CHAPTER 2

LITERATURE REVIEW

In the early days of educational opportunity, according to Cronbach (1990), the prestigious eastern U.S. colleges had little need for entrance examinations because their students came mostly from nearby schools. Knowing the standards of these schools, the colleges had no problem in accepting the applicants. When in the 1920s they sought applicants from the entire nation, the colleges found transcripts from distant high schools difficult to interpret. For a fairer comparison, the Scholastic Aptitude Test (SAT) was established. This was a placement test – to gain admission into these colleges. Therefore, we can say that placement tests are a better way to discriminate students of their prior knowledge at entry point.

Although diagnostic tests usually help to identify problem areas, they seldom provide reasons for difficulties and cannot prescribe solutions to overcome them (Ebel and Frisbie, 1991). A major challenge to the teacher is to synthesize the information of the students' entering behaviour so that instructional strategies and materials can be selected to optimise students' learning conditions. Oller (1979) claims that when the focus of language testing research is on learner characteristics, this challenge can be met. The tests themselves may be viewed as a procedure to elicit data to be analysed, in an attempt to diagnose specific aspects of learner development.

Diagnostic tests are readily available nowadays and can be used as a placement test, depending again on the purpose of the tester for using the test. Nevertheless, the
reliability of the test is much doubted as the sample sitting for these tests differ from the group that the test was developed for. They may vary in terms of culture and background.

Language tests are used for a variety of purposes, grouped in two categories (Bachman, 1991). Firstly, results of language tests may be used to make inferences about student’s language abilities or to make predictions about their capacity for using language to perform future tasks in contexts outside the test itself. Secondly, decisions (placements, diagnosis, selection, grading etc.) may be made about the students on the basis of what teachers infer from test scores about their levels of ability or their capacity for non test language use.

When the focus is on the tests themselves, questions of validity, reliability, practicality and instructional value must be considered.

In following Spolsky’s shift towards the ‘integrative-sociolinguistics’ view of language testing (Spolsky, 1979), it is more reliable and valid if a test is constructed collectively by expertise of language teachers, linguists, psychometricians, sociologists and anthropologists. The focus will be on authentic language use (sociolinguistic perspective) and on how to measure this use through language tests (integrative techniques). Although Spolsky professes idealism in language testing, the amount of time and cost involved to work collaboratively does not seem very practical for tests at school level.

According to Ebel and Frisbie (1991) and Sax (1989), the test-retest method is an indicator to the consistency of students in the achievement of the same test. If results
prove to be of the same or somewhat similar, then the test is concluded to be a reliable test. Murphy and Davidshofer (1998) say that this test must be administered to the same group of individuals at both occasions at an appropriate interval and the correlation between scores on the first test and the retest is used to estimate reliability of the test. In addition, Anastasi (1988) equates the term reliability to consistency. Test reliability is the consistency of scores obtained by the same persons when retested with the identical test or with an equivalent form of the test. She says that the retest reliability shows the extent to which scores on a test can be generalised over different occasions; if the reliability is higher, then the scores are not influenced by the random daily changes in the condition of the test takers or of the testing environment. One way of avoiding the difficulties encountered in test-retest reliability is through the use of alternate forms of the test. The correlation between the scores obtained on the two forms represents the reliability coefficient of the test. This coefficient thus combines two types of reliability – which is a more useful measure for evaluating tests.

Meanwhile, Murphy and Davidshofer (1998) are concerned with carry over effects particularly if the interval between the test and retest is short. People tend to remember questions and answers which could affect their answers the second time around. The experience of taking the test itself can change a person’s score which is referred to as reactivity.

The most important kind of reliability estimate for achievement tests would be equivalence reliability, claim Mehrens and Lehmann (1991). One can infer from the responses to a specific set of items the degree to which a person has mastered the
essential skills. For multiple skill tests, it is also essential to have data on the reliabilities of the sub-tests and difference score. As most achievement tests are designed to fit the curriculum, and students learn those materials in differing amounts and rates, it is not expected that these scores will remain constant. Hence, long-range stability coefficients would be rather meaningless. As such, for this study, a equivalent test is set for the retest session and is administered within a week from the first test. They also feel that the reliability will be higher when a test is given to a heterogeneous group, which is true about the sample for this study.

Thorndike and Hagen (1977) add that the use of two parallel test forms provides a very sound basis for estimating the precision of a psychological or educational test. The correlation between two parallel forms, usually administered with a lapse of several days or weeks in between, represents the preferred procedure for estimating reliability. Anastasi (1988) also emphasised that in the development of parallel forms, the test should contain the same number of items and the items should be expressed in the same form and should cover same type of content. The range and level of difficulty of items should be equal. Instructions, time limit, examples, format and all other aspects of the test must likewise be checked for equivalence. This study attempts to follow suit except that 30% of the items are repeated from the first test (Form I). To add, Ebel and Frisbie (1991) said that if an alternate forms of a test are given to a group, then the correlation between scores of both forms provides an estimate of score reliability. A high reliability estimate proves that the two forms can be used interchangeably as measuring the same traits. However, a low estimate shows that the two set of items probably are not sampling the
content domain very well.

This research will also attempt to help the non-statistically trained reader to understand the meaning of correlation enough to appreciate some of the interesting findings.

CORRELATION

Correlation is a statement of the strength of the association between two or possibly more variables (Sprinthall and Sprinthall, 1990). Though correlational statements do not imply causation, they can be used to make predictions. It is a way to assess reliability. If a certain test is given to a group of subjects on two separate occasions, and if those individuals who score high on the test the first time also score high the second time and those who score low the first time also score low the second time, the two sets of measurements are said to correlate and the test is considered reliable, because the test is yielding consistent scores.

Simple correlation, also known as Pearson product-moment correlation (Pearson r) is the degree of relationship between two variables. Correlation between two tests (test-retest) will tell how accurately or how consistently the test is measuring whatever it does measure. These correlations are called reliability coefficients. More than nine-tenths of all correlation coefficient that are reported in psychological and educational literature lie between −0.20 and +0.60, even though these limits include less than half of all possible correlation coefficients (Kurtz and Mayo, 1979).
INTERNAL CONSISTENCY

Besides Pearson r, Cronbach Alpha is another reliability estimate. Pearson r is used to correlate items between two tests (inter-test reliability) whereas Cronbach Alpha is used to correlate items within the same test (intra-test reliability), otherwise known as internal consistency. Internal consistency is the precision and consistency of test scores on one administration of a test across test items.

Internal consistency reliability is a totally independent concept from test-retest reliability. Alpha measures the extent to which item responses obtained at the same time correlate highly (as well as the number of items). It makes no assumptions about what one would obtain at a different time (test-test reliability) according to Bernstein (1995).

One may argue that when a high Cronbach Alpha indicates a high degree of internal consistency, the test or the survey must be uni-dimensional rather than multi-dimensional. Thus, there is no need to further investigate its sub-scales. Actually it is a common misconception. If a test is uni-dimensional, then it will show internal consistency. But if a test is internally consistent, it does not necessarily entail one construct (Gardner, 1995). Cronbach Alpha is a measure of common variance shared by test items. The Cronbach Alpha could be high when each test item shares variance with at least some other items; it does not have to share variance with all other items.

In improving reliability or validity estimates, Glover and Bruning (1987) suggest that the primary problems as related to the following should be identified: unclear directions given to the person being measured, lack of objectivity in scoring, and poor sampling of the person's behaviour. To remedy the problem, a change in procedure
should be proposed.

**SCORING**

After considering validity, reliability, practicality and instructional value in construction of tests, these tests are administered to the sample. Then rating is done – to give a score for the correct answers. There are many methods of rating and it is much easier if it is an objective test as there is only one right answer. If it is a subjective test, then discrepancy will surface. Anastasi (1988) seems to have a solution to this problem. She suggests that rater reliability can be found by having a sample of test papers independently rated by 2 examiners. The two scores by each test taker are then correlated in the usual way, and the coefficient that is obtained is a measure of rater reliability. This type of reliability is commonly computed when subjectively scored instruments are employed in research. As for this study, part of the recommendation is considered. 2 examiners will rate the essays independently but the mean score will be taken as the score for the student. If the discrepancy is too great (3 marks and above) then a third examiner will be asked to rate the essay.

In rating essay questions, Gronlund (1985) finds the global method (points given in comparison to an ideal answer holistically) is quicker than the analytical method (points assigned for desired features). Analytical method will be good when it serves as assistance to the grader in giving feedback. Since this research does not give feedback to students, the global method was used. The global method is also thought as a better method because the purpose of the placement test is to rank the students and stream them into homogeneous groups, which is what this method does.
SUPPLEMENTARY READING

All this while, the literature reviewed are relevant to the instrument, that is the test itself. At this juncture, a supplementary psychological reading was done with regards to the performance of the test takers. The need for this arose since the reliability of the instrument was also dependent on the test takers’ conditions. Anxiety was one of the elements noted in the observation during the administration of the tests. Anxiety is defined as an unpleasant emotional state or condition which is characterised by the subjective feelings of tension, apprehension and worry and by activation or arousal of the autonomic nervous system (Spielberger, 1972, pp. 482). However, the Drive Analysis Theory argues that increasing drive energized all habits, both correct and incorrect (Muller, 1992). When the task is easy, the correct response was dominant at the outset. Although both correct and incorrect responses are strengthened by the increasing drive, the correct performance would increase as drive increases. If the tasks were difficult, drive increments would energize already dominant error tendencies early in learning. Therefore, anxious subjects perform better than low anxiety subjects on easy tasks, but would be worse than low-anxiety subjects on difficult tasks.

Generally this study is based on an article about the implementation of the CELT at the Faculty of International Studies at Kyoritsu Women’s University, Japan. According to Mills, Swain and Weschler (1996), the purpose of the article was to explain why it was initially decided to separate the students into levels, how the placement was achieved and report some observations based on the results of the test. Prior to the implementation of CELT as a placement, students were separated randomly into classes.
The teachers were faced with a dilemma: teach at a more proficient level, but the rest of the class would be lost and frustrated; teach at a slower pace and others would be bored and feel unchallenged; or teach to the middle and nobody is satisfied. In the Japanese culture, the better students would hesitate to display their knowledge in front of the less competent ones which weaker students were afraid of appearing stupid if they tried and made mistakes. This problem was more evident in the speaking classes. Thus the teachers came to a conclusion in resorting to CELT as a placement test, helping them to discriminate students according to their ability. Since implementing this programme, the teachers have found that it was much easier to select the appropriate teaching methodologies. The pace can be adjusted and more attention can be given to students who are shy to ask for help. Evaluation is made more accurately and efficiently as the on-going tests are finely calibrated. Classroom management became more simplified as true learning environment was created. The most noticeable difference was the students’ motivation to learn English.

This results from the above mentioned article has inspired this study. However, the analysis for the research is done with regards to an adaptation from the CELT, that is the EPT at the PPP Training Centre.