topic relevance and content, so I have subsumed the two criteria under content.

Figure 4.5

% Positive Judgements, No judgements & Negative Judgements for Protocol 1 (L2)

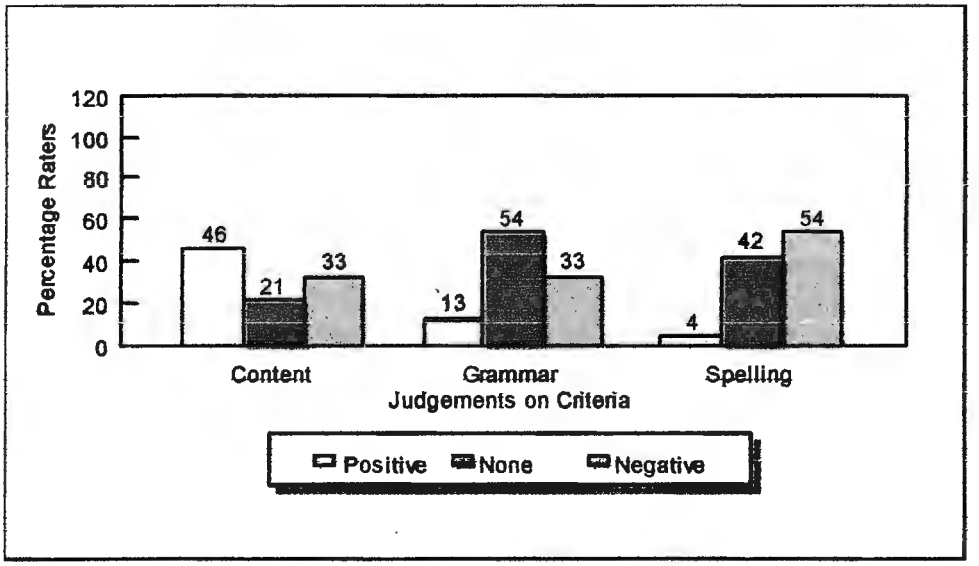


Figure 4.6

% of Positive Judgements, No judgements & Negative Judgements for Protocol 2 (L1)

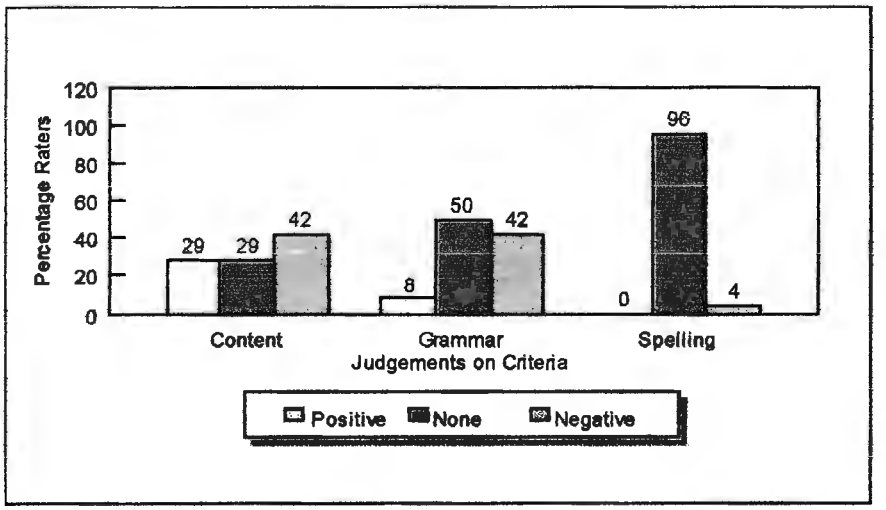


Figure 4.7

Percentage of Negative Judgements for Protocol 1 (L2)

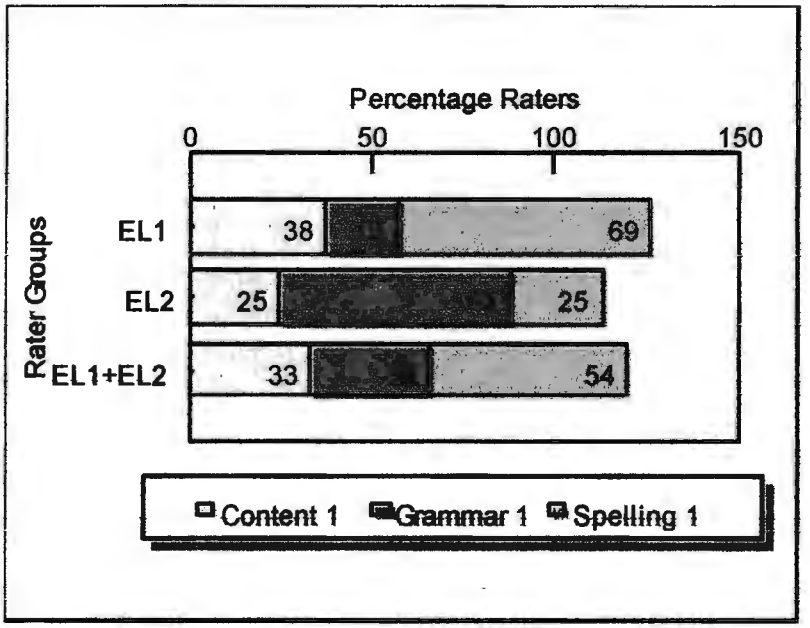
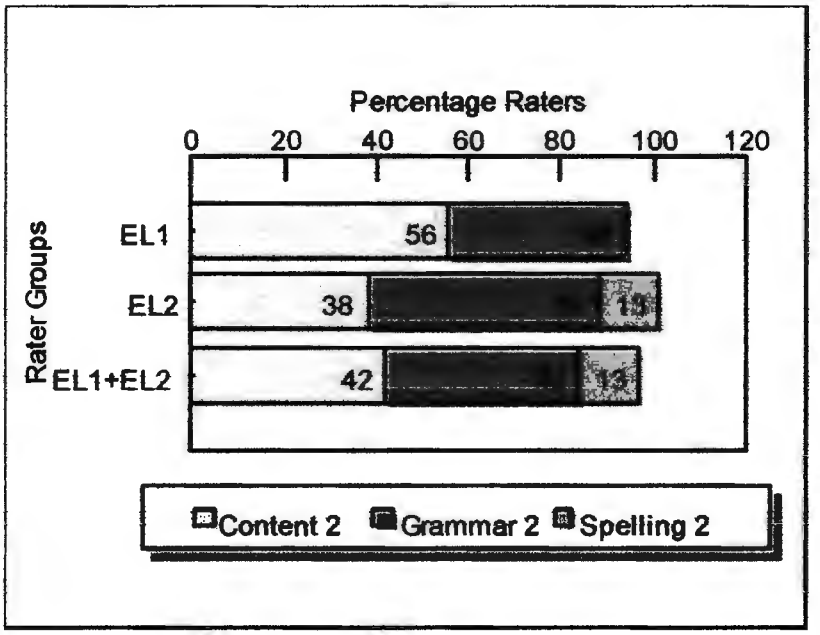


Figure 4.8

Percentage of Negative Judgements for Protocol 2 (L1)



4.8.1.2 Discussion of the NAETE workshop results

If no judgement was given on a particular criterion - indicated by "None" in Figures 4.5 and 4.6 - I assumed that the judgement for the unmentioned criterion was not negative or that the errors were not serious enough to warrant a specific mention. The fact that raters don't mention specific criteria is just as revealing as positive and negative judgements, for if one rater doesn't worry about spelling, for example, and another does, this could have a significant effect on the rating, and could even mean the difference between a pass or a fail.

I now compare the negative judgements of the EL1 and EL2 raters, which are shown in Figures 4.7 and 4.8. Before I do I would like to highlight that there were 16 EL1 participants but only eight EL2 participants at the NAETE workshop. This proportion of EL1 to EL2 educators of teachers of English is not indicative of South Africa as a whole, because there are far more EL2 than EL1 educators of English in South Africa. (I do not have precise data on this matter). The reason why the NAETE conference of 1996 had this unrepresentative proportion is possibly due to the fact that the conference was held in the Eastern Cape where there are fewer EL2 than EL1 educators of English, and also fewer tertiary institutions that cater for student teachers of English than exist in other areas such Gauteng (the Johannesburg-Pretoria region).

The overall picture of Protocol 1 (Figure 4.7) shows that 33% of all the participants (EL1 plus EL2) gave negative comments on content and grammar while 54% considered spelling to be a significant problem. There was a substantial difference between the negative judgements of EL1 and EL2 on grammar (19% and 63%, respectively) and on spelling (69% and 25%, respectively), where the judgements of EL1 are almost the reverse of EL2: what EL1 considers to be grammatical errors, EL2 considers to be spelling errors. It would have been interesting to find out which errors in the protocols raters considered spelling errors and which ones grammatical errors. Consider the highlighted errors in Protocol 1 (the whole protocol is repeated for easy

reference, where I have highlighted different kinds of errors, either in bold, italicised or underlined):

> If you cover a book you need several things such as a brown cover, a plastic cover and selotape ect. First you open your **couver** and put the book on the **corver**. You *folled* the cover onto the book and cut it with the sicor and folled it again. You stick the cover with the selotape so that it *mast* not come out of the book. Same *aplies* to when you cover with a plastic cover. Then you book is **corved** well.

One should not equate "spelling" with "deviant form". I judged the three italicised errors *folled and *aplies as spelling errors and *mast as a phonological error. The other deviant word forms are more difficult to specify. Are the different deviant forms of "cover" to be labelled as spelling or phonological errors? Compare these forms with the following deviant forms from Oller's[50] chapter on *dictation* tests:

rope - *robe
expected - *espected
ranch - *ransh
something - *somsing, *some think

Phonological errors could be of two kinds: (1) the deviant form mirrors the manner in which a word is pronounced, e.g. Bantu, Afrikaans and English speakers in South Africa generally have distinctive pronunciations: Bantu speakers usually pronounce "something" like *somsing (see Oller's list of errors above) and (2) the deviant form resembles another word in the language: rope is written as *robe. Oller's phonological distortions are examples from dictation tests. In such a case, phonological errors could be initiated by the presenter of the dictation, for example, a presenter who is a Tswana speaker could very well pronounce "something" as *somsing. In such a case, the presenter should be penalised not the test taker! What occurs in such a situation is the direct "transfer of training" where the teacher. presenter is the direct cause of the error. In the case of an essay, however, *somsing would be caused by the learner.

---

50 Oller, J.W., Jr. *Language tests at school*, 1979, p.279.

The deviant forms of "cover" in Protocol 1 could be (interlanguage?) variations on a phonological theme. Thus, the deviant forms of "cover" need not be spelling errors as many of the NAETE raters maintained. For example, the underlined error *you in the last line could be a morphosyntactic error (the possessive "your" is required), a spelling error or a phonological error.

In Protocol 2 (Figure 4.8) there are hardly any incorrect word forms, and thus little possibility of confusing spelling errors with grammatical (morphological and phonological errors). Only one out of the eight EL2 raters (13%) mentions spelling errors. Most of the errors in Protocol 2 are punctuation errors, which are judged to be "grammatical" errors by most in the EL1 and EL2 groups. I don't find any morphological or phonological errors. Here is Protocol 2 (L1) for easy reference:

> You need a roll of paper cover or plastic cover, A pair of scissors some sellotape. You put the book on the paper or Plastic and cut the length it is better if about 5 cm of cover was left from the book. You cut it into strips You fold the cover over the book. You then put strip of sellotape to keep **them** down. Then you put plasitic paper over it and stick it down. Then you can put your name and standard.

The punctuation errors are serious, while "left from the book" and "cut into strips" affect the coherence, to a certain extent. The pronoun "them" (in bold) is not a grammatical error because it agrees with "strips" in the previous sentence (not with "strip" in the same sentence). So, on my view there is only one grammatical error - the omission of "a" between "put" and "strip" in the third last sentence, but no spelling errors. There was a substantial difference in negative judgements between EL1 and EL2 on content; 56% and 38%, respectively. The overall picture is disconcerting, where there are 42% negative judgements on content and grammar.

So far, nothing has been said about the relationship between individual scores and judgements. Scores by themselves do not provide detailed information on the level of

proficiency that the scores represent. Similar scores between raters do not necessarily mean similar judgements, and also, different scores between raters do not necessarily mean different judgements. A few examples are provided from Protocols 1 and 2. The individual scores and detailed judgements of the 24 raters appear in Tables A (Protocol 1) and B (Protocol 2) of the appendix.

In Protocol 1 (L2) the following judgements went together with the same scores: A score of 3 for one rater represented "meaningless cloudy" (Rater E3) and for another rater the same score of 3 meant "misspelled many words but not to bad" (this was Rater E5, who was excluded from the main analysis because he/she was the fifth member of Group E, which had been reduced to four members). Many of the misspelled words in the L2 protocol were actually different forms of the one word "cover". A score of 5 for C4 represented " Topic deviates. Content sequence satisfactory. Major grammatical. errors detracts from coherence", but for D1 the same score represented "Only one great fault is spelling, quite distracting." D3, who awarded a score of 6, states: "the learner belongs to an elite group".

Consider the following examples from Protocol 2 (L1). E2 awarded a score of 5 and remarked: "General reluctance to give extremely high or low marks". E2's score for *Protocol 1* (L1) was 7, which seems to contradict the reluctance to give extremely high or low marks: unless a score of 7 is not an "extremely" high mark in E2's eyes. If so, one doesn't know what to make of E2's remark that a score of 5, which E2 gave for Protocol 2, steers a middle path between an "extremely low" and an "extremely high" score.

A few other examples from Protocol 2. Some raters attached more importance than others to the segment "cut into strips". Consider the remarks of the following raters, which all contained the phrase "cut into strips". They all awarded a score of 5 and commented only on content. They were all EL1 speakers:

D1: Less accurate. Difficult to understand "cut into strips".

E2: Unclear explanation. Cut what into strips?

F1: Fairly clear, except for "cut it into strips".

F2: Left out important details such as opening the book; "cut into strips" is confusing.

D2: Cohesion bad, e.g. "cut it into strips", but fairly coherent, not too many errors.

D1, E2 and F2 made a big issue of "cut into strips", which in their eyes made the content inadequate, while F1 and D2 made overall positive comments on content (16 of the 24 raters had positive overall comments on the content of this protocol). F1's comment seems the most reasonable because the fact that "cut into strips" is not in the correct sequence doesn't have a significant affect on *coherence* (this is not, I would think, a *cohesion* problem, as D2 states), because when one reads the sentence that follows this segment it is obvious that one is talking about cutting the sellotape into strips, not the paper used to cover the book- nor the book.

One may argue that owing to the fact that there are no data on which words in the protocols individual raters judged to be spelling or grammatical errors, there is no reason to believe that my judgements would be better than other people's. I would never claim them to be better, especially after so much reference in this study to the subjectivity of human beings. What is certain is that raters can't agree on what is a spelling error and what is a grammatical error

In the appendix, the differences between the NAETE raters are shown in Tables A and B. These differences are worrisome. Even more so when they are compared with the NAETE raters' answers to the questions on moderation that were given in the questionnaire at the NAETE workshop. (The questions appear in Table C of the appendix together with the responses of individual raters). A few individual rater judgements are now discussed.

In the questionnaire, 14 of the 24 participants stated - Question J(i) - that in their workplace they never found any significant difference between their ratings and those of their colleagues. Of the 7 raters who said that they did find significant differences in the workplace, only four found this a problem - Question J(ii). As far as the participation in moderation workshops was concerned (Question L), seven of the 24 stated that they had never participated in a moderation workshop. Of the 17 remaining raters, 11 commented on whether these moderation workshops resulted in any improvement. Of these 11 raters, four said that there was a great improvement, six said that there was fair improvement, one said that there was negligible improvement, and one said that there was no noticeable improvement.

4.8.1.3 Implications of interrater unreliability

Interrater reliability consists of two major kinds of judgements: (1) the order of priority for individual raters of performance criteria (criteria such as grammatical accuracy, factual relevance and spelling) and (2) the agreement between raters on the ratings that should be awarded if or when agreement is reached on what importance to attach to different criteria. This is also a validity issue. So, in interrater reliability one is not only interested in scores but what these scores represent. For example, if raters give the same low score for a protocol, but for completely different reasons, e.g. because (1) the spelling or (2) the grammar was bad or (3) because the writer was off the topic, the scores would be statistically reliable, but not valid because there would be no agreement on the purpose of the test. A test is said to be used for a valid purpose when the tester knows *what* is being tested. However, if testers can't agree on what that *what* is, i.e. if there is no interrater reliability, there can be no validity. So, validity and reliability are two sides of the same coin.

The variability in attention that raters pay to different criteria is a general problem in all kinds of educational institutions where "lecturers [or teachers] vary from penalising students heavily for mechanical and grammatical errors to looking through the

linguistic surface and marking on content and organisation."[51] There are different learning styles, teaching styles and also different rating styles. One rater, as indeed one learner or one teacher, may be mainly interested in the big picture, in coherence, while another may be mainly interested in systematicity and structure. Moderation workshops don't seem to be able to effect a truce in these "style wars".[52]

In the research situation it is possible to have more than one rater, even four. Four raters would be a rare luxury outside a research situation. Most testing situations are not research situations but teaching situations where often only one rater is available, where moderation workshops are seldom or never held, as shown in Table C of the appendix. One may argue that the reason teachers/raters don't have moderation workshops or have them seldom is that - as many of the participants said - they didn't find any significant difference in the ratings between their colleagues in the workplace, which would, therefore explain why moderation workshops were seldom held.

Raters at this conference workshop did not previously come together to discuss the protocols that they were asked to judge individually. One may argue that they should have done so. I would imagine, however, that educators of English teachers, even if they did not confer beforehand on assessment procedures, should nevertheless be in gross agreement on whether the grammar, content or spelling of a protocol on such a simple topic with such simple structures was good or bad. The fact that (1) they didn't agree on these fundamentals in the NAETE workshop, (2) many of them said in the questionnaire that they had little disagreement with their colleagues, (3) the majority held few or no moderation workshops, reveals an unsatisfactory situation. The big

---

51 Bock, M. 'Teaching grammar in context', in Angélil-Carter, S. (ed.). *Access to success: Literacy in academic contexts,* 1998).

52 (1) Dreyer, C. 'Teacher-student style wars in South Africa: The silent battle.' *System,* 26: 115-126 (1995).

(2) Oxford, R.L., Ehrman, M. and Lavine, R.Z. 'Style wars: Teacher- student style conflicts in the language classroom', in Magnan, S.S. (ed.). *Challenges in the 1990s for College Foreign Language Programs,* 1991.

question is how to deal with the problem in the normal situation where there is only one rater.

Actually, the idea of using only one rater may be a better idea than using several. To elaborate, I bring the difficulty of distinguishing between spelling and grammatical errors back into the discussion. According to Oller, errors of judgement in distinguishing between spelling, on the one hand, and morphology and phonology, on the other, are not substantial enough to affect reliability.[53] Ingram disagrees and maintains that "it is often a matter of judgement whether, for example, an error is merely spelling (to be disregarded) or phonological or grammatical."[54]

The crucial issue, therefore, as far as *reliability* is concerned, is perhaps not the difficulty a rater has in deciding how to categorise errors, but that one rater's judgements often differs from another's. If different interpretations on what is a spelling error and what is a grammatical error affect the reliability, the use of one rater would ensure more consistency in judgements when problematic items need to be distinguished within these three categories. The use of one rater is justified on the grounds that

> *there are cases where it is difficult to decide whether an error is really a spelling problem or is indicative of some other difficulty besides mere spelling. In such cases, for instance, 'pleshure', 'teast' for 'taste', 'ridding' for 'riding', 'fainaly' for 'finally', 'moust' for 'must', 'whit' for 'with' and similar instances, perhaps it is best to be consistently lenient or consistently stringent in scoring. In either case it is a matter of judgement for the scorer.*[55]

---

53 Oller, J.W., Jr. *Language tests at school*, 1979, p.279.

54 Ingram, E. 'Assessing proficiency: An overview on some aspects of testing', in Hyltenstam, K. and Pienemann, M. *Modelling and Assessing second language acquisition*, 1985, p.244.

55 Oller, J.W., Jr. *Language tests at school*, 1979, p.279.

Without consistency in *judgements*, there would be no consistency in scoring, i.e. scores will be arbitrary, which is one of the reasons why psychometric measurement has a bad name among many "real-life" testers.

If one takes into account the gargantuan problems of rater subjectivity, it may be better to use one rater to mark a specific group of test takers rather than several, so that if we can't improve interrater consistency to any significant extent, we can at least try to make sure that the same rater marks all the protocols of the group he or she teaches. But then, as we know, we cannot be sure that the rater won't mark differently before breakfast (a good or bad one) than after.

4.8.1.4 Conclusion to the NAETE workshop on interrater reliability

When it comes to doing research on raters' judgements, the problem of subjectivity can become very complex. For example, my research into interrater reliability was based on the judgements of other people's judgements (the raters discussed above). Thus, my judgement is a verbalisation (a fourth level of an interpretation) of an observation (the third level of interpretation) of other people's verbalisations (a second level of interpretation) of their observations (the first level of interpretation). When one adds a fifth, a sixth and more levels (an assessment of an assessment, of an assessment, etc.) hermeneutics can so easily become trapped in hermetic "webs of beliefs".[56] Raters (and raters of raters) are in danger of following a circular route to control what is very difficult or perhaps impossible to control, namely, subjectivity.

Learners may fail because they don't learn, or because they lack the academic ability, or because they are politically or economically suppressed, and for many other reasons. In my experience many fail *and* pass because of the luck of the draw - a "strict" rater or a "lenient" rater.

---

[56] Quine, W. and Ullian, J. *The web of belief* 1970, quoted in Moore, R. 'How science educators construe student writing', in Angélil-Carter, S. (ed.). *Access to success: Literacy in academic contexts*, 1998, p.83.

A point of interest: the L2 learner obtained a Joint Matriculation Board matriculation in 1992 with a C aggregate and a D for English. (Not many at MHS achieved a C aggregate). The L1 learner passed Grade 11 in 1991 with high marks in English and aggregate and then left MHS, because his family left Mmabatho.

4.9 Summary of Chapter 4

The statistical results were reported. High correlations were found between the tests and there was a substantial difference between the L1 and L2 groups. Reasons were given for not treating the L1 and L2 groups as separate populations in the correlational analysis. The dangers of comparing tests were also discussed.

Singled out for special attention was interrater reliability. The lack of interrater reliability is arguably the greatest problem in assessment because it is often the cause, though indirectly, of student failure - and success! It is on the issue of interrater reliability that matters of validity and reliability come to a head, because it brings together in a poignant, and often humbling, way what is being (mis)measured, and how it is (mis)measured. The next chapter deals with the battery of proficiency tests as predictors of academic achievement.

CHAPTER 5

The Prediction of Academic achievement

5.1 Introduction

There are three sections in this chapter:

(1) A correlational analysis and multiple regression analysis of predictions from Grade 7 to Grade 11. The Grade 12 results, which were externally examined, were only available as symbols, so I couldn't do a correlational analysis and multiple regression analysis that included the Grade 12 results. The exclusion of the Grade 12 results does not detract from the meaningfulness of these two analyses, because everyone in the sample who passed Grade 11 also passed Grade 12; even if it meant repeating Grade 11.

(2) Frequency distributions and discussion that give a detailed analysis of the predictive validity of the different English proficiency tests.

(3) General discussion of language proficiency as a predictor of academic achievement.

Although it is possible that annual predictions between English proficiency and academic achievement would yield higher correlations than long-term predictions, the tests in this study seek to find out what chance Grade 7 learners who entered MHS in 1987 would have of eventually obtaining a Joint Matriculation Board (JMB) matriculation exemption. For this reason I have not used subsequent English proficiency tests at MHS as predictors of academic achievement.

There exists a substantial difference between the L1 and L2 groups on all of the tests. Any single test, therefore, discriminates well between the L1 and L2 groups. What is important for establishing the construct validity of a test is not the equivalence in scores between *tests*, but whether a test discriminates well between weak and strong

test takers. The details in the predictions won't be exactly the same for all combinations of tests in the battery or for all the individual tests, but what is important is that the pattern of predictions would be similar, i.e. a clear distinction would be made between learners with low proficiency and high proficiency. Some tests in the test battery discriminate better than others between the L1 and L2 groups. This will become clearer in the analysis of the predictive validity of the individual tests.

5.2 Correlational analysis and multiple regressions of the predictions

In the correlational analysis, the criterion variables are GRADE 7, GRADE 9 and GRADE 11, which are the end-of-year aggregates for the respective grades.

TABLE 5.1

Grade 7 to Grade 11 Correlational Analysis of the Prediction of Academic Achievement (Aggregate) with Five English Proficiency Tests as Predictors ($p < .01$)

CRITERION	PREDICTORS				
	DICT	ESSAY	CLOZE	GRAM	ER
GRADE 7 AGGREGATE (N=75)	.64	.63	.59	.62	.54
GRADE 9 AGGREGATE (N=43<41)[1]	.33	.18	.34	.28	.15
GRADE 11 AGGREGATE (N=26<24)[2]	.12	.25	.30	.23	.33
[1] *N=43 for DICT, ESSAY and CLOZE; N=41 for GRAM and ER*					
[2] *N = 26 for DICT, ESSAY and CLOZE; N= 24 for GRAM and ER*					

The following observations are important:

1. The correlations for the GRADE 7 predictions are all significant. A validity coefficient under .40 would not be regarded as significant.[57] The correlation between GRAM and GRADE 7 is interesting. One might expect that ESSAY should correlate

[57] Cronbach, L.J. *Essentials of psychological testing*, 1970, p.126.

much higher than GRAM with GRADE 7, owing to the fact that examinations consist of a lot of writing. The difference between the ESSAY/GRADE 7 and GRAM/ GRADE 7 correlations, indeed between all the Grade 7 correlations, is not large enough to reject the possibility that a significant part of the difference is due to measurement error. Having said that, it might be thought that a knowledge of grammar by itself should not correlate highly with academic achievement. However, GRAM did not test grammar *by itself*, because the test takers did not have to learn a specific list of grammatical items (which is common in traditional second language classrooms) and then write the test to prove what they had *achieved*; often a rote affair. On the contrary, they were given a *proficiency* test, i.e. a test based on knowledge that was part of their (consistent, if not immutable) grammatical competence. Under such conditions there is no reason why a grammar test should not correlate highly with academic achievement, if they both involve understanding and not merely rote-learning.

2. The general pattern of the predictions from Grades 7 to 11 shows that the correlations become progressively lower, which means that English proficiency (tested in Grade 7) ceases to be a valid predictor after Grade 7.

3. The progressive decrease in the correlations may be due to any one or a combination of the following factors:

(i) The decrease in the size of the sample.

(ii) The narrowing of the range of scores as failures are progressively pruned from the system.

(iii) Different evaluation procedures from year to year and from teacher to teacher.

(vi) Changes in intellectual skills and/or subject knowledge and/or motivation.

In sum, high correlations exist between English proficiency and GRADE 7, but these correlations progressively decrease between GRADE 7 and GRADE 11. The correlations suggest that English proficiency, tested in GRADE 7, ceases to be a valid predictor of academic achievement beyond Grade 7. Shortly (section 5.3), I shall

show, with the help of frequency distributions, that such an interpretation would be inaccurate. For the moment, I move on to the multiple regressions.

The above correlation matrix only shows correlations between single tests and the criterion (aggregates of the different grades). Table 5.2 shows a multiple regression analysis of the Grade 7, Grade 9 and Grade 11 predictions. Only tests that made a contribution of more than .2% are included.

TABLE 5.2

Stepwise Multiple Regression Analysis of Predictions

Grade 7 to Grade 11 (N=75)

Predictors (Tests)	*Criterion* (Aggregate)	*R*	R^2	*Cumulative* R^2	*Sig. level*
	GRADE 7 (N=75)				
1. DICT		.6368	.4055	.4055 (40.6%)	.0808
2. GRAM		.6242	.0401	.4456 (44.6%)	.1863
3. ESSAY		.6337	.0130	.4587 (45.9%)	.1890
	GRADE 9 (N=41)				
1. DICT		.3231	.1043	.1043 (10.4%)	.2705
2. GRAM		.2777	.0022	.1065 (10.7%)	.7650
	GRADE 11 (N=24)				
1. ER		.3344	.1119	.1119 (11.2%)	.4033
2. CLOZE		.2946	.0054	.1173 (11.7%)	.7215

In the Grade 7 predictions DICT and GRAM are the first two measures that enter into the regression and they account for almost all of the common variance between the predictors (the proficiency tests) and the criterion (the aggregates of academic achievement in the different grades). DICT and GRAM are the only tests to feature in the Grade 9 regressions. This means that DICT and GRAM are better predictors than

ESSAY, CLOZE and ER. The issue, it may be argued, is not only a statistical one but also a *face validity* issue: a clash between practicality and acceptability. Henning et al.'s dilemma is a good example of the problem.[58] Although their educational context is the senior year in secondary school in Egypt where their tests are concerned with assessing academic potential for university access, the issues they deal with are universal in language testing. In their correlations with the Grand Total of their test battery, Henning et al. found that the highest correlation with Composition was with Error Identification (.76). They subsequently maintain that "Error Identification may serve as an indirect measure of composition writing ability".[59]

Although Henning et al. maintain that Reading Comprehension "like Listening Comprehension is of little psychometric value in predicting general proficiency", they concede that they must include reading in order for their battery to "find acceptance"[60]. Accordingly, they replace their Error Identification test with Reading Comprehension.[61] The psychometric "posture"[62] had not reckoned with face validity.

5.3 Frequency distributions of the predictions and data analysis

One may argue that DICT and GRAM are good short-term predictors (Grade 7) but that (1) that the correlations with later grades (Grades 9) are too low and (2) DICT and

---

58 Henning, G.A., Ghawaby, S.M., Saadalla, W.Z., El-Rifai, M.A., Hannallah, R.K. and Mattar, M. S. 'Comprehensive assessment of language proficiency and achievement among learners of English as a foreign language.' *TESOL Quarterly*, 15 (4), 457-466 (1981).

59 Ibid., 462.

60 Ibid., p.464.

61 One other reason Henning et al. give for rejecting their Error Identification test is that it had low reliability (Ibid., p.464). The 50-item reliability coefficient (KR-20) of their Error Identification test was .71. My argument would be that to obtain such high validity coefficients in spite of the low reliability of the tests shows what a practical test their Error Identification could be if improvements were made.

62 Lantolf, J.P. and Frawley, W. 'Proficiency: Understanding the construct.' *Studies in Second Language Acquisition (SLLA)*, 10 (2), 181-195 (1988), p.81.

GRAM don't even appear in the Grade 11 regressions. The frequency distributions of the predictions provide a clearer picture of the predictive validity of the individual tests. Histograms are used to show the details.

Before presenting the frequency distributions I need to explain what I mean by the term "a good predictor". In the multiple regression analysis above it was shown that some of the tests were better predictors than others. This had nothing to do with the actual scores of the different tests (i.e. whether some tests had higher scores than other tests) because multiple regression, which is based on correlation is only concerned with how scores covary, i.e. "go together". The primary issue in correlation is not actual scores but

> *the tendency for examinees to perform proportionately well or poorly in relation to one another...In a word it is the variance in test scores, not the mean of a certain group or the score of a particular subject on a particular task that is the main issue.*[63]

Correlation is not concerned with individual scores. A very high correlation does not mean that scores on two tests are either both high or both low. One test could have low marks, the other high, and there could still be a high correlation between them. That is what Oller's definition of correlation above explains. Thus, it would not be correct to maintain that a positive correlation means (1) an individual who achieves a high score on a variable (X) also achieves a high score on another variable (Y) or (2) an individual who achieves a low score on a variable (X) also achieves a low score on another variable (Y). Firstly, correlation is a group statistic; secondly a group may obtain high scores on one test and low scores on another test, yet there may still be a high correlation between the two tests. For example, suppose the scores on one test are 90, 80, 70, 60, 50, and the scores on another test are 45, 40, 35, 30, 25, respectively. There would still be a positive correlation, in fact a perfect correlation of 1, between the two tests, in spite of the fact that the first test has much higher scores than the

---

[63] Oller, J.W., Jr. *Language tests at school*, 1979, p.272.

second test. This is so because the test takers had the tendency to perform *proportionately* well. In this case the scores went together perfectly, e.g. 90 is twice 45, 80 is twice 40, 70 is twice 35, etc.

The question is whether one should assume that if there are low correlations, as was the case in the regression analysis of the Grade 9 and Grade 11 predictions, that there is no point in pursuing the predictive validity of the English proficiency tests any further. This would be a wrong deduction. This is where frequency distributions come in. A frequency distribution is more revealing than correlations (and multiple regressions, which are correlations) because the latter "cuts blindly through the thicket of complexity"[64] and accordingly may conceal meaningful details, especially when correlations are low. To get a clearer predictive picture one should also use frequency distributions.[65] Frequency distributions ignore "the tendency for examinees to perform proportionately well or poorly in relation to one another" (Oller above), which is the concern of correlation, and concentrates *instead* on the more general comparisons. In This study focuses on the ranges of scores *within* individual tests *between* the L1 and L2 groups.

In the predictions the *relatively* higher ranges should predict success whereas relatively lower ranges should predict failure, no matter what the value of the ranges. For example, if the ranges of scores in a group are 0-29 and 30-39, respectively, the 0-29 range should predict more failures or less passes than the 30-39 range. The "higher" achievers on such a test (30-39 range) should have a greater chance of success than those in the 0-29 range.

The frequency distributions of the tests are discussed in the following order (1) error recognition (ER), (2) essay (ESSAY), (3) cloze (CLOZE), (4) mixed grammar (GRAM) and (5) dictation (DICT). Of the original sample of 86 subjects there were

---

64 Herrenstein, J. *IQ in the meritocracy*, 1973, p.23.

65 Eysenck, H.J. Réponse à quelques réflexions naïves sur l'interprétation des coéfficients de corrélation. *Revue de Psychologie Appliquée*, 34 (2), 111-114 (1983).

five who left the school before Grade 11, after passing a grade. The remaining 81 subjects did the essay, cloze and dictation tests, while 75 of these 81 did the error recognition and grammar tests. In the multiple regression analysis a sample of 75 was used for all of the tests because each of the variables had to be the same length. In the frequency distributions, where I am far more interested in the comparison between the *L1 and L2 groups* on *each* test than in a comparison between tests (test scores, in this case), it is not necessary to have the same sample sizes. So, for the frequency distributions, the sample size for *(1) ESSAY, CLOZE and DICT is 81, where L1=45 and L2=36, and for (2) ER and GRAM is 75, where L1=39 and L2=36.* (See the "Grand Total" row of Table 5.3 below)..

To reiterate an important point: it doesn't matter if the L1 and L2 groups are different sizes (as long as they are not substantially different). In the frequency distributions, I am mainly interested in comparing the *proportion* of passes/ failures between the L1 and L2 *groups* on *each* test, and not in comparing test scores, which, as pointed out earlier (section 4.7), can be a dangerous exercise.

I now examine the pass rate of the subjects in the sample, who were also the 1987 entrants to Grade 7 at MHS. For each of the tests the following is provided: (1) the results for the whole sample (L1 + L2 groups) and (2) a comparison between the L1 and L2 groups.

In the summary of the predictions in Table 5.3 below there are three kinds of data for each test: (1) The number of Grade 12 passes (indicated as "Pass 7-12"), (2) failed Grades 7, 8 or 9 (indicated as "Fail 7-9"), and (3) failed Grades 10 or 12 (indicated as "Fail 10-12"). (Recall that all those who passed Grade 11 passed Grade 12). The results include subjects who failed a grade and passed a year later on the second attempt.

TABLE 5.3

Summary of Predictions Grade 7 to Grade 12

	CIOZE, DICT & ESSAY			ER & GRAM		
	All	*L1*	*L2*	*All*	*L1*	*L2*
Pass 7-12	**41**	**28**[1]	**13**	**37**	**24**	**13**
Fail 7-12	**40**	**17**	**23**	**38**	**15**	**23**
Grand Total	*81*	*45*	*36*	*75*	*39*	*36*
Fail 7-9	32	12	20	30	10	20
Fail 10-12[2]	8	5	3	8	5	3

[1] Included in this total are two pupils who left MHS after passing Grade 11 with very high marks. These would have in all probability passed Grade 12.

[2] Recall that nobody failed Grade 12, because all those who passed Grade 11 passed Grade 12.

Before I deal with the predictive validity of the English proficiency tests, a summary is provided of the Grade 12 pass rate of the L1 and L2 groups.

There were 41 passes and 40 failures of Grade 12. Of the 40 failures there were 32 who failed Grades 7, 8 or 9 and left MHS. The remaining eight subjects ("Fail 10-12") are those who failed Grades 10 or 11 and left the school. Table 5.4 shows the total matriculation exemptions for the L1 and L2 groups.

Table 5.4

Detailed Analysis of the Matric Pass Rate (N=81)

	Ordinary Pass	Matric Exemption	Total Passes	Total Failures	Grand Total	% Matric Exemption
L1 Group	1	27	28[1]	17	45	60.0%
L2 group	5	8	13	23	36	22.2%
Total	6	35	*41*	*40*	***81***	**43.2%**
% of 81 (L1+L2)	7.4%	**43.2%**	50.6%	49.4%	100%	-

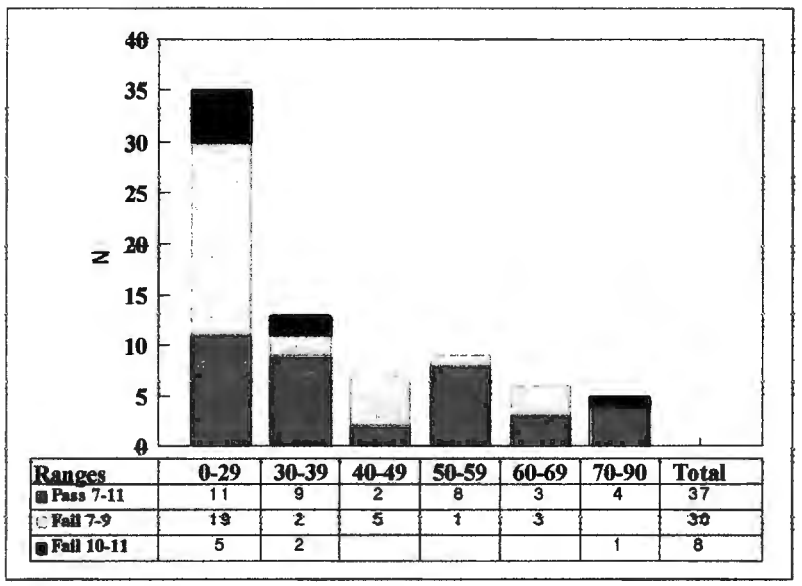
[1] The two pupils who passed Grade 11 with high aggregates are included because they would have without doubt have obtained a matriculation exemption (see Note 1 in Table 5.3). Recall that that there were originally 86 subjects but five left before Grade 11 after passing their respective grades.

Although 13 of the 36 L2 subjects passed Grade 12, only eight of the 13 obtained a matriculation exemption. Thus, only 22.2% of the 36 L2s obtained a matriculation exemption. In contrast, 27 of the 28 L1 pupils who passed Grade 12 obtained a matriculation exemption. Thus, 60% of the 45 L1s obtained a matriculation exemption.

The histograms and discussion of the data are provided below. To reiterate: the tests are discussed in the following order (1) error recognition (ER), (2) essay (ESSAY), (3) cloze (CLOZE), (4) mixed grammar (GRAM) and (5) dictation (DICT).

1. Error Recognition Test (ER) Predictions of Grades 7 to 12 Pass Rate: Figures 5.1 to 5.3.

Figure 5.1. ER Whole Sample (N=75)

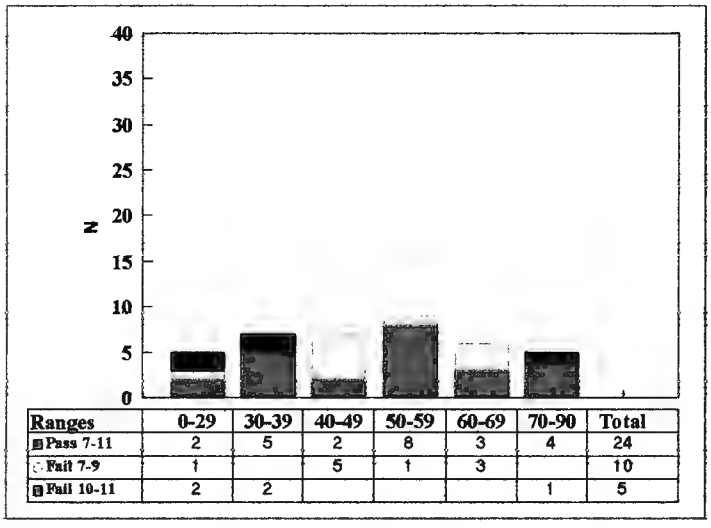


Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
Pass 7-11	11	9	2	8	3	4	37
Fail 7-9	19	2	5	1	3		30
Fail 10-11	5	2				1	8

The 0-29 range is a good predictor (24 failures out of 35). The 50-59 range is a very good predictor (8 passes out of 9). However, owing to the fact that the 60-69 range is not a good predictor, one cannot consider the 50-59 range on its own as a good

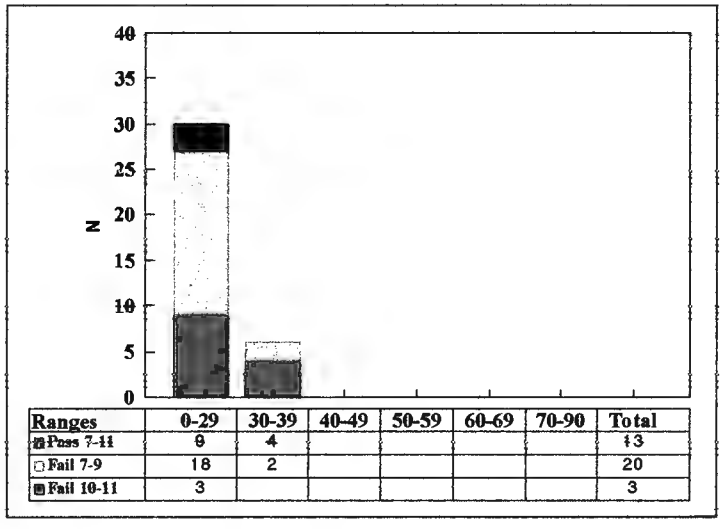
predictor because the logic of prediction is that the higher the range (e.g. 60-69) the better should be the prediction of success, but as shown the 60-69 range is a poorer predictor of success than the 50-59 range. In such a case one could pool the three upper ranges and make the broad observation that a score over 50 is a good predictor (15 passes out of 20).

Figure 5.2. ER L1 Group (N=39)



Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
Pass 7-11	2	5	2	8	3	4	24
Fail 7-9	1		5	1	3		10
Fail 10-11	2	2				1	5

Figure 5.3. ER L2 Group (N=36)

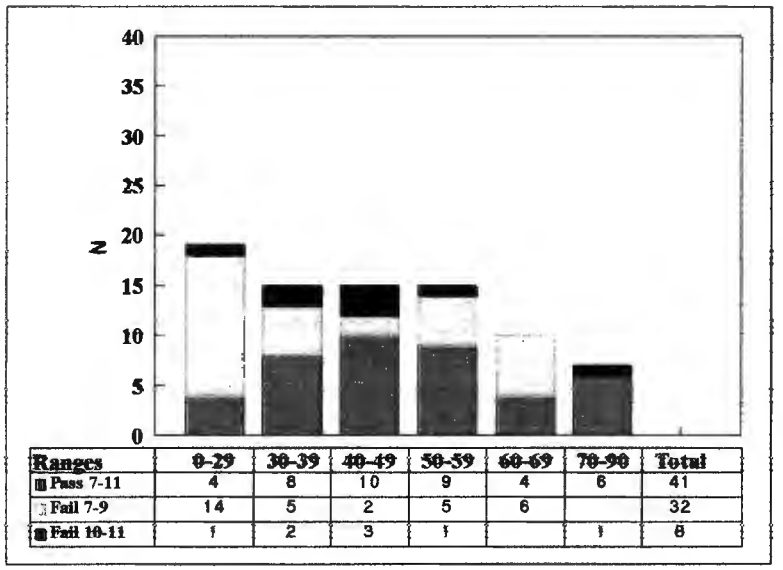


Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
Pass 7-11	9	4					13
Fail 7-9	18	2					20
Fail 10-11	3						3

Owing to the fact that the 0-29 range is occupied mostly by the L2 subjects, it comes as no surprise that most of the failures are L2 subjects. The L2 group has a narrow spread of scores where the 0-29 range is a good predictor (21 failures out of 30). The L1 group has a very wide spread (0-90), where a score over 50 is a good predictor. In sum, ER for the sample as a whole is a good predictor of success in the 50-90 range and good predictor of failure in the 0-29 range.

2. Essay Test (ESSAY) Predictions of Grades 7 to 12 Pass Rate (N=81): Figures 5.4 to 5.6.

FIGURE 5.4. ESSAY Whole Sample (N=81)

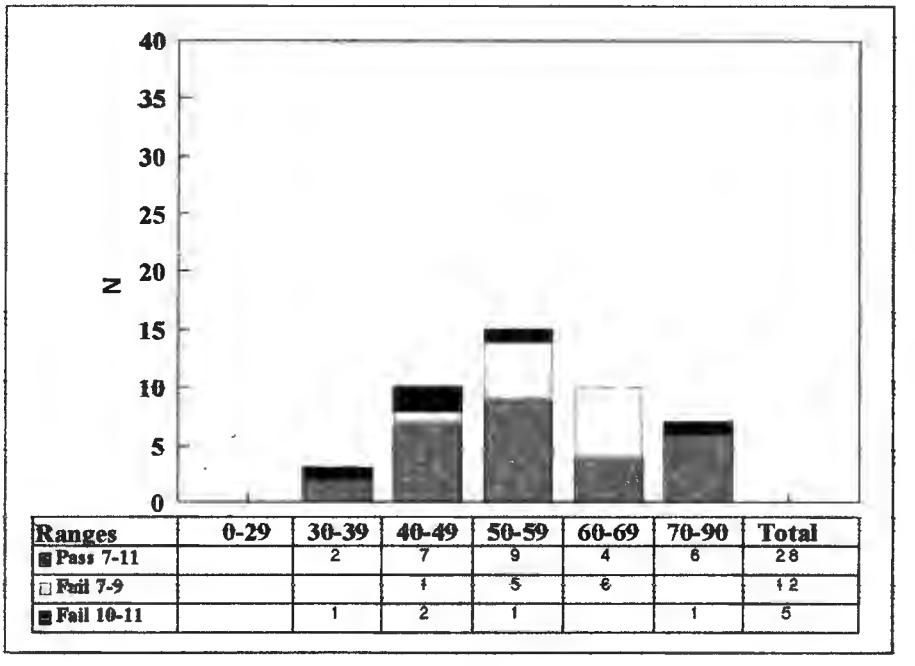


Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
Pass 7-11	4	8	10	9	4	6	41
Fail 7-9	14	5	2	5	6		32
Fail 10-11	1	2	3	1		1	8

The two extremes of the distribution (0-29 and 70-90) are very good predictors, and the 40-59 range is a good predictor (19 passes out of 30). The problem is that the 60-69 range is a poor predictor, which would mean that the 40-59 range ceases to be a good predictor, because somebody in the 60-69 range has a strong chance of failing, while somebody in the 40-59 range has a strong chance of passing. The reason why the 60-69 range is a poor predictor could be the following: writing ability changes

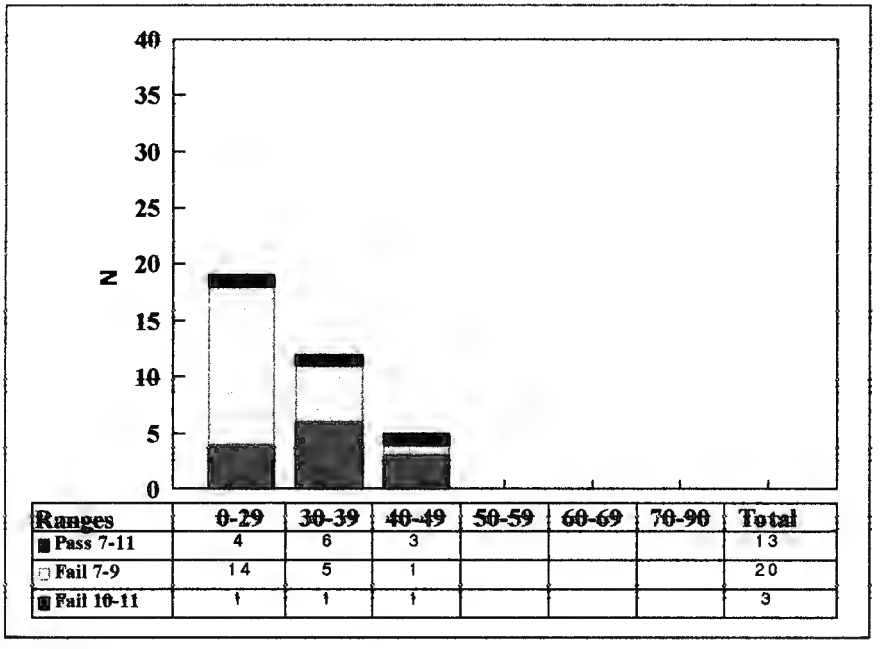
through the grades and there is wide scope for improvement. This applies to the 30-39 and 40-49 ranges as well. However, if one has very high proficiency (70-90 range) or very low proficiency (0-29 range), one might have reached the maximum level of one's potential. *This observation applies to all the tests in the study.* It is understandable that if one obtains a score of 80%-90% that one seldom obtains higher scores even if one does improve through the grades. This is probably due to the practice of not awarding scores higher than 80%-90%. In the case of the 0-29 range, learners usually don't obtain higher marks because they don't improve. The data become more informative when the sample is split into the L1 and L2 groups.

Figure 5.5. ESSAY L1 Group (N=45)



Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
Pass 7-11		2	7	9	4	6	28
Fail 7-9			1	5	6		12
Fail 10-11		1	2	1		1	5

Figure 5.6. ESSAY L2 Group (N=36)



Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
Pass 7-11	4	6	3				13
Fail 7-9	14	5	1				20
Fail 10-11	1	1	1				3

The level of writing ability in Grade 7 can progressively improve or regress through the grades so that a learner with a score in the 60-69 range in an essay proficiency test at the beginning of Grade 7 could score significantly higher or lower two or three grades later in an essay achievement test. If the assumption is that there is a causal connection between writing ability and academic achievement, we could conclude that several L1 failures in the 60-69 range regressed in their writing ability, while several L2 passes in the 0-39 ranges progressed in their writing ability. One also has to take into account the very important fact that even if those in the 60-69 range did not regress in their writing ability, it is possible that the cause of failure could have been a lack of academic ability or of the will to learn, or some other cause. (I discuss the theory of the relationship between language proficiency and academic achievement in more detail after the analysis of the predictions).

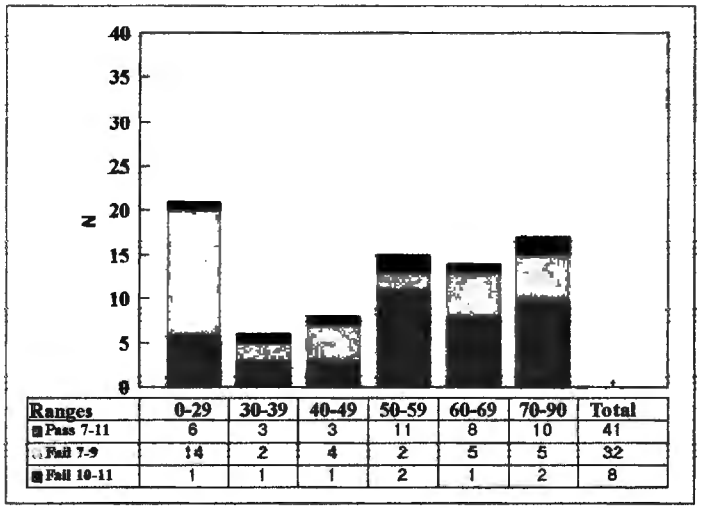
If we compare ER (Figure 5.1) with ESSAY (Figure 5.4), the ER 50-69 ranges (which belong to the L1 group) are good predictors of success (11 out of 15) whereas the ESSAY 50-69 ranges are poor predictors of success (13 passes out of 25). Care must

be taken, however, not to make spurious comparisons between the tests, because the 50-69 ranges for ER may not be the same level of difficulty as the 50-69 ranges for ESSAY. Thus, 50% on one test does not mean the same thing as 50% on another test (see section 4.7). What one should look at *first* is the relative pass rate in the different ranges within each test. Comparisons can then be made between the tests with the understanding that each test has its own criteria for determining the level of difficulty. Then there would be more clarity on what is meant by saying that ER is a better predictor than ESSAY in the 50-69 ranges (see section 4.7).

With regard to the 0-29 range of ESSAY (Figure 5.6), only L2 subjects occupy this range. Almost all the failures in this range (15 failures out of 19) dropped out very early: 11 failed in Grade 7, three failed in Grade 8 and one failed in Grade 11.

2. Cloze Test (CLOZE) Predictions of Grades 7 to 12 Pass Rate: Figures 5.7 to 5.9.

Figure 5.7. CLOZE Whole Sample (N=81)



Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
Pass 7-11	6	3	3	11	8	10	41
Fail 7-9	14	2	4	2	5	5	32
Fail 10-11	1	1	1	2	1	2	8

First the broad picture. CLOZE is a good predictor in the 0-49 range (23 failures out of 35) and the 50-90 range (29 passes out of 46). We home in on the individual ranges. The very low range (0-29) is the best predictor (15 failures out of 21).

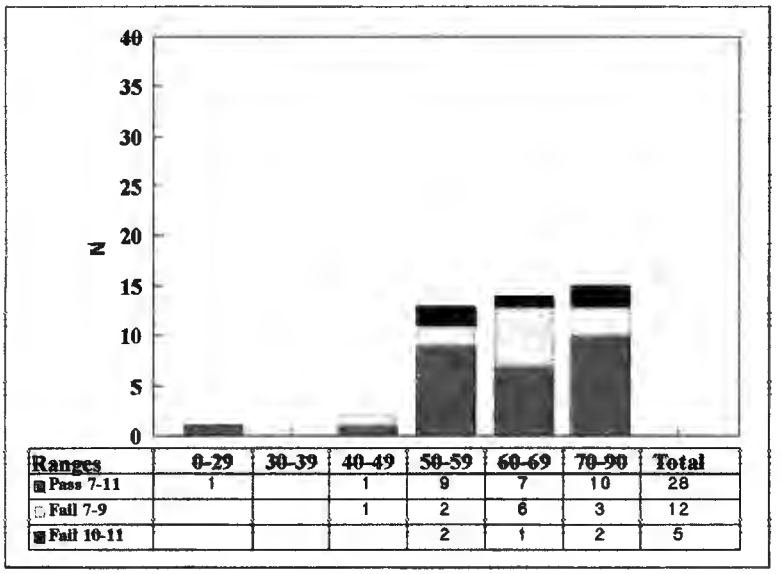
In the 60-90 range, six out of 14 failed, which shows that a substantial number of subjects in this range were "at risk".[66] These results, however, are for long-term prediction and Pienaar's tests are concerned with short-term prediction, i.e. predicting one year ahead. So, for Pienaar's purposes he probably would only look at the end of Grade 7 predictions and not be prepared to make any predictions beyond that grade. With regard to the 13 (out of 31) failures in the 60-90 range, 11 failed *later* than Grade 7. In the 0-39 ranges, 14 of the 18 failures occurred in Grade 7, two in Grade 8 and two in Grade 11.

Thus, in the 60-90 range, *Pienaar's tests are good short-term predictors, but not good long-term predictors.* Consider the L1 and L2 groups below. The L1 50-59 range is a good predictor but it has to be seen in relation to the 60-69 range. If the 60-69 range is a poor predictor, the good predictions of the 50-59 range are not useful by themselves because the higher the range the better the predictions should be. The majority of passes are in the 50-90 ranges, which belong to the L1 group. The L1 group has higher scores (on all the tests) because they, as a group, are expected to be better at English than the L2 group. The L1 and L2 groups are shown below.

---

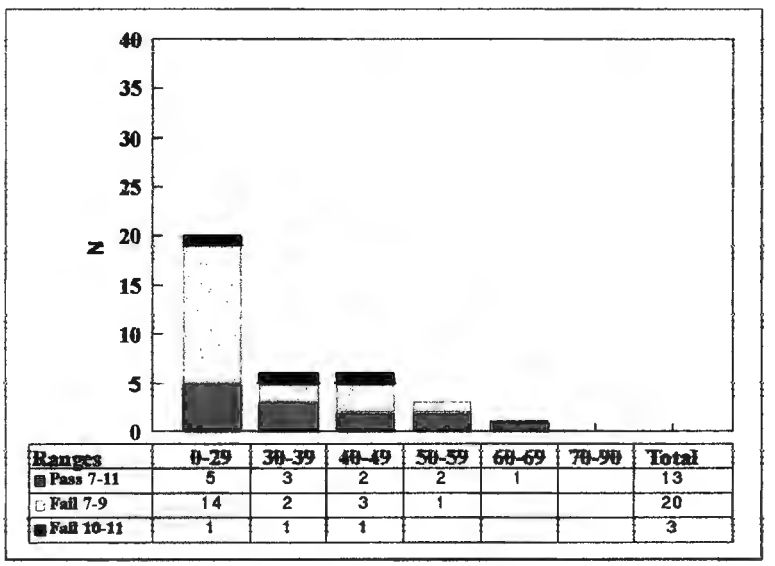
[66] Pienaar, P. *Reading for meaning: A pilot survey of (silent) reading standards in Bophuthatswana*, 1984, p.21.

Figure 5.8. CLOZE L1 Group (N=45)



Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
Pass 7-11	1		1	9	7	10	28
Fail 7-9			1	2	6	3	12
Fail 10-11				2	1	2	5

Figure 5.9. CLOZE L2 Group (N=36)



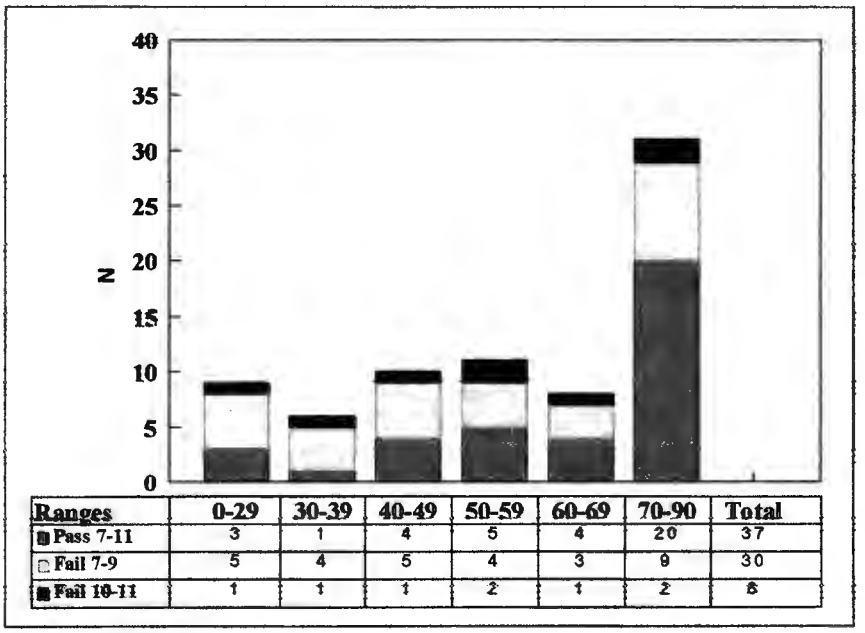
Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
Pass 7-11	5	3	2	2	1		13
Fail 7-9	14	2	3	1			20
Fail 10-11	1	1	1				3

To reiterate: a possible reason why CLOZE is not a good long-term predictor in the 60-69 range is because CLOZE does not discriminate in the long term as well as the other tests between low and high academic achievers. This does not mean that the cloze tests are less related to academic achievement than the other tests. It could mean that reading skills do not develop in tandem with general academic skills, and so it would not be possible to detect any normative pattern in the relationship between

reading skills (or cloze skills, if one objects to cloze being equated with reading) and academic achievement. (Similarly, in the case of writing skills). The good short-term and poor long-term predictions of CLOZE in this study fit in well with Pienaar's rationale that his tests are valid for short-term predictions only. Annual cloze tests might have produced better predictions, and that was the reason why Pienaar's tests are graded in "Steps" from Grade 3 to Grade 12 and beyond to Grade 12+.

4. Grammar Test (GRAM) Predictions of Grades 7 to 12 Pass Rate: Figures 5.10 to 5.12

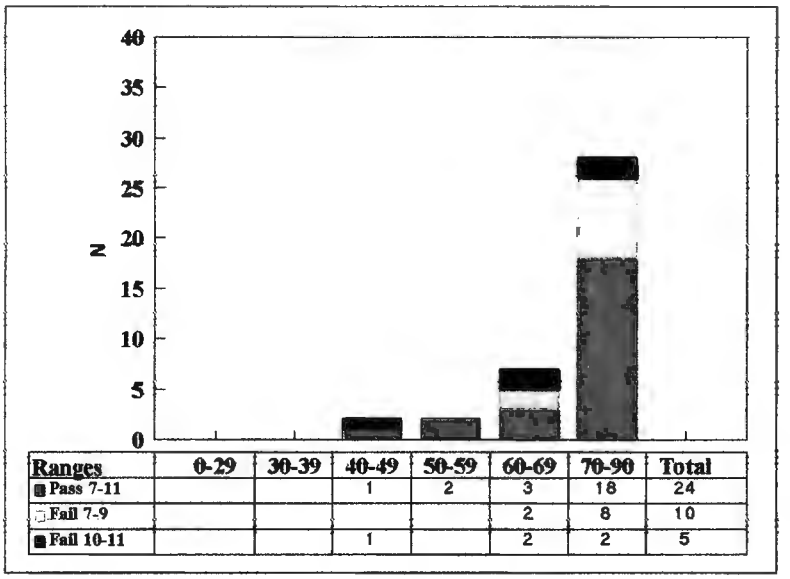
Figure 5.10. GRAM Whole Sample (N=75)



Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
Pass 7-11	3	1	4	5	4	20	37
Fail 7-9	5	4	5	4	3	9	30
Fail 10-11	1	1	1	2	1	2	8

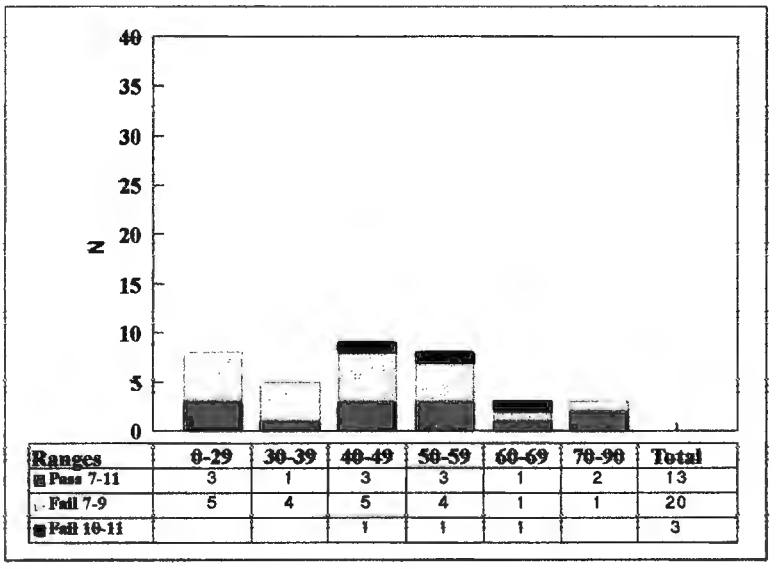
The 0-39 ranges are a good predictor (11 failures out of 15) and the 70-90 range is a good predictor (20 passes out of 31). Compare the L1 and L2 groups of GRAM.

Figure 5.11 GRAM L1 Group (N=39



Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
Pass 7-11			1	2	3	18	24
Fail 7-9					2	8	10
Fail 10-11			1		2	2	5

Figure 5.12. GRAM L2 Group (N=36)



Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
Pass 7-11	3	1	3	3	1	2	13
Fail 7-9	5	4	5	4	1	1	20
Fail 10-11			1	1	1		3

GRAM was much easier than ER (Figure 5.1), hence the presence of more subjects in the 70-90 range of GRAM than in the 70-90 range of ER. Because most of the 70-90 range of GRAM is occupied by the L1 group it is unsurprising that most of the passes in this range belong to the L1 group. In the L2 group one can make broad predictions where a score between 0-59 is a good predictor of failure (20 out of 30).

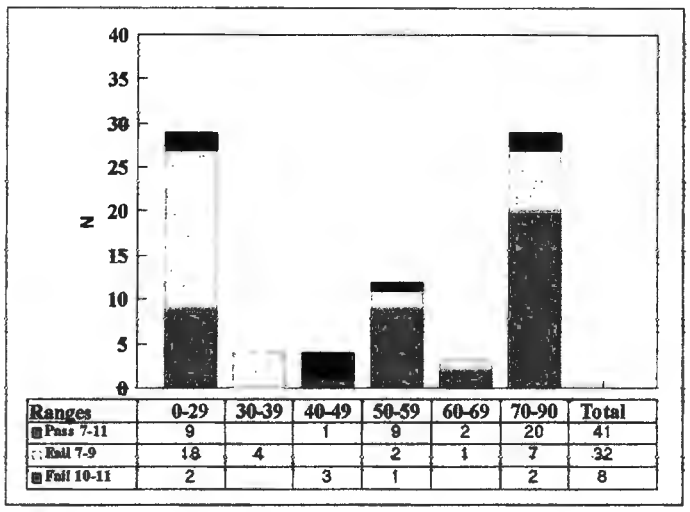
The GRAM scores are higher, in both the L1 and L2 groups, than the scores of the other tests. This, however, is not important in the "*group*-differences" approach to construct validity, because it is not the absolute difference between the scores of tests (i.e. the equivalence in scores) that is important, but the relative difference in each test between the L1 and L2 groups. (See the end of section 2.8.3 for the group-differences approach to construct validity).

As mentioned, GRAM is less difficult than the other tests. Difficulty, however, is a relative concept. "Low" proficiency subjects will always score *relatively* lower than "high" proficiency subjects, *even if the test is difficult for both groups*. The ER scores in the L2 group are very low (all L2 scores were in the 0-39 range; see Figure 5.3) relative to a 50% cut-off point, yet, 13 out of 36 passed Grade 12. ER is a very good predictor of failure not because the scores were very low in absolute terms (i.e. much lower than a designated cut-off point) but because whatever the scores of ER would have been it could predict that a *relatively* low score (under 40) was a good predictor of failure, and a relatively high score, which in this case was a score over 49, was a good predictor of success. An important point is that the whatever the scores of "high" proficiency learners, "low" proficiency learners will have *relatively* much lower scores. So, for the prediction of academic achievement, it doesn't really matter what the scores are in absolute terms as long as they discriminate in such a way that relatively high scores predict success and relatively low scores predict failure. Some of the tests do this better than others, which makes them better predictors.

Whatever the cut-off score for a test, it should be decided in terms of a norm. This is what norm-referenced testing is all about. So, if someone gets (a very low) 30% or a (very high) 80% on a test, it is a "normal" reaction - whether one knows anything about statistical norms or not - to ask: "What did the others get? In other words, "What was the norm?"

5. Dictation Test (DICT) Predictions of Grades 7 to 12 Pass Rate: Figures 5.13 to 5.15

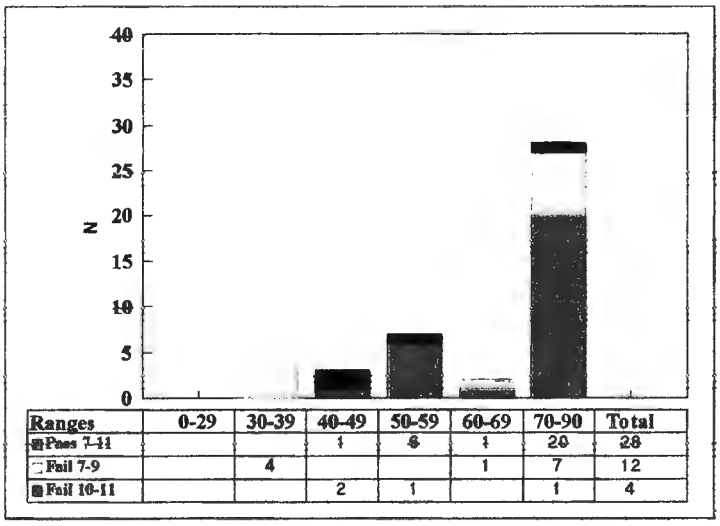
Figure 5.13. DICT Whole Sample (N=81)



Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
Pass 7-11	9		1	9	2	20	41
Fail 7-9	18	4		2	1	7	32
Fail 10-11	2		3	1		2	8

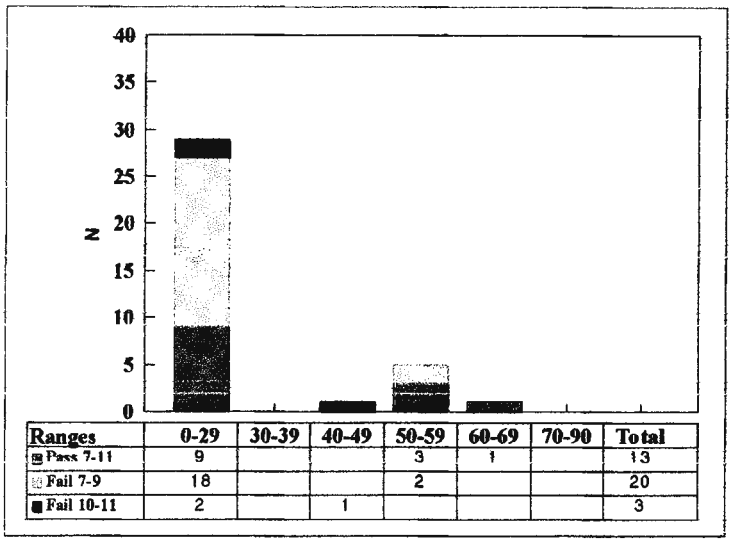
The 0-49 range is a very good predictor of failure (27 out of 37), and the 50-90 range is a very good predictor of success (31 out of 44). Most of the failures occurred in Grade 7 and Grade 8: of the 18 failures in the 0-29 range who failed between Grades 7 and 9, 14 failed Grade 7 and four failed Grade 8. The histograms below show a radical difference between the L1 and L2 groups.

Figure 5.14. DICT L1 Group (N=45)



Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
Pass 7-11			1	6	1	20	28
Fail 7-9		4			1	7	12
Fail 10-11			2	1		1	4

Figure 5.15. DICT L2 Group (N=36)



Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
Pass 7-11	9			3	1		13
Fail 7-9	18			2			20
Fail 10-11	2		1				3

The L1 group did much better in DICT than in CLOZE (Figure 5.8). Why is CLOZE more difficult than DICT for the L1 group when the passages in CLOZE and DICT belong to the same level ("Step 2" of Pienaar's tests)? The reason is that in CLOZE one has to *produce* the correct item, i.e. perform one's competence, without any help. In DICT the sounds produced by the presenter evoke what is stored in one's head, and so it is easier to produce. If one has low competence, as was the case for the majority of the L2 group (20 failures out of 29 in the 0-29 range), the clearest presenter cannot help. The question is whether this competence can progressively improve In this regard, it is noteworthy that nine of the 29 L2s in the 0-29 range passed Grade 12. This means either: (1) their listening and writing skills improved - after all, dictation tests listening and writing, or (2) listening skills are irrelevant to academic performance. As far as the latter is concerned, the research shows (see the review of the literature on dictation in section 3.3.5.2) that listening skills correlate highly with reading and writing. This study has also shown high correlations between dictation tests, cloze tests and essay tests. So it seems that listening skills are a manifestation of global proficiency.

DICT is the best predictor of all the tests. Broadly, there is a clear demarcation between the 0%-49% range, which is a very good predictor of failure (27 out of 37), and the 50-90 range which is a very good predictor of success (31 out of 44). DICT shows that the sample of subjects consists of two radically different levels in the ability to do dictation tests. The fact that the lowest range predicts failure very well and the highest range predicts success very well demonstrates the good predictive validity of DICT.

5.4 General discussion on language proficiency tests as predictors of academic achievement

There are two distinct claims with regard to the role of language proficiency in academic achievement:

- Language proficiency is a prerequisite for academic achievement, i.e. low language proficiency leads to academic failure.
- High language proficiency causes academic success.

The first claim relates to the conditions which make much of academic learning possible, while the second claim maintains that a high level of language proficiency leads to academic success. The data of the study provided statistical evidence for the first claim. With regard to the second claim, the data showed that learners with very high language proficiency are generally successful. I discuss the above two claims and the findings in the light of the literature of language proficiency as a predictor and cause of academic achievement.

Many research studies indicate that the kind of (foreign) language proficiency tapped by tests such as TOEFL, which have consisted - until recent years - of multiple-choice

tests, is a poor predictor of academic achievement.[67] TOEFL is also considered a poor predictor of communicative proficiency[68]. If the findings of these aforementioned studies are right, one would have to make sense of the findings of

> *seven concurrent validity studies in the United States between 1965 and 1968 [where] the correlations (Pearson r) between TOEFL and various tests, including the American Language Institute Tests of Proficiency, the "cloze" tests, and tests developed at various universities, varied between 0.79 and 0.89...When TOEFL scores were compared with teacher ratings, scores on written themes, or judgements of students' ability to pursue regular academic courses, the Pearson r ranged from 0.73 to 0.79.*[69]

Surely "students' ability to pursue regular academic courses" (Gue and Holdaway above) is what communicative proficiency in the school situation is mainly about. There appears to be conflicting findings on the issue: those of Gue and Holdaway above, on the one hand, and those of authors such as Van Lier[70] and Hale et al.[71], on the other. (Hale et al. will be discussed shortly). Perhaps the conflict is only apparent in that Gue and Holdaway's "ability to pursue regular academic courses" is not the same as the ability to succeed in an academic programme. Gue and Holdaway's context is the potential for success, not its inevitability. Thus, it is possible that TOEFL can assess whether an applicant is able to pursue an academic programme, yet it may still not be a good predictor of academic performance. In other words, if one's knowledge of the medium of instruction is poor, one will generally not be able to succeed in an academic programme. But if one's knowledge of the medium of

---

67 (1) Hale, G.A., Stansfield, C.W. and Duran, R.P. *TESOL Research Report 16*, 1984.
(2) Mulligan, A. C. Evaluating foreign credentials. *College and University*, 41, 307-313 (1966).
(3) Upshur, J.A. 'English language tests and predictions of academic success', in Wigglesworth, D.C. (ed.). *Selected conference papers of the Association of Teachers of English as a Second Language*. Los Altos, California, National Association for foreign Student Affairs (NAFSA) Studies and Papers, English Language Series 13, 85-93 (1967).
(4) Van Lier, L. *The classroom and the language learner*, 1988.

68 Van Lier, L., ibid., p. 233.

69 Gue, L. and Holdaway. 'English proficiency tests as predictors of success in graduate studies in education.' *Language Learning*, 23, 89-103 (1973), p.92.

70 Hale, G.A., Stansfield, C.W. and Duran, R.P. *TESOL Research Report 16*, 1984.

71 Van Lier, L. *The classroom and the language learner*, 1988.

instruction is good, this does not mean that one will necessarily succeed. Academic success depends on much more than proficiency in the medium of instruction. We take a closer look at Hale, Stansfield and Duran's position:

> *TOEFL scores are not appropriate predictors of future grades. Admissions decisions presumably should be made by first examining a student's past academic record and then using TOEFL scores to help determine whether the student has the necessary proficiency to do the required academic work. Yet TOEFL does provide predictive information of a sort. If the role of English proficiency in different programs of study were to be determined, guidelines could be established concerning the meaning of TOEFL scores for students at different levels of the programs of study.*[72]

The following appears to be Hale et al.'s argument: although TOEFL is a poor predictor of academic achievement, it nevertheless plays a secondary role in ascertaining the potential to achieve, but the primary predictive role should be given to past academic achievement. The implication appears to be that besides the observation that TOEFL multiple-choice tests are poorer predictors than past academic achievement, another observation is that all language proficiency tests are poor predictors of academic achievement, because if language tests were good predictors there would be no need to use a student's past academic record to predict future grades.

According to the several reports in Hale et al.[73], correlations between language proficiency tests and achievement tests, even when measured within the same year, have been found to be so low that there does not seem to be any meaningful relationship between language proficiency and academic achievement. Consider the following finding that "correlations with GPA [Grade Point Average] were lower for the second semester than the first [which] may be due to improvement in student's

---

72 Hale, G.A., Stansfield, C.W. and Duran, R.P., 1984.

73 For example, ibid, pp.115 and 177.

English skills during the first semester, which tends to reduce the role of language ability in determining academic success".[74]

Hale et al.'s reasoning seems to be the following: "English skills" (as measured by TOEFL) do not develop in tandem with general academic skills, and so it would not be possible to detect any normative pattern in the relationship between "English skills" (i.e. English proficiency) and academic achievement. Although the low correlations in Hale et al. suggest that language proficiency tests are poor predictors of academic success, these correlations do assess, according to Hale et al.[75], the minimum level of language proficiency required for academic success, which makes language proficiency tests valid predictors of academic failure. In other words, if learners have high English proficiency this does not necessarily indicate that they will be academically successful, but if they have limited English proficiency one can predict with greater certainty that they will be academically unsuccessful.

An important consideration in the assessment of levels of language proficiency is the minimum level that has just been mentioned. (See Carroll's "ability" in section 2.2, point (3) on the minimum, or "liminal", level). In trying to set a minimum level, one is concerned with what the individual can do in terms of established criteria. But, as I continue to emphasise, what the individual can do cannot be separated from what others can do.

5.5 The reliability and predictive validity of the Grade 6 reports of previous schools

5.5.1 Introduction

This section examines the reliability of the Grade 6 reports of former schools in order to (1) further substantiate the validity of the English proficiency tests as predictors of academic achievement, (2) substantiate the reliability of MHS's achievement

74 Hale, G.A., Stansfield, C.W. and Duran, R.P., 1984, pp.178-179.
75 Ibid., p.198.

measures, and (3) investigate Hypothesis 3 of the study, namely, *many of the former school reports (Grade 6 reports) that were used as criteria for admission to MHS were not valid predictors of academic achievement.* This hypothesis was based on the observation that many of the entrants with high Grade 6 report scores of former schools did not get beyond Grade 9 at MHS, which suggests that the Grade 6 scores were inflated and accordingly unreliable. The investigation involves using the English proficiency tests, Grade 6 achievement and Grade 7 achievement to predict the Grade 12 pass rate.

5.5.2 Historical background

The former DET (Department of Education and Training) Grade 12 ("matric") results have been disappointing for a number of years, and usually required substantial adjustments upwards. There was a decline in the DET pass rate from 48% in 1985 to 41% in 1991 to 38% in 1993.[76] The results have not been much better since the dissolution of the DET after the democratic elections of 1994. The overall South African Grade 12 pass rate of 1995 was 55,2%, which was almost 3% lower than 1994.[77] The university exemption rate for the whole country for 1995 was 17,9% for all races: 78% for Indians, 55% for whites, and 11% for blacks.[78] The overall South African 1997 Grade 12 results were worse than 1995, with an average 47% pass rate for the country, with some provinces as low as 35% and 45%.

Various reasons for the low pass rate have been suggested in academia and the media. Reasons given in academia are: (1) Bantu Education[79], which is claimed to be the direct cause of the low level of English proficiency among teachers and the low level

---

[76] Calitz, F. 'So what went wrong with the matric class of 97?' *Sunday Times (South Africa)*, January 11, 1998, p.14.

[77] St. Leger, C. 'Radical steps proposed for education after matric results shock.' *Sunday Times (South Africa)*, December 31, 1995a, p.1.

[78] St. Leger, C. 'Depressing statistics from years of turmoil.' Sunday Times (South Africa), December 31, 1995b, p.4.

[79] Hartshorne, K. 'Language policy in African Education in South Africa, 1910-1985', in Young, D. (ed.). *Bridging the gap*, 1987.

of English proficiency among learners, (2) the medium of instruction and learning from Grade 5 onwards (English) is a language which is non-cognate to the learner's first language[80], and (3) low academic ability.[81] Reasons given in the media are: irrelevance of the contemporary school system to real life, absence of a culture of learning and teaching, an impoverished primary school and preschool background, a pass-one-pass-all mentality, demoralisation and disillusionment of teachers, irresponsibility of teachers, poor administration by the Minister of Education and by the provinces, lack of commitment from the business sector, strikes encouraged by teachers' trade unions, and a general breakdown in society.

Another probable cause for low matriculation pass rate is indiscriminate advancement through the grades. According to Calitz[82] and Educamus[83], the educational casualty figures would have been much higher if automatic promotions, or indiscriminate advancement, did not occur in individual schools from one grade to the next. The report "Investigation into the causes of the unsatisfactory 1989 Std 10 [Grade 12] results" states:

> *At some of the schools visited, there was the view that it was not necessary to have condoned marks/results approved officially. A decision was taken by the school or teacher on whether a pupil should pass or fail and the marks*

---

80 Mascher, D. *The disintegration of an education system based on a non-cognate medium of instruction.* Paper presented at the South African Applied Linguistics Conference, University of the Witwatersrand, July 1991.

81 (1) Gamaroff, R. 'Solutions to academic failure: The cognitive and cultural realities of English as the medium of instruction among black ESL learners.' *Per Linguam*, 11 (2), 15-33 (1995c).

2) _____ 'Abilities, access and that bell curve.' Grewar, A. (ed.). *Proceedings of the South African Association of Academic Development "Towards meaningful access to tertiaty education"*, 1996b.

(3) _____ 'Language as a deep semiotic system and fluid intelligence in language proficiency.' *South African Journal of Linguistics*, 15 (1), 11-17 (1997b).

82 Calitz, F. 'So what went wrong with the matric class of 97?' *Sunday Times (South Africa)*, January 11, 1998, p.14.

83 Educamus. *Editorial: Internal promotions*, 36 (9), 3 (1990).

> *were adjusted accordingly. No criteria existed in terms of which marks were condoned*[84].

This random condoning of marks, which resulted in random promotions, started, Edcumas maintains, in the Primary school. Educamus (1990:3) cites the following data to substantiate its claim:

> *In November 1988, approximately 84 per cent of the pupils in Primary School passed their final examinations, as did approximately 66 per cent of the pupils in Std 6 to Std 9 [Grade 8 to Grade 11]. The pass rate for Std 10, however, was only 40,6 per cent. The sharp decrease in the pass rate of the Standard 10 [Grade 12] pupils in comparison with the rest of the standards is an indication that promotions in the lower standards leave much to be desired.*

These statistics indicate that much appears rosy in the garden until harvest time - the matric examination. One would think that a Grade 11 pass (an internal examination) would generally imply a Grade 12 pass (an external examination), *if* the Grade 11 result was a true reflection of the learner's worth. As far as I am aware there exists no empirical evidence to substantiate the claim that indiscriminate advancement occurred through the Grades at DET schools. The next section describes an empirical investigation of this issue.

5.5.3 An examination of the Grade 6 reports

The English proficiency tests, Grade 6 and Grade 7 aggregates are used to predict the Grade 12 pass rate.

Table 5.5 contains a summary of the predictions. In each cell the number *inside brackets* represents the total *failures* and the number *outside brackets* represents the total *passes*. The total number of subjects in each cell would be the total inside brackets plus the total outside brackets. The 60-90 range appears in bold because this range reveals very clearly the problem under investigation.

[84] Educamus, ibid.

TABLE 5.5

Grade 12 Pass Rate of L1 and L2 Groups with Three Predictors

N=69 (L1:N=38 ; L2:N=31)

RANGES (%)	PROF		AGGR7		AGGR6	
	L1	L2	L1	L2	L1	L2
0-39		7(21)		(2)	(1)	
40-59	7(7)	2(1)	3(8)	3(16)	1(2)	1(4)
60-90	**18(6)**		**22(5)**	**6(4)**	**24(10)**	**8(18)**
Total	25(13)	9(22)	25(13)	9(22)	25(13)	9(22)

For more clarity, histograms based on Table 5.5 are shown on the next two pages.

Table 5.5 shows that all three predictors are good predictors in the L1 group. Recall that the vast majority of L1 Grade 6 reports are those of CM Primary School. The L2 Grade 6 reports show a contrary picture. As shown, 26 of the 31 in the L2 group obtained an AGGR6 score in the 60-90 range, yet only eight of these 26 passed Grade 12 (see Figure 5.21 for more clarity). These L2 Grade 6 reports belonged to the "other schools", most of which were DET schools. The following histograms of the L1 and L2 groups provide a clear overview of the predictions.

Figure 5.16. L1 PROF as a Predictor of Grade 12

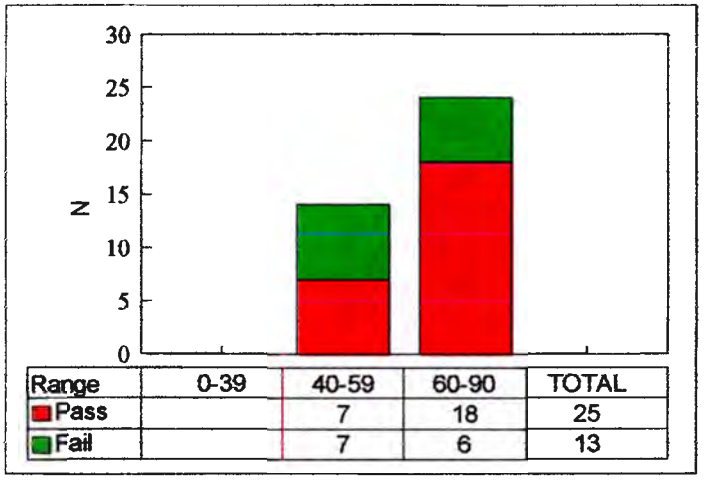
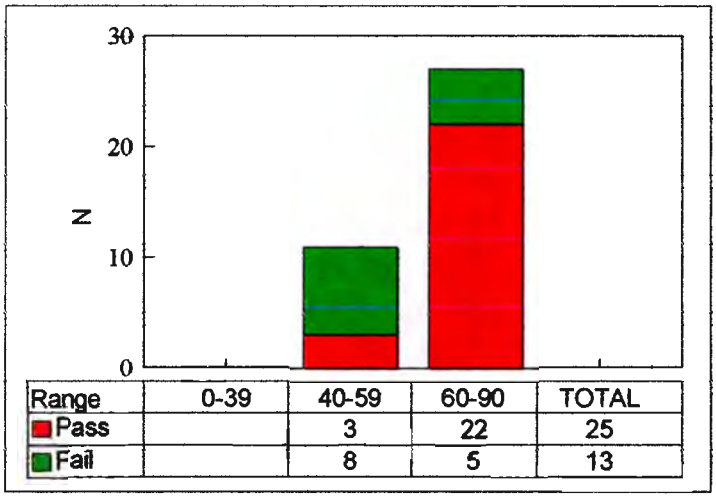
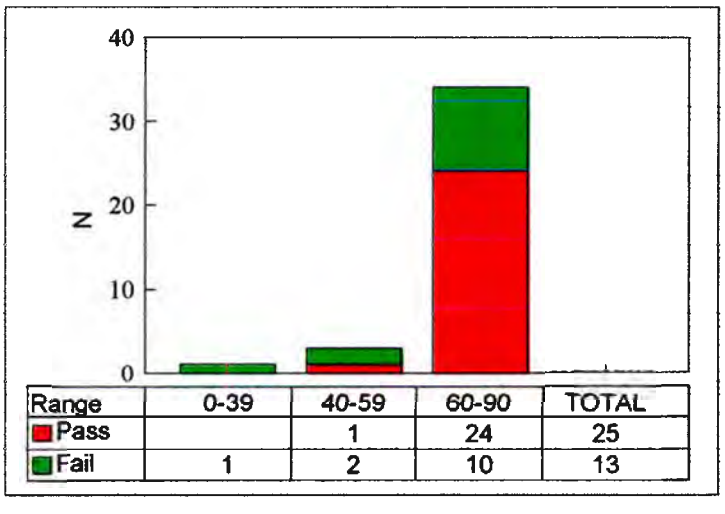


Figure 5.17. L1 AGGR7 as a Predictor of Grade 12



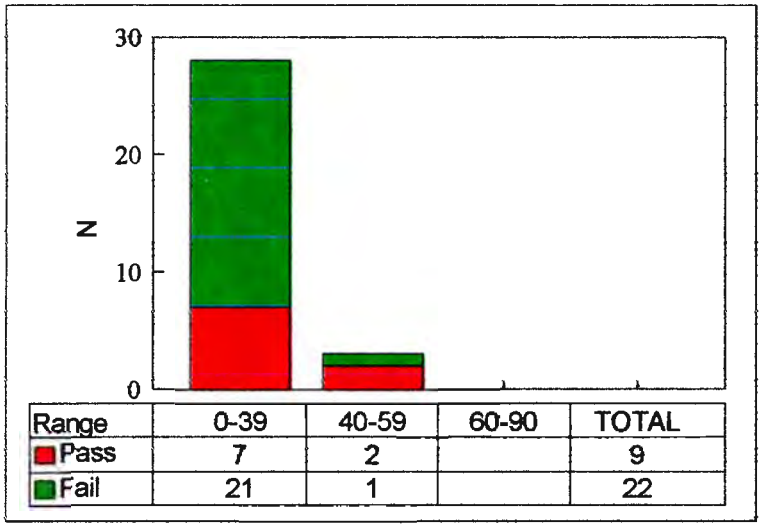
Range	0-39	40-59	60-90	TOTAL
Pass		3	22	25
Fail		8	5	13

Figure 5.18. L1 AGGR6 as a Predictor of Grade 12



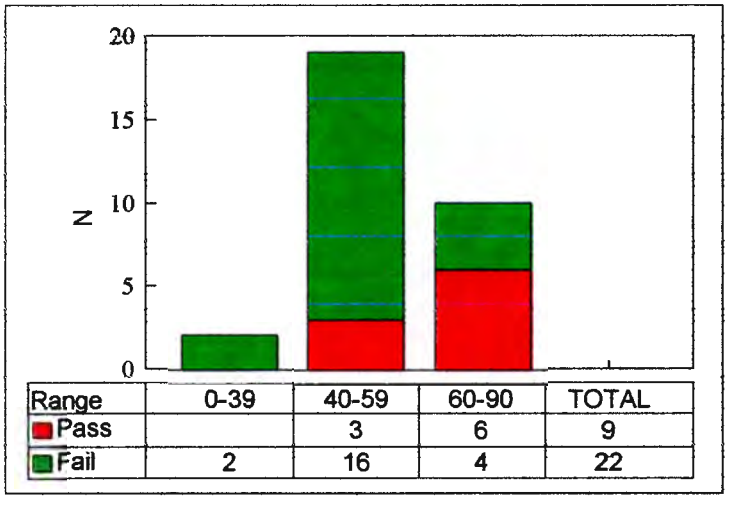
Range	0-39	40-59	60-90	TOTAL
Pass		1	24	25
Fail	1	2	10	13

Figure 5.19. L2 PROF as a Predictor of Grade 12



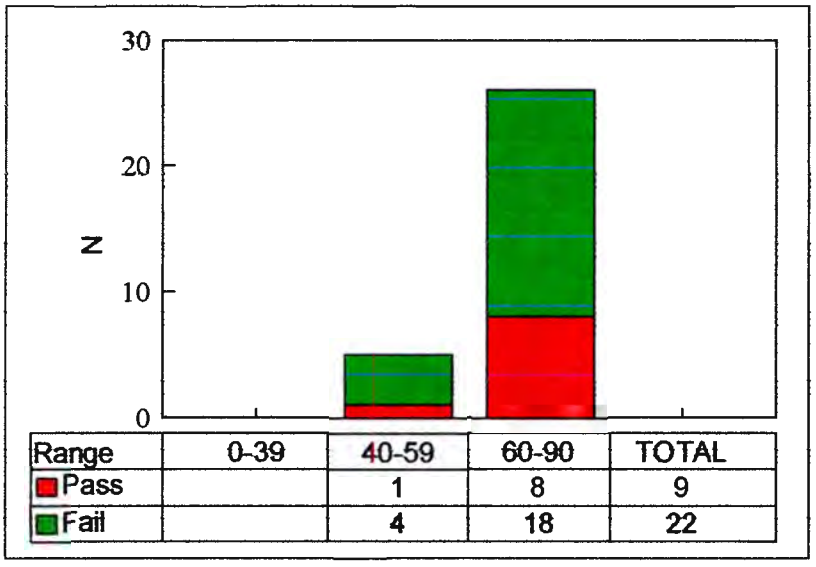
Range	0-39	40-59	60-90	TOTAL
Pass	7	2		9
Fail	21	1		22

Figure 5.20. L2 AGGR7 as a Predictor of Grade 12



Range	0-39	40-59	60-90	TOTAL
Pass		3	6	9
Fail	2	16	4	22

Figure 5.21. L2 AGGR6 as a Predictor of Grade 12



Range	0-39	40-59	60-90	TOTAL
Pass		1	8	9
Fail		4	18	22

I focus on the 60-90 range because it reveals very clearly the problem under investigation. The 60-90 range should be a good predictor of success. All three predictors of the L1 group in the 60-90 range were good predictors:

60-90 Range

1. In L1 PROF, 18 out of 24 passed. (Figure 5.16).
2. In L1 AGGR7, 22 out of 27 passed. (Figure 5.17).
3. In L1 AGGR6, 24 out of 34 passed. (Figure 5.18).

The L2 Grade 6 reports show a contrary picture:

1. In L2 PROF there were no scores in the 60-90 range. (Figure 5.19).
2. In L2 AGGR7, there were only 10 subjects in the 60-90 range of whom 6 passed. Figure 20).
3. In L2 AGGR6, there were 26 subjects in the 60-90 range of whom only eight passed. (Figure 21).

As shown, the vast majority of both the L1 and the L2 groups obtained AGGR6 scores in the 60 - 90 range. The question is how to explain the fact that in the L1 group the 60-90 range of ENG6 is a good predictor, while in the L2 group, the same range of ENG6 is a poor predictor. A probable explanation is that L2 ENG6 was inflated. Important is the observation that very few subjects in both groups who passed Grade 9 failed Grade 12. Thus, the period between Grade 7 and Grade 9 is a crucial period.

5.6 Summary of the findings of the study and their generalisability

5.6.1 Summary of the findings

This study tested three null hypotheses. These are repeated here:

Hypothesis 1. Discrete-point tests and/or integrative tests are *not* valid measures of levels of language proficiency.

Hypothesis 2. Discrete-point tests and/or integrative tests are *not* valid long-term predictors of academic achievement.

Hypothesis 3. Many of the reports (Grade 6) from former schools that were used as criteria for admission to MHS were *not* valid predictors of academic achievement.

With regard to **Hypothesis 1**:
All the tests of the battery were valid measures of levels of language proficiency, where there was a clear distinction between the L1 and the L2 groups. The null hypothesis is, therefore, rejected.

With regard to **Hypothesis 2**:

In the choice of the best tests for the prediction of academic achievement (the pass rate, in this case), the idea was to choose tests that discriminated well between strong and weak learners.

What I wanted to test was the long-term predictive validity of the tests. DICT (Figure 5.13) does the best job: a score over of 50% predicts a pass rate of 31 out of 44 subjects, and score under 50% predicts a failure rate of 27 out of 37 (these data include all 81 subjects). The second best test is ER (Figure 5.1: a score over 50% predicts a pass rate of 15 out of 20 subjects, and a score of under 30% predicts a failure rate of 11 out of 35 (these data include 55 of the 81 subjects). The next best test is GRAM (5.10): a score over 70% predicts a pass rate of 20 out of 31 subjects, and a score of under 40% predicts a failure rate of 11 out of 15 (these data include 46 of the 81 subjects). *Recall that in the multiple regression analysis (Table 5.2), ER was found to be a poor predictor.* The frequency distributions show otherwise. CLOZE and ESSAY were found to be good *short-term* predictors.

Hypothesis 2 is therefore partly rejected, because not all of the tests, e.g. CLOZE and GRAM, were good long-term predictors.

This does not necessarily mean at all that CLOZE and ESSAY are unrelated to long-term academic achievement. It could mean that writing and reading skills do not develop in tandem with general academic skills, and so it would not be possible to detect any normative pattern in the relationship between reading skills (CLOZE) and writing skills (ESSAY), on the one hand, and academic achievement, on the other (see section 5.4). Annual essay and cloze tests might have produced much better predictions, and that is why, with regard to CLOZE, (1) Pienaar's tests are graded in "Steps" from Grade 3 to Grade 12 and beyond to Grade 12+ and (2) CLOZE was a better short-term predictor than a long-term predictor.

If the reason why CLOZE (specifically the 60-90 range) is not a good long-term predictor is because reading skills do not develop through the grades in tandem with general academic skills (see previous paragraph), an interesting hypothesis to investigate is whether listening skills, as tested in DICT (a good long-term predictor), *do* develop in tandem with academic achievement.

Owing to the fact that some tests were relatively more difficult than others for the whole sample, e.g. ER was more difficult than GRAM, one cannot decide in an *absolute* sense that the best tests would be those that are good predictors in the lowest ranges of 0-39 and the highest ranges of 60-90. As I have pointed out, the "high" and "low" ranges are relative terms and refer to different scores depending on the difficulty of the test. For example, in ER (Figure 5.1) more than half the subjects were in the 0-39 range while in GRAM (Figure 5.10) only 15 subjects were in this range.

Owing to the fact that the Grade 7 sample of subjects described in this study is becoming representative of many schools in South Africa, it might be interesting to use at other schools some of the tests in this study or similar tests to predict academic achievement. The predictions of the whole sample in each test provide the best guide in this regard because it is highly unlikely in the new politics of "multicultural settings" that one would overtly categorise levels of proficiency in terms of labels such as L1 and L2, whether one means by these labels mother tongue and non-mother tongue, respectively, or as I have used the terms, namely, English First Language as a subject and English Second Language as a subject, respectively. (See section 6.2).

With regard to **Hypothesis 3**:

Recall that this hypothesis stated that many of the reports (Grade 6) from former schools that were used as criteria for admission to MHS were *not* valid predictors of academic achievement (as shown by the pass rate). It was found that the L1 Grade reports were valid predictors of academic achievement but that the L2 Grade reports

were not valid predictors of academic achievement. The reason why the L2 Grade 6 reports were not valid predictors was probably because the Grade 6 report scores were inflated, and, hence, unreliable. The null hypothesis is rejected, therefore, only for the L1 group, because the L2 reports were not valid predictors of academic achievement.

5.6.2 The generalisability of the findings

I now discuss the generalisability of the findings, specifically the predictions. With regard to the L1 group (i.e. those who took English as a First Language subject), the majority of these came from a CM Primary School that had English as the medium of instruction from Grade 1. Such primary schools were rare in the North West Province at the time this research was conducted, and so one could not have replicated the research or generalised the findings of this study to the wider population of the North West, that is, not until recent years. Since the 1994 elections, however, the situation has radically changed in that there are now many former privileged, or "white" schools that have opened their doors to disadvantaged, usually black, learners. Accordingly, the sample in this study of a mixture of learners with a wide range of English proficiency is becoming common in the urban areas of South Africa.

I would like to focus on the L2 group, which is, in any case, is more pertinent to this study. The argument is based on the fact that the Hypothesis 3 was rejected as far as the L2 group was concerned, that is, the L2 Grade 6 reports were not valid predictors of academic achievement - long-term or short-term. The reason: the Grade 6 DET reports were unreliable.

One might concede that the study, which showed the high failure rate of DET learners has internal validity, i.e. it is valid for learners at MHS, but of added interest is the generalisability of the findings.[85] In other words, do the subjects described in this study represent a population outside MHS?

---

[85] Pilliner, A.E.G. *Experiment in educational research*, 1973, p.43.

According to the results of Pienaar's[86] cloze tests, the vast majority of learners at DET schools were "at risk". There is no evidence from Pienaar's research or the research of this study, or other research[87] that the situation has improved.

A large proportion of the subjects in this study who had been described as disadvantaged, namely, those belonging to the L2 group, had low English proficiency - as measured by the English proficiency tests - and did not get beyond Grade 9 at the school. This supports the general view that an initial (in this case Grade 7) low level of English proficiency will result in educational failure. Also, a very high level of English will probably result in educational success.

The vast majority of these L2 learners originated from DET schools. Recall that they were admitted to the school on the basis of their high Grade 6 report scores. The fact that there was such a high dropout rate among them and that most of them obtained low English proficiency scores on the tests strongly suggested that the DET Grade 6 reports were inflated. If inflated marks occurred at the DET schools in this study, as I have argued, it is possible that this also occurred in other DET schools as well. The inflation of marks is probably the reason for indiscriminate advancement, or "grade creep". The high failure rate of DET learners at MHS shows that there are major assessment problems in former DET schools.

---

[86] Pienaar, P. *Reading for meaning: A pilot survey of (silent) reading standards in Bophuthatswana*, 1984.

[87] (1) Macdonald, C.A. *English language skills evaluation (A final report of the Threshold Project), Report Soling-17*, 1990a.

(2) Macdonald, C.A. *Crossing the threshold into standard three in black education: The consolidated main report of the Threshold Project.* Pretoria, Human Sciences Research Council, 1990b).

(3) In a recent project conducted by Makoni (personal communication) - in collaboration with the National Language Project - Grade 7 second language (ex-DET) speakers scored 21% on an English proficiency test administered to Grade 3 first language speakers who scored 75%. (See the end of section 6.2).

It might be risky to make judgements about the academic quality of the individual DET schools from which the L2 group originated, owing to the fact that in most cases not more than one or two L2 subjects originated from the same school. Nevertheless, it does seem to be more than coincidence that those L2 subjects at MHS who failed between Grade 7 and Grade 12 and had inflated Grade 6 reports, *were chosen from their respective schools as examples of academic excellence*, and were thus judged competent enough to enter a school that offered the JMB syllabus or a syllabus of an equivalent standard, which, as pointed out earlier, was considered to be of a much higher standard than other syllabuses such as that of the National Senior Certificate offered by the DET.

If these schools had chosen their mediocre achievers instead of their supposedly achievers to sit the entrance tests of MHS, it would not be possible to make valid inferences, because one wouldn't be able to tell whether the "bright" learners at these schools would have done better at MHS. All the other learners from these schools that had not been selected couldn't have been better than the ones selected. If those selected - the "high" achievers - were to fail at MHS, which they did, it would be logical to infer that the rest of the learners at these schools would probably have also failed at MHS, because they would probably also have had low English proficiency and limited academic ability.

The crux of the matter is this: Very few disadvantaged children who entered MHS in Grade 7 were able to fulfil the demands of the JMB syllabus and to benefit from the relatively enriching academic facilities offered at MHS. This was not only true of the specific sample in this study, but also - according to my more than seven years experience at MHS - of many learners who had attended MHS since its inception in 1980. A change of environment to a more advantaged setting such as the one that existed at MHS seemed, in many cases, to have had little significant effect on academic performance. This is driven home by the high failure rate among subjects with low scores on the English proficiency tests.

It would be incorrect, however, to infer that learners who obtain low scores in tests and examinations necessarily have a disadvantaged background, because deciding who is disadvantaged or not, depends on more than tests - admission tests or other kinds of tests. Valid decisions in this regard should also be based on the educator's personal knowledge of the kind of background that learners come from. In general, well-trained and experienced educators are able to make valid judgements about whether learners in their school are "disadvantaged".

I end this chapter with a mention of the role of background, specifically the role of English input. The role of input in the form of early exposure to English at home and the use of English as the medium of instruction from Grade 1 has been shown in other research to be a better predictor of English achievement than language proficiency tests such as TOEFL.[88] In this study, subjects who had had early exposure to English and had used English as the medium of instruction from Grade 1 (most of these were the CM Primary entrants) performed much better, on average, than the L2 group, who, in most cases, probably did not have early exposure to English, but all had English as the medium of instruction only from Grade 5 (officially, that is, because the medium of instruction rarely began in earnest in Grade 5).

Also the less the language distance between English and the mother tongue of learners, the greater the possibility of achieving in English later on. Thus, speakers of Romance languages are likely to learn English faster than speakers of Tswana, Xhosa, Russian and Japanese.[89] Caution is required in making such predictions based merely on language *structure*, because there are many other factors involved such as cognitive and cultural factors. For example, Russian is structurally further from English than Xhosa, yet Russian culture has more in common with English culture - owing to the fact that they are both more "Western" than Xhosa culture.

---

[88] Wilhelm, K.H. 'Use of an expert system to predict language learning success.' *System*, 25 (3), 317-334 (1997).

[89] Ibid., pp.325-326.

It must be emphasised that although the Grade 12 pass rate for the L1 group was much higher than the L2 group, it is still worrying that 17 out of 45 L1 subjects did not get as far as Grade 12. This shows that there is far more to academic achievement than early exposure to the medium of instruction, as I have mentioned on several occasions.

5.6 Summary of Chapter 5

This chapter consisted of an examination of the predictive validity of the English proficiency tests. High correlations were found between the tests and GRADE 7 (aggregate), but these correlations decreased sharply after Grade 7. The correlations suggested that English proficiency that was tested at the beginning of Grade 7 ceased to be a valid predictor of academic achievement beyond Grade 7. Frequency distributions were far more revealing and showed that English proficiency was found to be a good predictor of academic achievement in the low and very high ranges. DICT, ER and GRAM were found to be the best predictors.

The Grade 6 reports were also examined and it was found that the DET reports were inflated and consequently were poor predictors of academic achievement.

What is important in this investigation is not the equivalence or non-equivalence of scores between tests but the relative difference between high ability subjects (in this case the L1 group) and low ability subjects (in this case the L2 group). This is an important factor in the construct validity of the tests that have been used. It was also argued that although ESSAY and CLOZE were not good *long-term* predictors this did not mean that they were unrelated to long-term academic achievement.

CHAPTER 6

Implications for Testing and Conclusions

6.1 Introduction

In previous chapters I dealt with the validity, reliability and practicability of using discrete-point tests and integrative tests for the assessment of language proficiency and for the prediction of academic achievement. I argued that tests do not have to be direct to be authentic and that in proficiency testing one can justifiably use indirect tests such as grammar tests, cloze tests and dictation tests to predict communicative, or real-life, language ability. This chapter contextualises the issues, raised in previous chapters, within language testing in South Africa. Although the specific language discussed is English, some of the ideas discussed are applicable to other languages in South Africa.

Much has been written in South Africa on language testing in curriculum development in recent years.[1] In the section on "testing" in their article on curriculum development in language teaching, Barkhuizen and Gough remark that "language planners often underestimate the power of testing and examining."[2] (Their article is about language teaching with implications for language testing).

If language planners underestimate this power, they surely must be out of touch with what teachers think. A group of planners who has forefronted assessment issues is the group involved in the "Language assessment and national qualifications framework"[3] that is closely linked to "outcomes- based" education. (This document is discussed in section 6.4 where I shall examine the dramatic changes envisioned by the onset of "outcomes-based"

---

1 (1) Barkhuizen, G. and Gough, D. 'Language curriculum development in South Africa: What place for English?' *TESOL Quarterly*, 30 (3), 453-461 (1995).
(2) HSRC. *Ways of seeing the National Qualifications Framework*, 1995.
(3) HSRC. *Language assessment and the National Qualifications Framework*, 1996.

2 Barkhuizen, G. and Gough, D., ibid., p.464.

3 HSRC, *Language assessment and the National Qualifications Framework*, 1996.

education as contained in the "Language assessment and national qualifications framework" proceedings).

English assessment - as with the assessment of other languages in South Africa - is very complex, where languages are assessed as first, second and third languages, at standard grade and higher grade, where there are many different English syllabuses, and where independent and provincial organisations set their own assessment procedures.

Barkhuizen and Gough[4] raise many issues that will determine the future pattern of language assessment in South Africa. I single out three issues: The first, which has to do with the "levels of proficiency" issue, is whether language teaching and assessment should maintain the L1/L2 distinction, which is a controversial issue in South Africa as well as internationally. The second is the connection between norm-referenced tests and "negotiating the task-demands".[5] The third, and perhaps thorniest issue, involves the problems of establishing rater reliability when only one rater is available, which is the normal situation in a teaching context.

Although Barkhuizen and Gough do not explicitly mention scores or the forms that future tests may take, the second and third issues have much to do with whether language tests and/or their scores will count or even be used in the new education dispensation in South Africa. The first issue is discussed in section 6.2, the second in sections 6.3 and 6.4, the third also in section 6.4, and the fourth in section 6.5.

6.2 The L1/L2 and native speaker/non-native speaker distinctions.

The sample of subjects in the empirical investigation described in the previous two chapters consisted of a culturally, ethnically and linguistically diverse group with a wide

---

[4] Barkhuizen, G. and Gough, D. 'Language curriculum development in South Africa: What place for English?' *TESOL Quarterly*, 30 (3), 453-461 (1995), p.465.

[5] Macdonald, C.A. *English language skills evaluation (A final report of the Threshold Project)*, 1990a.

Macdonald, C.A. *Crossing the threshold into standard three in black education: The consolidated main report of the Threshold Project*, 1990b.

The sample of subjects in the empirical investigation described in the previous two chapters consisted of a culturally, ethnically and linguistically diverse group with a wide spread of English proficiency that is progressively becoming the norm in South African urban schools, previously "white" schools. MHS has already had 19 years experience dealing with linguistic, cultural and educational problems, which are only now beginning to surface in many schools in South Africa. The School was one of the few schools in the North West Province that had English-mother-tongue speakers, Bantu-mother-tongue speakers, Afrikaans-mother-tongue-speakers and mother-tongue speakers of other languages in the same classroom (except for the language subjects), where a large part of such a class contained learners with a relatively low level of English proficiency. In contemporary schools even the English Language class contains a hybrid of what were formerly referred to as L1 and L2 speakers. This is becoming the norm in urban schools. Accordingly, one of the major contemporary problems is how to teach (and test) the same English syllabus in the same classroom to learners with a wide range of language proficiency. This is particularly pertinent to English Second Language (ESL) learners, who are of special interest in this study. In this section I want to focus on the *levels* issue, which has been dealt with statistically in previous chapters, in order to examine how it affects ESL learners.

A major feature of this study is the problematic distinction between L1 and L2 *levels* of language proficiency. In this study L1 and L2 learners referred to learners at MHS who took English as a First Language or as a Second Language, respectively. In this section I elaborate on the L1/L2 distinction because of its central importance in this study. I include in the discussion the controversial notions of "mother-tongue speaker" and "native speaker".

Owing to the fact that there is an increasing uncertainty of what L1 language ability means, the literature is sharply divided on the issue.[6] This problem is indicative of the

---

[6] (1) Davies, A. *The native speaker in applied linguistics*, 1991.
(2) _____ 'Proficiency or the native speaker: What are we trying to achieve in ELT?', in Cook, G. and Seidlhofer, B. *Principle and practice in applied linguistics: Studies in honour of*

much larger problem of the indeterminacy of language (and hence also of epistemology) itself. "There is no perfect hypothetical language, to which the languages we have are clumsy approximations."[7]

The distinction between L1 and L2 has two descriptive connotations: (1) L1 is acquired before L2. L1 need not be the mother tongue, because the mother tongue may not end up as L1; (2) The L1 learner is stronger than the L2 learner, where the L2 learner is unlikely (the weak interpretation) or incapable (the strong interpretation) of reaching the L1 level.[8]

There is a tendency in South Africa to disembarrass education of the distinction between L1 and L2 because it refers demeaningly, it is argued, to "the level of competence of the speaker" and not to "curriculum strategies to meet the needs of the learners and the dictates of the learning context".[9] This tendency is no where clearer than in the recent thought behind language assessment and the National Qualifications Framework.[10] The L1/L2 dichotomy for many progressive theorists in South Africa has given rise to a "wide range of discriminatory language requirements and discriminatory assessment instruments."[11] I first examine South African views on the issue and then relate them to views outside South Africa.

Young advocates that the "apartheid" labels of ESL, L1 and L2 be discarded.[12] In support of Young's rejection of the apartheid labels of L1 and L2, others argue that it is

---

H.G. Widdowson, 1995).

(3) Makoni, S. "Language and identity in Southern Africa", in De la Gorgendiere, L, King, K and Vaughan, S. (eds.). *Ethnicity in Africa: Roots, meanings and implications*, 1996.

(4) Paikeday, T.M. *The native speaker is dead!*, 1985.

7 Harris, R. *The language myth*, 1981.

8 Musker, P. and Nomvete, S. 'Standards and levels in language assessment', in HSRC. *Language assessment and the National Qualifications Framework*, 1996.

9 Burroughs, E., Vieyra-King, M. and Witthaus, G. 'The assessment of language outcomes in ABET: Implications of an approach', in HSRC, *Language assessment and the National Qualifications Framework*, 1996, p.77.

10 HSRC. *Language assessment and the National Qualifications Framework*, 1996.

11 Musker, P. and Nomvete, S. 'Standards and levels in language assessment', in HSRC. *Language assessment and the National Qualifications Framework*, 1996, p.65.

12 Young, D. 'English for what and for whom and when?' *Language Projects Review*, 3 (2), 8 (1988).

discriminatory (in the worst sense) to compare levels of proficiency between L1 and L2 learners.[13] Burroughs et al. don't object to the comparison between *levels*: one couldn't do so without making nonsense of assessment, because without the differences in ability between learners there would be no construct validity. What they do object to is the *sociopolitical* meaning of attaching the labels of L1 and L2 to different levels.[14] Burroughs et al.'s view is similar to Young's above.

Young suggests that there be a uniform matriculation examination in English, graded in levels of difficulty. He also suggests that these levels of difficulty "be selected by candidates on the basis of their known competence in English"[15]. This does not mean that matriculation candidates' choices of exam level will not be guided by their teachers. Young's suggestion that one choose one's own level of difficulty has been happening at MHS not only for the higher grades but from Grade 7 onwards. Recall that the Grade 7 sample of the study were allowed to decide themselves whether they wanted to take English First Language or English Second Language at the School. In later grades it was possible for teachers to intervene and advise those weak at English to move from L1 to L2: advice that was invariably heeded. Rarely, if ever, did learners move from L2 to L1, even if they were good at English.

Planners of new policies of language assessment maintain that "the separation between first and second language is based on a positivistic construct of language."[16] What might

---

13 (1) African National Congress. *ANC policy guidelines for a democratic South Africa*, 1992).

(2) Burroughs, E., Vieyra-King, M. and Witthaus, G. 'The assessment of language outcomes in ABET: Implications of an approach', in HSRC, *Language assessment and the National Qualifications Framework*, 1996.

(3) Forrest, F. *A language curriculum framework for compulsory general education*, 1992.

14 Rampton, B.H. 'Displacing the 'native speaker': expertise, affiliation and inheritance.' *English Language Teaching Journal*, 44 (2), 97-101 (1990).

15 Young, D. 'English for what and for whom and when?' *Language Projects Review*, 3 (2), 8 (1988).

See also Young, D. 'The role and status of the first language in education in a multilingual society', in Heugh, K., Siegruhn, A., Pluddemann, S. (eds.), *Multilingual education in South Africa*, 1995.

16 HSRC, ibid. p.103.

explain the unpopularity of these labels among some South African researchers is that they regard "positivism" and apartheid as kissing cousins!

Positivism is not only linked to racism but to colonialism and sexism as well. Edidin maintains that this objectivist "one true story" approach is linked to analytical philosophy, to European masculine thought, or masculine thought in general, and to European colonial thought, e.g. Euro-American, Euro-Australian, Euro-African thought.[17] There is a danger, which Edidin is aware of, that "female" and "indigenous" thinking could be interpreted as lacking "analytical" clout. The "one true story" of logic and mathematics is not only accessible to male Europeans. This may explain, in part, the unpopularity among many researchers of statistical measurement in the human sciences. Of course, there are also many researchers who combine qualitative and quantitative methods.

How does Makoni's "boundaries" metaphor (mentioned two paragraphs earlier) of separating (Bantu) languages, which he is opposed to, fit in with his recommendation that the boundary between L1 and L2 should be maintained? (Makoni, personal communication) In the case of the separation of languages and dialects, this, to Makoni, it would seem, is an artificial separation without linguistic justification, whereas in the case of the L1/L2 distinction, this is a real distinction that has pedagogical justification.

Young maintains that the labels "L1 and "L2" segregate "English into separate learning 'boxes'" and asks: "While the ESL label is in keeping with international trends in most countries where English is taught and learned, is it not perhaps time that we, in South Africa, begin to consider how socially and politically divisive it is to continue using the ESL label?"[18] Yet, ESL is like any label that distinguishes between those who have more and those who have less. It segregates - nationally and internationally. If the rest of the world, in spite of the shortcomings of the ESL label, finds the label useful, couldn't we, a

---

17 Edidin, A. 'Eternal verities: timeless truth, ahistorical standards, and the one true story.' *American Philosophical quarterly*, 34 (2), 259-271 (1997), p.259.

18 Young, D. 'English for what and for whom and when?' *Language Projects Review*, 3 (2), 8 (1988).

decade after Young asked his question, become part of the international community, which has had, and continues to have, its fair share of racial-ethnic-linguistic discrimination? Perhaps the label "primary" is preferable to "L1" and so one could get rid of the notion of "first language". But if "L1" becomes a "primary" language, L2 could be seen as a "secondary" language, which, if not more discriminatory than "second" language, could belittle the value of the "second" language (L2). We could use "additional" language, which has a less discriminatory ring than "primary" language. Rampton[19] advocates a different set of terms. (Rampton is discussed later).

The sensitive issue in South Africa is that the labels ESL and L2 are indicative of a particular type of social life. The question is whether the policy of learning an African language first is not a tactic employed to delay the inevitable moment that one has to enter the English (or American) life-style (Makoni, personal communication).

The tendency in "National Qualifications Framework" thinking is that L2 levels should be based on the L1 paradigm and so one should calibrate levels of L1 or native proficiency to determine as precisely as possible what L2 learners are to aspire to.[20] This means that L1 and L2 learners could, indeed should, be given the same **proficiency** tests, as was done in this study. This does not mean that L1 and L2 teaching syllabuses should be the same. Indeed, there are very good reasons for keeping L1 and L2 syllabuses apart. I discuss this issue.

The term "apart" in the South African context is rank with negative connotations. If syllabuses must be kept apart, does this mean that L1 and L2 learners must be kept apart in the language classroom? From a practical point of view "separate syllabuses" means "separate classrooms". Barkhuizen goes even further by advocating not only separate L1

---

[19] Rampton, B.H. 'Displacing the 'native speaker': expertise, affiliation and inheritance.' *English Language Teaching Journal*, 44 (2), 97-101 (1990).

[20] Alexander, N. 1996. 'Drawing the issues together: In the context of language education policy, in HSRC.' *Language assessment and the National Qualifications Framework*, 1996, pp.105-106.

and L2 classrooms but separate L1 and L2 (ESL) Departments in secondary education.[21] Barkhuizen states: "The different aims of the second and first language syllabuses and the different approaches used to achieve these aims provide a further rationale for the establishment of an independent ESL Department."[22] Thus, Barkhuizen takes for granted that there should be separate L1 and L2 *classrooms*. No one would dispute that English teaching involves two different kinds of methodology: L1 and L2 methodology. But it is not necessary or pedagogically sound to (1) divide teachers into "L1" teachers and "L2" teachers and to go even further by (2) erecting departmental boundaries between them. One must not forget, however, that most English teachers in South Africa are not English-mother-tongue speakers, and would therefore have difficulty in teaching English as a First Language.

I look closer at the idea of "separate syllabuses/classrooms" for L1 and l2 learners. Barkhuizen examined L1 and L2 syllabuses in South African departments of Education and discovered that the weight assigned to the four languages "skills" is very different in the L1 and L2 situations.[23] Three examples:

(1) There is no listening skill section in the L1 syllabus.

(2) The oral skill is approached in different ways in the L1 and L2 syllabus. In the L2 syllabus the emphasis is on "articulating and pronouncing words in an acceptable manner" in "ways appropriate to circumstances and situation". In the L1 syllabus, however, the aim is "speak fluently, distinctly, with ease and enjoyment, and acquire poise and confidence in communicating."

(3) The reading of literature receives different emphases in the two syllabuses. Literature in the junior school counts up to 50% in the L1 syllabus, while in the L2 syllabus it only counts 15%. At the matriculation level there is more than 15% literature in the L2 syllabus but still far less than in the L1 syllabus. Furthermore, it's not just a question of the amount but the approach that is different in the two syllabuses.

---

21 Barkhuizen, G. 'Proposal for an independent English Second Language Department at Mmabatho High School.' *English Language Teaching Centre (ELTIC) Reporter*, 16 (1), 25-32. Johannesburg, 1991).

22 Ibid., p.30.

23 Barkhuizen, G., ibid.

In all three of the differences between L1 and L2 syllabuses, what stands out is that L1 learners are able to focus much more on the message, while L2 learners spend much of their time focusing on the medium. A caution: It may seem that L2 learners are more aware of structure than L1 users. But I think that the latter is the case only with BICS, not with CALP, i.e. not with formal reading, formal writing and formal listening and formal speaking. But let us limit ourselves to the modes (or "skills") of reading and writing, which carry more promotional weight in educational institutions. In academic reading and writing L1 users are also L1 learners, and the only way to write well is to be conscious of structure, i.e. of metalanguage. Jeffery suggests that we should open the case for grammar wider:

> *[T]he most elementary reading and writing presuppose word and sentence at least; and progress is awkward without noun, verb, number, tense, phrase, clause, and so on. These categories come naturally to nobody, and with difficulty to some, for they are all artificial abstractions from the flow of speech. Nobody speaks in words and sentences; therefore everybody needs to be taught about them...Inchoate thought has to be organised in order to make your meaning clear to your readers (and yourself), and that takes skill in arranging the units of WL [written language] into structures.*[24]

The metalingual function (i.e. the focus on the code or grammar) presides over the whole process of language, from idea to phoneme to message. Metalingual ability is more than the ability to manipulate a stock of static structures; it is an ability to manipulate language/ideas in dynamic ways. This ability is indispensable in the development of academic language proficiency. And crucially, the "higher" forms of monitoring (Krashen's "learning"/"Monitoring") is "test language".[25]

Thus, in the understanding and production of academic discourse, the focus often falls on grammar (the linguistic elements), which is not only concerned with keeping the

---

[24] Jeffery, C.D. 'The case for grammar: Opening it wider.' *South African Journal of Higher Education (SAJHE)*, Special edition (1990), p.120.

[25] Wald, B. 'A sociolinguistic perspective on Cummins' current framework for relating language proficiency to academic achievement', in Rivera, C. *Language proficiency and academic achievement*, 1984, p.50.

clothes of discourse clean. Du Toit and Orr describe a piece of writing containing grammatical errors as "clothes covered with food stains and dirt, with perhaps a few missing buttons".[26] Grammar is wider than this in that it is necessary for the effective communication of "potential meaning".[27] The metalingual function, therefore, especially in writing, plays a crucial role in academic discourse for both L1 and L2 learners.

Having said that both L1 and L2 learners require metalingual know-how to function effectively in academic situations does not mean that L1 and L2 learners should necessarily be taught in the same classroom. I would still go along with the idea that learning materials should be specifically designed to cater for second-language situations[28], and the only feasible way to teach such a syllabus would be in a separate space such as a classroom.[29]

The investigation in this study showed a radical difference between learners who take the subject English as a First Language and as a Second Language. How does this impact on the L1/L2 distinction? Is the distinction real or imagined? In a recent project conducted by Makoni (personal communication) - in collaboration with the National Language Project - Grade 7 second language speakers scored 21% in an English proficiency test administered to Grade 3 first language speakers who scored 75%. The tests used were traditional tests such as cloze, dictation and grammar tests, as was used in this study. In the light of these results, Makoni asks whether the integration of ESL learners with first language speakers (in the same classroom) is covert integration, which could turn out to be more lethal than the overt segregation of the apartheid era.

Barkhuizen changed his view by jettisoning not only the idea of separate English Departments, which was the main thrust of his 1991 article but also the idea of separate

---

26 Du Toit, A. and Orr, M. *Achiever's Handbook*, 1989, p.199.

27 Halliday, M.A.K. *Learning how to mean*, 1975.

28 Musker, P. and Nomvete, S. 'Standards and levels in language assessment', in HSRC. *Language assessment and the National Qualifications Framework*, 1996, p.67.

29 Peirce, B. 'On language difference and democracy.' *Language Projects Review*, 6, 21-24 (1991).

L1 and L2 syllabuses in favour of "multicultural settings".[30] Owing to the political changes in South Africa, what was acceptable in 1991 became unacceptable a year later, and so Barkhuizen met the new challenge of "what needs to be done"[31] under a different political and educational dispensation.

The idea of separate L1 and L2 syllabuses remains a good one, and so does the idea of multicultural settings, which takes into account the extremely important issue that language cannot be separated from culture.[32] It seems that it is difficult to implement both ideas simultaneously because it would mean that the idea of multicultural settings would be relegated to "social" language while the idea of separate L1 and L2 classrooms would hog "academic" language, which would exacerbate the racial-ethnic divide.

There have been some efforts to counteract this tendency of relegating "minorities" to second-class citizens, e.g. (1) the mainstreaming of ESL learners[33] as in the "Whole-language approach"[34] and the "Whole-school Approach"[35] and attention to different cognitive styles.[36]

---

30 (1) Barkhuizen, G. 'Teaching English in multilingual settings (TEMLS): What needs to be done.' *Journal for Language Teaching*, 26 (4), 53-68 (1992).
(2) Barkhuizen, G. 'Preparing teachers to teach in multilingual settings. Current approaches to the teaching of English for academic purposes: A critical appraisal.' *Proceedings (Part 1) of the South African Applied Linguistics Association conference 'Our multilingual society: Supporting the reality.* (University of Port Elizabeth, 1993)
(3) Barkhuizen, G. 'Using English in the South African classroom.' *Per Linguam*, 12 (1), 34-47 (1996).

31 Barkhuizen, G. 'Teaching English in multilingual settings (TEMLS): What needs to be done.' *Journal for Language Teaching*, 26 (4), 53-68 (1992).

32 The connection between language and culture is obviously very important. Unfortunately, I cannot discuss this issue in this study.

33 Clegg, J. (ed.). *Mainstreaming ESL: Case studies in integrating students into the mainstream curriculum*, 1996.

34 Westbrook, L. and Bergquist-Moody, S. 'A Whole-language approach to mainstreaming', in Clegg, J. (ed.). *Mainstreaming ESL: Case studies in integrating students into the mainstream curriculum*, 1996.

35 Reid, J. and Kitegawa, N. 'A Whole-school Approach to mainstreaming: The Rose Avenue ESL/D project', in Clegg, J. (ed.) *Mainstreaming ESL: Case studies in integrating students into the mainstream curriculum*, 1996.

36 Skehan, P. *A cognitive approach to language learning*, 1998.

The main thrust of mainstreaming is that the needs of ESL learners are not merely linguistic.[37] Communicative competence, therefore, should be regarded as only one aspect of the wide objective of academic achievement. What is required is a curriculum transformation that should start with a revamping of psycholinguistic courses for language teachers, which should include a heavy component on multicultural problems in the ESL classroom[38], but which should at the same time balance such an approach with an emphasis on those aspects - of which there are many - that are common to all cultures seeking entry and success in the international marketplace.

It is not only terms such as "first" and "second" language that are becoming unpopular in South Africa. Terms such as "native" language (which in South Africa meant "black") and "mother" tongue, which feminist movements find sexually discriminatory, are also problematic. James[39] lumps together "native speakerism", sexism and racism. In the last few paragraphs that remain of this section, I describe a few non-South African views on terms used to describe speakers of a language.

The "whole mystique of a native speaker"[40] who uses his/her mother tongue implies five things, which have been hotly contested[41]:

1. A particular language is inherited through birth into a particular social group.
2. If you inherit a language, you can speak it well.
3. One is or isn't a native/mother-tongue speaker.
4. A native speaker has a comprehensive grasp of the inherited language.

---

37 Clegg, J. (ed.). *Mainstreaming ESL: Case studies in integrating students into the mainstream curriculum*, 1996, p.3.

38 Wong, S. 'Curriculum transformation: A psycholinguistics course for prospective teachers of ESOL K-12', in Alatis, E. Straehle, C.A., Gellenberger, B. and Ronkin, M. (eds.). *Georgetown University Round Table on Languages and Linguistics*, 1995.

39 James, C. 'Don't shoot my dodo: on the resilience of contrastive and error analysis.' *International Review of Applied Linguistics*, 32 (3), 179-200 (1994), p.192.

40 Kachru, B.B. (ed.). *The Other Tongue: English across cultures*, 1982, p.7.

41 It was Christopherson (1973) who was among the first researchers to question these notions mentioned by Rampton. [Christopherson, P. *Second-language learning*, 1973].

5. Being a citizen of a country is analogous to being a native speaker of one mother tongue.

Rampton[42] proposes the following terms to replace terms such as "native", "mother tongue", "first language" and "second language"[43]:

1. Language expertise, i.e. the level of proficiency. An important issue in assessing expertise would be the models of language ability that one would use to decide on an acceptable or minimum level of expertise. But this is not a new issue, even if the terminology is new.

2. Language affiliation, which is concerned with the affective relationship of a learner towards a language.

3. Language inheritance. Membership of an ethnic group does not automatically mean that the language of the ethnic group has been automatically inherited. For example, it is not rare in South Africa that preschool and primary school learners change there "mother tongue" by entering another ethnic group.

Davies maintains: "In terms of ultimate attainment the post-pubertal second language learner may, exceptionally, attain native speaker levels of proficiency and therefore be indistinguishable from the native speaker."[44] Paikeday[45] maintains that only one exception proves that there is *no* rule, and, therefore, there is no such person as a native speaker - "the native speaker is dead![46]: accordingly, one can only legitimately speak of degrees of competence of language use, as one would about any other skill, e.g. rowing boats or mowing lawns.[47]

---

42 Rampton, B.H. 'Displacing the 'native speaker': expertise, affiliation and inheritance.' *English Language Teaching Journal*, 44 (2), 97-101 (1990).

43 See also Leung, C. Harris, R. and Rampton, B. *The idealised native-speaker, reified ethnicities and classroom realities: Contemporary issues in TESOL*, 1997.

44 Davies, A. 'Proficiency or the native speaker: What are we trying to achieve in ELT?', in Cook, G. and Seidlhofer, B. *Principle and practice in applied linguistics: Studies in honour of H.G. Widdowson*, 1995, p.145.

45 Paikeday, T.M. *The native speaker is dead!*, 1985.

46 Ibid.

47 Ibid., p.11.

In contrast to Paikeday, Medgyes maintains that the exceptions do not disprove the rule but prove it, and that, therefore, native speakers of a language are recognised as such,[48] even after taking into account some exceptions, or, as Quirk in Paikeday[49] put it, the "fuzzy edges".

I agree with Medgyes that "mother-tongue" speakers (the language one uses in one's early childhood) and "first language" speakers (the language one knows best) are usually identifiable.[50] "In [Medgyes] experience, liberal-minded researchers often shut their eyes to the glaring differences between natives and non-natives."[51] As far as the worth of a native speaker over a non-native speaker is concerned, Medgyes, who is specifically focusing on language teachers, points out that non-native speakers and native speakers each have their respective contributions to make to teaching.[52] For example, a proficient non-native speaker can be a better learner model than a native speaker, owing to the fact that the non-native speaker has personal experience of learning the language in question. Accordingly, many L2 learners have a higher degree of awareness about the target language than L1 learners have about their own. The native speaker, on the other hand, can be a better language model, i.e. as far as such things as prosody is concerned. Thus, it is false to assume that the more proficiency one has, the more effective one will be in the classroom. There are also other factors in teaching (and learning) that have nothing to do with the native/nonnative issue, namely, academic ability, which is closely related to Cognitive and Academic Language Proficiency (CALP; see section 2.6). So, it's nonsensical to ask "Who's worth more: a native or a non-native?"[53]

---

48 Medgyes, P. 'Native or non-native: who's worth more?' *ELT Journal*, 46 (4), 340-348 (1992).

49 Ibid., p.7.

50 There are exceptions where a person can have (1) more than one first language (2) low competence in a mother tongue, or (3) no first language, i.e. no language that one knows well, e.g. a "replacement" language, which was described in earlier chapters.

51 Medgyes, P. ibid., p.343.

52 Medgyes, P., ibid., pp.346-347.

53 Megyes, P., ibid., p.347.

The "elitist" may argue that "not all members of a linguistic community are equal in linguistic knowledge"[54] about their common native language, and so it is possible to have some native or L1 speakers who know more than others about the language. Harris argues against this "elitist" view. For Harris, language is indeterminate because knowledge as such is indeterminate. Accordingly, for Harris, there are no "job-secure" words.[55] But, if language (and ergo knowledge) is so insecure that one cannot define anything - approximately, to be sure; that is, if there is no native or L1 or L2 "bulls-eye"[56] in language, there can be no truth at all - approximate to be sure. For McArthur, Standard English is a "fuzzy subset of a very large set of Englishes"[57] and for Lass[58] English changes as we speak it. Standard English, in Lass's scenario - and possibly native language, L1 and L2 as well - becomes, in fuzzy-set theory, a subset of the fuzz that was.

6.3 Negotiating the task-demands and the "Threshold Project"

During this decade an increasing number of educationists and psychologists are arguing that the problem of education can be met by investigating motives, goals and conditions of learning/teaching in terms of Soviet activity theory.[59] What is commonly called activity

---

54 Harris, R. *The language myth*, 1981, p.171.

55 Ibid., p.175.

56 Russell, B. *The analysis of mind*, 1921, pp.197-198 in Harris, 1985, p.170.

57 McArthur, T. *English as a world language, as an African language, and as a South African language*. Paper presented at the English Academy of Southern Africa conference *"English at the turn of the Millennium"*, Johannesburg College of Education, 14-16 September, 1988. See also McArthur, T. The English language or the English languages? in Bolton, W.F. and Crystal, D. *The English language*, 1987.

58 Lass, R. *"English" - Talk at Will radio programme on SAfm*, 26 February, 1998.

59 (1) Campbell, C.M. *Learning and development: an investigation of neo-Piagetian theory of cognitive growth*. Master of Arts thesis, University of Natal, 1985.

(2) Clayton, E. 'Scaffold: Graded support or gibbet? The acquisition of terminology-concepts in a scientific discipline.' *Proceedings of the South African Association for Academic Development (SAAAD) Conference*, Technikon Free State, Bloemfontein, 29 November - 1 December, 1995.

(3) Donato, R. and McCormick, D. 'A sociocultural perspective on language learning strategies: The role of mediation.' *The Modern Language Journal*, 78 (4), 453-464 (1994).

(4) Macdonald, C.A. *Crossing the threshold into standard three in black education: The consolidated main report of the Threshold Project*, 1990b.

(5) Macdonald, C.A. *Reasoning skills and the curriculum*. Report Soling, 18, 1990c).

theory is the unique and self-consciously independent nature of the Soviet cultural-historical research tradition which is referred to simply as "activity theory". The latter involves an approach of mediation through negotiating task demands. In such a scenario, learners and teachers negotiate how, and sometimes what, should be taught.

In Soviet activity theory, one thinks, learns, creates through the historical process of sociocultural interaction[60] where the mediator plays the primordial role in cognitive and educational development.

Activity theory in Soviet psychology has become, "a dominant intellectual force for Western Researchers"[61] because of its belief that the intermental (social mind) is what gives substance to the intramental (individual mind). It is the intramental relationship between a mediator (e.g. a teacher) and a learner that is the grist of this activity. Outside of activity (or action), the psychological individual is reduced to a physiological and morphological husk. Outside of society, knowledge construction is not possible. On these presuppositions of Soviet activity theory are based "progressive" theories of education such as outcomes-based education.[62]

Soviet activity theory emphasises the activity of the individual in the world.[63] This activity consists of momentary conflicts which have to be surmounted in order to attain a higher

---

(6) Moore, R., Paxton, M., Scott, I. and Thesen, L. 'Language development initiatives and their policy contexts', in Angélil-Carter, S. (ed.). *Access to success: Literacy in academic contexts*, 1998.

(7) Wallace, B. and Adams, H. 'A framework for language.' *Bua!*, 10 (1), 16-17 (1995).

60 (1) Vygotsky, L.S. *Mind in society: The development of higher psychological processes.* (Edited by Michael Cole, Vera John-Steiner, Sylvia Scribner and Ellen Souberman, 1978.

(2) Vygotsky, L. and Luria, A. Tool and symbol in child development, in Van der Veer, R. and Valsiner, J. *The Vygotsky reader*, 1994.

61 Minick, N.J. L.S. *Vygotsky and Soviet activity theory: New perspectives on the relationship between mind and society.* PhD thesis, NorthWestern University, 1985, p.iii.

62 Murray, S. 'Exploring the possibilities of using an outcomes-based approach in English teacher education.' *Southern African Journal of Applied Language Studies*, 5 (2), 21-37 (1997), p.28.

63 Minick, N.J. L.S. *Vygotsky and Soviet activity theory: New perspectives on the relationship between mind and society.* PhD thesis, NorthWestern University, 1985). p.24.

level of equilibrium. It is the mediator's role to help the learner surmount these conflicts.[64] The distance (i.e. the difference) between the resultant development and the potential development is referred to as the "zone of proximal development" (ZPD)[65]. ZPD theory is based on the Piagetian notion that "[a]ll development is composed of momentary conflicts and incompatibilities which must be overcome to reach a higher level of equilibrium".[66] If the teacher presents conflict, and is also mindful to provide the resources for the child to surmount it, then development will occur. The answer to many learning problems, for activity theorists, lies in the mindful interventions generated by the teacher. This evokes issues such as learning styles, learner-centred and teacher-centred learning, task-based learning and critical language awareness.

In contrast to the strong social emphasis of language tasks in authors such as Gee[67], Hymes[68], Lee[69] and Van Lier[70], attention has been given by other authors, e.g. Cummins[71]; O'Malley[72], Oller and Perkins[73] and Saville-Troike[74] to the relation between language

---

64 The emphasis on the individual seems misplaced in Soviet psychology that was dominated by the *one true science* of the USSR. It is and isn't. In the Soviet state, the individual did have a place, which Vygotsky exploited to the full, but at the same time the individual had to *know his/her place*, i.e. the state was both the primordial cause and the ultimate goal of individual consciousness and conscience. For this reason, Soviet science lacked an authentic philosophical and anthropological base, namely the capacity to give free reign to the scientific imagination.

65 Vygotsky, L.S. *Mind in society: The development of higher psychological processes*, 1978, p.86.

66 Piaget, J. and Inhelder, B. *The psychology of the child*, 1969, p.78.

67 Gee. J. *Social linguistics and literacies: Ideology in discourses*. (London, The Falmer Press, 1991).

68 Hymes, D. 'On communicative competence', in Pride, J.B. and Holmes, J. (eds.). *Sociolinguistics*. (Harmondsworth, Penguin, 1972).

69 Lee, D. *Competing discourses: Perspective and ideology in language*, 1992.

70 Van Lier, L. *The classroom and the language learner*, 1988.

71 Cummins, J. 'Wanted: A theoretical framework for relating language proficiency to academic achievement among bilingual students', in Rivera, C. (ed.). *Language proficiency and academic achievement*, 1984.

72 O'Malley, J.M. 'The cognitive academic language learning approach (CALLA).' *Journal of Multilingual and multicultural development*, 9 (1 and 2), 43-60 (1988).

73 Oller, J.W., Jr. and Perkins, K. *Language in education: Testing the tests*, 1978.

74 Saville-Troike, M. 'What really matters in second language learning for academic achievement.' *TESOL Quarterly*, 18 (2), 199-219 (1984).

proficiency and the problem-solving abilities involved in academic performance. Prabhu's[75] "task-based" teaching is concerned with language in a problem-solving context. "'Communicative' competence, in the sense of an ability to achieve social situational appropriacy, is not seen [by Prabhu] as a relevant objective."[76]

The main philosophical influence in the "Threshold Project"[77] has been Soviet activity theory. The "Threshold Project", which has as its educational cohort primary schools in the North West Province has had a strong influence on recent attitudes towards testing in South Africa. The "Threshold Project" has also had a strong influence on secondary and tertiary education.[78] The philosophical underpinnings of the "Threshold Project" is heavily Vygotskian. Vygotsky has had an important impact on the thinking behind recent developments in language and education initiatives and policy .[79]

Macdonald[80] mentions the HSRC's norm-referenced test[81] which serves as a diagnostic tool for learners entering Grades 4, 5 and 6. She cites from the HSRC data that the average score of learners from these three standards were 22%, 44% and 66%, respectively, where the HSRC recommends that the test be converted into a criterion-

---

75 Prabhu, N.S. Task-based language teaching and its implications for testing, in Yeld, N. *Communicative language testing.* Report on British Council Course 559 offered at Lancaster University from 8 September to 20 September 1985, (published in) 1986, p.17.

76 Yeld, ibid., p.17.

77 Macdonald, C.A. *English language skills evaluation (A final report of the Threshold Project),* 1990a.

Macdonald, C.A. *Crossing the threshold into standard three in black education: The consolidated main report of the Threshold Project,* 1990b.

Macdonald, C.A. *Reasoning skills and the curriculum,* 1990c.

78 Gamaroff, R. *Activity theory, mediation and intelligence in learning.* Tenth World Congress of the Comparative Education Society (WCCES), Cape Town, July 12-17, 1998d.

79 Moore, R., Paxton, M., Scott, I. and Thesen, L. 'Language development initiatives and their policy contexts', in Angélil-Carter, S. (ed.). *Access to success: Literacy in academic contexts,* 1998, p.12.

80 Macdonald, C.A. *English language skills evaluation (A final report of the Threshold Project), Report Soling-17,* 1990a, p.46.

81 As I mentioned in section 1.2, a "norm-referenced" and "criterion-referenced", test or any "referenced" test is not an intrinsic characteristic of the test but the use to which the test is put that, or the use to which the test "refers". The same goes for validity. Tests aren't, strictly speaking, valid: the purpose of the test is what is valid.

referenced test, where Grade 4 learners should score a minimum of 80% to gain admission to Grade 5. Macdonald objects to the HSRC's "old" paradigm of psychometric testing because the majority of prospective Grade 5 learners would probably get far less than 80% on such a test, which would mean rejection for admission to Grade 5. The predicament, maintains Macdonald, would then be what to do with these unsuccessful learners.[82] This is possibly the reason why schools inflate marks: to increase the pass rate. Macdonald has highlighted a major problem: if one does not allow poor performers to sit exams and by so doing hope to discourage indiscriminate advancement through the grades[83], the problem is what to do with those who are not permitted to sit exams - and those who fail when permitted. (In 1997, 200, 000 South Africans failed Grade 12: a 47% pass rate).

Macdonald's second objection is that the causal or correlational link between language proficiency and academic achievement is not clear.

> *[T]he most difficult connection to make is that between different aspects of English communicative competence and their relation - causal or correlational - to formal school learning through EMI [English as a medium of instruction]. If one is able to set up these relationships in a reasoned way - and nobody to our knowledge has gone very far in this task - then the significance of the current test scores would be absolutely transparent. There is a way through this conundrum, and that is to change the nature of the question.*[84]

The implication is that some aspects of language proficiency, especially the highly "pragmatic" aspects that come close to mirroring real life, are, as discussed earlier, difficult to measure accurately. This is because the more subjective the test, the more unreliably it is liable to be measured. When academic achievement is brought into the

---

[82] Recall the results of Makoni's recent project (mentioned above) where Grade 7 second language speakers scored 21% in an English proficiency test administered to Grade 3 first language speakers who scored 75%. The problem here as well would be what to do with these low proficiency learners. Put them in the same classroom with first first-language speakers? This would not be a good educational idea.

[83] Bundy, C. *'Talk at Will' radio programme on SAFM* (6 January, 1998).

[84] Macdonald, C.A. *English language skills evaluation (A final report of the Threshold Project)*, 1990a, p.42.

picture, the *content* "housed" by the *language* becomes specialised, and consequently, it becomes even more difficult to distinguish between content and language. Unfortunately, changing the nature of the question, as Macdonald suggests (above), cannot negotiate "a way through the conundrum", but merely obfuscates it.

Also, consider Macdonald's remark (from the same quotation) that "the most difficult connection to make is that between different aspects of English communicative competence and their relation - *causal or correlational* - to formal school learning through EMI". Compare Macdonald's appraisal above with Ochner's:

> *[L]ess debatable, is the obvious fact that research design and the 'real' world only sometimes covary. We trade off internal for external validity, or vice versa; either way, we obtain in our experiments important results only from those small, and trivial, bits of human reality that allow a reductive analysis.*[85]

Owing to the complexities involved, both Macdonald and Ochsner are correct in so far as it is difficult to clarify the *causal* relation between language proficiency and school learning (Macdonald) or between experimental results and the "real" world (Ochsner). The *correlational* relation, however, is not difficult to clarify, because correlation has nothing to do with ontological, epistemological or causal relations. Correlation merely tells us how much two variables *covary*. Ochsner's connotation of "covary" (in his quotation above) is causal. But in statistics, "covary" says nothing about causes, it is merely another term for "correlate". Macdonald and Ochsner seem to be using statistical terms, e.g. *correlation* (Macdonald) and *covary* (Ochsner), in a non-statistical way, which confuses matters.

A third problem for Macdonald is that

> *doing things in such a post hoc way [i.e. the HSRC's psychometric tests] would fail to force us into analyzing the nature of the learning that the*

---

[85] Ochsner, R. 'A poetics of second-language learning.' *Language Learning*, 29 (1), 53-80 (1979), p.58.

> *child has to be able to meaningfully participate in...we would have described a test and some external criteria and identified children through the use of these - but we would have failed to explain what it is the children have to be able to do.*[86]

Macdonald (above) is contrasting the "post hoc" psychometric paradigm of the HSRC which "fail[s] to explain what it is the children have to be able to do" with her "negotiating the task-demands", which she claims *does* explain what children have to be able to do. Which raises *the* question: What is a real, authentic, natural task? For Macdonald the answer to this question lies in "negotiating the task-demands". In terms of this sociocultural perspective, knowledge is not meant to be a transmission of knowledge from teacher to learner but a co-construction of knowledge brought about through a mutual negotiation between teacher and learner of the task demands.[87]

(It has become inane to accuse any educational theory of not being interactionist (enough) between teacher and learner. That is not a debate anymore. What is a debate - *the* debate; no matter how "old" it is - is the contribution of innate, fixed, individual attributes versus the contribution of sociocultural forces to learning[88]).

Macdonald's solution to her three problems mentioned above is to replace the "outdated and rigid modes of curriculum development in South Africa"[89] such as psychometric measurement (norm-referenced and criterion-referenced tests) and the general ability of communicative proficiency with "negotiating the task-demands"[90], which involves "going

---

86 Macdonald, C.A. *English language skills evaluation (A final report of the Threshold Project)*, 1990a. p.46.

87 Chang-Wells, G.L.M. and Wells, G. 'Dynamics of discourse: Literacy and the construction of knowledge', in Forman, E.A., Minick, N. and Stone, C.A. (eds.). *Contexts for learning: Sociocultural dynamics in children's development*, 1993, p.59.

88 (1) Tooby, J. and Cosmides, L. 'Psychological foundations of culture', in Barkow, J.H., Cosmides, L., and Tooby, J. (eds.). *The adapted mind: Evolutionary psychology and the generation of culture*, 1992.

(2) Pinker, S. *The language instinct*, 1995.

89 Macdonald, C.A. *English language skills evaluation (A final report of the Threshold Project*, 1990a, p.46.

90 Ibid., p.46.

from one situation (and knowledge domain) to another to see how the curriculum in its broadest sense has been constituted, and which aspects are negotiable."[91] Examples of such tasks-demands are[92]:

- Following a simple set of instructions for carrying out a task.
- Showing command of a range of vocabulary in semantic clusters from across the curriculum.
- Solving problems involving logical connectives.
- Being able to show comprehension of simple stories and information books.

In the "negotiating the task-demands" approach, authenticity would not only be text or content authenticity, but also learner authenticity. where the content should be relevant to the learner's life-experiences.[93] In such activities (1) meaning would be primary, (2) there is a communication problem to solve, (3) the completion of the task has some priority, and (3) the task is assessed in terms of outcome.[94]

Macdonald's tasks are similar to Van der Walt's "communicative" tasks, which are often based on a central theme, where all questions are related to this theme: test questions such as (1) information gap, (2) task-dependancy, where information generated on one task is used to complete another task, and (3) testing cohesion (using linking words) and coherence (rewriting dialogues in the right order). However, I don't see why these "communicative" tasks cannot also be done using discrete-point and/or integrative *formats*, and using norm-referenced, statistically assessed tests. "After all, if evaluation ["assessment" in this study] implies a comparison with some kind of ideal performance [mother tongue?], is it not essentially normative?" (Alderson in Yeld[95]; my brackets).

---

91 Macdonald, C.A. *English language skills evaluation (A final report of the Threshold Project*, 1990a, p.46.

92 Ibid., p.47.

93 (1) Breen, M.P. 'Authenticity in the language classroom.' *Applied Linguistics*, 6 (1), 60-70 (1985).
(2) Lee, W.Y. 'Authenticity revisited.' *English Language Teaching Journal*, 49 (4), 323-328 (1995).

94 Skehan, P. *A cognitive approach to language learning*, 1998, p.95.

95 Yeld, N. *Communicative language testing.* Report on British Council Course 559 offered at Lancaster University from 8 September to 20 September 1985, (published in) 1986, p.39.

At the heart of this discussion is the problem of authenticity, of "real-life". The problem is, of course, knowing what a real-life task, or test, *is*; not merely looks like. Canale and Swain[96] make a distinction between "competence-orientated tests" and "actual performance". They call their performance tests "tasks", which they distinguish from competence "tests": "tasks correspond more directly to language use where an integration of these skills is required with little time to reflect on and monitor language input and output".[97] Yet, they still believe that tests, specifically discrete-point tests, are useful in measuring communicative competence.[98] They are probably referring to discrete-point *formats*. I suggest that many discrete-point formats would fit the bill of Canale and Swain's "tasks", if the latter are equivalent to Macdonald's list of "task-demands" mentioned above.

If by some good fortune we discovered what an "authentic" task was (and, accordingly, wasn't), it still doesn't follow that it is necessary to do "authentic" tasks in order to prove that we are proficient to do them, because communicative tasks, e.g. Macdonald's "task-demands" above, can be tested successfully through discrete- point tests.[99] Alderson, who is more cautious, maintains that we do not yet know what communicative tests are.[100] It doesn't seem wise, therefore, to try and separate - as Macdonald suggests - (general) communicative proficiency from a task-demand such as "showing command of a range of vocabulary (in semantic clusters) from across the curriculum."[101] After all, the most demanding part of "negotiating the task-demands" is often the (general) communicative proficiency part, especially for limited English proficiency learners. Low language proficiency students often have more problems with general background

96 Canale, M. and Swain, M. 'Theoretical bases of communicative approaches to second language teaching and testing.' *Applied Linguistics*, 1 (1), 1-47 (1980), p.34.

97 Ibid.

98 Ibid.

99 Politzer, R.L. and McGroarty, M. 'A discrete-point test of communicative competence.' *International Review of Applied Linguistics*, 21 (3), 179-191 (1983).

100 Alderson, J.C. 'Who needs jam?', in Hughes, A. and Porter, D. *Current developments in language testing*, 1983, p.90.

101 Macdonald, C.A. *English language skills evaluation (A final report of the Threshold Project*, 1990a:47.

knowledge than with new knowledge. For this reason a radical separation should not be made between a Language for Specific Purposes task and a general proficiency task, because the harder part is often the *general language proficiency* part, especially for low English proficiency learners.

Theory, so far, has not been of much help in saving our knowledge and beliefs from the abyss of ignorance. We can't get more explicit than a test. (A test is "the most explicit form of description, on the basis of which the tester comes clean about his/her ideas."[102] [See the introduction to the study]). There are limitations in the degree of explicitness that one can reach. This does not mean that we should stop measuring until we've decided what we are measuring, and let language theory take the lead; rather, we test the best we can with the knowledge that we already have , and take care not to yoke up with fortuitous and often circuitous bandwagons. Thus, it is difficult to prove that the HSRC's psychometric paradigm is worse than Macdonald's "negotiating the task-demands". Macdonald's argument, as mentioned earlier, is that the HSRC tests do not tap what learners "have to be able to do", which is also the motto of outcomes-based education[103] that is to be discussed in the next section. The problem is that the connection between the *activity* of doing "old" paradigm tests, such as those used by the HSRC, and the "new" paradigm *activity* of "negotiating task-demands" is far from clear. The reason why the connection is not clear is because there never was a disconnection in the first place. What we regard as an unclear connection is in fact an artificial disconnection. The kind of planning involved in "new' paradigm activities, i.e. "outcomes" seems to be not so different from what some theorists believe is the kind of planning also involved in quantitative research, specifically, ***know what you are going to do and stick to it***[104]. I elaborate on this issue in the next section.

---

102 Davies, A. *Principles of language testing*, 1990, p.2.

103 Spady, W. 'Outcomes-based education: An international perspective', in Gultig, J., Lubisi, C., Parker, B. and Wedekind, V. *Understanding outcomes- based education: Teaching and assessment in South Africa*, 1998, p.24.

104 Magnan, S.S. *Review of Creswell, J.W. 1994. Research design: qualitative and quantitative approaches*, 1997, p.256.

6.4 Outcomes-based Education (OBE) and Competence-based Education Training (CBET)

The "negotiating the task-demands" approach is closely related to the OBE of the South African "National Qualifications Framework"[105] (NQF), which originated out of the Independent Examinations Board's contribution to the development of outcomes-based assessment.[106] CBET is similar to OBE and so these will be treated synonymously in this discussion.

The NQF is considered by the HSRC as the most important educational project to have been formulated in South Africa.[107] The NQF, which is concerned with formal qualifications and levels of learning, claims to have overhauled concepts such as competence, performance, ability and assessment. Assessment in the NQF wants to move away from summative evaluation towards formative evaluation, where the emphasis is on "outcomes", i.e. the continuous assessment of task criteria.[108] Based on the NQF[109] is another important document, the "Language assessment and National Qualifications Framework"[110], which was the product of the first conference in South Africa to begin to examine standards and qualifications in depth and was concerned with bringing language policy and assessment policy together.[111] I first discuss the notions of competencies and abilities in CBET and then move on to CBET's notions of language assessment.

---

105 HSRC. *Ways of seeing the National Qualifications Framework*, 1995.

106 French, E. and Rensburg, I. 'Introductory comments: Language assessment and the NQF', in HSRC. *Language assessment and the National Qualifications Framework*, 1996, p.5.

107 (1) Engelbrecht, S. and Schuring, G. 'The NQF: Challenges in the language field, in HSRC.' *Language assessment and the National Qualifications Framework*, 1996.

(2) Mclean, D. 'Language education and the national qualifications framework: An introduction to competency-based education and training', in HSRC. *Language assessment and the National Qualifications Framework*, 1996.

108 Cress, K. Reassessing assessment. *The Teacher*, 1 (7), 9-10, 1996, p.9.

109 HSRC. *Ways of seeing the National Qualifications Framework*, 1995.

110 HSRC. *Language assessment and the National Qualifications Framework*, 1996.

111 Musker, P. and Nomvete, S. 'Standards and levels in language assessment', in HSRC. *Language assessment and the National Qualifications Framework*, 1996.

In CBET terminology the central concept of "competence" subsumes "competency", "competence", "outcome", "ability" and "capability".[112] As Mclean points out, the jargon keeps changing its meaning.[113] Consider the following definitions of the NQF[114] definitions of the above terms. I also give the NQF's definition of assessment:

(1) Ability: is a generic term for the mental and physical processes that people use, such as communication, decision-making, problem-solving and using tools. Abilities are developed through engaging with *knowledge* (declarative and procedural) and *activities* in a context. Abilities cannot be directly assessed: rather, *assessment* is carried out indirectly via the performance of tasks which rely on *abilities* for their completion.

(2) Capability: the expression of generic abilities as they relate to specific content areas, context and value frameworks.

(3) Competence: the capacity for continuous *performance* within specified ranges and contexts resulting from integration of a number of capabilities.

(4) Assessment: the process of determining *capability* which is carried out by observing and evaluating *performances*. There are different ways in which assessment can be carried out.

The NQF has dropped the "controversial" and "ambiguous" terms of "competencies" and "skills" and retained "competence" to describe "overall proficiency", while retaining "capability" to describe the "learning outcome"[115].

In the "narrow interpretation of competence", any "performance" should be either directly observable or measurable.[116] The NQF theorists maintain that mental performance falls outside the ambit of the "narrow" view of performance. Accordingly, thinking or mental

---

112 Mclean, D. 'Language education and the national qualifications framework: An introduction to competency-based education and training', in HSRC. *Language assessment and the National Qualifications Framework*, 1996.

113 Ibid.

114 (1) HSRC. *Ways of seeing the National Qualifications Framework*, 1995, pp.1-2.
(2) Mclean, D., ibid., p.31.

115 HSRC, ibid., p.39.

116 Ibid.

performance must belong to the NQF's wider view.[117] But, consider the traditional "narrow" definition of a test, or some other behaviour, as the operationalisation of the invisible construct. Surely some form of behaviour is required to perform one's competence, where competence is what one thinks in one's head. Undoubtedly one does or should do a lot of thinking, i.e. planning performance before production, before one lets it all hang out in performance, and, thus, in a sense, one is "performing" in one's head. This, however, is not performance proper.

One could argue that the narrow view of performance only considers the product, not the process. But, if it is true that products are performances external to the mind, so are processes. Both "products" and "processes" are *products.* If they weren't, we wouldn't be able to study or talk about them. The difference between "products" and "processes" is this: "products" are *end,* or final, products, while "processes" are intermediate products. The "process" shows the *development* towards the end product of the performance; the "product" shows the end product itself of the performance, which is the ultimate goal of the "process". Terms such as transitional competence or interlanguage should not fool us into believing that only fixed competence is a product. The term process (or processing) implies movement, and, of course, when the mind processes it also progresses. The mind has to also stop, stand still and "stare" (at its intermediate products) otherwise it will lose control of its own introspections. The product/process paradox is at least as old as ancient Greek philosophy - Parmenides (there is no movement, i.e. nothing but [material] Being) versus Heraclitos (there's nothing but movement, i.e. nothing but [obviously material] becoming[118,119].

Process assessment and product assessment are purported to be concerned with the gathering of radically different information. According to Scriven processes are formative

---

117 Ibid.

118 Windelband, W. *A history of philosophy, Vol. I,* 1958, p.

119 Parmenides maintained that generation, change, destruction, and motion are all illusions. The idea that change is illusion is more ancient than Parmenides and goes back to the much more ancient Indian philosophy. Heraklitos went to the other extreme: you can't, he maintained, swim in the same stream twice; for at least two reasons. The stream is everchanging; and so are you.

and goal free, while products are summative and orientated towards a specific goal.[120] This is also the view expressed by participants at the OBET[121] conference who maintained that traditional forms of assessment test the product not the process.[122]

The claim is that formative assessment is flexible because it takes into account the "contexts and conditions of performance" while summative assessment is rigid because it does not.[123] "Abilities", on such a view are flexible, and not the fixed entities of traditional psychology, as in Carroll[124] (see section 2.2). Yet, flexibility/rigidity are not either/or notions but complementary in the sense that one has to have a (rigid) framework of fixed psychological "attributes" within which growth and development, and the study of this growth and development, can take place. As discussed in section 2.2, abilities must (1) have fixed components (which are biologically based), (2) they must be consistent, i.e. not one-off performances, and (3) they must also be variable among humans, i.e. some humans are more able or less able than others.

Besner mentions two "unfortunate" developments that have arisen out of the "stigmatization of the product": the first has been "an uneasy deferral of the concrete, the practical and the substantial...and a corresponding valorization of the means of writing - of what goes on before products are produced"; the second has been the misperception that the product approach is traditional and therefore outdated.[125] The process came to be identified with humanism and the product with positivism. As Kinneavy[126] (cited in Besner[127]) points out, the (practical) product should be an important focus in modern

---

120 Scriven, M. 'The methodology of educational evaluation', in Tyler, R.W., Gagne, R.M. and Scriven, M. (eds.). *Perspectives of curriculum evaluation*, 1967.

121 CBET is the same concept as OBET (Outcomes-Based Education and Training).

122 HSRC. *Language assessment and the National Qualifications Framework*. 1996, p.101.

123 HSRC. *Ways of seeing the National Qualifications Framework*, 1995, p.39.

124 Carroll, J.B. *Human cognitive abilities: A survey of factor analytic studies*, 1993.

125 Besner, N. 'Process against product: A real opposition?' *English Quarterly*, 18 (3), 9-16 (1985), p.9.

126 Kinneavy, J. 'Restoring the humanities: The return of rhetoric from exile', in Murphy, J.J. (ed.). *The rhetorical tradition and modern writing*, 1982.

127 Besner, ibid., p.16.

humanistic thought, as it was in the ancient humanistic "rhetoric" of Isocrates and Cicero. For Hopkins the product/process debate was never an issue because the "text as product" was never in opposition to the process (of writing).[128]

The NQF definitions, except perhaps for *assessment,* are quite different from the traditional definitions that I have discussed in Chapter 2. It seems that the NQF's definitions will predominate in South Africa, e.g. in teacher training programmes, and so they will require getting used to. Gamble puts it more strongly: "The NQF will have a huge impact, because it can be seen in practice how formal assessment changes learner roles."[129] Gamble warns, however, that the "NQF cannot deliver total conceptual clarity to suit everyone's agenda."[130] After all, a test, any test, is an instrument for measuring learning *outcomes.* Calling mutton lamb may change the conception but not (really) change what is real (see my acknowledgement to the UCT philosopher, Andrew Murray, at the beginning of the study).

It is the "conceptual clarity" aspect (Gamble above) that I would like to say a little more about. OBE has four key principles: (1) clarity of focus, (2) expanded opportunity, (3) high expectations and (4) design down.[131] The two most important are (1) and (4), which. hang together, as shown in the next paragraph.

The clarity of focus principle "makes a clear picture of the desired outcome, the starting point of the curriculum, teaching, and assessment planning and implementation, all of

---

128 Hopkins, A. 'Review of Robinson, P. Academic writing: Process or product.' The British Council: Modern English Publications. *English Language Teaching Journal*, 44 (1), 239-240 (1990), p.239.

129 Gamble, J. 'Drawing the issues together', in HSRC. *Language assessment and the National Qualifications Framework*, 1996, p.107.

130 Ibid., p.108.

131 Spady, W. 'Outcomes-based education: An international perspective', in Gultig, J., Lubisi, C., Parker, B. and Wedekind, V. *Understanding outcomes- based education: Teaching and assessment in South Africa*, 1998, p.27.

which must *perfectly match* (or align with) the targeted outcome."[132] This evokes Magnan's[133] description of the differences between quantitative and qualitative research. According to Magnan the two basic differences between quantitative and qualitative research are (1) their relation to theory: the former is deductive while the latter evokes theoretical questions from the research situation, and (2) quantitative research defines terms earlier on while qualitative research allows meaning to develop from observation.[134]

It is arguable whether these are the distinguishing features of these two kinds of research, because these two features, I would think, belong to both quantitative and qualitative research. With regard to (1), all research , qualitative and quantitative, is based on theory (presuppositions) from which one makes deductions. With regard to (2), the idea is that one does not decide all in one shot what one is going to do, but rather one discovers, *inductively*, the direction one's research is going to take as one goes along. This is often done in qualitative research, as Magnan points out. There is no cogent reason, however, why quantitative researchers cannot discover or create what they are going to do, as more light is shed on the path one is travelling on or if one appears at a fork in the road.

According to Spady, whose opinions figure large in South African OBE, educators (specifically curriculum designers) must be "perfectly" clear about what they want the learner to be able to do and, accordingly, *they should "design down" (Spady's point 4 above) from the outcome they want to achieve.*[135] No change in midstream, in case this upsets the predesigned plan. This doesn't seem to be different from the rigid view that Magnan has of quantitative measurement.

It is hard to understand how anyone (learner or teacher) can be creative in OBE or in any kind of research if one cannot change in midstream and let the current of creativity,

---

132 Ibid.

133 Magnan, S.S. *Review of Creswell, J.W. 1994. Research design: qualitative and quantitative approaches*, 1997, p.256.

134 Magnan, S.S. *Review of Creswell, J.W. 1994. Research design: qualitative and quantitative approaches*, 1997, p.256.

135 Spady, W., ibid., p.27.

guided by broad objectives, take control. Surely, "progressive" education is supposed to be about learner-centredness. A design- down job of this thesis would not have yielded the same (interesting?) results, because I didn't have the final design of the study clearly in my head from the outset, even if it was largely quantitatively based. My initial intention was to do only the language proficiency tests. But it was only a month later (February, 1987) that I decided to use these tests to predict academic achievement. It *dawned* on me that the **whole point of language proficiency tests in the academic education situation is to predict academic achievement**. I observed that there has not been a PhD that has tackled in detail both language proficiency testing and its ability to predict academic achievement, and so I thought I would have a go. Gargantuan as the task was, I plodded on in the face of heavy criticisms. One applied linguist objected that the prediction of academic achievement had no place in an applied linguistics thesis. Another made a moral issue of it and found it "appalling" that I did not have a clear picture of the desired outcome from the start. Such attacks from "means-end" theorists, who are often appointed as mentors, reviewers and examiners, would be reason enough for some to give up that PhD and try better things like saving their marriage.

Arjun[136] seems to be right when he says that OBE is not very different from a "means-end" model, where the end - even if it is claimed that the end is based on broad "outcomes"- predetermines the means. The so-called new "process" paradigm of OBE is not very different from the old "product" paradigm of specified objectives: if it were different, according to Arjun, it would have little chance of being accepted by most teachers in South Africa. Exactly so.

Recall Macdonald's criticisms mentioned earlier and the criticisms of the proponents of NQF that tests such as the HSRC tests do not tap the process, i.e. what learners "have to be able to do."[137] The problem is that the contrast between the "old" paradigm activity of

---

136 Arjun, P. 'An evaluation of the proposed new curricula for schools in relation to Kuhn's conception of paradigms and paradigm shifts.' *South African Journal of Higher Education*, 12 (1), 20-26 (1998), p.25.

137 Macdonald, C.A. *English language skills evaluation (A final report of the Threshold Project)*, 1990a, p.46

doing tests, such as those used by the HSRC, and the "new" paradigm activity of "negotiating task-demands", which is highly critical of psychometrics, is far from clear. In this regard the contribution of Mclean[138] to assessment in competency-based education and training (CBET) is significant. At the core of CBET assessment is criterion-referenced assessment. Recall that norm-referenced assessment consists in assigning a grade by comparing learners with one another.

Criterion-referenced testing, in contrast, measures (in theory) an individual's performance against a set of criteria that is *supposedly* independent of the performance of other learners on these same criteria. Mclean states: "There is an increasing challenge to this distinction, based on research which demonstrates that forms of criterion-referenced assessment often finesse standards by using some sort of normative comparison...or that norm-referenced assessment often uses implicit disciplinary or other criteria."[139] (Recall the quotation from Rowntree; page 15).

The same idea is echoed in Gipps (in Mclean[140]): "We need to move the debate away from false dichotomies: criterion-referenced assessment versus norm-referenced assessment."[141] In spite of Gipps' warning, there have recently been shifts from norm-referenced notions such as validity, reliability and generalisability toward "trustworthiness" and "authenticity".[142] Yet, the "trustworthiness of score meaning"[143] - to understate the case - is not at all (in Messick, at least) antithetical to validity, reliability and generalisability.

---

138 Mclean, D. 'Language education and the national qualifications framework: An introduction to competency-based education and training', in HSRC. *Language assessment and the National Qualifications Framework*, 1996.

139 Ibid., p.46.

140 Mclean, ibid., p.46.

141 Gipps, C. *Beyond testing: Towards a theory of educational*, 1994, p.163.

142 Mclean, ibid., p.47.

143 Messick, S. *Meaning and values in test validation: The science and ethics of measurement*, 1988, p.19.

6.5 Rater consistency, or reliability

Besides the problem of dichotomising criteria and norms, there is the more serious problem of rater consistency (see section 4.8ff).

Testers try to solve the problem by using more than one rater, ideally four.[144] But in most testing situations only one, at most, two raters, are available. When it comes to doing *research* on rater' judgements, the problem of subjectivity can become much more complex and lock the whole assessment enterprise into a closed system.[145]

This problem is arguably the most obdurate bugbear in assessment. It was for this reason that I took my tests out of the School to shed more light on the problem of rater unreliability in the wider educational context. As emphasised in this study, the two macro-issues in testing are validity (what we are testing) and reliability (how we are testing). Reliability, i.e. objectivity, involves all of the following "how" questions of which rater reliability is, though important, only one.[146]

1. How many items should be included in a test?
2. How can one increase the representativeness of a test?
3. How many different kinds of tests should be used?
4. How can one ensure that there are sufficient raters?

The problem of rater reliability is how to be as fair as possible in the allocation of scores and judgements. This problem seems to take precedence over all other issues in testing. This is understandable because assessment is the last and most crucial stage in the syllabus. (Most learners only protest about poor teaching, unclear exam questions, etc. if

---

144 Computerised assessment is sometimes used to reach a high degree of consensus. But the problem of reliability in the assessment of "subject" tests such as essay tests remain. Because computers are limited in sussing out coherence.

145 Tucker, S.A. and Dempsey, J.V. A semiotic model for program evaluation. *The American Journal of Semiotics*, 8 (4), 73-103 (1991), p.77.

146 Weir, C.J. *Understanding and developing language tests*, 1993, pp.35-37.

they fail). Although rater reliability has been discussed at length in previous pages, something needs to be said concerning solutions to the problem.

As I showed earlier in my discussion of the NAETE educators of teachers of English, a large number of these do not consult with their colleagues on rating procedures. Hopefully, these lone raters do give their student teachers the following hoary advice:

1. At the beginning of an academic year, all the teachers in the department can rate a few students' assignments and discuss the criteria they used and the marks they awarded. This exercise done repeatedly over a period of time might increase interrater reliability.
2. Before a major test or an exam, questions set by individual teachers can be discussed (formally and/or informally) by the whole department in terms of:
   - clarity of the questions
   - length of questions, and
   - number of marks to be awarded for specific questions or sections.

The person who prepares the question should also give a memorandum to others in the department to demonstrate the criteria by which answers will be evaluated.

If one had to follow these procedures one would hope that objectivity would be increased. Yet, it seems that even if one does consult with colleagues (as would be the case with literary or music or culinary critics) hard problems remain. What is worrisome in the assessment of written output is that in spite of discussions and workshops on establishing common criteria, there remain large differences in the relative weight raters attach to the different criteria, e.g. linguistic structure, content and organisation. Raters may differ not only on the relative weight to attach to different criteria, but on the nature of each criterion. For example, with regard to "content" (facts), raters may have different views about the relative importance of specific supporting ideas in a particular topic. The

problem is trying to distinguish between language proficiency, academic skills and the mastery of content in academic performance. We have a dilemma: on the one hand, it is recommended that we test language ability and nothing else: "In language testing we're not normally interested in knowing whether students are creative, imaginative, or even intelligent, have wide general knowledge, or have good reasons for the opinions they happen to hold."[147] On the other hand, it is difficult, perhaps impossible, to separate language-specific cognitive structures from general problem-solving abilities[148] or from world knowledge.[149]

Language assessment is analogous to measuring pain: "How much does it hurt"[150] when you break a leg - or fail an important exam? On the other hand, as I mentioned earlier (towards the end of section 2.5), an optometrist can tell pretty accurately in five minutes whether someone needs specs.[151] Language in one sense is as mysterious, as subjective, as poetic! as pain, but in another sense it is as prosaic and objective as saying ABC in front of an eye-chart (for those with acceptable eyesight). This paradox is the grist of the validity/reliability conundrum, and perhaps of the conundrum of knowledge itself. As far as testing itself is concerned, the relationship between "subjective" and "objective" tests is one manifestation of the relationship between interpretation and the world "out there". Yeld, describing Allen's point of view, maintains that

> *although it is possible to have objective scoring, there is no such thing as an objective test. For example, all decisions surrounding test design and administration of a test are subjective. Since in [Allen's] view, there is no*

---

147 Hughes, A. *Testing for language teachers*, 1989, p.82.

148 (1) Bley-Vroman, R. 'The logical problem of foreign language learning.' *Linguistic Analysis*, 20 (1-2), 3-49 (1990).
(2) Vollmer, H.J. 'The structure of foreign language competence', in Hughes, A. and Porter, D. (eds.). *Current developments in language testing*, 1983, p.22

149 (1) Aitchison, J. *Words in the mind: An introduction to the mental lexicon.* 1987.
(2) Hudson, R. *Word grammar*, 1984, p.34.
(3) Taylor, J.R. *Linguistic categorization*, 1989, p.81ff.

150 Spolsky, B. *Measured words*, 1995, p.320.

151 Eye-tests are at the moment the worry of many South Africans, because the whole population has to submit to a compulsory eye-test after which a (compulsory) new driver's licence will be issued.

"absolute dichotomy" between objective and subjective tests, the criticism of subjective tests are not felt to be valid.[152]

6.6 Paradigm lost: paradigm regained?

It is the "positivistic" paradigm of reducing humans to objects or numbers that activity theorists, task-based theorists and outcomes-based theorists are opposed to. Positivism and anti-positivism are two irreconcilable beliefs, it seems. According to Nunan, "[u]nderpinning quantitative research is the positivistic notion that the basic function of research is to uncover facts and truths which are independent of the researcher.

Qualitative researchers question the notion of an objective reality."[153] Nunan's dichotomy may indeed reflect an authentic opposition between different kinds of world view, namely the quantitative/objective view and the qualitative/subjective view. The basic difference between quantitative methods and qualitative methods is simply this: the former reduces the data to numbers, the latter doesn't. An important debate is whether it is legitimate and useful to "crunch" human behaviour into numbers.

Grotjahn maintains that the qualitative-quantitative dichotomy is a crude over-simplification.[154] Nunan concedes this point yet "still believe[s] that the distinction is a real one, and that the two 'pure' paradigms are underpinned by quite different conceptions of the nature and status of knowledge."[155]

Choosing a paradigm is to a large extent a philosophical exercise based on value judgements. Paradigms are "incorrigible", i.e. they cannot be proven right or wrong by any empirical criterion. Some rational decisions have to be made, but these decisions are

---

152 Yeld, N. *Communicative language testing.* Report on British Council Course 559 offered at Lancaster University from 8 September to 20 September 1985, (published in) 1986, p.32

153 Nunan, D. *Research methods in language learning*, 1992, p.20.

154 Grotjahn, R. 'On the methodological basis of introspective methods', in Faerch, C. and Kasper, G. (eds.). *Introspection in second language research*, 1987.

155 Nunan, D. *Research methods in language learning*, 1992, p.5.

enmeshed in all sorts of subjective considerations where data, i.e. particulars, are often driven by theory, i.e. generalisations.[156]

In science one tries to be aware of the limitations of one's own paradigm. In Kuhn's words,

> *one of the things a scientific community acquires with a paradigm is a criterion for choosing problems that, while the paradigm is taken for granted, can be assumed to have solutions. To a great extent these are the only problems that the community will admit as scientific or encourage its members to undertake. Other problems, including many that had previously been standard are rejected as metaphysical, as the concern of another discipline, or sometimes as just too problematic to be worth the time...One of the reasons why normal science seems to progress so rapidly is that its practitioners concentrate on problems that only their own lack of ingenuity should keep them from solving.*[157]

The philosophy of science is saddled with two contrasting paradigms: the empiricist/ objective/reductionist paradigm and the ethnographical/subjective/holistic paradigm. The first paradigm, the "standard account"[158], involves putting questions directly to Nature and letting it answer: this is the paradigm of empiricist, or normal, science and the Age of Enlightenment, characteristic of modern European thought. This paradigm is based on three assumptions: (i) naive realism, i.e. the reality of objects are separate from observation, (ii) the existence of a universal scientific language, and (iii) the correspondence theory of truth, i.e. propositions about the world are true if they correspond to what is out there. Theories about the world, in this paradigm, must be inferred from observation.

An alternative paradigm, the "seamless web"[159], provides different answers to those offered by the first paradigm. This alternative paradigm has various sectarian aliases: naturalistic; inductivist, postpositivistic, ethnographical, phenomenological, subjective,

---

156 Hesse, M. *Revolutions and reconstructions in the philosophy of science*, 1980, p.187.
157 Kuhn, T. S. *The structure of scientific revolutions*. 2nd Edition., 1970, p.37.
158 Hesse, M. *Revolutions and reconstructions in the philosophy of science*, 1980, p.7.
159 Hughes, T.P. 'The seamless web: Technology, science, etcetera, etcetera.' *Social Studies of Science*, 16 (2), 281-292 (1986), p.292.

qualitative, hermeneutic, humanistic[160] and actor-network.[161] The "seamless web" protagonists accuse reductionists, of which psychometricians are prime examples, of tearing things from their context. Obviously, assessment should not be purely psychometrically based. For some researchers, the question is how much can or must psychometrics be used, for others the question is whether psychometrics should be used at all. And for some others: "If chemists juggled their basic units like we do, their laboratories would blow up."[162]

This strong antagonism is not an idiosyncratic phenomenon. Murray, who, although opposed to norm-referenced testing, is more accommodating. Murray[163] quotes Macken and Slade who state that "we believe that an effective language assessment program must be linguistically principled, explicit, criterion-referenced, and must inform different types of assessment, including diagnostic, formative and summative assessment."[164] Macken and Slade, and Murray, therefore, do not reject summative methods, i.e. quantitative measurement.[165]

While Americans deride the United Kingdom's lack of concern with empirical data, the United Kingdom, according to Alderson (in Yeld[166]), scorns American approaches to test

---

160 Lincoln, Y.S. and Guba, E.G. *Naturalistic enquiry*, 1985, p.7.

161 Mackenzie, D. *Knowing machines: Essays on technical change*, 1996.

162 Ochsner, R. 'A poetics of second-language learning.' *Language Learning*, 29 (1), 53-80 (1979, p.58.

163 Murray, S. 'Exploring the possibilities of using an outcomes-based approach in English teacher education.' *Southern African Journal of Applied Language Studies*, 5 (2), 21-37 (1997), p.28.

164 Macken, M. and Slade, D. Assessment: A foundation for effective learning in the school context, in Cope, W. and Kalantzis, M. (eds.). *The powers of literacy: A genre approach to teaching writing*, 1993, pp.205-206.

165 If, as part of an assessment programme, one wants to *predict* language *proficiency* or use language proficiency tests to predict academic achievement, I suggest that such predictions do not require a major focus on diagnostic testing or formative assessment *per se* as indicated in Murray. This is not to say that diagnostic and formative issues should be excluded from language proficiency tests, but only that we should keep in mind the different purposes of proficiency, diagnostic, aptitude and achievement tests.

166 Yeld, N. *Communicative language testing.* Report on British Council Course 559 offered at Lancaster University from 8 September to 20 September 1985, (published in) 1986, p.38

content, their naive view of language and their obsessive concern with data. Such derision and scorn is unfortunate. Hopefully, this study helped overcome this "atlantic" divide.

South Africa also is experiencing an opposition to quantitative measurement. What seems to be occurring in South Africa is an effort to downplay psychometric measurement, which is linked to the resistance to the unpopular notion of the one-off discrete-point test:

> *The psychometric, discrete-point testing tradition still has its adherents today; for example, the Initial Evaluation Test (designed by the HSRC) for black children in Stds 2, 3 and 4 is a discrete point test that is used as a crucial instrument in evaluating the success of an innovative language programme this year. From the point of view of testing functional language competence, the discrete point approach does represent some advance on the first tradition in that it explicitly addresses the pupil's ability to perform a number of specified tasks in the L2. However, the discrete point test cannot of its very nature measure the learner's ability to comprehend or produce holistically a larger or more natural corpus of language material that represented by individual- element questions. However, discrete point testing may be expanded to include lengthier, more naturalistic "real-life" exercises which have been tailored to isolate specific items at different points. In this case, we may be approaching more holistic testing, but at the price of using texts varying in authenticity.*[167]

It is incorrect to equate, firstly, "psychometric" with "discrete-point", and secondly, as I have argued, "authentic" tests with "real-life" exercises or tasks. With regard to psychometrics and discrete-point testing, in the early 60's, psychometric measurement was equated with Lado's[168] discrete-point tests. It is this psychometric- structuralist" position of Lado that Spolsky[169] is referring to in his rejection of psychometrics. It is this narrow Spolskyan view of psychometrics that Macdonald identifies with psychometrics in general.

---

167 Macdonald, C.A. *English language skills evaluation (A final report of the Threshold Project)*, 1990a. p.4

168 Lado, R. *Language testing*, 1961.

169 Spolsky, B. 'Approaches to language testing', in Spolsky, B (ed.). *Advances in Language Testing Series, 2.* (Arlington, Virginia. Center for Applied Linguistics, 1978).

Spolsky, B. 'The limits of authenticity in language testing.' *Language Testing*, 2, 31-40 (1985).

Macdonald maintains that the "old" psychometric paradigm has been superseded by a "new", "illuminative" paradigm which uses "naturalistic" "historical-developmental" and "socio-cultural" methods.[170] She rejects the "methods and instruments" of validity - face, content and construct validity - because it is not concerned with real-life language; with naturalness.[171] The question is whether naturalness in human knowledge, and in the testing of that knowledge, is possible without idealisation, i.e. without "unnatural" tests. It isn't. Having said that, all we can aim for is trying to make language (tests) as natural as possible.

Such opposition to psychometric measurement is regrettable, yet understandable in the light of the differences in world views and philosophical traditions that exist among researchers and policy-makers. These differences can be classified under two distinct "identity kits"[172], the quantitative and qualitative.

In the last two decades testing theories and tests have been radically transformed so that "discrete-point" and "integrative tests" have become extremely old hat for many testers, where the emphasis is on testing that is "communicative", or "criterion-referenced", or "performance", or "alternative", or "authentic" [173] or satisifies task demands[174], or is "task-based"[175]. These characteristics of testing, it is claimed, shed more light on the construct of language proficiency. Yet, in spite of decades of attempts to define it, the

---

170 Macdonald, C.A. *English language skills evaluation (A final report of the Threshold Project)*, 1990a. p.20,

171 Macdonald, C.A. *Crossing the threshold into standard three in black education: The consolidated main report of the Threshold Project.* 1990b. p.24.

172 Gee. J. *Social linguistics and literacies: Ideology in discourses*, 1991, p.142.

173 (1) Douglas, D. 'Developments in language testing.' *Annual Review of Applied Linguistics*, 15:167-187 (1995).

(2) Hamayan, E.V. 'Approaches to alternative assessment.' *Annual Review of Applied Linguistics*, 212-226 (1995).

(3) MacNamara, T.F. *Measuring second language performance*, 1996.

174 Macdonald, C.A. *English language skills evaluation (A final report of the Threshold Project)*, 1990a.

175 Skehan, P. *A cognitive approach to language learning*, 1998.

how[176] and the why[177] of language proficiency remains a conundrum. Although we may no longer stand before an "abyss of ignorance"[178] and may be able to agree with Alderson (cited in Douglas[179]) that language testing has "come of age", there are still many problems in language testing, the greatest one arguably being the problem of rater reliability. This does not mean that one should stop measuring until one has decided what we are measuring. One does the best one can by taking into consideration cogent views of language proficiency - modern views and dated ones.

In this study I used traditional discrete-point and integrative tests. These kinds of tests are certainly not dated. For example, one of the test specifications in Alderson, Clapham and Hall[180] is that "tasks" should be "discrete point, integrative, simulated 'authentic', objectively assessable" (A different tune from Alderson[181] discussed earlier, who refrained from using terms such as "discrete-point" and "integrative). These test specifications dovetail with the notion that although these tests do not mirror life, they are nevertheless "good dirty methods [of testing] overall proficiency".[182]

Several modern and state-of-the-art courses in language teaching far from avoiding tests such as multiple choice grammar tests, cloze tests and dictation tests, make them the main ingredients of a test battery. For example, Ur's module on testing lists the following "elicitation techniques"[183] (i.e. tests) :

---

176 Bachman, L.F. *Fundamental considerations in language testing*, 1990b.

177 Davies, A. *Principles of language testing*, 1990.

178 Alderson, J.C. 'Who needs jam?', in Hughes, A. and Porter, D. *Current developments in language testing*, 1983. p.90.

179 Douglas, D. 'Developments in language testing.' *Annual Review of Applied Linguistics*, 15:167-187 (1995), p.176..

180 Alderson, J.C., Clapham, C. and Wall, D. *Language test construction and evaluation*, 1995.

181 Alderson, J.C. 'The cloze procedure and proficiency in English as a foreign language.' *TESOL Quarterly*, 13, 219-227 (1979).

182 Bonheim, H. *Roundtable on language testing.* European Society of the Study of English (ESSE) conference, Debrecen, Hungary, September, 1997.

183 Ur, P. *A course in language teaching: practice and theory*, 1996. pp.33-45.

1. Short simple questions and answers.
2. True/false.
3. Multiple-choice.
4. Gap-filling and completion of sentences.
5. Matching items in one column to items in another.
6. Dictation.
7. Cloze.
8. Transformation of sentences, e.g. change from present to past tense.
9. Rewriting sentences using a few different words.
10. Translation.

As mentioned earlier (pages 8 and 9), the primary consideration for Ur is practicability.

The practicability of tests becomes crucial in modern education institutions, where classes are getting progressively larger. But I would think that Ur's idea of a battery of effective tests is not representative of contemporary views. More representative of contemporary views on what tests should look like are the "communicative" tests of Van der Walt[184] described above.

According to Alderson, Clapham and Hall[185], a test has a life cycle of 12 to 15 years. Yet, there is no scientific reason why some tests should only have a life cycle of 15 years, and not of 100 years. For example, Bloor et al.'s [186] multiple-choice tests that were used in this study are, in my opinion, still as valid as they ever were, because the grammar tested in these tests and the uses to which they are put will remain with the academic community long into the future. Even more so today, owing the fact that the boundary between grammar (syntax and morphology), lexis (which together comprise, in modern linguistics,

---

184 Van der Walt, J.L. 'Some characteristics of communicative tests.' *National Association of Educators and Teachers of English*, 9, 47-51 (1994).

185 Alderson, J.C., Clapham, C. and Wall, D. *Language test construction and evaluation*, 1995. p.227.

186 Bloor, M., Bloor, T., Forrest, R., Laird, E. and Relton, H. *Objective tests in English as a foreign language*, 1970.

lexico-grammar) and pragmatics is becoming increasingly blurred. Analogously, the distinction between artificial and natural learning, and, accordingly, between artificial and natural testing, is also becoming more blurred: an issue that has been dealt with at length in the study.

Lass maintains that by the time a book on English has been written it is out of date.[187] This is partly true about spoken language, but highly exaggerated with regard to written language, especially as far as the grammar of written academic language is concerned. For, if books and knowledge quickly become outdated, then there is no point in writing unless one is interested in the history of ideas or in excavating the historical sites and sedimentations of writing, which is the wont of deconstruction[188], constructivism and chaos theory, all of which have been received with alacrity by avant garde applied linguists[189] in the hope that they will add some much needed spice to applied linguistics. What we have to ask before deciding to jettison "old" paradigms for "illuminative" ones is,

> *what was the real need for the new test? Where was the evidence, rather than the opinion, that the old test was ineffective, past its prime, ready to pass on to greener pastures? What, in particular, did the users of the test -*

---

187 Lass, R. *"English" - Talk at Will radio programme on SAfm*, 26 February, 1998.

188 Gamaroff, R. Can the deconstructive tour (surprisingly) translate us anywhere? *Journal of Literary Studies*, 13 (3/4), 397-415 (1997c).

189 Larsen-Freeman [Larsen-Freeman, D. 'Chaos/complexity science and second language acquisition.' *Applied Linguistics*, 18 (2), 141-165 (1997)], who acknowledges that she is danger of being carried away to a certain extent by "chaos/complexity" science, states: "As I write this sentence, and as you read it, we are changing English". (p.149) This is true, but is it significantly true? Larsen-Freeman's rapprochement between linguistics and modern science launches language acquisition theory to exciting and giddy heights. But as yet, it hasn't, from a practical point of view of language in use, improved upon the Chomskyan notion of "creativity" and "generation". Without doubt, I have "created" or "generated" sentences that I have never heard or written before, and will probably not write again (these sentences). The basic grammar and lexis, though, hasn't changed since I wrote them or isn't changing as I write them, and these will be with me and many others for a long time to come. Readers know what I mean (I think), even if they don't agree with my appraisal of Larsen-Freeman. This is not to say that native speakers (if they are not all dead, that is!) agree on all points of grammar and lexis in the language that they are native speakers of (see section 6.2). Reductionism, according to Larsen-Freeman (ibid., p.151), is the attempt to reduce the variable, the chaotic to the fixed. There is another kind of reductionism: the attempt to reduce the relatively fixed to the absolutely chaotic.

> *students, the sponsors, the receiving institutions - feel or know about the need for rebirth?*[190]

Even if there exists a strong psychometric justification for using indirect tests as predictors of global language proficiency, "communicative" testers will argue that indirect tests are not authentic, because they do not test real life. The point is that language testing is *closely* related to language teaching, but they are not the same thing. If I suggest that all the discrete-point tests and integrative tests used in this investigation are useful for testing language *proficiency* (which is concerned with what one knows at a specific point in time), I do not of course mean that more direct methods are not effective in improving *achievement* (what one is specifically taught). If one was concerned with finding how a "language" syllabus (which, by definition, has achievement as its goal) relates to "content" syllabuses, then indirect language tests may be ineffective, but not necessarily so.

What is occurring in South Africa is an effort to downplay statistical (psychometric) measurement, which is linked to the resistance to the unpopular notion of the one-off test and the preference for process-oriented measures as described in OBE and CBET. For example, Docking contrasts the "rigorous and detailed management of competency development" with the "'loosely' defined evidence which is 'doctored' and legitimated through statistical procedures on the other (traditional teaching and assessment)."[191] As I have argued in this study, I find it hard to understand how one can establish any principles of language testing without some - indeed, a large - recourse to norms, no matter how "authentic" the task is claimed to be. And norms imply statistical measurement.

The rejection of statistical measurement by some OBE and CBET theorists in the name of restoring individuality to learning is misguided and is consequently having a negative influence on language testing in South Africa.

---

[190] Alderson, J.C., Clapham, C. and Wall, D. *Language test construction and evaluation*, 1995, p.228.

[191] Docking, R. 'Competency-based curricula - the big picture.' *Prospect*, 9 (2), (1994), p.15.

The main problem in human evaluation is how to assess human abilities in an individual, i.e. authentic, real way. One is conscious of the danger that

> *group statistics may falsify the facts of individual speech, since individuals having a given phenomenon always present or always absent are lost in group statistics among the hordes where uses of the phenomenon more obviously reflect the Great Bell Curve in the sky.*[192]

The "Great Bell Curve" as many mathematicians and statisticians are aware, straddles both earth and sky, and is thus an important concept in the study of human constructs, specifically the constructs of human abilities. As was discussed earlier (section 2.2), an important point in the study of human abilities is that the variability in ability between individuals obeys a "bell-curve" distribution. The "bell-curve" or "normal" distribution is the foundational principle of psychometrics, and a foundational principle in this study as well, which is heavily concerned with psychometrics, and consequently with the relationship (1) between individuals in a group, and (2) between groups. The difficulty in research is trying to be both group-orientated and individual-oriented.

Whatever the inadequacies of statistics, the best argument for its usefulness is the fact that much of academic evaluation ultimately ends up as a score, and if that is the brutish fact of the matter, we might as well try and measure this score properly. Having said that, it is undeniable that "true ethnography demands as much training skill"[193] as statistical measurement and it is possible to use the concepts of objectivity and reliability of the empirical analytical paradigm in an individual way.[194] What is important is that quantitative and qualitative researchers both realise that each has a crucial - and

---

192 Bailey, C.J. 'The state of no-state linguistics.' *Annual review of anthropology*, 5, 93-106 (1976), p.97-98. Cited in Nicholas, H. and Meisel, J.M. 'Second language acquisition: The state of the art', in Felix, S.W. and Wode, H. (eds.). *Language development at the crossroads: Papers from the Interdisciplinary Conference on Language Acquisition at Passau*, 1983, p.82. (Nicholas and Meisel's context is second language acquisition).

193 Nunan, D. *Syllabus design*, 1988, p.53.

194 Oskowitz, B. 'Preparing researchers for a qualitative investigation of a particularly sensitive nature: Reflections from the field.' *South African Journal of Psychology*, 27 (2), 83-88 (1997), p.83.

complementary - contribution to make to the human sciences[195], where the "accumulation of data is at best the humble soil in which the tree of knowledge can grow".[196]

Being mindful of Hesse's caveat that "if all theories are dangerous and likely to be superseded, so are the present theories in terms of which the inductivist judges the past"[197], we need to be careful in any claims we may have to ultimate truth (a good example was set by Spolsky[198] in his attenuation of his strong negative attitude towards psychometrics mentioned earlier) because the search for truth is a never-ending path towards understanding and stability of meanings, which is indispensable for individual freedom and social equilibrium.

During the last two decades there have been attempts towards making educational research more "human", and through these attempts has sprung the conflict between the orthodox scientific and objective methods of experimental research and statistical analysis, on the one hand, and "new paradigm research"[199], on the other.

Below is a summary of the salient features of "new paradigm research"[200]:

1. There is too much "quantophrenia" going on. The emphasis should fall on human significance, not on statistical significance. Researchers should become involved in the human side of the phenomenon under study, because the person behind the data can often upset the neat statistics. This means that people should not be reduced to variables or to operational definitions in order to be manipulated into a research design. The paradox of knowledge is that it is impossible to understand the bits - of culture, theory, language, etc. - unless we understand how the whole fits together in its function; and

---

195 Huysamen, G.K. 'Parallels between qualitative research and sequentially performed quantitative research.' *South African Journal of Psychology*, 27 (1), 1-8, (1997).

196 Lorenz, K. 'On the biology of learning', in Kagan, J. *On the biology of learning*, 1969, p.77.

197 Hesse, M. *Revolutions and reconstructions in the philosophy of science*, 1980, p.5.

198 Spolsky, B. *Measured words*, 1995.

199 Reason, P. and Rowan, J. (eds.). *Human enquiry: A source book of New Paradigm Research*, 1981.

200 Ibid., pp.xiv-xvi.

without an understanding of the structure of the discrete bits, we won't be able to understand how the whole works.[201] Which echoes the UCH and the discrete-point/ integrative controversy, and Oller's famous paragraph (see page 22):

> *[N]ot only is some sort of global factor dependent for its existence on the differentiated components which comprise it, but in their turn, the components are meaningfully differentiated only in relation to the larger purpose(s) to which all of them in some integrated (integrative? - original brackets) fashion contribute.* [202]

2. Care must be taken not to make outlandish generalisations from unrepresentative samples.

3. Safe, respectable research should be avoided.

4. Fear of victimisation may cause the researcher to pick only those bits of research that will impress and please.

5. Science requires the humility to change one's views in the light of better theories or new observations.

Reason and Rowan's[203] view is that statistical (quantitative, objective) research and "human" (qualitative, subjective) research are complementary. This is the view expressed in this study as well.

Rutherford, like Reason and Rowan, expresses his misgivings about the danger of reducing humans to objects. He quotes the physicist Niels Bohr: "Isolated material particles are abstractions, their properties being definable and observable only through their interaction with other systems."[204] Rutherford's point is that the *testing* of humans cannot be isolated into parts. But surely language can only be tested through its parts: small parts (sentences and parts of sentences) or big parts ("pragmatic" language). *How*

---

201 Rorty, R. *Philosophy and the mirror of nature*, 1980, p.319.

202 Oller, J.W., Jr. 'A consensus for the 80s', in *Issues in language testing research*, 1983, p.36.

203 Reason, P. and Rowan, J. (eds.). *Human enquiry: A source book of New Paradigm Research*, 1981.

204 Rutherford, W.E. *Second language grammar: Learning and teaching*, 1987, p.65.

the parts and the whole interact is what we're not so clear about, which is the basic problem not only of testing, but also of knowledge itself.

Both the quantitative and qualitative paradigms describe and prescribe, but the tendency in qualitative research is more towards the descriptive, as in ethnographical research, while quantitative research tends to be more prescriptive (e.g. assigning cut-off scores). Yet, in order to make moral, political and economic sense of assessment, both paradigms (a "holodigm"!) are necessary. The reduction of one paradigm leads to the reduction of the other. Accordingly, the suggestion that quantitative measurement and traditional forms of testing be replaced by qualitative methods such as "negotiating the task-demands" is untenable. This is not to say that task-demands do not have a valuable place. The "old paradigm" of traditional testing and "negotiating the task-demands" could quite easily complement each other.

The problem in all kinds of assessment, whether formative, summative, quantitative or qualitative, is not only how to assess individual people, but how to assess individual criteria (e.g. grammar, content). Criteria are abstractions when separated from the functional language of life. But if one does not reduce life to constructs, or to criteria, or to tasks, or to guidelines, there is the strong possibility that learning (and teaching), and thus the study of and writing about language proficiency, or about anything, would not bear much fruit. The problem is how to take humpty-dumpty apart without getting egg all over one's face.

6.7 Conclusion of the study

In this study I have tried to:

(1) Show that integrative *or* discrete-point tests are valid measures of levels of proficiency and valid predictors of academic achievement.

(2) Elucidate some of the problems encountered in the psychometric assessment of second language proficiency and the prediction of academic achievement.

(3) Describe some of the difficulties in constructing a satisfactory - and satisfying - theory of the relationship between "real-life" tests and "old paradigm" tests. Ultimately, the coupling of "communicative" tests or activities exclusively to "real-life" tests or activities has been a lamentable "distraction".[205]

The main issue in educational testing is how to measure, and accurately, individual differences within language-specific abilities and academic abilities; how to recognise performance, which has to do with the setting of valid standards, i.e. with what one considers relevant to fulfilling the purposes of education. It is on this issue of relevance that people differ, and a major part of that relevance is the sociopolitical dimension. In South Africa, admission tests, placement tests and promotion tests now play second fiddle to the more pressing desire for sociopolitical transformation. In this regard, a ruling (April 1996) by the newly constituted Constitutional Court of South Africa has prohibited the use of admission tests at government-aided schools, which means that "language competency to determine admission may now not be used".[206] The argument is that particular interest groups were using such tests to "ensure a homogeneous cultural character in public schools"[207] to the detriment of the majority of the population.

In OBE the emphasis is on co-operative learning. The danger is that co-operative learning may lead to co-operative testing. There's nothing wrong with doing exercises in class together, but there is a whole lot that is questionable about doing tests together.[208] After all, groups don't have skills and abilities; individuals do. Individuals go for job interviews; not groups. I am reminded of a delegate's remark at a South African Academic Development conference at the University of Fort Hare in 1995 in a session on co-operative learning. He remarked that students at his university would only be interested in co-operative learning if there were to be co-operative testing. It is possible to design

---

205 Widdowson, H.G. 'Skills, abilities, and contexts of reality.' *Annual Review of Applied Linguistics*, 18, 323-333 (1998), p.33.1

206 Rickard, C. '"Racist" school tests outlawed.' *Sunday Times*, April 7, 1996, p.4.

207 Ibid.

208 Alderson, J.C., Clapham, C. and Wall, D. *Language test construction and evaluation*, 1995, p.42.

good tests that yield a mixture of group and individual scores, as long as the group scores are not an excuse to hide the inadequacies of individuals, which is difficult to avoid in group scores. But ultimately it is the individual score that must carry the most weight, because a group is made up of individuals who, when interviewed for most jobs, e.g. teachers, managers, laboratory technicians, clerks, must be able to deliver the "individual" goods.

Sebatane[209] questions whether learners should fail at all. He argues that if learners fail this would indicate that they have been denied their "basic" educational rights ("basic" for Sebatane means any of the grades from Grade 1 to Grade 10). Dreyer[210], in a similar vein, argues that grading, i.e. the "wad-ja-get" grading game[211], causes people to fail.

As Sebatane and Dreyer would have it, there should be less talk about individual abilities and more talk about potential that can develop into ability under the correct mediation/ intervention. The onus, therefore, falls on the teacher to ensure that this potential flowers into ability.

The views of Sebatane and Dreyer are diametrically opposed to the "teach to the test" mentality. Yet, if one doesn't teach to the test (in this very un-ideal world) most learners will not take their work seriously. *The majority of learners are not interested in the enlightenment notion of the love of knowledge: they simply want to know what they have to learn to pass.* Makoni (personal communication) argues for a model of curriculum development that begins with language testing and then moves backwards to syllabus/ curriculum design. I don't see how the syllabus/curriculum can operate in any other way.

---

209 Sebatane, E.M. *Assessment policy strategies, implementation and impact: A global perspective.* Tenth World Congress of the Comparative Education Society (WCCES), Cape Town, July 12-17, 1988.

210 Dreyer, C. *Testing: The reason why pupils fail.* National Association of Educators of Teachers of English (NAETE) conference " Training teachers for the South African context, Potchefstroom College of Education, September 17-18, 1998.

211 Kirschenbaum, H., Simon, S.B. and Napier, R.W. *Wad-ja-get?: The grading game in American Education*, 1971.

In the recent documentation of the South African Employment Bill, the term ability has been replaced by potential, the reasoning being that if someone doesn't have ability this doesn't mean that one does not have potential. If aptitude tests are very difficult to construct, the construction of "potential" tests must be gargantuan, if not impossible. I would think, the popularity of the notion of potential among South African policy-makers, for whom tests that involve grading of any kind - whether in education and the workplace - are becoming increasingly unpopular.

The scope of this study did not allow an in-depth discussion of the role of social, psychological, cultural and political factors in language testing. This does not mean that they are not important, for they can reveal much about the *causal* connections between language proficiency and academic performance. In attempting to find equitable methods of assessment, one should be sensitive to the upheavals that may result from trying to impose a Euro-American sociocultural system on the black populations of South Africa. For example, one of the puzzles encountered by English-mother-tongue (usually white) teachers of ESL and other academic subjects is the general bewilderment rather than resistance of black learners when confronted by the cultural demands of white society. What these learners seem to regard as central is not cognitive growth, reasoning, or logic, but rather the social adjustments needed to cope with learning a different language and culture.[212] This does not mean that there is no logic in Africa! Feelings and emotions play a determining role in the learning process. Accordingly, feelings of cultural anomie (estrangement) should not be pushed aside in the rush to develop new curricula, where testing plays such an important role.[213]

South African educationists such as Alexander (in Singh) insist on fundamental changes. Alexander (in Singh[214]) suggests that future policy calls for a coupling of "antiracist/sexist/

[212] Cazden, C.B., John, V.P. and Hymes, D. *Functions of language in the classroom*, 1985, p.xxxi.

[213] Gamaroff, R. 'Solutions to academic failure: The cognitive and cultural realities of English as the medium of instruction among black ESL learners.' *Per Linguam*, 11 (2), 15-33 (1995c), p.15.

[214] Singh, M. 'Universities: The wave of transformation.' *Centre for Scientific Development (CSD) Bulletin*, 4 (7). (Pretoria, Centre for Scientific Development (CSD), 1992, p.2.

elitist/classist/authoritarian/conformist educational practices" to "new methodologies/ syllabi/ways of assessment/attitudes towards language". Although Alexander's considerations should be given serious thought in the restructuring of methodologies, one should also give serious consideration to individual differences in cognitive/academic ability. It would be unwise to ascribe all causes of the inability to learn and do tests and examinations to environmental factors such as bad teaching, poverty, politics and the incompatibility of a target culture's cognitive styles, or *tricks*. What should become a primary focus in education is the rainbow of individual differences manifested in the ability to overcome intellectual, social, cultural and political constraints that hinder academic development and learning.

Assessment is a highly politicised domain, where educationists, in their endeavour to improve evaluation systems, often come into conflict firstly, with politicians who may have little knowledge or interest in educational research, especially research that does not resonate with their political views, and secondly, with other educationists who may have contrary - often radically different - views on what should or can be changed. The question is how to make up for past inequities so that disadvantaged learners can move towards a "pedagogy of possibility".[215] The difficulty is finding educationists and politicians - and applied linguists and cross-cultural psychologists - who know what is possible, and who, accordingly, are able to make judgements that are motivated by scientific facts and realistic compassion and not by a short-sighted sense of democracy, or by the desire for retribution, political control, or financial gain. In South Africa, educational issues in the context of evaluation, e.g. admission tests, placement tests and promotion tests, are beginning to play second fiddle to the more imperious need for sociopolitical transformation. Also frowned upon is the "L2" label. A common outcomes approach, in my view, does not mean that the L1/L2 distinction as well as the tests that cater for these two kinds of learners should get lost in the nooks and crannies of multiculturalism. Let's not change useful labels, but our thinking dispositions.[216] The

---

215 Peirce, B. 1989. 'Towards a pedagogy of possibility in the teaching of English internationally: People's English in South Africa.' *TESOL Quarterly*, 23 (3), 401-420 (1989).

216 Perkins, D., Jay, E. and Tishman, S. 'New conceptions of thinking: From ontology to

danger is that in the frantic resistance to differences, or diversity, we scuttle sensible principles of pedagogy.

The desire for common approaches and the concomitant resistance to diversity in education that seems to follow, emerge from the conviction that a "high degree of cultural homogeneity is important to the stability, and perhaps the moral integrity, of a political community".[217] This is understandable in a country where one group had dominated other groups on a vast institutional scale. That is why much is at stake in testing, where assessments have to be made about levels of ability, where judgements - often the occasion, and sometimes the cause, of much distress - have to be made about whether somebody should be admitted to an education programme or to a job, or promoted to a higher level. Within the sociopolitical and multi-lingual-cultural-racial-ethnic context of South Africa, these judgements assume an intense poignancy.

There is bound to be a clash between the reality of the lived experience, of living language, of the ethical value of each individual and the requirement of sorting individuals into groups in order to decide in a top-down fashion who gets and does what. Objectivity, i.e. understanding and being understood, is what all humans seek, whether scientists, teachers, learners or postmodernists. Being reduced to objects is what humans abhor. *That* is the reliability-validity problem and the problem that this study tried to illuminate.

Protagonists of "task-based" learning might have liked a more original study in assessment, which should have started where this one finishes, namely, with the question of task demands which learners have to cope with in education, where such a study should have devised assessment tasks which attempt to mirror such tasks, and then compared learners' performances on such tasks with their academic achievements so that

---

education.' *Educational Psychologist*, 28 (1), 76 (1993), p.76.

[217] Cloete, G.N, Muller, J., M.W. Makgoba, Kong. D.E. (eds.). *Knowledge, identity and curriculum transformation in Africa*, 1997, p.108.

one would have got a clearer idea of "what really matters in second language learning".[218] But this was not the main focus of this study. In this study I tried to show the enduring value of "old" elicitation techniques and of the quantitative methods that are used to assess output. Undoubtedly, significant features are lost when only quantitative measures are used. That is why I have advocated that quantitative and qualitative measures be used together. Section 4.8.1ff provided an example of how this is done..

What I have attempted to do in this study is not simply to defend "old paradigm" research but to, if not recreate it, rethink, or re-imagine, it, on the basis of resources from an old paradigm. What is undoubtedly true is that I have applied this old paradigm to an educational context fraught with political and social controversy; a context that could only be touched on in this study.

6.8 Summary of Chapter 6

This study was based on "old paradigm" research, which I found to be a valuable tool in language testing. In spite of their loss in popularity, "old paradigm" norm-referenced tests such as "discrete-point" tests (all of which are indirect tests) and "integrative" tests (mostly indirect tests) have not lost any of their validity, reliability or practicality. The findings were summarised and their implications were discussed in terms of improving testing, where various South African initiatives were discussed. The main issue in educational testing is how to recognise performance, which has to do with the setting of valid standards. Valid standards should be concerned with fulfilling the relevant purposes of education. The mother of questions is: What is relevant? With regard to the theoretical and practical focus of this study, which has been the re-assessment of "old paradigm" testing, the relevant issue in the *predictive* validity of *proficiency* tests is finding tests that will do the best job, even if this means resurrecting - and if this be too strong, then transforming - the old paradigm.

---

[218] Saville-Troike, M. 'What really matters in second language learning for academic achievement.' *TESOL Quarterly*, 18 (2), 199-219 (1984).

Although it is true that language knowledge is much more than something external and autonomously measurable, in order to test this knowledge some form of idealisation is required, even in "authentic" language research, which claims to address life's real problems. I reiterate what I said at the beginning of section 3.1.1, no sample of tests can adequately represent the vast variability of language, nor does it have to. The reason why this is so is "because of the generative nature of language which acts as its own creative source".[219] It is this generative and creative nature of language that in literary theory is the grist of the concept of *intertextuality* where no text can be completely cut off from other texts, historically or geographically. Analogously in testing theory, tests cannot be cut off from the history and geography of their *intertestuality*.

The importance of consistency in assessment cannot be stressed enough. Statistics can assist one in this regard, taking into account that the data still have to be interpreted against the *generally agreed upon* (and there's the rub) theoretical underpinnings of language learning. In this endeavour, one hopes to find a middle way between (1) flogging the so-called dead horse of modernism/positivism - which has been represented in this study by psychometric testing, and (2) putting all one's money on the postmodern horse of humanism, which has been represented in this study by the negotiation of the task-demands of "real-life".

The main issue is not whether to go back to the test or not, but to test - and teach - fairly. Zero testing means zero learning: for most learners, and perhaps for most teachers as well. *That* is real life, and so forth.

---

219 Davies, A. *Principles of language testing*, 1990, p.3.

Bibliography

African National Congress. *ANC policy guidelines for a democratic South Africa.* (Johannesburg, ANC, 1992).

Aitchison, J. *Words in the mind: An introduction to the mental lexicon.* (Oxford, Basil Blackwell, 1987).

Alatis, J.E. (ed.). *Georgetown University Round Table on Languages and Linguistics.* (Washington, D.C., Georgetown University Press, 1992).

Alderson, J.C. *A study of the cloze procedure with native and non-native speakers of English.* Unpublished Ph.D. Dissertation. (Edinburgh, University of Edinburgh, 1978).

Alderson, J.C. 'The cloze procedure and proficiency in English as a foreign language.' *TESOL Quarterly*, 13, 219-227 (1979).

Alderson, J.C. 'Native and nonnative speaker performance on cloze tests'. *Language Learning*, 30 (1), 59-77 (1980).

Alderson, J.C. 'Report of the discussion on general language proficiency', in Alderson, J.C. and Hughes, A. *Issues in language testing: ELT Documents III.* (The British Council, 1981a).

Alderson, J.C. 'Report of the discussion on Communicative Language Testing', in Alderson, J.C. and Hughes, A. *Issues in language testing: ELT Documents III.* (The British Council, 1981b).

Alderson, J.C. 'Reaction to the Morrow paper, in Alderson, J.C. and Hughes, A. *Issues in language testing: ELT Documents III.* (The British Council, 1981c).

Alderson, J.C. 'Who needs jam?', in Hughes, A. and Porter, D. *Current developments in language testing.* (London, Academic Press, 1983).

Alderson, J.C. 'The cloze procedure and proficiency in English as a foreign language', in Oller, J.W., Jr. (ed). *Issues in language testing research.* (Rowley, Massachusetts, Newbury Publishers, 1983a).

Alderson, J.C. and Clapham, C. 'Applied linguistics and language testing: A case study of the ELTS test.' *Applied Linguistics*, 13 (2), 149-167 (1992).

Alderson, J.C., Clapham, C. and Wall, D. *Language test construction and evaluation.* (Cambridge, Cambridge University Press, 1995).

Alderson, J.C. and Hughes, A. *Issues in language testing: ELT Documents III.* (The British Council, 1981).

Alexander, N. 1996. 'Drawing the issues together: In the context of language education policy, in HSRC.' *Language assessment and the National Qualifications Framework.* (Pretoria, Human Science Research Council Publishers, 1996).

Allen, D. 'Testing written communication', in Yeld, N. *Communicative language testing*: Report on British Council Course 559 offered at Lancaster University from 8 September to 20 September 1985. (Cape Town, University of Cape Town. 1986).

Allright, R.L. 'Research on teacher behaviour and the curriculum', in Burt, K. and Dulay, H.C. *New directions in second language learning, teaching and bilingual education.(TESOL,* Washington, D.C., 1976).

American Psychological Association. *Standards of educational and psychological measurement.* (Washington D.C., American Psychological Association, 1974).

Anderson, J. *Psycholinguistic experiments in foreign language testing.* (Queensland, University of Queensland Press, 1976).

Arjun, P. 'An evaluation of the proposed new curricula for schools in relation to Kuhn's conception of paradigms and paradigm shifts.' *South African Journal of Higher Education,* 12 (1), 20-26 (1998).

Ashworth, P. and Saxon, J. 1990. On competence. Journal of Further and Higher education, 14 (2), 23.

Atkinson, M., Kilby, D. and Roca, I. *Foundations of general linguistics.* (London, George, Allen and Unwin, 1982).

Bacheller, F. 'Communicative effectiveness as predicted by judgements of the severity of learner errors in dictations', in Oller, J.W., Jr. and Perkins, K. (eds.). *Research in language testing.* (Rowley, Massachusetts, Newbury House, 1980).

Bachman, L.F. 'The trait structure of cloze test scores.' *TESOL Quarterly,* 16 (1), 61-70 (1982).

Bachman, L.F. 'The development and use of criterion-referenced tests of language ability in language program evaluation', in Johnson, R.K (ed.). *The second language curriculum*. (Cambridge (USA), Cambridge University Press, 1989).

Bachman, L.F. 'Assessment and evaluation. *Annual Review of Applied Linguistics (1989)*, 10, 210-226 (1990a).

Bachman, L.F. *Fundamental considerations in language testing*. (Oxford, Oxford University Press, 1990b).

Bachman, L.F. and Clark, J.L.D. The measurement of foreign/second language proficiency.' *American Academy of the Political and Social Science Annals*, 490, 20-33 (1987).

Bachman, L.F. and Palmer, A.S. *Language testing in practice*. (Oxford, Oxford University Press, 1996).

Bailey, C.J. 'The state of no-state linguistics.' *Annual review of anthropology*, 5, 93-106 (1976).

Barkhuizen, G. 'Proposal for an independent English Second Language Department at Mmabatho High School.' *English Language Teaching Centre (ELTIC) Reporter*, 16 (1), 25-32. Johannesburg, 1991).

Barkhuizen, G. 'Teaching English in multilingual settings (TEMLS): What needs to be done.' *Journal for Language Teaching*, 26 (4), 53-68 (1992).

Barkhuizen, G. 'Preparing teachers to teach in multilingual settings. Current approaches to the teaching of English for academic purposes: A critical appraisal.' *Proceedings (Part 1) of the South African Applied Linguistics Association conference 'Our multilingual society: Supporting the reality*. (University of Port Elizabeth, 1993).

Barkhuizen, G. 'Using English in the South African classroom.' *Per Linguam*, 12 (1), 34-47 (1996).

Barkhuizen, G. and Gough, D. 'Language curriculum development in South Africa: What place for English?' *TESOL Quarterly*, 30 (3), 453-461 (1995).

Barry, D.M., Cahill, S., Chamberlain, J.C., Reinecke, S. and Roux. . *Past and future approaches to language test development with examples*. (Pretoria, Human Sciences Research Council, undated).

Bennett, A. and Slaughter, H. 'A sociolinguistic/discourse approach to the description of the communicative competence of linguistic minority children', in *Rivera, C. The ethnographical/ sociolinguistic approach to language proficiency assessment.* (Clevedon, Avon, Multilingual Matters Ltd., 1983).

Bereiter, C. and Scardemalia, M. 'Does learning to write have to be so difficult?', in Freedman, A., Pringle, I. and Yalden, J. *Learning to write: First language/second language.* (London, Longman, 1983).

Bernstein, B. *Class, codes and control.* (St. Albins, Paladin, 1971).

Besner, N. 'Process against product: A real opposition?' *English Quarterly*, 18 (3), 9-16 (1985).

Bhaba, H. *The location of culture.* (London, Routledge, 1994).

Bialystok, E. 'A theoretical model of second language learning.' *Language Learning*, 28 (1), 69-84 (1978).

Biggs, C. 'In a word, meaning', in Crystal, S. (ed.) *Linguistic controversies.* (London, Edward Arnold, 1982).

Bloor, M., Bloor, T., Forrest, R., Laird, E. and Relton, H. *Objective tests in English as a foreign language.* (London, Macmillan, 1970).

Bley-Vroman, R. 'The logical problem of foreign language learning.' *Linguistic analysis*, 20 (1-2), 3-49 (1990).

Bock, M. 'Teaching grammar in context', in Angélil-Carter, S. (ed.). *Access to success: Literacy in academic contexts.* (Cape Town, University of Cape Town Press, 1998).

Bolinger, D. 'The atomization of language.' *Language*, 41, 555-573 (1965).

Bonheim, H. *Roundtable on language testing.* European Society of the Study of English (ESSE) conference, Debrecen, Hungary, September, 1997).

Bormuth, J. 'Mean word depth as a predictor of comprehension difficulty.' *Journal of Educational Research*, 15, 226-231 (1964).

Botha, H.L. and Cilliers, C.D. 'Programme for educationally disadvantaged pupils in South Africa: A multi-disciplinary approach.' *South African Journal of education*, 13 (2), 55-60 (1993).

Bott, D. and Satithyudhakarn, V. 'Dictation: Easy and accurate evaluation of "Co-co".' *English Teaching Forum*, 24 (3), 42-44 (1986).

Bradbury, J., Damerell, C., Jackson, F. and Searle, R. 'ESL issues arising from the "Teach-test-teach" programme', in Chick, K (ed.). *Searching for relevance: Contextual issues in applied linguistics*. (South African Applied Linguistics Association (SAALA), 1990).

Breen, M.P. 'Authenticity in the langauge classroom.' *Applied Linguistics*, 6 (1), 60-70 (1985).

Bridgeman, B. *Essays and multiple-choice tests as predictors of college freshman GPA*. (ETS Research Report, 1991).

Brière, E. 'Are we really measuring proficiency with our foreign language tests?' *Foreign Language Annals*, 4, 385-91 (1971).

Brown, A. 'The effect of rater variables in the development of an occupation-specific language performance test.' *Language Testing*, 12, 1-15 (1995).

Brown, J.D. 'A closer look at cloze: Validity and reliability', in Oller, J.W., Jr. (ed). *Issues in language testing research*. (Rowley, Massachusetts, Newbury Publishers, 1983).

Brown, J.D. *Understanding research in second language learning*. (Cambridge, USA, Cambridge University Press, 1988).

Brown, J.D. 'Language programme evaluation: A synthesis of existing possibilities', in Johnson, R.K. (ed.). *The second language curriculum*. (Cambridge, USA, Cambridge University Press, 1989).

Brown, J.D. 'Statistics as a foreign language - Part 2: More things to consider in reading statistical language studies.' *TESOL Quarterly*, 26, (4), 629-664 (1992).

Brown, J.D. 'Language program evaluation: Decisions, problems and solutions.' *Annual Review of Applied Linguistics*, 15, 227-248 (1995).

Brown, K. *Linguistics today*. (Suffolk, Fontana Paperbacks, 1984).

Bridges, P. 'Transferable skills: A philosophical perspective.' *Studies in Higher Education*, 18 (1), 43-52 (1993).

Brumfit, C.J. Notional syllabuses revisited: A response. *Applied Linguistics*, 2 (1), 90-92 (1981).

Bundy, C. *'Talk at Will' radio programme on SAFM* (6 January, 1998).

Burroughs, E., Vieyra-King, M. and Witthaus, G. 'The assessment of language outcomes in ABET: Implications of an approach', in HSRC, *Language assessment and the National Qualifications Framework.* (Pretoria, Human Science Research Council Publishers, 1996).

Butzkamm, W. 'Review of H. Hammerly, "Fluency and accuracy: Toward balance in language teaching and learning"'. *System*, 20 (4), 545-548 (1992).

Byrnes, H. and Canale, M. (eds.). *Defining and developing proficiency: Guidelines, implementations and concepts.* (Lincolnwood, Chicago, Illinois, National Textbook Company. In conjunction with the American Council on the Teaching of Foreign Languages, 1987).

Calitz, F. 'So what went wrong with the matric class of 97?' *Sunday Times (South Africa)*, January 11, 1998.

Campbell, C.M. *Learning and development: an investigation of neo-Piagetian theory of cognitive growth.* Master of Arts thesis, University of Natal, 1985.

Canale, M. 'On some dimensions of language proficiency', in Oller, J.W., Jr. (ed). *Issues in language testing research.* (Rowley, Massachusetts, Newbury Publishers, 1983).

Canale, M. 'The measurement of communicative competence.' *Annual Review of Applied Linguistics*, 8, 67-84 (1988).

Canale, M. and Swain, M. 'Theoretical bases of communicative approaches to second language teaching and testing.' *Applied Linguistics*, 1 (1), 1-47 (1980).

Carmines, G. and Zeller, A. *Reliability and validity assessment.* (Beverley Hills, California, Sage Publications, 1979).

Carroll, B.J. *Testing communicative performance.* (Oxford, Pergamon, 1980).

Carroll, J.B. *Fundamental considerations in testing for English language proficiency of foreign language students.* (Washington, D.C., Center for Applied Linguistics, 1961).

Carroll, J.B. 'The psychology of language testing', in Davies, A. (ed.). *Language testing symposium.* (London, Oxford University Press, 1968).

Carroll, J.B. 'Psychometric approaches to the study of language abilities', in Fillmore, C.J., Kempler, D. and Wang, S. (eds.). *Individual differences in language ability and language behaviour.* (New York, Academic Press, 1979).

Carroll, J.B. 'Psychometric theory and language testing', in Oller, J.W., Jr. (ed). *Issues in language testing research.* (Rowley, Massachusetts, Newbury Publishers, 1983).

Carroll, J.B. *Human cognitive abilities: A survey of factor analytic studies.* (Cambridge, Cambridge University Press, 1993).

Cassirer, E. *Language and myth.* (Davies Publications, 1946).

Cattell, R.B. *Measuring intelligence with culture-fair tests.* (Institute of Personality and Ability Testing, 1973).

Cazden, C.B., John, V.P. and Hymes, D. *Functions of language in the classroom.* (Illinois, Waverley, 1985).

Chang-Wells, G.L.M. and Wells, G. 'Dynamics of discourse: Literacy and the construction of knowledge', in Forman, E.A., Minick, N. and Stone, C.A. (eds.). *Contexts for learning: Sociocultural dynamics in children's development.* New York. Oxford University Press, 1993).

Child, J. 'Proficiency and performance in language testing.' *Applied Linguistic Theory*, 4 (1/2), 19-54 (1993).

Chomsky, N. *Syntactic structures.* (The Hague, Mouton, 1957).

Chomsky, N. *Aspects of the theory of syntax.* (Cambridge, Massachusetts, M.I.T. Press, 1965).

Chomsky, N. 'Recent contributions to the theory of innate ideas.' *Synthese,* 17, 2-11 (1967).

Chomsky, N. *Reflections on language.* (New York, Pantheon Books, 1975).

Chomsky, N. *Language and the problem of knowledge.* Cambridge, Massachusetts, MIT Press, 1988).

Christopherson, P. *Second-language learning.* (Harmondsworth, Penguin, 1973).

Clark, J.L.D. 'Theoretical and technical considerations in oral proficiency testing', in Jones, R.L. and Spolsky, B. (eds.). *Testing language proficiency.* (Arlington, Virginia, Center for Applied Linguistics, 1975).

Clark, J.L.D. 'Psychometric considerations in language testing', in Spolsky, B. (ed.). *Approaches to language testing: Advances in language testing*, Series 2. (Arlington, Virginia, Center for Applied Linguistics, 1978).

Clark, J.L.D. Language testing: Past and current status - Directions for the future. *Language testing*, 64 (4), 431-443 (1983).

Clayton, E. 'Scaffold: Graded support or gibbet? The acquisition of terminology-concepts in a scientific discipline.' *Proceedings of the South African Association for Academic Development (SAAAD) Conference*, Technikon Free State, Bloemfontein, 29 November - 1 December, 1995.

Clegg, J. (ed.). *Mainstreaming ESL: Case studies in integrating students into the mainstream curriculum.* (Clevedon, Multilingual Matters Ltd., 1996).

Cloete, G.N, Muller, J., M.W. Makgoba, Kong. D.E. (eds.). *Knowledge, identity and curriculum transformation in Africa.* (Johannesburg, Maskew Miller Longman, 1997).

Cobb, P., Wood, T. and Yackel, E. 'Discourse, mathematical thinking, and classroom practice', in Forman, E.A., Minick, N. and Stone, C.A. (eds.). *Contexts for learning: Sociocultural dynamics in children's development.* (New York, Oxford University Press, 1993).

Coe, R.M. 'An apology for form; or, Who took the form out of the process.' *College English*, 49 (1), 13-28 (1987).

Collier, V.P. 'Age and rate of acquisition of second language for academic purposes.' *TESOL Quarterly*, 21 (4), 617-641 (1987).

Collier, V.P. 'Second-language acquisition for school: Academic, cognitive, sociocultural, and linguistic processes', in Alatis, E. Straehle, C.A., Gellenberger, B. and Ronkin, M. (eds.). *Georgetown University Round Table on Languages and Linguistics* (1995).

Corder, S.P. *Error analysis and interlanguage.* (Oxford, Oxford University Press, 1981).

Cress, K. Reassessing assessment. *The Teacher*, 1 (7), 9-10. Johannesburg, Teacher Trust and Mail and Guardian, 1996.

Creswell, J.W. *Research design: qualitative and quantitative approaches.* (Thousand Oaks, CA, Sage, 1994).

Christie, R. 'Avoid grade creep.' *Mail and Guardian*, 16-20 January, 1998, p.22.

Cronbach, L.J. *Essentials of psychological testing.* (New York, Harper and Row, 1970).

Cumming, A., Shi, L and So, S. 'Learning to do research on language teaching and learning: Graduate apprenticeships.' *System*, 25 (3), 425-433 (1997).

Cummins, J. 'Linguistic interdependence and the educational development of bilingual children.' *Review of Educational Research*, 49, 222-51 (1979).

Cummins, J. 'The cross-lingual dimensions of language proficiency: Implications for bilingual education and the optimal age issue.' *TESOL Quarterly*, 14 (2), 175-87 (1980).

Cummins, J. 'Language proficiency and academic achievement', in Oller, J.W., Jr. (ed). *Issues in language testing research.* (Rowley, Massachusetts, Newbury Publishers, 1983).

Cummins, J. 'Wanted: A theoretical framework for relating language proficiency to academic achievement among bilingual students', in Rivera, C. (ed.). *Language proficiency and academic achievement.* (Multilingual Matters 10. Clevedon, Multilingual Matters Ltd., 1984).

Cummins, J. 'Interdependence of first- and second-language proficiency in bilingual children', in Bialystok, E. *(ed.). Language processing in bilingual children.* (Cambridge, Cambridge University Press, 1991).

Cziko, G.A. Psychometric and edumetric approaches to testing: Implications and applications. *Applied Linguistics*, 2 (1), 27-44 (1981).

Cziko, G.A. 'Improving the psychometric, criterion-referenced, and practical qualities of integrative testing.' *TESOL Quarterly*, 16 (3), 367-379 (1982).

Cziko, G.A. 'Some problems with empirically-based models of communicative competence.' *Applied Linguistics*, 5 (1), 23-37 (1984).

Dandonoli, P. 1987. 'ACTFL's current research in proficiency testing', in Byrnes, H. and Canale, M. (eds.). *Defining and developing proficiency: Guidelines, implementations and concepts.* (Lincolnwood (Chicago), Illinois, National Textbook Company. In conjunction with the American Council on the Teaching of Foreign Languages, 1987).

Darnell, D.K. *The development of an English language proficiency test of foreign students using a clozenthropy procedure: Final Report.* (Boulder, University of Colorado, 1968).

Davidson, F. 'Norms appropriacy of achievement tests: Spanish-speaking children and English children's norms.' *Language Testing*, 11, 83-95 (1994).

Davies, A. *Language testing symposium: A psycholinguistic approach.* (London, Oxford University Press, 1968).

Davies, A. 'The construction of language tests', in Allen, J.P.B. and Davies, A. (eds.). *The Edinburgh course in applied linguistics, Vol. 4: Testing and experimental methods.* (London, Oxford University Press, 1977).

Davies, A. 'Language testing; survey article.' *Language teaching and linguistics abstracts*, 23 (4), part 2, 215-231 (1978).

Davies, A. *Principles of language testing.* (Oxford, Basil Blackwell Ltd., 1990).

Davies, A. *The native speaker in applied linguistics.* (Edinburgh, Edinburgh University Press, 1991).

Davies, A. 'Proficiency or the native speaker: What are we trying to achieve in ELT?', in Cook, G. and Seidlhofer, B. *Principle and practice in applied linguistics: Studies in honour of H.G. Widdowson.* (Oxford, Oxford University Press, 1995).

De Villiers, A. 'Disadvantaged students: Analysing the zone of proximal development.' *South African Journal of Higher Education*, 10, 135-139 (1996).

Diller, K.C. (ed.). *Individual differences and universals in language learning aptitude.* (Rowley, Massachusetts, Newbury House, 1981).

Docking, R. 'ompetency-based curricula - the big picture.' *Prospect*, 9 (2), 15 (1994).

Donato, R. and McCormick, D. 'A sociocultural perspective on language learning strategies: The role of mediation.' *The Modern Language Journal*, 78 (4), 453-464 (1994).

Douglas, D. 'Developments in language testing.' *Annual Review of Applied Linguistics*, 15,167-187 (1995).

Dreyer, C. 'Teacher-student style wars in South Africa: The silent battle.' *System*, 26, 115-126 (1995).

Dreyer, C. *Testing: The reason why pupils fail.* National Association of Educators of Teachers of English (NAETE) conference " Training teachers for the South African context, Potchefstroom College of Education, September 17-18, 1998.

Du Toit, A. and Orr, M. *Achiever's Handbook.* (Johannesburg, Southern Book Publishers, 1989).

Ebel, R.L. 'Must all tests be valid?' *American Psychologist*, 16, 640-647 (1961).

Ebel, R.L. and Frisbie, D.A. *Essentials of educational measurement.* (5th ed.) (Englewood Cliffs, New Jersey, Prentice Hall, 1991).

Edidin, A. 'Eternal verities: timeless truth, ahistorical standards, and the one true story.' *American Philosophical quarterly*, 34 (2), 259-271 (1997).

Educamus. *Editorial: Internal promotions*, 36 (9), 3 (1990).

Engelbrecht, S. and Schuring, G. 'The NQF: Challenges in the language field, in HSRC.' *Language assessment and the National Qualifications Framework.* (Pretoria, Human Science Research Council Publishers, 1996).

Entwhistle, W.J. *Aspects of language.* (London, Faber and Faber, 1953).

Eysenck, H.J. Réponse à quelques réflexions naïves sur l'interprétation des coéfficients de corrélation. *Revue de Psychologie Appliquée*, 34 (2), 111-114 (1983).

Farhady, H. 'The disjunctive fallacy between discrete-point tests and integrative tests', in Oller, J.W., Jr. (ed). *Issues in language testing research.* (Rowley, Massachusetts, Newbury Publishers, 1983).

Farhady, H. New directions for ESL proficiency testing, in Oller, J.W., Jr. (ed). *Issues in language testing research.* (Rowley, Massachusetts, Newbury Publishers, 1983a).

Forrest, F. *A language curriculum framework for compulsory general education.* (Johannesburg, Centre for Educational Policy Development, 1992).

Fotos, S. 'The cloze test as an integrative measure of EFL proficiency: A substitute for essays on college entrance examinations.' *Language Learning*, 41 (3), 313-336 (1991).

French, E. and Rensburg, I. 'Introductory comments: Language assessment and the NQF', in HSRC. *Language assessment and the National Qualifications Framework.* (Pretoria, HSRC Publishers, 1996).

Froome, S. *Why Tommy isn't learning.* (London, Tom Stacey, 1970).

Gamaroff, R. *Native Language Transfer in Tswana Speaker's English.* Unpublished MA thesis. Potchefstroom: Potchefstroom University for Christian Higher Education, 1986.

Gamaroff, R. *The relation between culture-fair tests and second language Proficiency and achievement.* Paper presented at the South African Association of Language Teacher (SAALT) conference, University of Pretoria, July, 1991.

Gamaroff, R. 'Affirmative action and academic merit.' *Forum*, 1 (1) (1995a) (Journal of the World Council of Curriculum Instruction, Region 2, Africa South of the Sahara, Lagos.

Gamaroff, R. 'Deep Language, Intelligence and Language Proficiency in (Academic) Learning.' *Proceedings of the Linguistic Society of South Africa conference.* (University of Port Elizabeth, 1995b).

Gamaroff, R. 'Solutions to academic failure: The cognitive and cultural realities of English as the medium of instruction among black ESL learners.' *Per Linguam*, 11 (2), 15-33 (1995c).

Gamaroff, R. 'Critical language study as a solution to academic failure: The cognitive realities.' *Proceedings of the South African Academic Development (SAAAD) conference*, Free State Technikon, Bloemfontein, 29 November - 2 December, 1995d.

Gamaroff, R. 'Is the (unreal) tail wagging the (real) dog?: Understanding the construct of language proficiency.' *Per Linguam*, 12 (1), 48-58 (1996a).

Gamaroff, R. 'Abilities, access and that bell curve.' Grewar, A. (ed.). *Proceedings of the South African Association of Academic Development "Towards meaningful access to tertiary education".* (Alice, Academic Development Centre, Fort Hare, 1996b).

Gamaroff, R. *Workshop on quantitative measurement in language testing.* National Association of Educators of Teachers of English (NAETE) Conference, East London Teacher's Centre, September, 1996c.

Gamaroff, R. 'Paradigm lost, paradigm regained: Statistics in language testing.' *Journal of Language Teaching*, 31 (2), 131-139 (1997a).

Gamaroff, R. 'Language as a deep semiotic system and fluid intelligence in language proficiency.' *South African Journal of Linguistics*, 15 (1), 11-17 (1997b).

Gamaroff, R. Can the deconstructive tour (surprisingly) translate us anywhere? *Journal of Literary Studies*, 13 (3/4), 397-415 (1997c).

Gamaroff, R. 'Cloze tests as predictors of global language proficiency: A statistical analysis.' *South African Journal of Linguistics*, 16 (1), 7-15 (1998a).

Gamaroff, R. 'Language, content and skills in the testing of English for academic purposes.' *South African Journal of Higher Education*, 12 (1), 109-116 (1998b).

Gamaroff, R. *Interrater reliability, the bug of all bears: Report on the 1996 NAETE Workshop on quantitative measurement in language testing*. National Association of Educators of Teachers of English (NAETE) conference " Training teachers for the South African context, Potchefstroom College of Education, September 17-18, 1998c.

Gamaroff, R. *Activity theory, mediation and intelligence in learning.* Tenth World Congress of the Comparative Education Society (WCCES), Cape Town, July 12-17, 1998d.

Gamaroff, R. *Psychometrics and reductionism in language assessment.* Paper presented at the SAAAD/SAARDHE conference "Capacity-building for quality teaching and learning in further and higher education", University of Bloemfontein, 22-24 September, 1998e.

Gamaroff, R. 'The dictation test as a measure of communicative language proficiency.' *International Review of Applied Linguistics* (Forthcoming).

Gamaroff. R. (In Press). Rater reliability in language assessment: the bug of all bears. *System*, 28, 31-53.

Gamble, J. 'Drawing the issues together', in HSRC. *Language assessment and the National Qualifications Framework.* (Pretoria, Human Science Research Council Publishers, 1996).

Gardner, R.C. and Tremblay, P.F. 'On motivation: measurement and conceptual considerations.' *The Modern Language Journal*, 78 (4), 524-527 (1994).

Gee. J. *Social linguistics and literacies: Ideology in discourses.* (London, The Falmer Press, 1991).

Geyer, J.R. *Cloze Procedure as a predictor of comprehension in secondary social studies materials.* (Olympia, Washington, State Board for Community College Education, 1968).

Gipps, C. *Beyond testing: Towards a theory of educational assessment.* (London, Falmer Press, 1994).

Givon, T. *Understanding grammar.* (London, Academic Press, 1979).

Goodman, K.S. 'Analysis of oral reading miscues: Applied psycholinguistics.' *Reading Research Quarterly*, 5, 9-30 (1969).

Grotjahn, R. 'On the methodological basis of introspective methods', in Faerch, C. and Kasper, G. (eds.). *Introspection in second language research.* (Clevedon, Avon, Multilingual matters, 1987).

Gue, L. and Holdaway. 'English proficiency tests as predictors of success in graduate studies in education.' *Language Learning*, 23, 89-103 (1973).

Gultig, J., Lubisi, C., Parker, B. and Wedekind, V. *Understanding outcomes-based education: Teaching and assessment in South Africa.* (Oxford, Oxford University Press, 1998).

Hale, G.A., Stansfield, C.W. and Duran, R.P. *TESOL Research Report 16.* (Princeton, New Jersey, Educational Testing Service, 1984).

Halliday, M.A.K. *Learning how to mean.* (London, Arnold, 1975).

Halliday, M.A.K. *Spoken and written language.* (Oxford, Oxford University Press, 1989).

Hamayan, E.V. 'Approaches to alternative assessment.' *Annual Review of Applied Linguistics*, 212-226 (1995).

Hanania, E. and Shikhani, M. 'Interrelationships among three tests of language proficiency: Standardized ESL, cloze and writing.' TESOL Quarterly, 20, 97-109 (1986).

Harris, R. *The language myth.* (London, Duckworth, 1981).

Harrison, A. 'Communicative testing: Jam tomorrow?', in Hughes, A. and Porter, D. (eds.). *Current developments in language testing.* (London, Academic Press, 1983).

Hartog, P. and Rhodes, E.C. *The marks of examiners.* (New York, Macmillan, 1936).

Hartshorne, K. 'Language policy in African Education in South Africa, 1910-1985', in Young, D. (ed.). *Bridging the gap.* (Cape Town, Maskew Miller, 1987).

Hatch, E. *Second language acquisition: A book of readings.* (Rowley, Massachusetts, Newbury House, 1983).

Haussmann, N.C. *The testing of English mother-tongue competence by means of a multiple-choice test: An applied linguistics perspective.* Doctoral thesis. (Rand Afrikaans University, Johannesburg, 1992).

Henning, A. *A guide to language testing.* (Rowley, Massachusetts, Newbury House, 1987).

Henning, G.A., Ghawaby, S.M., Saadalla, W.Z., El-Rifai, M.A., Hannallah, R.K. and Mattar, M. S. 'Comprehensive assessment of language proficiency and achievement among learners of English as a foreign language.' *TESOL Quarterly*, 15 (4), 457-466 (1981).

Herrenstein, J. *IQ in the meritocracy.* (London, Allen Lane, 1973).

Hesse, M. *Revolutions and reconstructions in the philosophy of science.* Bloomington, Indiana University Press, 1980).

Hinofotis, F.B. 'Cloze as an alternative method of ESL placement and proficiency testing', in Oller, J.W. (Jr.) and Perkins, K. (eds.), in *Research in language testing.* (Rowley, Massachusetts, Newbury House, 1980).

Hoover, W. and Gough, P. 'The simple view of reading.' *Reading and writing: An interdisciplinary journal*, 2, 127-160, (1990).

Hopkins, A. 'Review of Robinson, P. Academic writing: Process or product.' The British Council, Modern English Publications. *English Language Teaching Journal*, 44 (1), 239-240 (1990).

Horowitz, R. and Samuels, S. 'Comprehending oral and written language: Critical contrasts for literacy and schooling', in Horowitz, R. and Samuels, S. (eds.). *Comprehending oral and written language.* (San Diego, Academic Press, 1987).

HSRC. *Language teaching: Report of the Committee of University Principals.* (Pretoria, Human Sciences Research Council, 1981).

HSRC. *Ways of seeing the National Qualifications Framework.* (Pretoria, Human Sciences Research Council, 1995).

HSRC. *Language assessment and the National Qualifications Framework.* (Pretoria, Human Science Research Council Publishers, 1996).

Hudson, R. *Word grammar.* (Oxford, Basil Blackwell Inc., 1984).

Hughes, A. 'Conversational cloze as a measure of oral ability.' *English Language Teaching Journal*, 35 (2), 161-168 (1981).

Hughes, A. *Testing for language teachers.* (Cambridge, Cambridge University Press, 1989).

Hughes, T.P. 'The seamless web: Technology, science, etcetera, etcetera.' *Social Studies of Science*, 16 (2), 281-292 (1986).

Huysamen, G.K. 'Parallels between qualitative research and sequentially performed quantitative research.' *South African Journal of Psychology*, 27 (1), 1-8, (1997).

Human, L. and Hofmeyer, K. *Black managers in South African organisations.* (Cape Town, Juta and Co. Ltd., 1985)

Hutchinson, T. and Waters, A. *English for special purposes: A learner-centred approach.* (Cambridge, Cambridge University Press, 1987).

Hyltenstam, K. and Pienemann, M. *Modelling and Assessing second language acquisition.* Clevedon, Avon, Multilingual Matters Ltd., 1985).

Hymes, D. 'On communicative competence', in Pride, J.B. and Holmes, J. (eds.). *Sociolinguistics.* (Harmondsworth, Penguin, 1972).

Ingram, E. *English Language Battery (ELBA).* (Edinburgh, Department of Linguistics, University of Edinburgh, 1964).

Ingram, E. 'Item analysis', in Davies, A. *Language testing symposium.* (London, Oxford University Press, 1968).

Ingram, E. 'English standards for foreign students.' *University of Edinburgh Bulletin*, 9, 4-5 (1973).

Ingram, E. 'Assessing proficiency: An overview on some aspects of testing', in Hyltenstam, K. and Pienemann, M. *Modelling and Assessing second language acquisition.* (Clevedon, Avon, Multilingual Matters Ltd., 1985).

Irvine, P., Atai, P. and Oller, J.W., Jr. 'Cloze, dictation, and the test of English as a foreign language.' *Language Learning*, 24 (2), 245-252 (1974).

Jackson, M .D. and McClelland, J. L. 'Processing determinants of reading speed.' *Journal of Experimental Psychology*, 108, 151-181 (1979).

Jacob, E. 'Studying Puerto Rican children's informal education at home', in Rivera, C. (ed.). *The ethnographical/sociolinguistic approach to language proficiency assessment.* (Clevedon, Avon, Multilingual Matters Ltd., 1983).

Jacobs, B. 'Neurobiological differentiation of primary and secondary language acquisition.' *Studies in Second Language Acquisition*, 33, 247-52 (1988).

Jacobs, B. and Schumann, J. Language acquisition and the neurosciences. *Applied Linguistics*, 13 (3), 282-301 (1992).

James, C. 'Don't shoot my dodo: on the resilience of contrastive and error analysis.' *International Review of Applied Linguistics*, 32 (3), 179-200 (1994).

Jeffery, C.D. 'The case for grammar: Opening it wider.' *South African Journal of Higher Education (SAJHE)*, Special edition (1990).

Johanson, L. 'Shocking low reading levels in many Bop schools.' *Matlhasedi*, 7 (1&2), 27 (1988), Mmabatho, Institute of Education, University of Bophuthatswana.

Johnson, R.K. 'The cloze procedure: New perspectives', in Read, J.A.S. (ed.). *Directions in language testing.* (Singapore, Singapore University Press, 1981).

Johnson, F.C. and Kin-Lin, C.W.L. 'The interdependence of teaching, testing, and instructional materials', in Read, J.A.S. (ed.). *Directions in language testing.* (Singapore, Singapore University Press, 1981).

Kachru, B.B. (ed.). *The Other Tongue: English across cultures.* (Oxford, Pergamon, 1982).

Kaczmarek, C.M. 'Scoring and rating essay tasks', in Oller, J.W. (Jr.) and Perkins, K. (eds.). *Research in language testing.* (Rowley, Massachusetts, Newbury House, 1980).

Kaplan, A. *The conduct of enquiry: Methodology for behavioral science.* (San Francisco, Chandler, 1964).

Kelly, P. 'Lexical ignorance: The main obstacle to listening comprehension with advanced foreign language learners.' *International Review of Applied Linguistics*, 29 (2), 134-149 (1991).

Kinneavy, J. 'Restoring the humanities: The return of rhetoric from exile', in Murphy, J.J. (ed.). *The rhetorical tradition and modern writing*. (New York, Modern Languages Association, 1982).

Kirschenbaum, H., Simon, S.B. and Napier, R.W. *Wad-ja-get?: The grading game in American Education*. (New York City: Hart Publishing Company, 1971).

Krashen, S. 'A response to McClaughlin, "The monitor model": Some methodological considerations.' *Language Learning*, 29 (1), 151-167 (1979).

Krashen, S. *Second language acquisition and second language learning.* (Oxford, Pergamon Press, 1981).

Krashen, S. and Terrell, T. *The natural approach: Language acquisition in the classroom*. (Hayward, California, Alemany Press, 1983).

Kuhn, T. S. *The structure of scientific revolutions*. 2nd Edition. (Chicago. University of Chicago Press, 1970).

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

Lakoff, G. *Women, fire and dangerous things.* (Chicago, University of Chicago Press, 1987).

Lamb, S. M. 'Semiotics of language and culture', in Fawcett, R.P., Halliday, M.A.K., Lamb, S.M. and Makkai, A. *The semiotics of culture and language*, Vol. 2. (London, Frances Pinter, 1984).

Lantolf, J.P. and Frawley, W. 'Proficiency: Understanding the construct.' *Studies in Second Language Acquisition (SLLA)*, 10 (2), 181-195 (1988).

Larsen-Freeman, D. 'From unity to diversity: Twenty-five years of language teaching methodology.' *English Teaching Forum*, 25 (4), 2-10 (1987).

Larsen-Freeman, D. 'Chaos/complexity science and second language acquisition.' *Applied Linguistics*, 18 (2), 141-165 (1997).

Larsen-Freeman, D. and Long, M.H. *An introduction to second language acquisition research*. (New York, Longman, 1990).

Lass, R. *"English" - Talk at Will radio programme on SAfm*, 26 February, 1998.

Laufer, B. 'Why are some words more difficult than others? - Some intralexical factors that affect the learning of words.' *International Review of Applied Linguistics*, 28 (4), 293-307 (1990).

Lazaraton, A. 'Qualitative research in applied linguistics: A progress report.' *TESOL Quarterly*, 29 (3), 455-471 (1995).

Lee, D. *Competing discourses: Perspective and ideology in language*. (London, Longman, 1992).

Lee, W.Y. 'Authenticity revisited.' *English Language Teaching Journal*, 49 (4), 323-328 (1995).

Leech, G. *Semantics*. (Harmondsworth, Middlesex, Penguin, 1981).

Leung, C. Harris, R. and Rampton, B. *The idealised native-speaker, reified ethnicities and classroom realities: Contemporary issues in TESOL*. (London, Thames Valley University, 1997).

Lincoln, Y.S. and Guba, E.G. *Naturalistic enquiry*. (Newbury Park, California, Sage Publications, 1985).

Loevinger, J. 'Objective tests as instruments of psychological theory', in Jackson, D.N. and Messick, S. *Problems in human assessment*. (Huntington, New York, Robert E. Krieger Publishing Company, 1967).

Lorenz, K. 'On the biology of learning', in Kagan, J. *On the biology of learning*. (New York, Harcourt, Brace and World, Inc., 1969).

Lumley, T. and McNamara, T. 'Rater characteristics and rater bias: implications for training', *Language Testing*, 12, 55-21 (1995).

Macdonald, C.A. *English language skills evaluation (A final report of the Threshold Project), Report Soling-17*. (Pretoria, Human Sciences Research Council, 1990a).

Macdonald, C.A. *Crossing the threshold into standard three in black education: The consolidated main report of the Threshold Project*. (Pretoria, Human Sciences Research Council, 1990b).

Macdonald, C.A. *Reasoning skills and the curriculum*. Report Soling, 18. (Pretoria. Human Sciences Research Council, 1990c).

Macken, M. and Slade, D. Assessment: A foundation for effective learning in the school context, in Cope, W. and Kalantzis, M. (eds.). *The powers of literacy: A genre approach to teaching writing*. (London, Falmer Press, 1993).

Mackenzie, D. *Knowing machines: Essays on technical change*. (MIT Press, Cambridge, Massachusetts, 1996).

Maclean, M. 'Using rational cloze for diagnostic testing in L1 and L2 reading.' *TESL Canada Journal*, 2, 53-63 (1984).

MacNamara, T.F. *Measuring second language performance*. (Harlow, Essex, Addison Wesley Longman Limited, 1996).

Magnan, S.S. *Review of Creswell, J.W. 1994. Research design: qualitative and quantitative approaches*. (Thousand Oaks, CA, Sage, 1997).

Makalela, J. *Institutionalized Black South African English (IBSAE)*. National Association of Educators of Teachers of English (NAETE) conference " Training teachers for the South African context, Potchefstroom College of Education, September 17-18, 1998.

Makoni, S. Some of the metaphors about language, in language planning course in South Africa: Boundaries, frontiers and commodification. *Per Linguam*, 11 (1), 25-34 (1995).

Makoni, S. "Language and identity in Southern Africa", in De la Gorgendiere, L, King, K and Vaughan, S. (eds.). *Ethnicity in Africa: Roots, meanings and implications*. (Centre of African Studies, University of Edinburgh, 1996).

Makoni. S. 'In the beginning was the missionaries' word. in Prah, K.K. (ed.). *Between distinction and extinction*. (1998).

Mandler, J.M. *Stories, scripts and scenes: Aspects of schema theory*. (Hillsdale, New Jersey. Lawrence Erlbaum Associates, 1984).

Markham, P. The rational deletion cloze and global comprehension in German. *Language Learning*, 35, 423-430 (1985).

Mascher, D. *The disintegration of an education system based on a non-cognate medium of instruction*. Paper presented at the South African Applied Linguistics Conference, University of the Witwatersrand, July 1991.

Mattar, M.S. 'Comprehensive assessment of language proficiency and achievement among learners of English as a foreign language.' *TESOL Quarterly*, 15 (4), 457-466 (1981).

McArthur, T. *English as a world language, as an African language, and as a South African language*. Paper presented at the English Academy of Southern Africa

conference *"English at the turn of the Millennium"*, Johannesburg College of Education, 14-16 September, 1988.

McArthur, T. The English language or the English languages? in Bolton, W.F. and Crystal, D. *The English language*, 1987.

Mcintyre, S.P. 'Language learning across the curriculum: A possible solution to poor results.' *Popagano*, 9 and10, June (1992).

Mclean, D. 'Language education and the national qualifications framework: An introduction to competency-based education and training', in HSRC. *Language assessment and the National Qualifications Framework.* (Pretoria, Human Science Research Council Publishers, 1996).

Medgyes, P. 'Native or non-native: who's worth more?' *ELT Journal*, 46 (4), 340-348 (1992).

Meisel, J.M. and Clahsen and Pienemann, M. On determining developmental stages in second language acquisition. *Studies in Second Language Acquisition*, 3, 109-135 (1981).

Messick, S. *Validity.* (Princeton, New Jersey, Educational Testing Service, 1987).

Messick, S. *Meaning and values in test validation: The science and ethics of measurement.* (Princeton, New Jersey, Educational Testing Service, 1988).

Messick, S. *Constructs and their vicissitudes in educational and psychological measurement.* (Princeton, New Jersey, Educational Testing Service, 1989a).

Messick, S. 'Validity', in Linn, R.L. (ed.). *Educational measurement.* Third Edition. (New York, American Council on Education, Macmillan Publishing Company, 1989b).

Millar, R. 'The pursuit of the impossible.' *Physics Education*, 23, 156-159 (1988).

Minick, N.J. L.S. *Vygotsky and Soviet activity theory: New perspectives on the relationship between mind and society.* PhD thesis, NorthWestern University, 1985.

Mitchell, G. and Fridjhon, P. 'Matriculation examinations and university performance.' *Bulletin for Academic Staff*, 9 (1), 28-43 (1988). University of Durban-Westville.

Mitchell, R. and Myles, F. *Second language acquisition.* (London, Arnold, 1998).

Moore, R. 'How science educators construe student writing', in Angélil-Carter, S. (ed.). *Access to success: Literacy in academic contexts.* (Cape Town, University of Cape Town Press, 1998).

Moore, R., Paxton, M., Scott, I. and Thesen, L. 'Language development initiatives and their policy contexts', in Angélil-Carter, S. (ed.). *Access to success: Literacy in academic contexts.* (Cape Town, University of Cape Town Press, 1998).

Moss, P. 'Can there be validity without reliability?' *Educational Researcher*, 23 (2), 5-12 (1994).

Morrissey, M. D. 'Toward a grammar of learner's errors.' *International Review of Applied Linguistics*, 21 (3), 193, 207 (1983).

Morrow, K. 'Communicative language testing: Revolution or evolution', in Alderson, J (ed.). *Issues in language testing.* (ELT Documents, The British Council, 1981).

Mullen, K. A. 'Evaluating writing proficiency in ESL', in Oller, J.W. (Jr) and Perkins, K, in *Research in language testing.* (Massachusetts. Newbury House, 1980).

Mulligan, A. C. Evaluating foreign credentials. *College and University*, 41, 307-313 (1966).

Munby, J. *Communicative syllabus design.* (Cambridge, Cambridge University Press, 1978).

Murray, S. 'Current approaches to the teaching of English for academic purposes: A critical appraisal.' *Proceedings (Part 1) of the South African Applied Linguistics Association conference "Our multilingual society: Supporting the reality"*, 28-30 June, University of Port Elizabeth, 1993.

Murray, S. 'Exploring the possibilities of using an outcomes-based approach in English teacher education.' *Southern African Journal of Applied Language Studies*, 5 (2), 21-37 (1997).

Musker, P. and Nomvete, S. 'Standards and levels in language assessment', in HSRC. *Language assessment and the National Qualifications Framework.* (Pretoria, Human Science Research Council Publishers, 1996).

Nicholas, H. and Meisel, J.M. 'Second language acquisition: The state of the art', in Felix, S.W. and Wode, H. (eds.). *Language development at the crossroads: Papers from the Interdisciplinary Conference on Language Acquisition at Passau.* (Tübingen, Gunter Narr Verlag, 1983).

Nunan, D. *Syllabus design.* (London. Oxford University Press, 1988).

Nunan, D. *Research methods in language learning.* (Cambridge University Press. Cambridge, New York, 1992).

Ochsner, R. 'A poetics of second-language learning.' *Language Learning,* 29 (1), 53-80 (1979).

O'Dell, F. *English as a foreign language: Intermediate examinations.* (London, Longman, 1986).

Olivier, A. 'The role of input in language development at tertiary level.' *South African Journal of Education,* 18 (1), 57-60 (1998).

Oller, J.W., Jr. 'Cloze tests of second language proficiency and what they measure.' *Language Learning,* 23 (1), 105-118 (1973).

Oller, J.W., Jr. 'Cloze, discourse, and approximations to English', in Burt, K. and Dulay, H.C. *New directions in second language learning, teaching and bilingual education.* (TESOL, Washington, D.C., 1976).

Oller, J.W., Jr. 'Evidence for a general language proficiency factor: An expectancy grammar.' *Die Neuren Sprachen,* 75, 165-174 (1976a).

Oller, J.W. Jr. 'How important is language proficiency to IQ and other educational tests', in Oller, J.W. (Jr.) and Perkins, K. (eds.). *Language in education: testing the tests.* (Rowley, Massachusetts, Newbury House, 1978).

Oller, J.W., Jr. *Language tests at school.* (London, Longman, 1979).

Oller, J.W., Jr. 'Language as intelligence.' *Language Learning,* 31 (2), 465-492 (1981).

Oller, J.W., Jr. 'A consensus for the 80s', in Oller, J.W., Jr. (ed.). *Issues in language testing research.* (Rowley, Massachusetts, Newbury Publishers, 1983).

Oller, J.W., Jr. '"g", "What is it?', in Hughes, A. and Porter, D. (eds.). *Current developments in language testing.* (London, Academic Press, 1983a).

Oller, J.W., Jr. *Issues in language testing research.* (Rowley, Massachusetts, Newbury Publishers, 1983b).

Oller, J.W., Jr. *Language and experience: Classic pragmaticism.* (Lanham, Maryland, University Press of America, 1989).

Oller, J.W., Jr. *Language and bilingualism: More tests of the tests.* (London and Toronto. Associated University Press, 1991).

Oller, J.W. Jr. 'Adding abstract to formal and content schemata: Results of recent work in Peircean semiotics.' *Applied Linguistics*, 16 (3), 274-306 (1995).

Oller, J.W., Jr. and Conrad, C. 'The cloze technique and ESL proficiency.' *Language Learning*, 21, 183-196 (1971).

Oller, J.W., Jr. and Kahn, F. Is there a global factor of language proficiency?, in Read, J.A.S. *Directions in language testing.* (Singapore, Singapore University Press, 1981).

Oller, J.W., Jr. and Perkins, K. *Language in education: Testing the tests.* (Rowley, Massachusetts, Newbury House, 1978).

Oller, J.W., Jr. and Perkins, K. 'Language proficiency as a source of variance in self-reported affective variables', in Oller, J.W. (Jr.) and Perkins, K. (eds.). *Language in education: testing the tests.* (Rowley, Massachusetts, Newbury House, 1978a).

Oller, J.W., Jr. and Perkins, K. (eds.). *Research in language testing.* (Rowley, Massachusetts, Newbury House, 1980).

Oller, J.W., Jr. and Richard-Amato, P. A. *Methods that work: A smorgasbord of ideas for language teachers.* (Cambridge, Massachusetts, Newbury House, 1983).

Olshtain, E., Shohamy, E., Kemp, J. and Chatow, R. 'Factors predicting success in EFL among culturally different learners.' *Language Learning*, 40 (1), 23-44 (1990).

Omaggio, A.C. *Teaching language in context: Proficiency-orientated instruction.* (Boston, Massachusetts, Henle and Henle, 1986).

O'Malley, J.M. 'The cognitive academic language learning approach (CALLA).' *Journal of Multilingual and multicultural development*, 9 (1&2), 43-60 (1988).

Oskowitz, B. 'Preparing researchers for a qualitative investigation of a particularly sensitive nature: Reflections from the field.' *South African Journal of Psychology*, 27 (2), 83-88 (1997).

Oxford, R.L., Ehrman, M. and Lavine, R.Z. 'Style wars: Teacher- student style conflicts in the language classroom', in Magnan, S.S. (ed.). *Challenges in the 1990s for College Foreign Language Programs.* (Boston, MA, Heinle and Heinle, 1991).

Paikeday, T.M. *The native speaker is dead!* (Toronto, Paikeday Publishing Inc., 1985)

Parlett, M. 'Illuminative evaluation', in Reason, P. and Rowan, J. (eds.). *Human enquiry: A sourcebook of New Paradigm Research.* (Chichester, Wiley, 1981).

Peirce, B. 1989. 'Towards a pedagogy of possibility in the teaching of English internationally: People's English in South Africa.' *TESOL Quarterly*, 23 (3), 401-420 (1989).

Peirce, B. 'On language difference and democracy.' *Language Projects Review*, 6, 21-24 (1991).

Perkins, K. 'On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability', *TESOL Quarterly*, 17 (4), 651-671 (1983).

Perkins, D., Jay, E. and Tishman, S. 'New conceptions of thinking: From ontology to education.' *Educational Psychologist*, 28 (1), 76 (1993).

Phillips, S. 'An ethnographic approach to bilingual language proficiency assessment', in Rivera, C. (ed.). *The ethnographical/sociolinguistic approach to language proficiency assessment.* (Clevedon, Avon, Multilingual Matters Ltd., 1983)

Piaget, J. and Inhelder, B. *The psychology of the child.* (New York, Basic Books, 1969).

Pienaar, P. *Reading for meaning: A pilot survey of (silent) reading standards in Bophuthatswana.* (Mmabatho, Institute of Education, University of Bophuthatswana [North West], 1984).

Pilliner, A.E.G. 'Subjective and objective testing', in Davies, A. (ed.). *Language testing symposium.* (London, Oxford University Press, 1968).

Pilliner, A.E.G. *Experiment in educational research.* (Buckinghamshire, The Open University, 1973).

Pinker, S. *The language instinct.* (London/New York. Penguin Books, 1995).

Politzer, R.L. and McGroarty, M. 'A discrete-point test of communicative competence.' *International Review of Applied Linguistics*, 21 (3), 179-191 (1983).

Popham, W.J. *Modern educational measurement.* (Englewood Cliffs, New Jersey, Prentice-Hall, 1981).

Porter, D. 'Cloze procedure and equivalence.' *Language Learning*, 28 (2), 333-41 (1978).

Porter, D. 'Assessing communicative proficiency: The search for validity', in Johnson, K. and Porter, D. (eds.). *Perspectives of communicative language teaching.* (London, Academic Press, Inc, 1983).

Quine, W. and Ullian, J. *The web of belief.* (New York, Random House, 1970).

Rampton, B.H. 'Displacing the 'native speaker': expertise, affiliation and inheritance.' *English Language Teaching Journal*, 44 (2), 97-101 (1990).

Rea, P. 'Language testing and the communicative language teaching curriculum', in Lee, Y.P. et al. *New directions in language testing.* (Oxford. Institute of English, 1985).

Read, J.A.S. *Directions in language testing.* (Singapore, Singapore University Press, 1981).

Reason, P. and Rowan, J. (eds.). *Human enquiry: A source book of New Paradigm Research.* (Chichester, Wiley, 1981).

Reid, J. and Kitegawa, N. 'A Whole-school Approach to mainstreaming: The Rose Avenue ESL/D project', in Clegg, J. (ed.). 1996. *Mainstreaming ESL: Case studies in integrating students into the mainstream curriculum.* (Clevedon, Multilingual Matters Ltd., 1996).

Rickard, C. '"Racist" school tests outlawed.' *Sunday Times* (South Africa), April 7, 1996, p.4.

Rivera, C. (ed.). *The ethnographical/sociolinguistic approach to language proficiency assessment.* (Clevedon, Avon, Multilingual Matters Ltd., 1983).

Rorty, R. *Philosophy and the mirror of nature.* (Princeton, Princeton University Press, 1980).

Rowntree, D. *Assessing students: How shall we know them.* (London, Harper and Row, Publishers, 1977).

Rowntree, D. *Statistics without tears.* (Harmondsworth, Penguin Books, 1981).

Rumelhart, D. E. *Introduction to human information processing.* (New York, John Wiley and Sons, 1977).

Rushton, J.P. *Race, evolution and behaviour.* (New Brunswick, Transaction Publishers, 1995).

Russell, B. *The analysis of mind.* (New York, 1921).

Rutherford, W.E. *Second language grammar: Learning and teaching.* (London, Longman, 1987).

Sammonds, P. 'Ethical issues and statistical work', in Burgess, R.G. (ed.). *The ethics of educational research.* (New York, Falmer Press, 1989).

Samuel, A.G. 'Phonemic restoration: Insights from a new methodology.' *Journal of Experimental Psychology*, 110, 474-494 (1981).

Santos, T. 'Professors' reactions to the academic writing of nonnative-speaking students.' *TESOL Quarterly*, 22 (1), 69-90 (1988).

Saussure, F. de. *Course in general linguistics.* (London, Fontana, 1916 [1974]).

Savignon, S.J. *Communicative competence: Theory and classroom practice.* (Reading, Mass. Addison-Wesley Publishing Company, 1983).

Savignon, S.J. 'Language, communication, social meaning, and social change: The challenge for teachers', in Alaltis, J.E. *Georgetown University Round Table on Languages and Linguistics.* (Washington, D.C., Georgetown University Press, 1992).

Saville-Troike, M. 'What really matters in second language learning for academic achievement.' *TESOL Quarterly*, 18 (2), 199-219 (1984).

Scriven, M. 'The methodology of educational evaluation', in Tyler, R.W., Gagne, R.M. and Scriven, M. (eds.). *Perspectives of curriculum evaluation.* (Chicago, Rand McNally, 1967).

Sebatane, E.M. *Assessment policy strategies, implementation and impact: A global perspective.* Tenth World Congress of the Comparative Education Society (WCCES), Cape Town, July 12-17, 1988.

Sehlapelo, M and Terre Blanche, M. 'Psychometric testing in South Africa: Views from above and below.' *Psychology in society* (PINS), 21, 49-59 (1996).

Selinker, L. 'Interlanguage.' *International Review of Applied Linguistics*, 10 (3), 219-231 (1972).

Simpson, G. J. 'Are matric marks relevant selection criteria for higher education?', in Blaquière, A. (ed.). Intercultural issues in teaching and learning. *Proceedings of the 1986 South African Association for Research and Development in Higher Education (SAARDE)*, 68-72. (University of Natal, Convening Committee, 1987).

Sinclair, M. 'Reading comprehension research and linguistic pragmatics: Mapping out some unrecognised interdisciplinary common ground.' *Stellenbosch Papers in Linguistics*, 28, 83-108 (1994).

Singh, M. 'Universities: The wave of transformation.' *Centre for Scientific Development (CSD) Bulletin*, 4 (7). (Pretoria, Centre for Scientific Development (CSD), 1992).

Skehan, P. *Individual differences in second language learning*. (London, Arnold, 1989).

Skehan, P. *A cognitive approach to language learning*. (Oxord, Oxford University Press, 1998).

Snow, R.E. ' The training of intellectual aptitude', in Detterman, D.K. and Sternberg, J. (eds.). *How and how much can intelligence be raised*. (Norwood, New Jersey, Ablex Publishing Corporation, 1982).

Spady, W. 'Outcomes-based education: An international perspective', in Gultig, J., Lubisi, C., Parker, B. and Wedekind, V. *Understanding outcomes- based education: Teaching and assessment in South Africa*. (Oxford, Oxford University Press, 1998).

Spolsky, B. 'Approaches to language testing', in Spolsky, B (ed.). *Advances in Language Testing Series, 2*. (Arlington, Virginia. Center for Applied Linguistics, 1978).

Spolsky, B. 'Some ethical questions about language testing', in Klein-Braley, C. and Stevenson, D.K. (eds.). *Practice and problems in language testing*, Vol. 1. (Frankfurt, Verlag Peter D. Lang, 1981).

Spolsky, B. 'The limits of authenticity in language testing.' *Language Testing*, 2, 31-40 (1985).

Spolsky, B. *Conditions for second language learning.* (Oxford, Oxford University Press, 1989).

Spolsky, B. *Measured words.* (Oxford, Oxford University Press, 1995).

Starfield, S. 'Contextualising language and study skills.' *South African Journal of Higher Education*, Special Edition (1990).

Starfield, S. and Kotechka, P. *Language and learning: The Academic Support Programme's intervention at the University of the Witwatersrand*, Paper presented at the South African Applied Linguistics Association Conference, July, 1991.

Stern, H. H. *Fundamental concepts of language teaching.* (Oxford, Oxford University Press, 1983).

Sternberg, R.J. *The psychologist's companion: A guide to scientific writing for students and researchers.* (Cambridge, New York, Cambridge University Press, 1993).

Stevenson, D.K.'Pop validity and performance testing', in Lee, Y., Fok, A., Lord, R. and Low, G. (eds.). *New directions in language testing.* (Oxford, Pergamon, 1985).

Stevenson, D.K. 'Authenticity, validity, and a tea party.' *Language Testing*, 2, 41-47 (1985a).

St. Leger, C. 'Radical steps proposed for education after matric results shock.' *Sunday Times (South Africa)*, December 31, 1995a.

St. Leger, C. 'Depressing statistics from years of turmoil.' *Sunday Times (South Africa)*, December 31, 1995b.

Stoker, D.J. *Statistical tables.* (Pretoria, Academica, 1974).

Strevens, P. *Teaching English as an international language.* (Oxford, Pergamon, 1980).

Stubbs, J. and Tucker, G. 'The cloze test as a measure of English proficiency.' *Modern Language Journal*, 58, 239-241 (1974).

Stump, T.A. 'Cloze and dictation tasks as predictors of intelligence and achievement scores', in Oller, J.W. (Jr.) and Perkins, K. (eds.). *Language in education: testing the tests.* (Rowley, Massachusetts, Newbury House, 1978).

Suenobu, M., Kanzaki, K., Yamane, S. and Young, R. 'Listening comprehension and the process of information acquisition by non-native speakers of English.' *International Review of Applied Linguistics*, 24 (3), 239-248 (1986).

Swain, S. 'Large-scale communicative language testing: A case study', in Lee, Y., Fok, A., Lord, R. and Low, G. (eds.). *New directions in language testing*. (Oxford, Institute of English, 1985).

Taylor, B.P. 'In search of real reality.' *TESOL Quarterly*, 16 (1), 29-43 (1982).

Taylor, J.R. *Linguistic categorization*. (Oxford, Oxford University Press, 1989).

Taylor, W. 'Cloze procedure: A new tool for measuring readability.' *Journalism Quarterly*, 30, 414-438 (1953).

Terre Blanche, M. 'Crash.' *South African Journal of Psychology*, 27 (2), 59-63 (1997).

Tooby, J. and Cosmides, L. 'Psychological foundations of culture', in Barkow, J.H., Cosmides, L., and Tooby, J. (eds.). *The adapted mind: Evolutionary psychology and the generation of culture*. (New York, Oxford University Press, 1992).

Tönnes-Schnier, F. and Scheibner-Herzig, G. 'Measuring communicative effectiveness through dictation.' *International Review of Applied Linguistics*, 26 (1), 35-43 (1988).

Tremblay, R.F. and Gardner, R.C. 'Expanding the motivation construct of language learning.' *The Modern Language Journal*, 79 (4), 505-518 (1995).

Tucker, S.A. and Dempsey, J.V. A semiotic model for program evaluation. *The American Journal of Semiotics*, 8 (4), 73-103 (1991).

Tyler, S.A. 'Post-modern ethnography: from document of the occult to occult document', in Clifford, J. and Marcus, G.E. (eds.). *Writing culture. The poetics and politics of ethnography*. (Berkeley, University of California, 1986).

Upshur, J.A. 'English language tests and predictions of academic success', in Wigglesworth, D.C. (ed.). *Selected conference papers of the Association of Teachers of English as a Second Language*. Los Altos, California, National Association for foreign Student Affairs (NAFSA) Studies and Papers, English Language Series 13, 85-93 (1967).

Upshur, J.A. 'Functional proficiency theory and a research role for language tests', in Brière, E. and Hinofotis, F.B. (eds.). *Concepts in language testing.* (Washington, TESOL, 1979).

Ur, P. *A course in language teaching: practice and theory.* (Cambridge, Cambridge University Press, 1996).

Valette, R.L. *Modern language testing: A handbook.* (New York, Harcourt Brace and World, 1969).

Van der Walt, J.L. 'Some characteristics of communicative tests.' *National Association of Educators and Teachers of English*, 9, 47-51 (1994).

Van der Walt, J. *The implications for language testing of IBSA [Institutionalised Black South African English].* National Association of Educators of Teachers of English (NAETE) conference " Training teachers for the South African context, Potchefstroom College of Education, September 17-18, 1998.

Van Lier, L. *The classroom and the language learner.* (New York, Longman, 1988).

Vellutino, F., Scanlon, D., Small, S. and Tanzman, M. 'The linguistic bases of reading ability: Converting written to oral language.' *Text*, 11, 99-133 (1991).

Vollmer, H.J. 'Why are we interested in general language proficiency?', in Alderson, J.C. and Hughes, A. *Issues in language testing: ELT Documents III.* (The British Council, 1981).

Vollmer, H.J. 'The structure of foreign language competence', in Hughes, A. and Porter, D. (eds.). *Current developments in language testing.* (London, Academic Press, 1983).

Vygotsky, L.S. *Mind in society: The development of higher psychological processes.* (Edited by Michael Cole, Vera John-Steiner, Sylvia Scribner and Ellen Souberman). (Cambridge, Massachusetts, Harvard University Press, 1978).

Vygotsky, L. and Luria, A. Tool and symbol in child development, in Van der Veer, R. and Valsiner, J. *The Vygotsky reader.* (Oxford, Blackwell, 1994).

Wald, B. 'A sociolinguistic perspective on Cummins' current framework for relating language proficiency to academic achievement', in Rivera, C. *Language proficiency and academic achievement.* (Multilingual Matters 10. Clevedon. Multilingual Matters Ltd., 1984).

Wallace, B. and Adams, H. 'A framework for language.' *Bua!* (National Language Project), 10 (1), 16-17 (1995).

Weaver, W.W. and Kingston, A.J. 'A factor analysis of the Cloze procedure and other measures of reading and language ability.' *Journal of Communication*, 13, 252-261 (1963).

Weintraub, S. 'The cloze procedure.' *The Reading Teacher*, 6, 21, 567, 569 (1968), 571, 607.

Weir, C.J. *Communicative language testing*. (Exeter, University of Exeter, 1988).

Weir, C.J. *Understanding and developing language tests.* (London, Prentice Hall, 1993).

Westbrook, L. and Bergquist-Moody, S. 'A Whole-language approach to mainstreaming', in Clegg, J. (ed.). *Mainstreaming ESL: Case studies in integrating students into the mainstream curriculum*. (Clevedon, Multilingual Matters Ltd., 1996).

Widdowson, H.G. 'The teaching of English through science', in Dakin, J., Tiffin, B. and Widdowson, H.G. *Language in education*. (London, Oxford University Press, 1968).

Widdowson, H.G. *Explorations in applied linguistics*. (Oxford, Oxford University Press, 1979).

Widdowson, H.G. 'New starts and different kinds of failure', in Freedman, A., Pringle, I. and Yalden, J. (eds.). *Learning to write: First language/Second language*. (London, Longman, 1983).

Widdowson, H.G. 'Knowledge of language and ability of use.' *Applied Linguistics*, 10 (2), 128-137 (1989).

Widdowson, H.G. *Aspects of language teaching*. (Oxford, Oxford University Press, 1990).

Widdowson, H.G. 'Communication, community and the problem of appropriate use', in Alatis, J.E. *Georgetown University Round Table on Languages and Linguistics*. (Washington, D.C., Georgetown University Press, 1992).

Widdowson, H.G. 'Skills, abilities, and contexts of reality.' *Annual Review of Applied Linguistics*, 18, 323-333 (1998).

Wilhelm, K.H. 'Use of an expert system to predict language learning success.' *System*, 25 (3), 317-334 (1997).

Wilkins, D.A. 'Notional syllabuses revisited.' *Applied Linguistics*, 2 (1), 83-89 (1981).

Windelband, W. *A history of philosophy, Vol. I.* (New York, Harper and Brothers, Publishers, 1958).

Wood, R. 'The agenda for educational measurement', in Nuttal, D. (ed.). *Assessing educational achievement.* (London, Falmer Press, 1988).

Woods, A., Fletcher, P. and Hughes, A. *Statistics in language Studies.* (London, Cambridge University Press, 1986).

Wong, S. 'Curriculum transformation: A psycholinguistics course for prospective teachers of ESOL K-12', in Alatis, E. Straehle, C.A., Gellenberger, B. and Ronkin, M. (eds.). *Georgetown University Round Table on Languages and Linguistics.* (Washington, D.C., Georgetown University Press, 1995).

Yeld, N. *Communicative language testing.* Report on British Council Course 559 offered at Lancaster University from 8 September to 20 September 1985. (Cape Town. University of Cape Town, 1986).

Yeld, N. 'Communicative language testing and validity.' *Journal of Language Teaching*, 21 (3), 69-82 (1987).

Young, D. 'A priority in language education: Language across the curriculum in black education', in Young, D. and Burns, R. (eds.). *Education at the crossroads.* (Rondebosch, University of Cape Town, 1987).

Young, D. 'English for what and for whom and when?' *Language Projects Review,* 3 (2), 8 (1988).

Young, D. 'The role and status of the first language in education in a multilingual society', in Heugh, K., Siegruhn, A., Pluddemann, S. (eds.). *Multilingual education in South Africa.* (Johannesburg, Heinemann, 1995).

Zaaiman, H. *Selecting students for mathematics and science: The challenge facing higher education in South Africa.* (Pretoria, Human Sciences Research Council Publishers, 1998).

Zughoul, M.R. and Kambal, O. K. 'Objective evaluation of EFL composition.' *International Review of Applied linguistics*, 21 (2), 87-103 (1983).

Appendix

Tables A and B show the scores and judgements of individual raters on Protocol 1 and Protocol 2, respectively. These tables have been divided into EL1 and EL2 sections, then sorted within the EL1 and EL2 sections on scores in ascending order so that the same scores appear together, which makes it easy to compare similar scores with their corresponding judgements. If the language in the L1 column is English then this is an EL1 speaker.

Table A

Scores and Judgements of Raters on Protocol 1

Raters	First language	Score	Raters' Judgements	Content	Grammar	Spelling
Protocol 1 - English First Language Raters (EL1)						
D4	English	3	Many spelling errors.			Negative
E3	English	4	Can understand in spite of errors. Facts given not clear and logical.	Negative		
F1	English	4	Some confusion about the folding procedure.	Negative		
F2	English and Afrikaans	4	Folding instructions confusing.	Negative		
F3	English	4	Imprecise instructions on how to cover a book.	Negative		
A2	English	5	Well visualised but inconsistent spelling.	Positive		Negative
A3	English	5	Satisfactory, but poor spelling and grammar.		Negative	Negative
B2	English	5	Logical structure but a spelling problem.	Positive		Negative
B3	English and Xhosa	5	Coherent and cohesive. Some spelling mistakes.	Positive		Negative
C2	English	5	Explicit and cohesive. Surface errors not affect meaning.	Positive	Negative	Negative
C4	English	5	Topic deviates. Content sequence satisfactory. Major grammatical. Errors detracts from coherence.	Negative	Negative	
D1	English	5	Only one great fault is spelling, quite distracting.	Positive	Positive	Negative
F4	English	5	Not enough details. Inconsistency of spelling.	Negative		Negative
B1	English	6	Lucid but main problem is spelling.	Positive		Negative
D2	English	6	Logically structured, spelling errors main problem.	Positive		Negative
E2	English	7	Clear logical, no serious grammatical errors, only spelling errors.	Positive	Positive	Negative
Protocol 1 - English Second Language Raters (EL2)						
C1	Sotho	3	Meaningless, cloudy.	Negative	Negative	
E4	Tswana	3	The student is relevant but the text is full of grammatical errors and inconsistent.	Positive	Negative	
A1	Ewe	4	Grammatical errors but adequate description.	Positive	Negative	
A4	Venda	4	Grammatical accuracy is a problem.		Negative	
E1	Xhosa	5	No comment			
B4	Zulu & Venda	5	Mechanics a problem, but understandable	Positive		Negative
C3	Xhosa	6	Topic not relevant. Any book is covered in this way. Content accurate. A few gr. errors but meaning not affected. Spelling inconsistent.	Negative	Negative	Negative
D3	Xhosa	6	Has good command of language. **This learner belongs to an "elite group".**		Positive	Positive

Table B

Scores and Judgements of Raters on Protocol 2

Rater	L1	Score	Raters' Judgements	Content	Grammar	Spelling
Protocol 2 - English First Language Raters (EL1)						
A2	English	5	Logical approach.	Positive		
A3	English	4	On topic but content confusing. Gr. inaccurate.	Negative	Negative	
B1	English	4	Muddled. Poor syntax and idiomatic usage.	Negative	Negative	
B2	English	4	Repetitive. Simple vocab, poor syntax.	Negative	Negative	
B3	English & Xhosa	4	Poor punctuation. Language is poor.		Negative	
C2	English	4	Lack of cohesion makes writing less explicit despite limited surface errors. Content interpretable.	Positive	Negative	
C4	English	5	Topic relevant. Content: missing propositions, little connection. Reasonable grammatical accuracy.	Negative	Positive	
D1	English	5	Less accurate. Difficult to understand. "cut into strips".	Negative		
D2	English	5	Cohesion bad, e.g. "cut it into strips", but fairly coherent, not too many errors.	Positive	Negative	
D4	English	4	Topic relevant, content meaningful and gr. better than 1.	Positive	Positive	
E2	English	5	Unclear explanation. Cut what into strips? **General reluctance to give extremely high or low marks.**	Negative		
E3	English	5	Can understand in spite of errors. Unclear and illogical.	Negative		
F1	English	5	Fairly clear, except for "cut it into strips".	Positive		
F2	English & Afrikaans	5	Left out important details such as opening the book; "cut into strips" is confusing.			
F3	English	4	Neither gives precise enough instructions to enable s.o. who does not know how to cover a book to cover one.			
F4	English	6	Quite good in terms of "understanding ability". Grammar not good.			
English Second Language Raters (EL2)						
A1	Ewe	5	Content fine.	Positive		
C1	Sotho	4	Errors affect meaning.		Negative	
E4	Tswana	8	-			
A4	Venda	3	Grammatical inaccuracy.		Negative	
C3	Xhosa	4	Topic not relevant. Any book is covered in this way. Content accurate. A few gr. Errors but meaning ok.	Positive	Negative	
D3	Xhosa	4	Very limited vocab. "Perhaps he is from the low income group."	Negative		
E1	Xhosa	5	Does not state clearly in opening sentence what he/she intends to do.	Negative		
B4	Zulu & Venda	4	Mechanics blocks meaning, imprecise but understandable. Wrong sequence.	Negative	Negative	Negative

Questionnaire on moderation workshops

The relevant questions (J to L) of the questionnaire and Table C containing the corresponding data are presented below:

J. (i) Do you find that one or more of your colleagues in the workplace evaluate(s) pupil/student protocols in such a way that your respective allocation of scores is significantly different? Yes...., No..... (ii) If, yes, Do you find this to be a serious problem? Yes...., No.....

K. Do you have moderation workshops/meetings with your colleagues? 1. Never.....; 2. Once annually.....; 3. More than once annually...., 3. More than twice annually.....

L. If your answer in the previous item is not "never", have you found that these moderation workshops/meetings at your institution have ironed out the assessment disparities between you and your colleagues? 1. There has been a great improvement.....; 2. A fair improvement.....; 3. A negligible improvement.....; 4. No noticeable improvement.....;

5. They're a waste of time.....

Table C

Raters Opinions on Moderation Workshops

			Question J(i)	*Question J(ii)*	*Question K*	*Question L*
	Place of Study	**Experience (years)**	**Significant difference between colleagues**	**Find L(i) to be a Problem**	**Moderation workshops**	**Is there any Improvement**
A1	University	20	No[1]		1	-
A2	Wits, Univ. of South Africa	12	No		Never	-
A3	University	16	Yes	Yes	1	2
A4	University	7	No		Never	-
B1	Natal	18	No		2	1
B2	Rhodes University	7	No		2+	2
B3	Univ. of Transkei	8	No		1+	3
B4	University	9	-		Never	-
C1	Lesotho	6	No		1	1
C2	Univ. of South Africa	12	No		2+	1
C3	Univ. of Fort Hare	18	No		Never	-
C4	Rhodes	10	No		2+	-
D1	Fort Hare	20	Yes	No[2]	1+	2
D2	South Africa & UK	7	Yes	No	2+	2
D3	Fort Hare	38	-		Never	-
D4	Venda	4	Yes	No	Never	-
E1	University	7	No	-	2+	2
E2	Potchefstroom	12	No	No	1	-
E3	College	4	No	-	1	-
E4	Rhodes	10	No	No	1	-
F1	Exeter	28	Yes	Yes	Never	-
F2	OFS, UCT, Cambridge	5	Yes	Yes	1	4
F3	Lancaster	20	Yes	Yes	2+	2
F4	Bangalore, UK	11	-	-	2+	-

1 *If the answer to J (i) is No, then no answer is required for J (ii).*

2 *It's odd that this rater and the next two would have no problem if they discovered significant differences in the ratings they gave the same student.*