"OLD PARADIGM" LANGUAGE PROFICIENCY TESTS AS PREDICTORS OF LONG-TERM ACADEMIC ACHIEVEMENT

Raphael Gamaroff

A dictation test, a cloze test and an error recognition test were given to a group of Grade 7 pupils at Mmabatho High School from different linguistic and cultural backgrounds, about half of whom had originated from former DET schools. These tests were subsequently examined as possible predictors of long-term academic achievement. The best predictor was the dictation test, followed by the error recognition test. The cloze test was not a strong predictor of academic potential of long-term success. The research also showed that many learners from former DET schools did not perform well in a school that uses a Joint Matriculation Board syllabus or its equivalent. This has implications for current policy in education where the distinction between L1 and L2 is eschewed in favour of "multicultural settings".

'n Dikteertoets, 'n clozetoets en 'n foutherkenningstoets is aan 'n groep graad 7-leerlinge van die Mbabatho Hoërskool gegee. Hierdie leerlinge is van verskillende taal- en kultuuragtergronde; bykans die helfte van hulle is afkomstig van die eertydse DOO-skole. Hierdie toetse is gevolglik ontleed as moontlike aanduiers van langtermyn-prestasie. Die beste aanduier was die dikteertoets, met die foutherkenningstoets in die tweede plek. Die clozetoets was nie 'n sterk aanduier van akademiese vermoë of langtermyn-sukses nie. Die navorsing het verder getoon dat baie leerders van voormalige DOO-skole nie so goed gevaar het in 'n skool wat gebruik maak van 'n leerplan van die Gesamentlike Matrikulasieraad of ekwivalent daarvan nie. Dit hou implikasies in vir die huidige opvoedingsbeleid waar die onderskeid tussen T1 en T2 ter syde gestel word ten gunste van "multikulturele omgewings".

1.   INTRODUCTION

This article examines the viability and practicality of using "old paradigm" English proficiency tests to predict long-term academic achievement. The tests used are "indirect" tests: error recognition, cloze and dictation tests.

The educational context of the study is Mmabatho High School in the North West Province of South Africa where I spent January 1980 to April 1987 as a teacher and researcher in ESL (Gamaroff, 1986). English is the medium of instruction at the School for all courses except the non-English language courses.

The School's main criterion for admission was previous academic achievement indicated on school reports. Pupils at the School wrote the JMB matriculation examination since it was regarded as a better predictor of tertiary academic achievement than the Senior Certificate examinations offered in the former Department of Education and Training (DET) and the

examinations offered by the other education departments in South Africa (Mitchell & Fridjhon, 1988; Peirce, 1990). On the demise of the JMB examination, Mmabatho High School changed to the Independent Examinations Board (IEB) syllabus in 1993.

Since its inception in 1980 until 1993 (the period which I am concerned with), the School had a low pass rate between Grade 7 and Grade 9, especially with regard to disadvantaged pupils. The description disadvantaged is used to refer to those who had suffered educational deprivation as a result of adverse social, economic or political conditions. It is well known that many pupils in former DET schools were disadvantaged.

In January 1987 I began the longitudinal investigation reported in this article, which extended over the period January 1987 to December 1993. I arranged with the Grade 7 teachers to present a battery of English proficiency tests to entrants to Grade 7 (Std 5) in January 1987 to assess the level of English proficiency of these entrants.

I left the School in April 1987 and subsequently undertook a longitudinal study where the tests were used to predict academic achievement from Grade 7 to Grade 12 (Std 10).

2. 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 is a major cause of academic failure.

- High language proficiency is a major cause of academic success.

Academic language proficiency is perceived as closely related to students' "ability to pursue regular academic courses" (Gue and Holdaway, 1973:92), but the ability to *pursue* regular academic study does not mean that students will *succeed* in academic study: language proficiency tests can assess whether a student has the language skills needed to pursue an academic programme, but they are not necessarily good predictors of academic success. Academic success depends on much more than proficiency in the medium of instruction.

In fact, 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 (Hale *et al.*, 1984:115,177). Correlations with GPA [Grade Point Average] were lower for the second semester than the first [which] may be due to improvement in students' English skills during the first semester, which tends to reduce the role of language ability in determining academic success" (Hale et al., 1984:178f). The inference to be drawn is that "English skills" (as measured by TOEFL [Teaching of English as a foreign language]) 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.

2

However, although language proficiency tests appear to be poor predictors of academic achievement, these correlations do assess, according to Hale *et al.* (1984:198) the minimum level of language proficiency required for academic success. These language proficiency tests are valid predictors of academic failure.

The research reported here, however, shows that a very high score (70%-90% range) in the English proficiency tests that I used, is a good predictor of academic success.

Here too an important consideration was the minimum level that has just been mentioned. In trying to set that minimum level, one is concerned with what the individual can do in terms of established criteria. In my tests I was concerned with long-term prediction of success rather than whether pupils had the potential to achieve academic success.

3.   THE ENGLISH PROFICIENCY TEST BATTERY

(See the appendix for full details of the tests).

The original English proficiency test battery consisted of:

-       Two cloze tests taken from Pienaar (1984:59,61), "*Reading for Meaning*", each consisting of ten items.

-       Two dictation tests. These were two restored cloze tests from Pienaar's "*Reading for meaning*" (Pienaar, 1984:58,62)

-       Two essay tests (devised by myself).

-       One error recognition test (Bloor *et al.,* Forrest *et al.,* 1970).

-       One mixed grammar test (Bloor *et al.,* 1970:35-40). These last two tests consisted of multiple-choice items and their scores were adjusted to reduce the effect as a factor.

For this article I have selected three of these tests:

(1)   Pienaar's (1984) cloze tests where chosen because of my interest in researching the use of these cloze tests to predict academic achievement (Gamaroff, forthcoming a).

(2)   The dictation tests were chosen, because they are two "unmutilated" passages of Pienaar's (1984) cloze tests, which made it possible for me to compare the predictive validity of the cloze tests with that of the dictation tests. The cloze passages are different from the passages used for the dictation tests, but they all both belong to the same level - Pienaar's Step 2 - and are regarded as parallel passages (insofar as it is possible to find parallel passages). 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 from Step 2 instead of four because I used the unmutilated form of the other two Step 2 passages for my dictation tests. I used Pienaar's Form B and Form D passages of Step 2 for the cloze tests and his Form A and Form C passages of

3

Step 2 for the dictation tests. Step 2 was intended for Grade 7 to 9 L2 learners and for Grade 5 and 6 L1 learners. Thus Step 2 should be much easier for L1 learners than for L2 learners. For Pienaar the labels "L1" and "L2" are equivalent to mother-tongue speaker and non-mother-tongue speaker, respectively. I use these labels differently, as I explain shortly in the description of the sample of subjects.

(3)   The error recognition test was chosen for two reasons. First, unlike the other two test methods, it is a multiple-choice "discrete-point" test, and, second, because it has been rejected, unfortunately, in the past as a predictor because it lacked acceptability i.e. face validity, in spite of the fact that it had high correlations with more acceptable tests such as composition tests. To elaborate: Henning *et al.* (1981:76) found that the highest correlation with composition was with "Error Identification". They consequently maintain that   Error Identification may serve as an indirect measure of composition writing ability" (Henning *et al.*, 1981:462). Although Henning *et al.* (1981:464) maintain that Reading Comprehension "like listening comprehension is of little psychometric value in predicting general proficiency", they concede that reading has to be included in their battery so that it will "find acceptance" (Henning *et al.*, 1981:464). Accordingly, they replaced their "Error Identification" test with a reading comprehension test. The psychometric "posture" (Lantolf & Frawley, 1988:81) had not reckoned with face validity.

I shall show that my error recognition test (which is exactly the same kind of test with very similar items as Henning et al.'s "Error Identification test) is a much better long-term predictor than the cloze ("reading for meaning") test, and should be used, in spite of it being unacceptable to "real-life" testers.

In the 60s and 70s a major issue in testing was whether it was possible to find the "One Best Test" (Alderson, 1981:190). In contemporary research one is looking for a small battery of valid, reliable and *practical* tests rather than the utopian "One Best Test". I believe that these three qualities are present in the tests (all indirect) in this investigation.

Rea (1985) gives the following reasons why indirect tests should be used:

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 sample only 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.

4. SUBJECTS

The sample of subjects (n=81) consists of mother-tongue speakers from diverse linguistic backgrounds, Tswana, Sotho, English, Afrikaans and some expatriates, e.g. Greek, Sri Lankan and Filipino. Two thirds were Tswana-mother-tongue speakers. The 81 subjects include those who failed a grade but succeeded on the second attempt. Two subjects with high English proficiency scores left the School after passing Grade 11 with very good results. These two subjects would in all probability have passed Grade 12 and so are included in the n=81 sample.

Subjects were divided into an L1 group and an L2 group. The L1 group is designated as those who had taken English First-Language as a subject (n=45), and the L2 group is designated as those who had taken English Second-Language as a subject (n=36)).

The 81 subjects originated from 31 schools. Most of the L1 group originated from Connie Minchin Primary School, Mmabatho, which was the main feeder School for L1 entrants to the School during 1980 to 1990. The rest of the L1 group came from a white school, a "coloured" school and a few DET schools. The L2 subjects came from 27 former DET schools and three church schools. *The L1 group was a mixture of mother-tongue and non-mother-tongue speakers of English*, while all the L2 group were Bantu-mother-tongue speakers of which only four were not Tswana-mother-tongue speakers.

The following conditions were the same for all the subjects (i.e. the L1 and L2 groups):

(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 that entrants to the School had received different treatment prior to entering the School, where some may have been disadvantaged, but that all subjects 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 my 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 fall into "L1" and "L2" groups.*

5. METHOD

For the purpose of test administration, subjects were divided into four groups and the tests were administered in four classrooms by four Grade 7 teachers, with a mixture of L1 and L2 subjects in each group. The error recognition test consisted of 50 items. The two cloze passages consisted of 10 "rational deletion" items. A rational deletion procedure differs from fixed

5

deletion in that specific items are selected, e.g. verbs, whereas in fixed deletion every nth word is deleted, for example, every 7th of 9th word. The raw scores of all three tests were converted to percentages.

In my dictation scoring procedure, one incorrect word was worth one point, and errors were subtracted from a possible score of 20 points. 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 it is presented in the same way to them all" (Alderson, *et al.*, 1995:57), i.e. using only one presenter. I used four presenters, one for each of the four groups. The presenters then repeated the process on a rotational basis so that each of them presented the two dictation tests to all four groups. An analysis of variance (ANOVA) was conducted to test for any significant difference between the first presentation of each of the four presenters. No statistically significant difference was found between the four different groups/presenters. The ANOVA was done on the *first* presentation of each presenter. Accordingly, presenters and groups were not confounded (Gamaroff, forthcoming b).

I marked all the tests. The question is whether the scoring procedures in these tests lack evidence of consistency of application because there was only one rater (myself). This was not a problem in the marking procedure of the dictation test because I did not have to worry about distinguishing between spelling and grammatical errors (Oller, 1979:276,282; Cziko, 1982; Alderson, Clapham & Hall, 1995:46) since only wrong forms of words, intrusions and omissions were considered. In the cloze tests special care was taken to ensure that all acceptable answers were considered. The error recognition test had only one possible answer. The answers to the latter two tests were provided by the test compilers.

6.   RESULTS

The results are reported in the following order:

-    The means and standard deviations of the tests for the L1 and L2 groups (Table 1).

-    A summary of the pass rate (Table 2).

-    Histograms of the pass rate, which give a clearer picture of Table 2.

In the rest of the investigation I use the term "test", which refers to (1) the average of the two tests (forms) in each of the test methods of CLOZE and DICT and (2) the ER test.

Care must be taken not to make spurious comparisons between the tests. For example, 50% on one test may not mean the same thing as 50% on another test. 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 level of difficulty and marking conditions. Two examples: (1) score of 50% on CLOZE would be in the "at risk zone" (Pienaar, 1984:19) but would be average for ER. (2) ER is a multiple-choice test and was consequently adjusted for guessing. (I suggest that adjustment for guessing gives a truer picture). The other tests, which were not multiple-choice tests, were not adjusted for guessing. If ER had not been adjusted for guessing this would not have significantly changed the substantial difference

between the L1 and L2 scores. In sum, it is the "relative difference in proficiencies" (Oller, 1979:394) between *learners* of high language ability (in this case the L1 group) and of low language 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.

The sample in Table 1 below comprises the subjects who did all the tests (n=80<86). Five of these left the School before Grade 11 after passing a grade and are therefore excluded from the Grade 7 to Grade 12 predictions; which leaves n=81 for the predictive investigation.

Table 1. Means and Standard deviations of the L1 and L2 Groups

L1 Group			L2 Group		
	M	SD		M	SD
CLOZE (N=49)	65	14	CLOZE (N=37)	26	16
DICT (N=49)	71	17	DICT (N=37)	16	19
ER (N=43)	50	18	ER (N=37)	12	11

The L1 and l2 groups have substantially different means and standard deviations, so it is not necessary to use a z-test to find out whether there is any statistically significant difference between the means of the two groups (Nunan, 1992:29).

In the summary of the predictions in Table 2 below there are three kinds of data for each test: those who (1) passed Grade 12 (indicated as "Pass 7-12"), (2) failed Grades 7, 8 or 9 (indicated as "Fail 7-9"), and (3) failed Grades 10 or 11 (indicated as "Fail 10-11"). The results take into account subjects who failed a grade and passed a year later on the second attempt. All those who passed Grade 11 passed Grade 12.

Table 2. Summary of Predictions

	CLOZE & DICT			ER & PROF		
	All	*L1*	*L2*	*All*	*L1*	*L2*
Pass 7-12	**41**	**28**	**13**	**37**	**24**	**13**
Fail 7-9	32	12	20	30	10	20
Fail 10-11	8	5	3	8	5	3
Total Fail	**40**	**17**	**23**	**38**	**15**	**23**
Grand Total	*81*	*45*	*36*	*75*	*39*	*36*

Figure 5.4 provides a picture of the extent to which the tests may be said to be long-term predictors following details. There were 41 passes and 40 failures between Grades 7 and 11. Of the 40 failures there were 32 who failed Grades 7, 8 or 9 and left the School. The remaining eight subjects are those who failed Grades 10 and 11 and left the school.
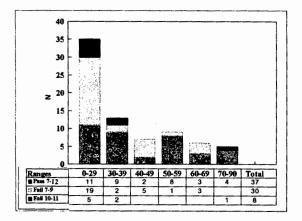
The relatively higher ranges should predict success whereas relatively lower ranges should predict failure. For example, if the ranges in a group are 0-29 and 30-39, respectively, the 0-29 range should predict more failures or less passes than the 30-39 range. The "higher" achievers on such a test (30-39 range) should have a greater chance of school success than those in the 0-29 range. The point of prediction is to find out whether *relatively* high and low scores predict success and failure, respectively.

Thus in this research a good predictor is one which can predict a pass or failure rate of two subjects out of three. Thus 20 out of 30 passes (or failures) is viewed as a good prediction.

The histograms of the individual tests show (1) frequency distributions of the test scores of the whole sample and of the L1 and L2 groups, and (2) how well each test predicts academic achievement. The predictions are discussed in the following order: error recognition (ER), cloze (CLOZE) and dictation (DICT). All 81 subjects did the cloze and dictation tests, but only 75 did the error recognition tests. I shall now discuss the individual tests as predictors of pass rate.

1. Error Recognition Test (ER) Predictions of Grades 7 to 12 Pass Rate

Figure 5.1. ER Whole Sample (n=75)

Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
■ Pass 7-12	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, since 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. If, however, one pools the three upper ranges, one could make the broad observation that a score over 50 is a good predictor (15 passes out of 20).
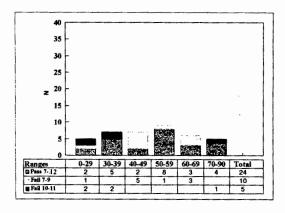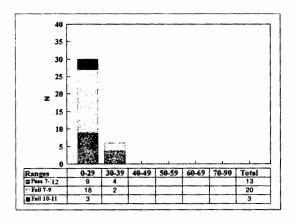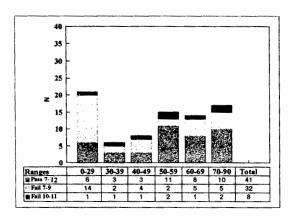
Figure 5.2. ER L1 Group (n=39)

Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
▫ Pass 7-12	2	5	2	8	3	4	24
· Fail 7-9	1		5	1	3		10
■ Fail 10-11	2	2				1	5

9

Figure 5.3. ER L2 Group (n=36)

Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
▣ Pass 7-12	9	4					13
⠂⠂ Fail 7-9	18	2					20
■ Fail 10-11	3						3

Since the 0-39 ranges are occupied mostly by 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 wide spread, where a score over 50% is a good predictor. To sum up, 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. Cloze Test (CLOZE) Predictions of Grades 7 to 12 Pass Rate: Figures 5.4 to 5.6

Figure 5.4. CLOZE Whole Sample (n=81)

Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
▣ Pass 7-12	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 quite a good predictor in the 50-90 range (29 passes out of 46). A close look at the individual ranges reveals that the very low range (0-29) is the best predictor (6 passes out of 21).

In the L1 graph, the 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.

According to Pienaar (1984:21) the 40-49 range is in the "at risk" zone. The 40-49 range in CLOZE, the whole sample (graph 5.4) bears his assertion out (five out of eight failures). The 60-90 range shows that a substantial number of subjects in this range were also at risk, because they failed. However, there is a fundamental difference between these results. My tests are for long-term prediction and Pienaar's tests are concerned with short-term prediction, i.e. predicting one year ahead. With this in mind, Pienaar would probably look only 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, all except two failed *later* than Grade 7. Thus, in the 60-90 range Pienaar's tests are good short-term predictors (of success), but not good long-term predictors.

In the 0-39 ranges, 14 of the 18 failures occurred in Grade 7, two in Grade 8 and two in Grade 11.

There was a similar pattern of failures in DICT (to be discussed shortly). Consider the L1 and L2 groups in Graphs 5.5 and 5.6. The majority of passes are in the 50-90 ranges, which belong to the L1 group. As one might have expected, the L1 group had higher test scores (on all the tests) than the L2 group.

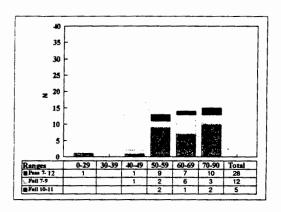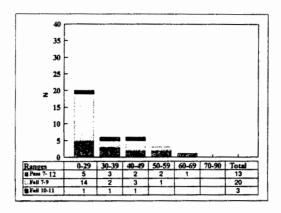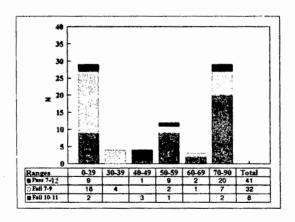Figure 5.5. CLOZE L1 Group (n=45)

Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
■ Pass 7- 12	1		1	9	7	10	28
Fail 7-9			1	2	6	3	12
■ Fail 10-11				2	1	2	5

Figure 5.6. CLOZE L2 Group (n=36)

Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
▪ Pass 7- 12	5	3	2	2	1		13
⸬ Fail 7-9	14	2	3	1			20
▪ Fail 10-11	1	1	1				3

In the discussion in Section 7 I shall suggest reasons why the two upper ranges (60-69 and 70-90) in CLOZE are not good long-term predictors.

3. Dictation Test (DICT) Predictions of Grades 7 to 12 Pass Rate: Figures 5.7 to 5.9

Figure 5.7. DICT Whole Sample (n=81)

Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
▪ Pass 7-12	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

12

four failed Grade 8. The histograms below show the radical difference between the L1 and L2 groups.
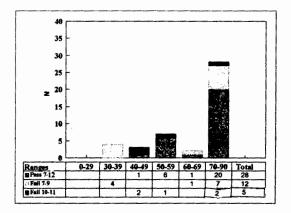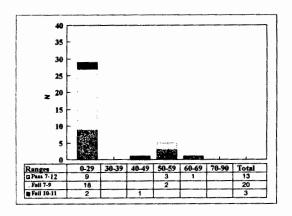
Figure 5.8. DICT L1 Group (n=45)

Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
▣ Pass 7-12			1	6	1	20	28
⠿ Fail 7-9		4			1	7	12
▪ Fail 10-11			2	1		2	5

Figure 5.9. DICT L2 Group (n=36)

Ranges	0-29	30-39	40-49	50-59	60-69	70-90	Total
▣ Pass 7-12	9			3	1		13
⠿ Fail 7-9	18			2			20
▪ Fail 10-11	2		1				3

7. DISCUSSION

The 0-29 range is a very good predictor in all the tests. Subjects who failed in this range dropped out before being at the school for two years. DICT is the best predictor of the three tests where a score over 50% predicts a pass rate of 31 out of 44 and a score under 50% predicts a failure rate

of 27 out of 37. The second best predictor is ER where a score over 50% predicts a pass rate of 15 out of 20 , and a score of under 30% predicts a failure rate of 24 out of 35. CLOZE is not a good long-term predictor of success as I shall show later.

In DICT many in the L2 group would probably not have understood the dictation no matter how clear the presentation. One might explain the difference in performance between the L1 and the 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 (Rumelhart, 1977; Samuel, 1981; Kelly, 1991). In DICT the words are highly predictable for the L1 group and therefore this group does not have to rely totally on the sound input. The opposite is the case for the L2 group where there is almost a total reliance on the bottom-up process of sound recognition (Kelly, 1991:136). 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 (Suenobu et al., 1986:244). 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 (Chomsky, 1975; Kelly, 1991:138; Laufer, 1990:294-95). Problems can occur in any of these areas.

What was said about information-processing in DICT applies to CLOZE as well. The difference between DICT and CLOZE is that test-takers can read the text in CLOZE, whereas in DICT one is limited to the temporal constraints of the presentation. So, why is CLOZE more difficult than DICT for the L1 group when the passages in CLOZE and DICT belong to the same level ("Step")? A possible answer to that in CLOZE one has to *produce* the correct item, i.e. perform one's competence without the kind of help that may be obtained in DICT. In DICT the sounds produced by the presenter may evoke what is stored in one's head, and so it may be easier to produce. If one has low competence, as was the case for the majority of the L2 group, the clearest presenter cannot help.

CLOZE is not a good long-term predictor in the 60-90 range because it does not discriminate, in the long term, as well as the other tests do 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 (see Section 2). 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+. The good short-term and poor long-term predictions of CLOZE fit in well with Pienaar's rationale.

If the reason why CLOZE (specifically the 60-90 range) is not a good long-term predictor is because reading skills as one kind of language skill may 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.

14

To sum up, the tests had a wide spread of scores and distinguished clearly between low-English-Proficiency subjects and high-English-proficiency subjects. Relatively high scores should predict success and relatively low scores should predict failure. The tests showed this quite well.

8. IMPLICATIONS

My sample 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. Mmabatho High School 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.

Although one might concede that the high failure rate reported at Mmabatho High School is of interest, of more interest is the generalisability of the findings (Pilliner, 1973:43). In other words, do the subjects described in this study represent a population outside the School? It is not possible to generalise the L1 findings because most of the subjects originated from only one School (Connie Minchin Primary).

Since however, the diversity of the Grade 7 sample of subjects described in this article is becoming representative of many schools in South Africa, it might be interesting to use some of the tests in this study or similar tests to predict academic achievement at other schools. 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. It is worth noting that educational and political arguments may lead to very different conclusions. Using the terms L1 and L2 to mean the same as I do here, Barkhuizen (1991) gives sound educational arguments for keeping L1 and L2 learners separate. However, from a different angle, the "multicultural settings" policy of the new South African curriculum, Barkhuizen (1992, 1993) reaches different conclusions and gives sound political arguments against keeping L1 and L2 learners separate.

With regard to the L2 group, some inferences can be made. A large proportion of the L2 subjects that had been described as disadvantaged had low English proficiency - as measured by the English proficiency tests - and dropped out before Grade 9 at the School. This supports the general view that an initial (in this case Grade 7) low level of English proficiency will probably

The crux of the matter is this: many disadvantaged children who entered the School in Grade 7 were unable to fulfil the demands of the JMB syllabus and to benefit from the relatively enriching academic facilities offered at the School. This was not only true of the specific sample in this study, but also of many learners who had attended the School since its inception in 1980. A change of environment to a more advantaged setting such as the one that exists at the School seemed, in many cases, to have little significant effect on academic performance. This is driven home by the high failure rate among the subjects with low scores on the English proficiency tests.

It would be incorrect to infer that learners who obtain low test scores 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 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 are "disadvantaged".

9. CONCLUSION

Communicative testers argue that tests such as those I have used are out of date and have been replaced by "real-life"-task demands. These "old paradigm" tests may not have the same degree of face validity as "real-life" tests but they certainly are no less "authentic" than "new paradigm" tests (Gamaroff, 1996, 1997). I would argue that robust tests such as dictation, cloze and error recognition should not be rejected in the highly mistaken belief that they are less authentic - and therefore of less value - than "communicative" tests.

In traditional language teaching, "applied linguistics" is largely concerned with second or foreign language teaching, and "educational linguistics" is largely concerned with mother tongue/first language teaching (Crystal, 1996:418). These definitions make it easier to understand an applied linguist's "advice" to me that an investigation such as the one in this article belongs to "education" and is, accordingly, not an applied linguistics topic. Yet, if language proficiency tests are not literally applied to the prediction of academic achievement, what is the *ultimate* good of much of applied or educational linguistics and much that is published in these fields?

REFERENCES

ALDERSON, JC. 1979. The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly*, 13:219-227.

ALDERSON, JC. 1981. Reaction to the Morrow paper. In: Alderson, JC. and A Hughes, *Issues in language testing: ELT Documents III.* London: The British Council.

ALDERSON, JC, C CLAPHAM AND D WALL. 1995. *Language test construction and evaluation.* Cambridge: CUP.

BACHMAN, LF. 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press.

BARKHUIZEN, G. 1991. Proposal for an independent English Second Language Department at Mmabatho High School. *English Language Teaching Centre (ELTIC) Reporter*, 16(1):25-32. Johannesburg.

BARKHUIZEN, G. 1992. Teaching English in multilingual settings (TEMLS): What needs to be done. *Journal for Language Teaching*, 26(4):53-68.

BARKHUIZEN, G. 1993. 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*, 28-30 June, University of Port Elizabeth.

BLOOR, M, T BLOOR, R FORREST, E LAIRD & H RELTON. 1970. *Objective tests in English as a foreign language*. London: Macmillan.

CRYSTAL, D. 1996. *The Cambridge encyclopedia of language* (Second Edition). New York: Cambridge University Press.

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

CHOMSKY, N. 1975. *Reflections on language*. New York: Pantheon Books.

CRONBACH, LJ. 1970. *Essentials of psychological testing*. New York: Harper & Row.

GAMAROFF, R. 1986. *Native language transfer in Tswana speaker's English*. Unpublished MA thesis, Potchefstroom University for Christian Higher Education.

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

GAMAROFF, R. 1997. Paradigm lost, paradigm regained: Statistics in language testing. *SAALT Journal*, 31(2):131-139.

GAMAROFF, R. (Forthcoming a). The cloze test as a measure of language proficiency: A statistical analysis. *South African Journal of Linguistics*, 16(1):7-15.

GAMAROFF, R. (Forthcoming b). The dictation test as a measure of communicative language proficiency. *International Review of Applied Linguistics*.

GUE, L. AND HOLDAWAY. 1973. English proficiency tests as predictors of success in graduate studies in education. *Language Learning*, 23:89-103.

HALE, GA, CW STANSFIELD AND RP DURAN. 1984. *TESOL Research Report 16.* Princeton, New Jersey: Educational Testing Service.

HENNING, A. 1987. *A guide to language testing.* Rowley, Massachusetts: Newbury House.

HENNING, GA, SM GHAWABY, WZ SAADALLA, MA EL-RIFAI, RK HANNALLAH, AND MS MATTAR. 1981. Comprehensive assessment of language proficiency and achievement among learners of English as a foreign language. *TESOL Quarterly*, 15(4):457-466.

HUGHES, A. 1989. *Testing for language teachers.* Cambridge: Cambridge University Press.

KELLY, P. 1991. Lexical ignorance: The main obstacle to listening comprehension with advanced foreign language learners. *International Review of Applied Linguistics*, 29(2):134-149.

LANTOLF, JP AND W FRAWLEY. 1988. Proficiency: Understanding the construct. *Studies in Second Language Acquisition (SLLA)*, 10(2):181-195.

LAUFER, B. 1990. 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.

MITCHELL, G AND P FRIDJHON. 1988. Matriculation examinations and university performance. *Bulletin for Academic Staff*, 9(1):28-43. University of Durban-Westville.

NUNAN, D. 1992. *Research methods in language learning.* Cambridge, New York: Cambridge University Press.

OLLER, JW, JR. 1973. Cloze tests of second language proficiency and what they measure. *Language Learning*, 23(1):105-118.

OLLER, JW, JR. 1976. Cloze, discourse, and approximations to English. In: Burt, K. & Dulay, H.C. *New directions in second language learning, teaching and bilingual education.* TESOL: Washington, D.C.

OLLER, JW, JR. 1979. *Language tests at school.* London: Longman.

OLLER, JW, JR. 1983. A consensus for the 80s. In: Oller, JW Jr. (Ed.). *Issues in language testing research.* Rowley, Massachusetts: Newbury Publishers.

OLLER, JW, JR. AND F KAHN. 1981. Is there a global factor of language proficiency? In: Read, J.A.S. *Directions in language testing.* Singapore: Singapore University Press.

PEIRCE, BN. 1990. Student writers, the DET syllabus, and matric marking: A critical evaluation. *Eltic Reporter*, 15:(3):3-11.

PIENAAR, P. 1984. *Reading for meaning: A pilot survey of (silent) reading standards in Bophuthatswana.* Mmabatho: Institute of Education, University of Bophuthatswana (University of the North West).

PILLINER, AEG. 1968. *Subjective and objective testing.*

POPHAM, WJ. 1981. *Modern educational measurement.* Englewood Cliffs, New Jersey: Prentice Hall.

REA, P. 1985. Language testing and the communicative language teaching curriculum. In: Lee, YP *et al.* 1985. *New directions in language testing.* Oxford. Institute of English.

RUMELHART, DE. 1977. *Introduction to human information processing.* New York: John Wiley & Sons.

SAMUEL, AG. 1981. Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology*, 110:474-494.

SIMPSON, GJ. 1987. Are matric marks relevant selection criteria for higher education? In: Blaquiere, A. (Ed.). *Intercultural issues in teaching and learning. Proceedings of the 1986 South African Association for Research and Development in Higher Education (SAARDE) Conference.* University of Natal, Convening Committee:68-72.

SUENOBU, M, K KANZAKI, S YAMANE AND R YOUNG. 1986. Listening comprehension and the process of information acquisition by non-native speakers of English. *International Review of Applied Linguistics*, 24(3), 239-248.

APPENDIX

Cloze passages

Cloze passage 1: Form B Step 2 (Pienaar, 1984:59):

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 __4__ 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 my 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).

Cloze passage 2: Form D Step 2 (Pienaar, 1984:61):

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 her 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 at once. My mother saw that I looked after him properly and he grew up into a beautiful grey Alsation. 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.

Dictation passages

Passage 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.

Passage 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!

ont

umer