Construct Validation of the Reading Subskills of the Boğaziçi University English Proficiency Test

145737

145737

Dissertation submitted to the

Graduate Institute of Social Sciences

in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Foreign Language Education

by

Aylin Ünaldı

Boğaziçi University

2004

The dissertation of Aylin Ünaldı

is approved by:

Professor Dr. Cem Alptekin (Committee Chairperson)

Professor Dr. Hüsnü Enginarlar

Assistant Professor Dr. Yasemin Bayyurt

Assistant Professor Dr. Belma Haznedar

Assistant Professor Dr. Gülcan Erçetin

July 2004

ACKNOWLEDGEMENTS

First, I wish to acknowledge gratitude to my advisor, Professor Cem Alptekin, for all the illuminating experience I have had with him. I consider him the embodiment of Sirat* in my context and working with him is as great a challenge as satisfaction. I wish to thank Professor Ayşe Akyel and Assistant Professor Gülcan Erçetin for the torchlight they shed in the moments of darkest blindness and the timely encouragement they gave me. I should also thank Professor Cyril Weir whose enlightening training sessions stimulated this study. Şadiye Ağım, Sumru Başaran, Melahat Behlil, Fatma Taşkent and Aylin Vartanyan, my dear friends and colleagues deserve generous thanks as without their resourcefulness and dedicated effort they put in testing, this research would have never materialised. I also feel indebted to Professor Eser Taylan, Deniz Atlı and Dr. Nilgün Fırat for the support they gave me through the painful experience of both working as an English teacher and doing research. My warmest thanks should go to my husband, Bahri Besimoğlu, who has always upheld me despite the unbearable hardships he himself has gone through in the course of his marriage with me, and of course to Hadiye Cangökçe, my soul mate, who could manage to keep faith in me even when I lost belief in myself. I heartily thank them for being in my life. I should also say that neither Bahri nor I will ever forget how Mom saved us many times from deadly hunger. I love her as much as I love her dishes!

* the imaginary bridge between paradise and hell, finer than a hair and sharper than a sword.

CURRICULUM VITAE

AYLİN ÜNALDI

Date of birth: 04. 03. 1966

Place of birth: Istanbul

EDUCATION

Years	**Institution**	**Degree gained**
1988 - 1993	Boğaziçi University Faculty of Education Department of Foreign Language Education	BA in Foreign Language Teaching
1993 - 1994	Boğaziçi University Faculty of Letters MA in Linguistics	Coursework completed
1994 - 1995	University of Reading Faculty of Letters Department of Linguistic Science	MA in Applied Linguistics/ Language Teaching
1996 - 2004	Boğaziçi University Faculty of Education Department of English Language Education	PhD in Foreign Language Education

WRITTEN WORKS

Ünaldı, A. (1995). Revision behaviours of student writers in English and Turkish: *accounting for the 'quality' of revision through Rhetorical Structure Theory.* Unpublished MA Thesis: University of Reading.

Daller, H., Treffers-Daller, J. and Ünaldı, A. (1996). The development of a Turkish C-Test. In J. A Coleman: University Language Testing and the C-Test; Proceedings of a Conference held at the University of Porsmouth. University of Porsmouth. Occasional Papers in Linguistics.

Tütüniş, B. and Ünaldı, A. (1997). Content based academic writing. Proceedings of the Conference held at Bilkent University: 'Quality Learning in English Medium Higher Education: The Challange of Content, Skills and Language'. Bilkent University.

Daller, H., Treffers-Daller, J., Ünaldı, A. and Yıldız, C. (2002). The development of a Turkish C-test. In J. A. Coleman, R. Grotjahn and U. Raatz (Eds.), University language testing and the C-test. Bochum: AKS-Verlag.

ABSTRACT

Construct Validation of the Reading Subskills of the Boğaziçi University English Proficiency Test

Aylin Ünaldı

The purpose of this study is to present construct validity evidence for the reading module of the Boğaziçi University English Language Proficiency Test (the BUEPT). Following the suggestions of Messick's (1989a) validity framework, the study provides evidence for *content*, *substantive*, *structural*, *generalisability* and *external aspects* of construct validity of the BUEPT reading module. Initially, a theoretically sound and practically applicable reading framework (Urquhart and Weir, 1998) that would ensure content relevance and representativeness was chosen and test specifications were developed based on that framework. The tests were piloted and the analysis of score distributions and item performance through classical test theory helped improve the technical quality of the tests minimising the construct irrelevant test variance. Expert judgement was taken using an analysis scheme based on Bachman et al.(1995) and verbal protocols of the test takers were analysed in order to investigate whether or not each item reflects the content defined by each dimension of the reading construct as defined in the framework. The factor structures of the tests were analysed using the Principal Component Analysis and the BUEPT reading module was compared to the IELTS reading test both in terms of content congruence and the correlation between them. The findings from these investigations provided substantial support for the validity of the score interpretations based on the BUEPT reading test. The study generally supports the soundness and applicability of the Urquhart and Weir's (1998) framework.

KISA ÖZET

Boğaziçi Üniversitesi İngilizce Dil Yeterliliği Sınavı Okuma Altbecerilerinin Yapı Geçerliliği Bulgulaması

Aylin Ünaldı

Bu çalışmanın amacı Boğaziçi Universtitesi İngilizce Dil Yeterliliği Sınavı'nın (BUEPT) yapı geçerliliğine kanıt sunmaktır. Sınavın yapı geçerliliği Messick'in (1989a) geçerlilik teorisine koşut olarak *içerik*, *esas*, *yapı*, *genellenebilirlik* ve *sınav dışı* ölçütlere kanıt sunularak incelenmiştir. Öncelikle, sınavın içerik uygunluğunu ve kapsamlılığını sağlayacak, teorik geçerliliği ve uygulanabilirliği olan bir okuma modeli (Urquhart ve Weir, 1998) seçilmiş ve sınav tanımlamaları bu modele dayanarak geliştirilmiştir. Sınavlara deneme uygulamaları yapılmıştır. Klasik test teorisi kullanılarak yapılan puan dağılımları incelemesi ve madde analizi sonucunda yapıyla ilgisiz test varyansı en aza indirilerek sınavların teknik kalitesi arttırılmıştır. Bachman ve diğerlerine (1995) dayalı olarak geliştirilen bir analiz cetveli yardımıyla uzman kanısı alınmış ve sınavdaki her maddenin modelde belirlendiği biçimde, okuma yapısının her boyutunun tanımladığı içeriği yansıtıp yansıtmadığı, sınavı alan bir grup öğrencinin verdiği sözel tutanaklar aracılığıyla incelenmiştir. Sınavların faktör yapısı, Ana Öğe Analizi yöntemi kullanılarak incelenmiş ve BUEPT okuma sınavı, hem içerik benzerliği hem de aralarındaki korelasyon bakımından IELTS'in okuma bölümüyle karşılaştırılmıştır. Yapılan bu incelemelerden elde edilen sonuçlar, BUEPT okuma sınavından alınacak puanlar üzerine yapılacak yorumların geçerliliğini önemli ölçüde desteklemiştir. Ayrıca, bu çalışma, Urquhart ve Weir'in (1998) okuma modelinin geçerliliği ve uygulanabilirliğine genelde destek vermiştir.

TABLE OF CONTENTS

CHAPTER 4: RESULTS AND DISCUSSION

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background to the Study

Assessment plays an essential role in language education and research. When language tests are in question there is always serious interest in reliability and validity. Test reliability, which is defined as the extent to which the results can be considered consistent or stable can be established through statistical analysis, i.e. by calculation of a reliability coefficient (Brown, 1996). Validity on the other hand, is assumed to be a collection of evidence from multiple sources of information, an 'evolving property' and 'a continuing process' (Messick, 1989a, 13). Messick defines validity as

> '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' (ibid, 13).

Cumming (1996, 1) underlines the importance of validation in 'arbitrating educational and linguistic policies, institutional decisions, pedagogical practices, as well as tenets of language theory and research', and Moss' (1992) survey makes it clear that 'construct validity', as a unitary concept basic to all the considerations of validity, has been recognised as central to all these issues. Messick (1989b) stresses that although it is faceted, validity is a unitary concept since all forms of evidence, whether relating to the relevance and representativeness of content or the correlation

between the test scores and criterion (and many other aspects of validity that are judged to be different types of validity), are in fact construct-related evidence and should not be considered as 'types' of validity. Bachman (1990a, 256) states that construct validation can be seen as verifying or falsifying a scientific theory relating to the abilities we want to measure. However, since what we measure on a test is the interaction between the ability and the test method used, and not the ability per se, the use of appropriate test methods is of crucial importance to arrive at correct *inferences* based on test scores. To generalise these inferences beyond the test situation, an appropriate theory of the ability in question should be operationalised in tests. Constructs of reading, for example, are operationalised through the texts, the tasks, the understandings of readers and inferences we make based on test scores. Therefore, how much our test reflects the theory and how adequately the theory reflects what is involved in reading will determine the construct validity of our test (Alderson, 2000).

In attempts to understand the nature of the reading construct, several theories and models have been proposed. Earlier models emphasise the sequential, linear nature of reading with lower subprocesses such as phoneme and word identification leading to higher processes such as syntax and sentence analysis resulting in text comprehension (bottom-up, text-driven models). With increasing understanding on the role of the reader and the importance of the background knowledge he/she brings in the reading process, theories emphasising top-down processes were proposed (e.g. schema-theoretic view). However, further research made it clear that reading is not a unidirectional process. On the contrary, processing in both dimensions proceeds simultaneously, in an interactive manner (interactive models).

Most of these models make reference to several skills or subprocesses that are thought to be employed in reading, suggesting that reading is a multidimensional, multicomponential process. No matter how unclear the research is on the issue, language tests make use of reading taxonomies with respect to the design of the test tasks and items. On the other hand, tests based on the unitary view of reading, which suggests that reading is an indivisible process, might tap on certain linguistic skills such as vocabulary and syntax and neglect others, thereby putting construct validity at risk.

Therefore, what we wish to include in our test is closely related with the theoretical framework underlying the test, and test specifications should make explicit the link between the theory and test through operational definitions (Alderson, et al. 1995; Bachman, 1990a; Urquhart and Weir, 1998). Comprehensive test specifications with a sound theoretical and empirical base and their implementation in test tasks as accurately as possible will enhance the likelihood of a 'construct-valid' instrument. Data from the use of such an instrument might also provide us with some understanding of the nature of the reading construct (Weir, et al. 2000).

1.2 Aims of the Present Research

This research aims at giving a detailed account of the construct validation study of the reading module of an English for Academic Purposes (EAP) proficiency test at the university level. The test in question is The Boğaziçi University English Language Proficiency Test (BUEPT) administered to all students upon entrance to the English medium Boğaziçi University in Turkey at the beginning of each

academic year in September. Those students whose test results indicate that their English level is not adequate are required to spend at least one semester in the university's School of Foreign Languages where they follow intensive English classes. They are required to take the exam at the end of the language training period and satisfy the lowest 60 % passing mark before they can pass on to their university studies proper. The students who can attain an average of 80% in first-semester achievement exams can take the proficiency test at the end of the first semester in January. On the other hand, the rest of the students take the test in June on completion of the two-semester prep school. The ones who fail the June test are required to follow an additional six-week remedial language program and attempt the August test.

The BUEPT versions as they were used until September 2000 were developed based on an extensive needs analysis as a skill-based academic English proficiency test by the Testing Office of the School of Foreign Languages under the supervision of Arthur Hughes in 1982 (Hughes 1988, 1989). The BUEPT had three main components: Listening, Reading and Writing. The listening part had two subcomponents; while-listening and note-taking; the reading test also had two subcomponents; scanning and detailed reading.[1] In the writing part, students were required to write two one-page expository essays.

[1] The listening section consisted of two ten-minute lectures and associated questions. In the while-listening part questions were given in advance and students were expected to respond to questions while listening to the lecture. During the second lecture, students took notes and answered the questions delivered after the lecture was over, using their notes. In the reading section, there were two long texts (approximately 3000 words). The first one had scanning and the other involved detailed reading tasks (see Hughes, 1988 for details).

After eighteen years of use, The BUEPT underwent revision to yield a better test constructed following the suggestions of recent research in EAP reading, language testing and validation. The researcher together with four other members of the Testing Office of the School of Foreign Languages of Boğaziçi University worked to produce test specifications and several test versions following a certain reading model under the supervision of Professor Cyril Weir with the support of the British Council, Istanbul. Three main components of the test were retained while subcomponents were revised to give both quantitatively and qualitatively different tests. The reading component was changed to involve five texts in three subcomponents; scanning, search reading and careful reading.[2] Five equivalent test versions which were produced following the same reading model and test construction principles were administered between September 2000 and 2001.[3]

1.3. Overview of Methodology

As mentioned above, the present research aims at providing evidence from multiple sources as to the construct validity of the reading component of the new Boğaziçi University Proficiency Test. Within the framework Messick (1989a) suggested, the study primarily involves considerations of the theoretical construct as embodied in the test. It gives a detailed account of the test specifications and how these are operationalised in the test tasks. Expert judgement is integrated in this analysis.

---

[2] In the listening part, the sections were retained as they were in the original version. However, the nature of the lectures was dramatically changed from a format close to dictation to more authentic lectures which were recorded live and edited keeping their natural format. Additionally, the first writing task was changed to graph interpretation. The second task was retained in its original format.

[3] Later versions of the BUEPT (starting from the January 2002 test) underwent further changes reflecting the approaches of newly appointed testing office members. Even though the test retains the same section titles it cannot be claimed that it still reflects the same approach discussed here.

Qualitative data from test takers' retrospection are incorporated into the analysis to cast light on the processes the test takers use while answering the test items. Statistical analyses on item quality and factorial make-up of four different versions of the test are discussed. The BUEPT September 2000 reading module was also compared with the reading module of a standardised English language test, namely, the IELTS's reading module. In order to make sound comparisons between the two tests, the comparability of the tests was analysed with a content analysis scheme adopted from Bachman et al. (1995).

1.4 Research Questions

As stated above, the present study aims at investigating the construct validity of the reading component of the BUEPT reading test, which includes scanning, skimming, search reading and careful reading at the global level components. Six research questions were formulated in order to investigate the five of the six aspects of construct validity framework suggested by Messick (1989a).

The first two research questions cover the content aspect of construct validity:

1) How is the construct defined and reflected in the test?
2) Do the experts agree on the operations measured by the test items as specified by the test writers?

The third research question corresponds to the substantive aspect of construct validity:

3) What are the operations utilised by the test takers to arrive at the correct answers?

The structural aspect of the construct validity is investigated by the fourth research question:

4) What are the dimensions of the reading construct measured by the test?

For the analysis of the generalisability aspect, the fifth research question is formulated:

5) Do the factor structures of the different versions of the test show similarities across versions?

Finally, the external aspect is investigated within the scope of the sixth research question:

6) What will be the relation between an established criterion measure and the test under investigation?

1.5 Overview of the Thesis

Following this introductory chapter, the thesis will present in Chapter 2 a review of the literature on the issues concerning construct validity and validation research, reading theories and taxonomies, and the controversy regarding the nature of the reading construct, with a detailed account of the reading model the test under investigation is based on.

In Chapter 3, the method used in the present study will be described. This chapter informs the reader of the development of test specifications and operations, empirical development of the test versions, statistical analyses used to investigate the research question, qualitative analyses of test performance and the comparison of the BUEPT

reading test with a criterion test. It is in this chapter that a detailed description of the validation research of the test will be given.

Chapter 4 will present the results and the discussion of the results concerning the issue of the componential nature of reading. Finally, in Chapter 5, a summary of the findings will be presented, the implications of the findings will be discussed, and the limitations of the present study will be considered with suggestions for future research to be put forward.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

It is discouraging to see how unclear the issues relating to the nature of the reading construct remain even though it has become test developers' motto to claim that 'a clear and explicit definition of language ability is essential to all language test development and use' (Bachman, 1990a, 3-4). In order to assess the construct- the ability we wish to test[1]- we need to know what the construct is, as Alderson (2000) states, but it is also prevalent in reading literature to assert that an unambiguous and impartial understanding of the reading construct has as yet remained impossible. Since the definition of a construct is closely related to the approach we choose to elicit the performance, the validity of our definition determines the validity of the test we use.

As mentioned in chapter 1, this study presents a systematic approach to the development of a construct valid EAP reading test with a priori and a posteriori procedures for test development and validation. To achieve construct validity in the test developed, primarily the construct of reading had to be defined both theoretically and operationally. This chapter discusses theoretical issues concerning construct validation and presents a review of literature on existing theories of the reading process (what happens during the reading process) and research literature concerning

[1] A construct is defined as a psychological concept which derives from a theory of the ability to be tested in Alderson (2000, 118).

the componentiality of the reading construct (whether reading can be divided into underlying skills for the purposes of teaching and testing). The last part of the chapter gives 'an expanded model of reading' on which the test in question is based.

2.2 Construct Validity

Bachman (1990a, 25-26) and Bachman and Palmer (1996, 21) define **validity** as a matter of the quality of the judgements made and test use on the basis of test scores. To repeat the definition of validity, Messick (1988, 33; 1989a, 13 and 1995a, 741)[2] states that validity is an overall, 'integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and the *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment'. Bachman also quotes from Standards for Educational and Psychological Testing:

> '... validity always refers to the degree which evidence supports the inferences that are made from the scores (the appropriateness, meaningfulnes, and usefulnes of the specific inferences)[3]. The inferences regarding specific uses of a test are validated, not the test itself.' (American Psychological Association, 1985 in Bachman 1990a, 237-238)

Therefore, validity does not derive from the content and the procedures of the test itself but is a function of the meaning of the test scores, how test takers and the context of assessment interact with the test (Bachman, 1990a; Messick, 1995a).

---

[2] See also Chapters 1 and 2 in Wainer and Braun (1988) for further discussions on validity.
[3] Parentheses added.

Construct validity for its part, is defined as 'the extent to which performance on tests is consistent with predictions we make on the basis of a theory of abilities, or constructs.' 'A construct (mental ability, a postulated attribute of people assumed to be reflected in test performance)[4] is defined in terms of a theory that specifies how it relates to other constructs and to observable performance' (Bachman, 1990a, 255). Therefore, construct validation can be seen as verifying or falsifying a scientific theory. In construct validation, constructs (the language abilities considered affecting test performance)[5] are defined theoretically and operationally, and these definitions are related to observations of behaviour; to test scores (op. cit.).

2.2.1 Messick's Framework

It is also important to note here that the current tendency in the field is to see validity as a unitary concept with content relevance, criterion relatedness and meaningfulness of construct being all complementary types of evidence feeding into adequate score interpretation and score use. Therefore, sources of validity evidence are unlimited: Investigating the relation of the content of the test with the domain of reference, individuals' responses' with the items and tasks, test scores' with other measures and background variables, and investigating the differences in these structures and processes over time, across groups and settings, and the impact of experimental interventions (teaching, treatment, manipulation, etc) provide evidence for validity. Intended and unintended social consequences of interpreting and using the test scores in particular ways are also pieces of evidence for validity (Messick 1989a, 16).

---

[4] cited from Carroll (1987) and Cronbach and Meehl (1955) in ibid.
[5] Bachman also points out that in defining the traits we intend to measure, we should be clear in distinguishing'traits' from 'test facets', which sometimes posits a very complex problem.

Messick (1989a, 20) in his seminal paper on validity discusses facets of validity in *progressive matrix* with a four-way classification (See Table 2.1).

Table 2.1 Facets of Validity as a Progressive Matrix (Messick, 1989b)

	Test Interpretation	**Test Use**
Evidential Basis	Construct Validity (CV)	CV+ Relevance/Utility (R/U)
Consequential Basis	CV+ Value Implications (VI)	CV+ R/U+ VI+ Social Consequences

This view conceptualises construct validity as one unitary concern, and it includes, beyond score meaning, relevance and utility, value implications, and social consequences within the boundaries of validity (Brown and Hudson 2002, 241).

Evidential basis for test interpretation is the empirical investigation of construct validity; this is explained by Messick as follows: 'the evidence and rationales supporting the trustworthiness of score interpretation in terms of explanatory concepts that account for both test performance and relationship with other variables' (1989a, 34). Convergent and discriminant evidence[6] should be available to check against the *construct underrepresentation* and *construct-irrelevant variance* in the test. Construct underrepresentation occurs when the test is too narrow and fails to include important dimensions or facets of the construct. In construct-irrelevant variance, the test contains excess reliable variance that is irrelevant to the construct. Two basic kinds of construct-irrelevant variance are *construct-irrelevant difficulty*, which occurs when aspects of the task that are not central to the construct make the test more difficult for some individuals or groups resulting in invalidly low scores,

[6] Convergent evidence comes from high correlations between similar measures that are assumed to tap on the same construct, and discriminant evidence is to show that these measures 'are not related to some other construct that could account for the intercorrelations' (Messick 1989a, 35).

and *construct-irrelevant easiness*, which appears when extraneous clues in item and test format help test takers to arrive at correct answers in ways irrelevant to the construct being assessed (as in the case with highly familiar test material). Construct-irrelevant easiness leads to invalidly high scores (ibid., 34-35, Messick 1995a, 742-743).

Evidential basis for test use involves appraisal of empirical investigations of both construct validity and the relevance of the scores to the applied purpose and the utility of the scores in the applied setting. Messick (1989a, 64) underlines that 'the unified concept of validity integrates considerations of content, criteria and consequences into a construct framework for testing rational hypotheses about theoretically relevant relationships, including those with an applied focus'. Construct-valid score meaning provides a rational basis to relate tests to criterion, to judge content relevance and representativeness, to anticipate potential testing outcomes and to detect possible side effects. Construct theories of performance domains and critical aspects of performance (designating knowledge, skills, cognitive processes, etc.) serve as a guide for test construction and evaluation, and also as a means for the appraisal of the meaning and adequacy of the criterion measures themselves. In applied settings, where tests are constructed and used, judgements of the relevance and representativeness of domain coverage and critical aspects of performance serve as criteria, and they are fundamentally value judgements. Judgements of content relevance and representativeness are critical for the development of test specifications to guide test development and for the evaluation of a test for its appropriateness to a specific purpose. However, the setting of test specifications and constructing items that are judged by experts to meet those

specifications is only a starting point. Whether items tap relevant knowledge or skill cannot be left to the judgements of the experts alone. Construct-related evidence of the type argued in evidential basis of test interpretation is also needed (ibid., 64-70).

Messick (1988, 41) stresses that the process of construct interpretation involves both 'theoretical context of implied relationships to other constructs (substantive or trait implications) and value context of implied relationships to good and bad, to desirable and undesirable attributes and behaviours'. Judgemental appraisals of the value implications of constructs and their associated measures provide c*onsequential basis of test interpretation.* According to Messick (1989a, 58-63), questions of score interpretations are closely linked with the questions of validity. Therefore, score-based inferences should be judged on the basis of their implications, not just in terms of what they entail but also in terms of what they make more likely. Values - meanings attached to attributes, actions and outcomes - can bias score-based inferences and actions, which in turn has consequences for individuals, institutions and society. However, this is not to say that science should be value-free; rather values should serve as subject matter for scientific investigation. Constructs, broader conceptual categories than test behaviours, bring a variety of value connotations in score interpretation from the value connotations of broader theories in which the construct is embedded, and from ideologies about the nature of humankind, society and science which determine our perceptions and actions. From this respect, it is important to chose consistent labels for constructs that capture as closely as possible the essence of the construct's theoretical meaning in terms of its salient value implications in that labels carry a range of implied theoretical and empirical referents. Theories might be dominated by different conceptions and attitudes about

the origins and development of a particular construct. They might vary in their emphasis on different determinants of the ability. The approaches to the study of the construct might differ in theoretical and methodological preferences conveying implied value commitments about the nature of the human being and human perfectibility. Moreover, ideologies give theories their perspective and influence theoretical conceptions and test interpretations in a way that goes beyond empirically grounded relationships in the construct theory. Ideological differences might result in different perceptions of value implications and test interpretation and use. In sum, scientific observations are theory-laden and theories are value-laden. Thus, scientific judgements such as those in test interpretation are value judgements. Exposing value assumptions of a construct theory and its links with ideology is a challenging task. However, if each construct theory is contrasted with alternatives, with antithetical counter-perspectives in a dialectical mode of examining and testing, they can be subjected to empirical grounding, policy debate or both. Value implications of test names, construct labels, theories and ideologies in terms of their relation to score-based inferences need to be supported empirically and justified rationally since these are socially relevant for score-based actions and they serve to link the construct to questions of social policy.

The last facet in Messick's framework is the *consequential basis of test use*, which is the appraisal of both potential and actual social consequences of the applied testing as an integral part of validity. The appropriateness of the intended testing purpose and all the possible occurrences of unintended outcomes and side effects are the issues of validity. For instance, sex or ethnic differences occurring as an adverse impact of a test used for selection would reduce the functional worth of the test if

these differences are not valid properties of the construct tapped by the test. In the same manner, the use of a particular test might lead to increased emphasis on memory rather than divergent thinking in learning and teaching, which is a case that might be evaluated as favourable or unfavourable depending on one's educational values. However, potential consequences of test use are virtually infinite and all of the critical possibilities especially unintended ones may not be identified. Once again, construct theory by articulating links between processes and outcomes, provides a rational basis for hypothesising potential outcomes and side effects (ibid., 86-87).

2.2.2 Aspects of Construct Validity

As mentioned before, validation as the most critical step in test development and use by which tests scores gain meaning is an on-going evaluative judgement process. The conceptualisation of validity given above interrelates considerations of content, criteria and consequences as fundamental aspects of a more comprehensive theory of construct validity that involves both score meaning and social values in test interpretation and use. As Benson (1998, 10) puts it, the process by which test scores gain meaning through construct validation is very similar to the way in which scientific theories are developed and evaluated. With observations and information from previous research, a theory of construct is formulated. In the theory, the relationships among the focal construct and other constructs are described and hypotheses and rival hypotheses are generated. These are tested and conclusions are drawn. The conclusions might require further observations of the behaviour, provide alternative explanations, indicate a revision, suggest additional hypotheses or support

the theory. In construct validation, the theory and the test are constantly evaluated and refined in an iterative manner. Cronbach (1989, in ibid) distinguishes between *weak* and *strong programs* of construct validation. The former relies primarily on empirical research through correlational studies. The latter is typified by the prominent role the theory plays preceding and guiding test development and validation. Three components of strong construct validation are determined as *substantive*, *structural* and *external* components (Nunally 1978, in ibid). The theoretical domain of the construct is specified and then operationally defined in terms of the observed variables (empirical domain) in the substantive component. As such, the empirical domain is a reflection of the theoretical domain. When the theoretical domain is well defined, the empirical domain will be easier to operationalise and facilitate developing measures of the construct. In the structural component, items are related to the structure of the construct by determining to what extent the observed variables relate to one another and to the construct. Such analysis usually involves correlational statistics. The external component refers to the questioning of whether or not the measures of a given construct relate in expected ways with measures of other constructs. Group differentiation studies in which groups putatively differing in terms of the construct are compared (differences might exist or can be created through manipulation) are typical of the studies of construct validation in external stage. Benson (1998) summarises the statistical and conceptual methods to obtain validity evidence in each stage of conducting a strong program of construct validation in a framework given in Appendix 2.1. Benson suggests this framework as a continuum - as opposed to three discrete stages - in which each stage leads to the next or reevaluated in the light of the evidence gathered. He also suggests that 'a strong program of validation would include all three stages and more

than one method in each stage' (1998, 15). Messick (1989a, 1995a, 1995b), on the other hand, suggests six distinct aspects of construct validation which would be an instrumental guideline in dealing with the complexities inherent in appraising the appropriateness, meaningfulness and usefulness of score inferences. In effect, these aspects are seen as general validity criteria or standards for all educational and psychological measurement (Messick, 1989a). They emphasise *content*, *substantive*, *structural*, *generalisability*, *external* and *consequential* aspects, which are briefed below as they are presented in Messick 1995a and 1995b.

1) Content relevance and representativeness: The *content* aspect of construct validity includes evidence of content relevance, representativeness and technical quality; specification of the knowledge, skills and other attributes revealed by the assessment tasks (specification of the boundaries of the construct domain to be tested). This can be done through job, task or curriculum analysis and especially through scientific inquiry into the nature of domain processes. Domain theory is the primary basis for the specification of the boundaries and structure of the construct to be assessed. Tasks selected for assessment should be both relevant to and representative of the domain. All important parts of the domain should be covered. The content relevance and representativeness of assessment tasks are traditionally evaluated by expert professional judgement.
2) Substantive theories, process models and process engagement: The *substantive* aspect refers to theoretical rationales and process modelling in identifying the domain processes to be revealed in assessment tasks and the observed consistencies in test responses. Substantive aspect goes beyond the content relevance and representativeness by adding the need for empirical evidence for

response consistencies or performance regularities reflecting domain processes. Think aloud protocols and eye movement analyses during task performance, correlation patterns among part scores can yield such evidence.

3) Scoring models as reflective of task and domain structure: The *structural* aspect refers to the consistency of scoring models with the structure of the construct domain. The theory of the construct domain, besides the selection or construction of relevant assessment tasks, should guide the rational development of construct-based scoring criteria and rubrics. Thus, the internal structure of the assessment should be consistent with the internal structure of the construct domain.
4) Generalisability and the boundaries of score meaning: The *generalisability* aspect refers to the extent to which score interpretations generalise to and across population groups, settings, tasks, time and raters. Generalisability depends on the degree of correlation of the assessed tasks with other tasks tapping the same construct.
5) Convergent and discriminant correlations with external variables: The *external* aspect of construct validity includes convergent and discriminant evidence from relationship of the assessment scores with other measures and nonassessment behaviours as implicit in the construct theory. This requires that constructs represented in the assessment rationally account for the external pattern of correlations, especially those between the assessment scores and criterion measures. Such empirical evidence confirms the utility of the scores for the intended purpose.
6) Consequences as validity evidence: The *consequential* aspect appraises value implications of score based interpretations and the rationale for evaluating the intended and unintended consequences of the interpretations. Social

consequences of testing may be positive or negative. Adverse consequences are likely to occur when scoring and interpretation are biased. Construct underrepresentation or construct-irrelevant variance are two major threats of test invalidity that may give way to negative impact on individuals or groups.

In sum, relevant evidence for testing the hypotheses can come from several sources: To quote Messick (1989a, 16),

> 'The basic sources of validity evidence are by no means unlimited. Indeed, if we ask where one might turn for such evidence, we find that there are only a half dozen or so distinct sorts. We can look at the content of the test in relation to the content of the domain of reference. We can probe the ways in which individuals respond to the items or tasks. We can examine relationships among responses to the tasks, items, or parts of the test, that is, the internal structure of test responses. We can survey relationships of the test scores with other measures and background variables, that is, the test's external structure. We can investigate *differences* in these test processes and structures over time, across groups and settings, and in response to experimental interventions – such as instructional or therapeutic treatment and manipulation of content, task requirements, or motivational conditions. Finally, we can trace the social consequences of interpreting and using the test scores in particular ways, scrutinizing not only the intended outcomes but also unintended side effects.'

Cronbach (1988) gives five different perspectives of validity to underline the significance of context in validation studies. Angoff (1988), in the same volume,

discusses the evolution of conceptions of validity. Moss (1992, 1995, 1998) reviews the consensus on validity conceptions emerging in the field. Johnson and Plake (1998) give a detailed analysis of shifts in theoretical validity perceptions as they are reflected in published test standards within the last 50 years. These studies, and many others which cannot be cited here, all conform to Messick's framework given above. Therefore, Messick's framework can be considered as the representative of modern view of validity, which has guided many arguments and practises in the measurement field. However, it should also be mentioned that the framework does not go without criticisms. Although Messick's perspective does probe value implications as a means of addressing the consequential basis of test interpretation and use, it is still considered to be highly tied to the positivistic/psychometric position, which requires a detached and neutral stance toward research and inquiry. Linn, et al. (1991), Baker, et al., (1993) and Linn (1994) propose 'the content quality' and 'transfer' criteria to possibly extend Messick's framework. Moss (1994, 1996) extends the discussion beyond the traditional-psychometric paradigm, incorporating the notions of performance assessment. She argues for a *hermeneutic* approach to validity, which emphasises the role of prior understandings, values and contextualised knowledge of the researcher in the analysis and interpretation. Moss (1994) questions the place of 'necessary but insufficient' reliability, potentially constraining innovativeness in education.[7] However, the unified view of validity as reflected in Messick's framework has guided substantial discussion and research both in education and measurement fields. It provides a systematic guideline for construct validation research and for the evaluation of the research findings. It has been influential in the field of language testing as well.

---

[7] See Messick (1994) for his arguments on performance assessments.

2.2.3 Validation in Language Testing

As validity is a central concern in educational measurement, it is of utmost importance in foreign language testing as well since definition of validity primarily determines what constitutes a good test. The implications of changing perspectives of validity have been reflected in several studies in the language testing field within the last decade. The most influential work among them is that of Bachman's (1990a). Bachman integrated AERA/APA/NCME Standards[8] and Messick's framework to discuss validation and validity evidence in communicative language testing. In line with Messick, he considered construct validity as principal. Bachman presented validation as a process through which a variety of evidence concerning test interpretation and use is gathered. For construct validation, he states that correlational evidence (in terms of test dimensionality), experimental evidence regarding the effects of treatment, analysis of test-taking processes, group differences and changes over time constitute support for evidential basis of validity. For consequential basis of validity, evidence that will be gathered should relate to construct validity or evidence that supports particular test interpretation, multiple perspectives of value systems from test takers, test developers and test users, practical usefulness of tests as well as the ethics and misuse of test and social consequences of test invalidity (Kunnan 1998a).

Chapelle and Douglas (1993) in their evaluation of the dimensions of communicative language testing in Canale's framework at the beginning of 1990's, underline the importance of the unified view of validity and multiple approaches to evidence

[8] The acronyms stand for American Educational Research Association, American Psychological Association and the National Council on Measurement in Education, respectively.

collection in attributing meaning to test scores. The collection of papers from the 14th Language Research Colloquium in Cumming and Berwick (1995) also reflects that several studies in language testing have taken the same approach to validity or address at least one aspect of the Messick's framework. Chapelle and Douglas' recent chapter in Bachman and Cohen (1998) emphasise the place of language tests as elicitation devices in second language research. Particularly, Chapelle (1998) stresses that construct definition and validation of tests used as elicitation devices should be evaluated since learners' performance on language tests is used to make inferences extending beyond the observed performance. As learners' interlanguage is of a changing nature (Bachman, 1998 and Tarone, 1998), researchers need reliable descriptions of language at its various stages of development. Among three approaches to construct definition (trait, behaviourist and interactionalist approaches), the interactionalist approach, emphasising performance as a sign of underlying trait but influenced by the context in which it occurs, has been supported empirically especially by the communicative competence perspective.[9]

Hamp-Lyons and Lynch (1998), analysing abstracts of papers presented at the Language Testing Research Colloquium (LTRC) since its beginning, investigated whether language testing research has been dominated by the traditional paradigm or whether alternatives in reliability and validity considerations have taken roots. Kunnan (1998a) and Chapelle (1999) review language assessment and testing research reflecting validity themes from various lines of inquiry from the post-1980

[9] Chapelle (1998, 44) specifically refers to Bachman's (1990a) general interactionalist construct definition of communicative language ability which includes both knowledge, or competence, and the capacity for implementing that competence in context.

period. Kunnan categorises research focus to fit into Messick's framework[10] and Chapelle discusses six aspects of validity which cover the essentials of the same framework.[11] The following sections will summarise exemplary research studies in chiefly but not exclusively second language reading assessment under the headings of Messick's four categories. The majority of the studies that will be mentioned below relate to evidential basis of test interpretation, that is, to certain aspects of construct validation. Obviously, construct validation with its various aspects and primacy for all the four facets of validity has an outstanding weight in language testing, too.

2.2.3.1 Evidential Basis of Test Interpretation

Kunnan (1998a, 3-5) categorises research that falls into Test Interpretation section under Evidential Basis into four areas: language proficiency components, test dimensionality, test validation process, and test development. It is obvious that studies aiming at one of the research focuses above would necessarily have recourse to the others since in many cases all four concerns are interdependent as suggested by the unified view of validity. The summary given below intends to illustrate recent research in the language testing field as it relates to these four categories briefly. Nonetheless, it is obviously far from being a complete survey.

The basic theme that overruns early discussions on language test validation was evidently the nature of language ability. Differing views of language ability – Oller's

[10] See Appendix 2.2.

[11] Chapelle (1999, 260-262) categorises the approaches to validation in language testing as 'content analysis, empirical item or task analysis, dimensionality analysis, relationships of test scores with other tests and behaviours, differences in test performance, testing consequences'.

(1976) unitary trait hypothesis and Canale and Swain's (1980) three-component model of communicative competence[12] together with Cummin's (1980, 1984 in Verhoeven, 1992) CALP (cognitive language proficiency) and BICS (basic interpersonal communicatice skills) dichotomy – triggered considerable discussion. One of the early works that drew on Canale and Swain's theoretical model of communicative competence and utilised confirmatory factor analysis was Bachman and Palmer's study (1982), in which the researchers identified more than one factor in the language proficiency data. More recently, Harley et al. (1990) investigated whether the three key components of language proficiency – grammatical, discourse and sociolinguistic competence – would emerge as distinct components of second language proficiency. They concluded that their findings provided only tentative support for the hypothesis that these constructs are distinguishable. Bachman (1990b) in discussing the methodology employed in the project pointed out that inaccuracies in the statistical methods and problematic operationalisations of the traits contributed to the inconsistencies in the findings. However, Canale and Swain's model has been useful for advancing the field's knowledge. Bachman (1989, 1990a, 2001) further developed Canale and Swain's (1980) framework to include two main components: organisational and pragmatic, the former consisting of grammatical and textual competence and the latter of illocutionary and sociolinguistic competence. Bachman's framework of communicative language ability has been very influential in the field in that it is seen as being comprehensive with measurable components and consistent with modern linguistic theory (Skehan, 1991). For example, Bachman et al. (1988) and Bachman et al. (1995) use this framework to evaluate the range of language abilities measured by Cambridge and TOEFL proficiency test batteries in

[12] The three components in the model are grammatical, sociolinguistic and strategic competence. Discourse competence was later added to the model.

order to reveal the comparability of the constructs of communicative language proficiency as they are measured by these tests. Wu and Stansfield (2001), as well as several examples given in Bachman and Palmer (1996), adopt the same framework in test development. Fouly et al. (1990) suggest that the notion that language abilities are divisible can be explained by two different hypotheses: the correlated-trait hypothesis, which states that separate traits underlie language test performance and these are correlated with each other, and the higher-order hypothesis, which states that these traits are separate and influenced by a single higher-order factor. Fouly et al. using improved statistical procedures found that both models were accounted for equally well by their data. In line with Messick (1989a), Fulcher (1998), too, emphasises the importance of models of language in validation of language tests on the grounds that they provide a framework for writing test specifications and a guide for test score interpretation. Fulcher investigates the validity of Widdowson's (1983, in ibid.) discourse model of communicative competence and performance as a basis for developing tests of reading. The study has been able to find empirical evidence for the highest level of processing in the model, i.e. procedural level. Yet, it has suggested that the levels of communicative capacity and linguistic knowledge may be underdefined in the model. Another study that focuses on the role of underlying performance capacities in second language performance is that of McNamara's (1996). McNamara (1995, 1996) discusses that present models of language ability should be broadened to address the interactional aspect of performance. Skehan (1998) also suggests that second language learning theory should increase its scope by incorporating more psycholinguistic perspectives, which emphasise information processing and cognitive abilities. Skehan (1998) and Skehan and Foster (2001) propose a 'dual-mode' model for task-based instruction and task-based performance

assessment, which would be an alternative to previous models that neglect the processing and contextual sides of language use, i.e. how abilities are mobilised in actual performance in certain contexts. Norris et al. (1998) exemplify a similar approach to language assessment, but one which focuses on task as the fundamental unit of analysis, which necessarily reduces the construct of interest to the performance on the task itself.[13] On the other hand, Chalhoub-Deville (1997),[14] in her review article, points out that L2 proficiency models are numerous and involve diverse components, a fact which reveals the lack of consensus among models in their representation of proficiency. She comments that seeking an ultimate model may be an unattainable goal, and therefore operational models which portray the construct at a contextual level can be more accessible and practical.[15] Recent examples to operational models may be those which are formed for TOEFL 2000 by Chapelle et al. (1997) and Enright et al. (2000). Such arguments on the nature of language ability and the adequacy of the theoretical models proposed are also valid for reading research in particular. These are discussed in section 2.3 in detail.

Related with the conception of language proficiency and its components, research on test dimensionality is also substantial to validity research. Recent advances in statistical tools have extended the limits of statistical analyses into the validity and especially dimensionality research. Henning et al. (1985), Henning (1992), McNamara (1991) and Choi and Bachman (1992) discuss and illustrate the potential of the Item Response Theory (IRT) for investigating aspects of validity of language

[13] See Bachman (2002) for the problems relating to task selection, generalisability and extrapolation in task-based language performance assessment.
[14] See also Alderson and Clapham (1992) for earlier comments on the lack of consensus among applied linguists as to the nature of language proficiency.
[15] See also Widdowson's (2001) comments on communicative competence models as they lead to discrete view of language components.

tests. Kunnan (1992), Bae and Bachman (1998), Brown and Ross (1998) and Lynch and McNamara (1998) employ 'Generalisability Theory' (G-theory)[16] and factor analysis to investigate the properties of several language tests. Lee (1996) demonstrates that FACET, a multifaceted IRT analysis, can account for item calibration and group comparisons effectively with EAP tests. Clapham and Corson (1997) devote a full section to the quantitative and qualitative validation of tests. Kunnan's (1998b) article on Structural Equation Modelling (SEM) is also informative in that it presents various applications of this advanced statistical method in language assessment research. The statistical methods mentioned above have been effectively used both in the componential analysis of the reading skill and item performance in reading tests, too.[17]

As second language testing increasingly makes use of tasks instead of objectively scored test items, rating scales are widely used in the assessment and reporting of language test performance. The nature of rating scales forms an important issue in the discussions of language proficiency as most rating scales are assumed to measure underlying language competence, the language construct or the constructs that form the language ability (e.g. the ACTFL). Therefore, their reliability and validity should be evaluated with careful consideration of the construct they purport to measure. Rating scales, which usually describe language behaviour in skill areas in terms of the degree to which a learner can achieve a language goal, vary in their content and specificity. Alderson (1991) categorises scales used in language proficiency assessment according to the purposes they fulfil. The ones that provide information about typical or likely behaviours of candidates at a given level are user-oriented

[16] See Bachman (1997) for extended discussion of Generalisability Theory.
[17] See section 2.3

scales. The ones that provide guidance for assessors who are rating performances such as writing and speaking assessment scales are assessor-oriented scales. The scales that serve the function of guiding the construction of tests as in the case with the ACTFL are construct-oriented scales. Alderson states that the users of rating scales need to be clear about the purpose of a scale since problems arise when the three functions are confused. He illustrates in the frame of the IELTS Revision Project that when performances described in the scales are not actually those that are elicited in the test, the validity of scales is brought into question. Bachman (1990a) and Bachman and Palmer (1996) raise questions about the generalisability and accuracy of the scales widely used in the field and suggest that scales should be based on a solid theoretical ability model. Brindley (1998a) also discusses that most rating scales are ad hoc productions describing actual performances of language users in a specific context rather than operationalisations of second language acquisition processes. As they stand, they may be helpful for reporting program outcomes, nevertheless the generalised descriptions of levels of language proficiency as in most rating scales are oversimplifications. There are several studies revealing the weaknesses of rating scales and conceptualisations of language performance underlying their descriptors. The ACTFL has been criticised widely in that respect. For example, Lee and Musumeci (1988) question the validity of the reading model given in the ACTFL and find no evidence in support of the three-way (text type, reading skill, task-based performance) hierarchical conceptualisation of the reading construct on which the scale is based. Fulcher (1996a) shows that statistical findings reported in Dandonoli and Henning (1990, in ibid.) to present evidence for the validity of the ACTFL are inadequate. Fulcher points out several methodological inaccuracies in Dandonoli and Henning and his factor analysis reveals counter

evidence for the validity claims for the ACTFL. Brindley (1998a) suggests that scales should be empirically developed informed by second language acquisition research and validated taking into consideration text, task and content characteristics along with psychological dimensions. Upshur and Turner (1995) also discuss the reliability and validity problems with rating scales but illustrate an effective procedure to develop an oral performance rating scale empirically. Their rating scale is based on binary questions (whether a certain characteristic exhibited by the test taker or not) tapping the behaviour at the boundaries of levels of performance. Fulcher (1996b) argues that detailed theoretical descriptions and qualitative evidence should underpin scale development and scales should be revised in the light of evidence in an iterative manner. He compares the occurrences in his data of the explanatory categories relating to fluency description (key features) to the descriptions in the band scales of the IELTS through discriminant analysis, the result of which helps refine definitions for bands in return. North and Schneider (1998) use the item-banking technique to produce a language proficiency scale. They pooled in descriptors from various scales and refined them through consultations with teachers. Their work is substantial in that it draws together theory, accumulated knowledge relating to proficiency scales, employs rigorous quantitative and qualitative analyses and is empirically tested across languages. Most importantly, it has formed the basis of the scales of the Common European Framework (Council of Europe, 2000), which is intended to form a common basis for the learning, teaching and assessment of languages across Europe.

Although most of the studies mentioned at the beginning of this section are related with test development and validation, if one has to exemplify test validation research

per se, Criper and Davies (1988), McNamara (1990), Shohamy and Inbar (1991), Brown (1991), Wall et al. (1991, 1994), Kunnan (1993), Cushing Weigle and Lynch (1995) and Cumming and Mellow (1995), Clapham and Alderson (1997), Fulcher (1997), Brindley (1998b), Guerrero (2000) can be listed. Particularly, in reading research, Bachman et al. (1988) present a detailed analysis of the range of reading abilities measured in Cambridge tests (Certificate of Proficiency in English-CPE and First Certificate in English-FCE) and the Test of English as a Foreign Language (TOEFL). The researchers evaluate the construct validity of the tests by investigating the content relevance and content coverage with reference to Bachman's (1990a) theoretical framework of communicative language ability (CLA). Their analysis also involves description of the test methods facets and tasks that are included in the tests and quantitative analysis of test performance. Bachman et al.'s 1988 study is a preliminary attempt to use CLA framework for the comparison of content across tests and to provide both systematic qualitative and quantitative evidence for construct validity.[18] Anderson et al. (1991) is the first study to combine information on test taking processes with information on item content and item-test performance in the construct validation of an EFL reading comprehension test. They used data gathered from retrospective think-aloud protocols of the test takers' reading and test taking strategies, data from content analysis of reading passages and questions, and traditional test performance statistics. Stemmer (1992) illustrates the use of introspective and retrospective techniques for construct validation of C-tests. Pierce (1992, 1994) gives a detailed account of the test development process for TOEFL reading comprehension items. The article reveals how test specialists' judgements are incorporated with statistical data at several levels in item development to ensure

[18] Bachman et al. (1995) more recently use the more refined form of the CLA framework for the comparison of full forms of FCE and TOEFL tests.

the construct validity of the reading test. Freedle and Kostin (1993a, b) analyse item difficulty of a large sample of the TOEFL reading comprehension items and examine whether text and text-related variables play a significant role in item difficulty. They suggest that their positive findings support their claim that multiple-choice reading items yield construct valid measures of comprehension. Schedl et al. (1996), checking the validity of TOEFL specifications, examine whether the TOEFL reading comprehension items categorised in test specifications as 'reasoning items' test any abilities not measured by other item types. While their findings did not support that the items in question measure a unique construct, all the test forms used in the study appeared two-dimensional. Further analysis suggested that the second factor may be related to passage content or position (end-of-test effect). Buck et al. (1997) investigate the attributes, i.e. the knowledge, skills and strategies that underlay the Test of English for International Communication (TOEIC) reading comprehension part. Suggesting that taking reading tests 'involves skills over and above those involved in simple reading' (ibid. 429) especially in multiple-choice format, the researchers generated a large list of attributes and examined the relationship between the attributes and test performance through rule-space analysis. The attributes being the component parts of the construct, Buck et al. suggest that the information that rule-space analysis provides helps in construct validation of tests. Clapham (1997) demonstrates the revision and validation of the IELTS reading module including drafting test specifications, content validation through incorporation of multiple expert judgements, and discussions of predictive and construct validity. Khalifa's (1997) study, too, is an example of construct validation of a reading test. The study illustrates how a test should be based and operationalised on a theoretical foundation, a reading framework in this case, and how experts' and test-takers' judgements are

incorporated into the investigations of construct validity. Through Principal Component Analysis, the study also provides evidence for the dimensionality of the reading construct. Another study that should be mentioned is that of Weir et al. (2000). Weir et al. present a methodical approach to test development and validation especially with its emphasis on a priori validation which involves target situation analysis, theoretical discussions and analysis of EAP reading course books and tests. The study also reveals the importance of the systematisation of test development process through methodical text selection and item writing. Considering the fact that much of the validation of a test must be established before the test is actually administered, it suggests that both qualitative and quantitative analyses must be done before the test is actually administered. In terms of the construct representativeness of tests, Alptekin (1999, 2000) also cautions that construct valid reading tests should reflect cognitive processes underlying reading and exclusion of items tapping processes at either micro or macro level would risk construct invalidity.

At the core of reading research and the validation investigation with reading tests is the issue of the effect of test method on the performance of test takers. Test method refers to 'the tasks used to elicit reading performance and encompasses both the content and format of the assessment instrument' (Riley and Lee 1996, 173). Research on test methods is essential to validation inquiry since it is well-established in the field that performance on language tests varies both according to an individual's language ability and to the characteristics of the test method (Bachman 1990a, 113). One main concern in relation to test method is the extent to which the type of response a test taker is supposed to give to a task in a reading test influences his or her interaction with the text. Bernhardt (1991, 193) suggests that a successful

assessment tool of reading ability must be integrative in nature and it must 'examine the extent to which a text actually communicates a coherent message to the reader'. In this respect, tests tapping on discrete points in a text cannot reflect the possible comprehension to the full extent. Criticising popular test methods such as cloze, multiple choice, direct content questions, she asserts that immediate recall protocols[19] are purer methods of measuring comprehension since there is no interference from the tester to the reader's reading process. Deville and Chalhoub-Deville (1993) also suggest that written recall protocol has advantages over traditional tests of reading since the task requires test takers to construct as complete an understanding of a passage as possible. Nevertheless, Alptekin (2004) deems recall protocols to be heavily dependent on explicit memory, which is often not part of our memory for texts and stories, and therefore not appropriate for the assessment of inferencing in reading. For Alptekin, it is also possible that comprehension in the L2 might be confounded with the production ability in the L2, or even L1 in those cases where the subjects use their L1 in recall protocols.

Secondly, as a constructive integrative task that might be an alternative to much criticised discrete point test items, summarising has received considerable attention in reading research. However, the problems with reliability prevent summarising to be used more widely. Cohen (1994) discusses problems with summarising (inconsistent responses, direct copying, unreliable scoring, scorer training, etc) and suggests (Cohen, 1993) that detailed instructions (specific guidelines) might help remove some of these problems. Interestingly, although he finds that guided instructions had positive effects on the summarising of foreign-language texts, the

[19] Recall protocol can be defined as the procedure of writing down whatever one remembers after reading a text without any recourse to the text, which is usually in the native language.

results were less clear with native-language texts (Hebrew). Wolf (1993) compares assessment tasks mostly used in comprehension research. She finds that among the three methods - multiple choice, open-ended questions and rational deletion cloze - her subjects could reflect their comprehension (post-reading) best in multiple choice tasks both in their native language and the foreign language. Her findings confirm the hypothesis that both assessment task type and the language in which subjects are tested can affect experimental findings. Huhta and Randell (1995) also compare several methods of reading comprehension test methods (i.e. conventional multiple choice, open-ended question, summary, multiple choice summary). They propose multiple choice summary[20] as a practical alternative to more laborious techniques that aim to measure the comprehension of main content of a text. However, the researchers report that the reliability problems associated with conventional summary tasks persist in multiple choice summary tasks, too. In line with Bernhardt (1991), Riley and Lee (1996, 172-173), too, mention the validity problems with discrete point items reflecting a fragmented, compartmentalised reading of a text and suggest that 'valid measures of reading comprehension consist of integrative tasks that reflect the constructive processes involved in reading'. They compare two global response modes, the recall and summary protocols and report that when the test takers were instructed to write summaries focusing on main ideas, their texts involved more main ideas written more coherently as opposed to free recall texts in which the test takers wrote down everything they can remember. Confirming Bensoussan and Kreindler (1990), they suggest that summarisation may be an effective technique to encourage students to read at a more global level and to form text base in a more efficient manner. Hudson (1996), reviewing recent advances in reading theory and Messick's

[20] In multiple choice summary, test takers are asked to select the best summary among alternatives by justifying their choice.

validity conception, suggests that there is a clear need for expanding beyond multiple choice format. According to Hudson, task based items requiring constructed response should be used in TOEFL 2000 in order to avoid construct underrepresentation. Kobayashi (2002) investigates test method (cloze, open-ended, summary writing) and text type effects (association, description, causation, problem-solution) on reading comprehension test performance. She finds interacting effects of both test method and text type on test takers' performance and suggests that test writers should take these factors into consideration to enhance the validity of their tests.

There are also several studies that focus on differential effects of taking tests on computers and feasibility of different test methods used in reading tests delivered through this medium. For example, Henning et al. (1993) investigate whether the computer adaptive open-ended testing of reading comprehension might prove advantageous over multiple-choice questions. It is suggested that only when the questions are scored for degree of correctness or incorrectness, the open-ended test format reveals good statistics. However, scoring items for the degree of correctness is not an easy process and requires frequent piloting and manual checking. Young et al. (1996), Chalhoub-Deville and Deville (1999) and Sawaki (2001) address various issues regarding the effects and comparability of different modes of presentation, content balancing, the nature of tasks and construct validation in computer adaptive tests. Issues in computer-adaptive testing of reading proficiency is a recent volume that consists of discussions of several theoretical aspects of computer adaptive tests (Chalhoub-Deville, 1999).

A number of studies have also looked into item characteristics that might affect item difficulty. Perkins and Brutten (1988) and Anderson et al. (1991) investigate the functioning of textually explicit, textually implicit and scriptially implicit items. The former found that the three item types are significantly different in terms of item discriminability but scriptially implicit items, which depend heavily on background knowledge, do not discriminate between test takers adequately. The latter, on the other hand, found no significant relationship between the test item types and the level of item difficulty. Furthermore, Perkins and Brutten (1992) compare factual, generalisation and inference reading comprehension questions assuming that these types of questions assess different skills and entail different processing depths. The researchers found significant differences in the means of three types of questions, however, contrary to their expectations, the mean for the generalisation questions was slightly lower than that for inference questions. Perkins (1992) finds a difference in the item difficulty of the questions according to the type of topical structure on which questions are based. Freedle and Kostin (1993a, b) identify twelve text and text-by-item (text and item interaction) variables, which they found to affect item difficulty of the multiple choice items in the reading part of the TOEFL. Their findings suggest that these variables provide independent predictive information as to the difficulty of the items, and support the construct validity assumption of TOEFL multiple-choice items. Bachman et al. (1996) compare the content characteristics of the reading items from six forms of FCE using Bachman's (1990a) framework and report some preliminary evidence for a relationship between content analysis ratings by trained specialists and statistical values of item difficulty and discrimination. Fortus et al. (1998) identify eight text-related and eighteen item-related variables that may account for the item difficulty in an EFL reading test which comprises sentence

completion, restatement and reading comprehension type of items. Their judges attained a high level of agreement when text and item related factors were evaluated jointly in relation to item difficulty. Rupp et al. (2001) investigate the item difficulty of computerised reading and listening items by combining two statistical analysis methods, multiple regression and classification and regression tree (CART). The researchers used Freedle and Kostin's (1993a, b) text and item-text interaction variables but synthesised the results from two statistical methods to understand the relative contributions and interrelations of variables that affect item difficulty. The researchers claim that combining two methodological perspectives provided a clearer support for their construct definitions.

Another issue related with test method is whether test takers use the operations or skills that test items purport to measure while responding to them. The issue is relevant in some of the work cited above and will be detailed in section 2.3.

2.2.3.2 Evidential Basis of Test Use

Studies trying to form an 'evidential basis for test interpretation', i.e. construct validation relating to score interpretation of tests abound in language testing and testing of reading in particular. However, a clear consensus as to the what reading construct is has not been reached yet. One obvious explanation for this is that assessing reading comprehension is problematic in that the process by which readers create meaning from texts is an invisible process. Besides the issues mentioned in the previous section, language testing researchers have looked into phenomena such as test taking processes, test taking strategies and test taker characteristics to have a

deeper understanding of the reading and test taking processes. These, according to Kunnan (1998a), form 'the evidential basis of test use' in Messick's framework.

Verbal protocols are widely used to supplement quantitative data in language testing for the judgements of validity of the instruments, especially to evaluate the skills or cognitive processes that test questions are purported to measure. Despite the criticism that they may be an incomplete reflection of actual internal processing, it is now generally believed that with the data from verbal reports, we can obtain more accurate, valid data on cognitive processes (e.g. Matsumoto, 1993). For example, Nevo (1989) used both introspective and retrospective reports to analyse the processing of reading comprehension tests in L1 and L2 in relation to the cognitive (contributory vs. noncontributory) strategies used by the test takers when responding to multiple choice test items. Anderson et al. (1991) note that information on the test taking process is a part of construct validation and investigate the test taking strategies as well as item content and item performance while Alderson (1990a, 1990b) uses qualitative data to argue against the multidivisible nature of the reading construct (See section 2.3). Gordon and Hanauer (1995) investigate the interrelationship between meaning construction and testing tasks using think-aloud data. They suggest that readers' processes are differentially affected by different reading tasks (multiple choice and open ended questions) since the tasks are additional sources of information which interact with the readers' development of the mental model in the course of test taking. Wijgh (1995) compares the reading strategies of students as they process different text types and evaluates whether they are desirable or not (as designated by the researcher) both through protocol analyses and observation. Storey (1997) analyses the cognitive processes underlying the

completion of discourse cloze tests through the think-aloud protocol, and he compares the observed cognitive processes used in the test completion with the ideal performance that is indicated by a model of reading process.

Closely linked with test taking processes, test taking strategies have received much attention in language testing research, too. Anderson (1991) gathered forty-seven processing strategies and, through verbal reports, analysed adult L2 readers' differential strategies as they responded to a standardised multiple choice reading test and as they read longer academic texts. Cohen (1998a, b) discusses strategies for taking a multiple choice reading comprehension test in detail. Purpura (1997, 1999) employs sophisticated statistical methods to analyse the relationship between test takers' cognitive and metacognitive strategy use as reported by test takers and their performance in a FCE test. Phakiti (2003) uses the questionnaire that is developed in Purpura's study to investigate the same issue; however, he supports his findings by additional retrospective data. Beyond the research cited, an interesting study in test taking strategy quest is Allan's (1992) test-wiseness instrument (TOTWESL). Allan notes that test-wiseness is an important source of test construct invalidity since scores of some learners may be influenced by skills which are not the focus of the test. He suggests that TOTWESL might be used for diagnostic purposes.

Test taker characteristics, the next issue to be discussed in relation to 'evidential basis of test use' cover a wide range of variables that may affect test performance therefore score interpretation. These variables are considered to form the nomothetic span of the construct whose investigation is required to secure generalisability. Alderson and Urquhart (1985) investigate the effects of students' academic discipline

on their reading test performance. Hammadou (1991), Clapham (1996) and Rigway (1997) are some studies exemplifying the investigation of the effect of background knowledge on reading comprehension. Kunnan (1994, 1995) analyses the relationship between cultural background and previous instruction and test performance. Among the studies that investigate differential item functioning (DIF) in language tests, Ryan and Bachman (1992) examine the differential functioning of TOEFL and FCE reading test items for test takers of equal ability but different native language and curricular backgrounds. Elder (1996) compares groups of Italian, Chinese and Greek native speakers in terms of score differences and DIF of Australian Language Certificate items. Brown (1999), too, investigates the relative contributions of different language background test takers to test variance as a part of a detailed study on the TOEFL. Sparks et al. (1998) examine the effect of native language skills and foreign language aptitude on foreign language grades in different level proficiency groups. Besides these, Chapell (1988) discusses field independence as a possible source of language variation and, therefore, language test performance. Alptekin (1991) also notes that individuals vary in their hemisphericity and they might be differentially affected by test tasks demanding differentially lateralised functions of cognition.

One important aspect of evidential basis of test use, which is not covered in Kunnan (1998a), is the fact that Messick (1989a) classifies 'content related evidence' within this category. Fulcher (1999a, 20) underlines that 'content validity' evidence should be understood as 'evidence for the relevance of the test to the specific and applied purpose and for the utility of the test in applied setting'. Messick (1989a, 41) states that 'content validity does not reside in the test, but in the judgement of experts'.

This opens up a whole discussion on EAP/ESP and the needs analysis studies that formed the basis of many EAP/ESP tests and, consequently, authenticity issues. Clapham (1993, 2000), Alderson (1999), Douglas (2000) and Hamp-Lyons and Lumley (2001) problematise such issues as whether EAP/ESP testing is justified, and whether authenticity is a validity argument or not. Since argumentation on such substantial issues needs extended discussions, as is the case with many testing issues, we leave the matter without further discussion here. However, it should be mentioned that expert judgements relating to content analysis of tests and test tasks also provide evidence for 'test use'. Studies reporting expert judgements have been mentioned elsewhere in the chapter (e.g. Bachman et al. 1988, Alderson and Lukmani 1989, Pierce 1992, Perkins and Brutten 1992, Upshur and Turner 1995). Alderson's 1993 chapter presents a thorough critique of using expert judgements in test validation.

2.2.3.3 Consequential Basis of Test Interpretation

Kunnan (1998a, 6) states that the studies focusing on obtaining 'feedback from test takers regarding tests they have taken' and 'feedback from university subject matter specialists' regarding test content/method and test appropriacy are the studies that fall into the 'consequential basis of test interpretation' category. Such studies are not many. The following are some examples. Lumley (1993a, 1993b) investigates teachers' perception of subskills of reading comprehension test items. Lewkowicz (1997) gets feedback on the authenticity of listening test tasks from native and nonnative experts. She also asks the test takers which task (multiple choice or integrated test task) indicated their ability to use the language in the real world,

finding widely differing views among both groups of respondents. Norton and Stein (1998) discuss how a reading passage might have an unexpected reaction from test takers because it had culturally controversial meaning for the readers which was not foreseen by the test developers. The case presented in the study is an example of the fact that tests and textual meanings are socially constructed and different perceptions of test developers and test takers might end in a fundamental validity paradox. Weir (2001) questions the formative value of formal testing in the eyes of stakeholders. Administrators, teachers and students were asked their views on the value of language testing in the classroom. His data suggests 'lack of feel-good' concerning testing. A perception study in reading research comes from Lin (2002). Lin investigates the EFL learners' perception of prior knowledge and its roles in reading comprehension. He reports that for lower proficiency subjects, the knowledge of vocabulary and syntax is perceived as important. However, as the language level of the subjects increases, they place less importance on linguistics knowledge. They rather deem conceptual and sociocultural knowledge as more important.

2.2.3.4 Consequential Basis of Test Use

Especially stimulated by Messick's framework, the concern for the consequential basis of test use expanded in the 1990s. Under the category of 'consequential basis of test use' issues relating to 'washback effect' and 'ethics and standards' and 'professionalisation of the field' are discussed.

Washback (backwash or impact) is defined as 'the influence that writers on language testing, syllabus design and language teaching believe a test will have on the

teaching that precedes it' (Alderson and Hamp-Lyons 1996, 280). The commonsensical understanding of it among most practitioners is simply that 'good tests have beneficial influence on teaching'. However, several studies have revealed that not always is there such a direct relationship between tests and teaching practices. Alderson and Wall (1993) point to the inadequacy of the notion that tests are powerful determiners of what happens in the classroom, since other factors such as teachers' competence, their understanding of the principles underlying the test, the school system may influence teaching as reflected in the researchers' Sri Lanka project. Wall (1996) revises the possible reasons why the Sri Lanka project had little impact on the methodology the teachers used from a broader perspective of 'innovation theory' and suggests that it may take a long time before any impact surfaces. It is also a fact that many factors relating to resources, society and the perceptions of practitioners and policy makers have a direct impact on the test being implemented in an educational setting. For example, Alderson and Hamp-Lyons (1996) observe TOEFL preparation courses and interview teachers and students to identify whether such a preparation course leads to negative washback only to find that teaching for the TOEFL might change according to teachers and their understanding of 'teaching for the test'. Shohamy et al. (1996) compare the washback effect of two different tests (Arabic SL and English FL) and note that impact might change according to whether tests are high or low stakes, and according to the status of the language being tested, and also to the purpose, format of the test and skills being tested in it. Saville (2000) underlines the distinction between 'washback' and 'impact' – washback being the effect of tests on language teaching and learning, and 'impact' being the effect on educational processes and society in general – and exemplifies a comprehensive research project that attempts

to deal with several aspects of the impact of IELTS. Bailey (1996) suggests that in order to have beneficial washback, the purpose of the test and meaning of the scores must be explicit for the stakeholders, tests should be authentic and reflect learning objectives, and should be based on a sound theory. Messick (1996, 242), once more underlining the essentiality of construct validity, maintains that 'a poor test may be associated with positive effects and a good test with negative effects', and such consequences should be related to 'good or bad educational practices apart from the quality of the test'. For Messick, 'one should concentrate first on minimising construct under-representation and construct-irrelevant difficulty in the assessment. That is, *rather than seeking washback as a sign of test validity, seek validity by design as a likely basis for washback.*' (ibid, 252; italics original). Therefore, unfavourable results associated with valid tests 'are not the test makers' responsibility' (ibid, 253).[21] However, Hamp-Lyons (1997) proposes that consequences of test use have impact on the society as a whole and this is not only the test makers' but also testing agencies', textbook publishers' and administrators' responsibility. Shohamy (1997) and Lynch (1997) caution about unethical uses of tests and also hold both testers and test users responsible for any negative consequence that a test may lead to. Working in the perspective of critical social theory and seeing tests as a means of political and social control, several researchers pondered on the ethical and unethical practices of language tests (e.g. Hill and Parry 1992, Tharu 1993, Lynch 1997, Shohamy 1997, Shohamy 2001, Spolsky 1997, Fulcher 1999b). 19th Language Research Colloquium (Kunnan, 2000) and a full section in Clapham and Corson (1997) are devoted to impact and fairness problems in relation to the intrusive nature of language testing and normative roles of tests as

[21] See also Davies (1997).

well as alternative assessment query. Two issues that should be mentioned in relation to these are 'standards' and calls for 'professionalism' in the field. In order to encourage the best practice in assessment and standardisation in the quality of tests, guidelines for right conduct and practice, 'standards' and 'codes of ethics' have been published (e.g. International Test Commission 2000, ILTA Code of Ethics 2000, ALTE Code of Practice and Quality 2002). These focus on the use of tests in accordance with and taking care of social contexts and also universal principles as well as fair practice and professional responsibilities of people and bodies involved in assessment. Alderson and Buck (1993) observe the lack of standards in the British language examination system. Bailey and Brown (1995) urge the need to develop a standard in teacher preparation in language assessment in line with the increased theoretical, practical and legal importance of language testing. Davies (1997) calls for the promotion of an institutionalised ethical milieu for professional activity, which would determine and disseminate the requirements of being a professional language tester through informing the public. Bachman (2000, 19) also underlines the importance of professionalisation of the field, that is, the training of language testing professionals and the development of standards of practice and mechanisms for their implementation and enforcement. He mentions that Messick's framework has been increasingly accepted as a paradigm in validation research, which has become central to language testing research. Bachman concludes that professionalisation and validation research are both essential to language testing and 'cannot proceed independently of each other' since 'the primary impetus for professionalisation is the need to establish standards for ethical conduct, which itself must be grounded in valid test use' (ibid., 23).

2.2.4 Summary

Section 2.2 has focused on the validity conceptualisations in measurement and discussed in detail Messick's (1989a) unitary view of validity, which brings together considerations of validity interpretations and consequences of test use. Centrality of construct validation has been discussed in relation to both education and language assessment fields. An overview of validation research in the field of language testing and specifically in the assessment of reading skill has been given with reference to exemplary research. The next section will focus on what is essential in the construct validity research of a reading test, namely, the reading construct itself. It will focus on reading theories and models both in first language and second/foreign language and discuss the details of a reading model that forms the basis for the operationalisation of the reading test under scrutiny.

2.3 The Nature of the Reading Construct

Constructs of reading as with any other construct are based on a theory of reading which accounts for any factor affecting the reading process or product.[22] For example, Grabe (2000) stresses that any comprehensive theory should account for linguistic, processing, learning, social, affective and motivational aspects of reading. Therefore, before discussing how reading ability can be tested, one should be clear about the factors involved in the process of reading. In attempting to understand what is involved in reading, language professionals have proposed a variety of definitions

[22] Alderson (2000, 3-4) defines 'process' as 'what we mean by reading proper: interaction between a reader and the text' and 'product' as 'the understandings they (readers) end up'.

and theories varying in the emphasis placed on text-based and reader-based variables.

No matter how familiar the concept of 'reading' sounds to any literate individual, it is to that extent difficult to define it. Definitions proposed are numerous and vary in details. Johnson (1983, in Khalifa 1997) gives a comprehensive definition of reading as:

> 'a complex behaviour which involves conscious and unconscious use of various strategies, including problem-solving strategies, to build a model of the meaning which the writer is assumed to have intended. The model is constructed using schematic knowledge structures and the various cue systems which the writer has given (e.g. words, syntax, macrostructure, social information) to generate hypotheses which are tested using various logical and pragmatic strategies' (in ibid., 5).

On the other hand, Urquhart and Weir (1998, 14) simply state that reading is 'dealing with language messages in written or printed form'. Bernhardt (1991, 5) lists several dictionary definitions of 'reading' and comments that 'no clearly stated, empirically supported, and theoretically unassailable definition' has been proposed. It is perhaps its elusive nature that makes reading one of the most researched and speculated upon subfields in cognitive psychology and applied linguistics. Various aspects of reading (nature, acquisition in first and second language, literacy, perceptual and cognitive aspects, and many others) have been researched and several reading models have been proposed. These models are traditionally classified in the literature according to

the emphasis they put on the 'level' of text processing, i.e. whether readers predominantly process a text at lower levels such as phoneme and word level (bottom-up models), or they start with text level predictions and confirm them as they move on (top-down models), or whether high-level and low-level processes interact as readers deal with texts (interactive approaches). Urquhart and Weir (1998) group these models as 'process models' and Grabe and Stoller (2002), as 'metaphorical models'. However, it seems that with recent research advances, this classification has become obsolete since reading has been recognised as an 'interactive' process for about two decades now. Indeed, most models mentioned below are perceived as suggesting some degree of interaction between 'levels' even though they emphasise a certain mode of direction in the process of reading (e.g. Goodman's top-down model suggesting a cyclical nature between phases of hypothesis forming).

Rayner and Pollatsek (1989, 25) mention that most reading models are 'general frameworks which provide some biases about which aspects of reading are really important'. Cognitive psychologists, in accordance with their professional interest, focus on and attempt to model the reading process as it is thought to take place in the human mind. They are interested in the details of how certain factors operate and usually provide empirical evidence for what are considered 'low level' processes such as word recognition, syntactic processing, etc. On the other hand, there are theorists for whom whether or not a factor such as grammar, world knowledge, etc. has a measurable effect on reading performance has been the focus of attention. The models that reflect the latter approach 'consist simply of areas of skills or knowledge thought to be involved in the (reading) process' (Urquhart and Weir, 1998, 46). They

attempt to 'model reading ability rather than the reading process and to identify a set of distinct and empirically isolable constituents' (Hoover and Tunmer; 1993, 4). The models proposed by the former group of researchers are briefly discussed below as 'process models' and the ones that are proposed by the theorists that might fall into the latter group are given as 'componential models'. There are apparent overlaps among the models and such a classification may not reflect clear-cut differences for the reason that all models must make reference both to 'processes' and 'components' (or 'factors') to a certain extent in dealing with the reading process. Obviously, some of these have cognitive focus and some might have a more applied linguistics inclination – besides the fact that the latter usually benefits largely from the former. However, as Rayner and Pollatsek (1989) put it, each model emphasises some aspect of reading. Thus, certain models of reading may focus on components involved in reading without much recourse to the psychological aspects or vice versa, and some may focus both on the process and the components involved in reading as is the case with Carr and Levy's (1990) model. Overlap in the categorisation of reading models seems inescapable. A brief review of certain reading models which have frequently been referred to in the literature is given below under two headings: process models and componential models.[23]

[23] There are obviously many other reading models that are not covered in this brief review (see Britton and Graesser (1996) for example). A complete coverage would be beyond the scope of this study.

2.3.1 Process Models

2.3.1.1 Bottom-up Models

Traditional views of reading have conceived of comprehension as a simple process of decoding symbols. Readers process a written text by beginning with lower linguistics units (letters, words, phrases, sentences) and working their way up to higher-level ones in a linear fashion. Each process builds upon prior subprocesses, but higher subprocesses cannot feed back into lower components (Alderson, 2000). Bottom-up models analyse reading as a process in which lower units are analysed and gradually added to higher units until the meaning is constructed through the application of syntactic and semantic rules. Comprehension takes place after this series of operations are complete with little influence from general world knowledge, contextual information or higher order processing strategies (Gough, 1972 in Khalifa, 1997 and LaBerge-Samuels, 1974; Carver, 1977-78 in Barnett, 1989, Rayner and Pollatsek, 1989, Grabe and Stoller, 2002). Gough, the often cited theorist, argues that fluent readers are quick in identifying letters reading serially, letter by letter. Phonemes are stored primarily in short-term memory until words and sentences are understood. However, he does not specify how each subprocess operates and how understanding takes place (Barnett, 1989).

Carver's **rauding theory** (1977-1978 in Carver, 1997 and 1998) with its emphasis on linear and unidirectional processing from letters to sounds and meaning, is among the bottom-up models as well. Rauding theory within the framework of the 'simple view' of reading, emphasises cognitive processes such as lexical access, semantic

encoding, sentence integrating, proposition integrating and idea remembering that will be in use successively in differing rates and purposes of reading.

Rayner and Pollatsek (1989), as Gough and Carver above, adopt an 'information processing' approach and place themselves within the bottom-up framework. Their model, however, allows 'some influences from top-down processes' (ibid., 26, 472). Reading processing sequence, according to Rayner and Pollatsek, begins during eye fixation with the initial encoding of the printed words after which lexical access takes place. Lexical access creates an auditory code (inner speech) and rules and analogies are activated automatically. This process may involve multiple lexical items simultaneously. As the lexical access is completed, meaning of fixed words is integrated into an ongoing text representation in working memory.

2.3.1.2 Top-down Models

Top-down approaches (knowledge-based or concept-driven processing, also known as **schema-theoretic** views of reading) have emphasised the importance of the reader and the knowledge he or she brings to the text. The reader uses schemata - networks of information stored in the mind which act as filters for incoming information (Alderson, 2000) - in order to make guesses about what might come next in the text, and picks up information to confirm or reject such guesses. In the schema-theoretic view, therefore, reading is seen as 'an active hypothesis-forming activity going stage by stage from semantic top to the formal linguistic bottom until the semantic representation of the text is reconstructed. Having once guessed the real sense on the top, an efficient reader need not analyse all bottom elements in the text such as

phonemic cues' (Uljin, 1980 in Khalifa, 1997). With schema theory, the reader is seen as bringing not only linguistic knowledge but also formal, content and cultural knowledge into the reading process (Goodman, 1967 in Grabe, 1991; Smith, 1979 in Smith, 1994). The most influential, and most criticised model in this trend is Goodman's '**hypothesis-testing model**', otherwise known as the 'psycholinguistic guessing game' model. He argues that readers make predictions about the grammatical structure in a text by the help of their linguistic knowledge and semantic concepts. Then, they confirm their predictions by sampling the print. Therefore, reading is an iterative process of hypothesising, sampling and confirming information based on background knowledge, expectations and sampling features from the text and context (Goodman, 1986, 1996 in Grabe and Stoller, 2002). Smith (1979 in Smith, 1994) too, stresses the purposeful and anticipatory nature of the reading process with the primacy of the prior knowledge the reader brings in the comprehension process. However, it has been shown by at least some research that good readers make use of the context less and their word recognition is fast and efficient (i.e. Stanovich, 1991).

2.3.1.3 Interactive models

Interactive models are currently considered to more adequately characterise the nature of the reading process. Interactive models are not unidimensional but cyclic: processing in both dimensions is expected to proceed simultaneously as well as interact and influence each other. Rumelhart (1977 in Alderson, 2000) incorporates feedback mechanisms that allow knowledge sources to interact with visual input. The reader arrives at a final hypothesis about the text by synthesising knowledge from

multiple sources (linguistic as well as world knowledge) which interact continuously and simultaneously. The reader processes the printed information by starting at lower levels and his or her expectations act downwards but simultaneously influence the processes at lower levels. Higher-level processing may take place before lower-level processing. Thus, a word may be understood before the sounds and letters are decoded (Carrell and Eisterhold, 1988).

Stanovich (1980 in Barnett, 1989; Khalifa, 1997 and Alderson, 2000) proposes an **interactive-compensatory model** of reading which attempts to account for individual differences in reading fluency. The degree of interaction among the components depends upon the knowledge deficit in individual components; strength in one component can compensate for a deficit in another. If there is deficiency at word recognition stage for a poor reader, for example, knowledge of the topic (top-down processing) may allow for compensation. A skilled reader, on the other hand, may focus on the words (bottom-up) to compensate for the deficiency in his/her topic knowledge. Based on additional evidence from eye movement studies, it is now acknowledged by even bottom-up theorists that efficient reading occurs when the reader expands processing capacity on higher level comprehension processes rather than on word recognition. This occurs via efficient decoding processes in a good reader (Stanovich, 1991, 21). It is the poor readers who use context more. Fluent readers read most words on a page (Perfetti, 1991, 1999). Samuel and Kamil (1988, 22,36) point out that 'a *weakness* in one source of knowledge results in heavier reliance on other sources of knowledge, regardless of their level in the processing hierarchy'. In recent studies, Stanovich and Stanovich (1999) and Stanovich (2000) indicate that fluent readers are efficient in word recognition, that automatic processes

do not interact but work independently (less-automatic processes do), and that interaction and compensation increase when difficulties are met.

Rumelhart (1977, in Carrell, 1988) is one of the first theorists to show that syntactic, semantic, lexical and orthographic information can influence our perceptions. Information from these sources converges on what he calls 'a pattern synthesiser' operating simultaneously and inter-dependently. With a series of examples on how higher order knowledge is used to disambiguate lower stage analysis (semantic knowledge influencing word perception, word knowledge influencing syntax, etc), Rumelhart accomodates many different types and directions of processing that take place in reading in his model. Later, Rumelhart and McClelland et al. (1986 in Barnett, 1989) expand on the original interactive model and propose **parallel distributed processing models** with which they attempt to explain how the human mind works. They suggest that 'information processing takes place through the interactions of a large number of simple processing elements called units, each sending excitatory and inhibitory signals to other units'. These units represent hypotheses about words, syntactic elements, etc. Interconnections among units form the constraints known to exist between the hypotheses (ibid., 27).

Just and Carpenter's (1980) model accounts for comprehension processes but base its arguments on the data from the eye movement analysis studies of readers. They have shown that time spent on a lexical item is directly related to the amount of time needed to process that word. Readers make longer pauses at points where processing loads are greater (e.g. content words, important clauses, ends of sentences). Just and Carpenter define five processes any of which can affect the processing of the other:

1. seeing the next word and extracting its physical features
2. seeing the word as a word and comparing it to the mental lexicon
3. assigning a case (e. g. nominative, objective) to the word
4. relating the word to the rest of the words
5. wrapping up the sentence when complete.

Just and Carpenter (1987) in their **Reader Model** also stress that the main processes in reading are fundamentally language comprehension though they see reading as a multicomponent skill that involves a large family of different tasks beginning with printed words, ending with the new knowledge that the reader acquires. The various levels of text, including words, phrases, sentences, and the whole text are operated on by some of the component processes of reading. The prominent level in reading is considered to be the lexical level, including encoding printed word and accessing its meaning in a mental dictionary (lexical access). Readers try to interpret each word of a text (immediacy of interpretation). Phrases and clauses are analysed at syntactic and semantic levels. In order to make sense of a text, the reader must construct a representation of the concepts and the situation to which the text is referring (referential representation). Component processes in reading are coordinated in time and can operate in parallel by using a common working memory (ibid., 23). A production system, which is central to the model, operates on the contents of memory and triggers necessary production rules for the integration of a text structure or inserting new elements in working memory (in 'recognise-act' cycles). During a production cycle, contents of memory are simultaneously assessed through an interaction of productions and the production conditions (Stanovich, 1996). Individual differences in language comprehension can be attributed to the variations

in total amount of activation in memory, which is responsible for processing and storage (Just and Carpenter, 1992).

The interactive view of reading process has been widely acknowledged in **second language reading** studies. The studies done by Coady (1979) and Bernhardt (1991) to name two demonstrate that researchers have been involved in identifying the interacting components of L2 reading. One of the first models of interactive reading in ESL is Coady's (1979) **psycholinguistic model**. Coady assumes that comprehension results from the interaction of conceptual abilities, background knowledge and process strategies. Individual process strategies are:

1. Phoneme-grapheme correspondences
2. Graheme-morphophoneme correspondences
3. Syllable-morpheme correspondences
4. Syntactic information (deep and surface)
5. Lexical meaning and contextual meaning
6. Cognitive Strategies
7. Affective mobilisers.

According to Coady's model, learners progress from reliance on concrete processing strategies (e.g. grapheme-phoneme correspondences) to more abstract strategies (e.g. contextual or lexical meaning). Coady also notes that a reader shifts processing strategies or changes the balance between them to match different types of texts or to accomplish different goals.

Bernhardt's (1986c in Barnett, 1989) **constructivist model** of second language reading includes 'text-based' and 'extra text-based' components. 'The reader recognises words and syntactic features, brings prior knowledge to the text, links the text elements together and thinks about how the reading process is working (metacognition)' (ibid, 47). Bernhardt (1991) later revises her model to include three components: language, literacy and world knowledge. The language component includes word structure, word meaning, syntax and morphology. The literacy component involves the reader's preferred level of understanding, goal setting and comprehension monitoring. Higher levels of literacy will enable the reader to deploy different strategies. According to Bernhardt, literacy includes knowing how to approach a text, why one approaches it and what to do with it. The world knowledge, on the other hand, involves background knowledge a reader possesses and uses to facilitate comprehension.

Carrell (1988) maintains that reading comprehension is characterised as involving an interaction of 'text-based' and 'knowledge based' processes (the latter indicating the reader's existing background) and the most efficient reading is a bidirectional combination of text-based and knowledge-based processes. Carrell and Eisterhold (1988, 79)[24] further revise the concept of background knowledge drawing a distinction between 'formal schemata' (background knowledge of the formal, rhetorical, organisational structures of different type of texts) and 'content schemata' (background knowledge of the content area of a text).

[24] See also Carrell (1990) and (1992).

In interactive approaches to reading in a second language, reading difficulties are attributed to background differences, language processing differences and social context differences (Grabe, 1991). When some reading processes are not automatised or when readers have insufficient command of the language poor reading behaviours could be the outcome. Grabe (ibid.) describes good reading behaviours as rapid (the reader needs to maintain the flow), purposeful, interactive (the reader makes use of background knowledge), comprehending (the reader expects to understand), flexible (the reader employs a range of strategies such as adjusting the reading speed, skimming ahead, considering titles, headings, text structure information, etc.) and gradually developing (fluent reading is the product of long term effort and gradual improvement). Automatic bottom-up processing (automatic perceptual/identification skills) is essential since it allows the readers to focus on higher level processing. According to Grabe, students have problems in reading when their low proficiency makes them word bound and they are not yet efficient in bottom-up processing; syntactic and vocabulary knowledge are critical components of efficient reading. Grabe (1991, 377) lists component skills as follows:

1. Automatic recognition skills
2. Vocabulary and structural knowledge
3. Formal discourse structure knowledge
4. Content/world background knowledge
5. Synthesis and evaluation skills
6. Metacognitive knowledge and skills monitoring.

Hudson (1991) summarises the current view of interactive models as follows:

> 'Reading involves the simultaneous application of elements such as context and purpose along with knowledge of grammar, content, vocabulary, discourse conventions, graphemic knowledge and metacognitive awareness in order to develop an appropriate meaning.' (ibid., 83)

Interactive models of reading are more comprehensive than top-down and bottom-up processes since they account for reader differences (degree of skill, level of language, background, metacognitive strategies, etc.), different purposes of reading (skimming, scanning, etc.) and differences among texts (discourse conventions, etc). However, Grabe and Stoller (2002) caution that the combination of useful ideas from bottom-up and top-down views to form an interactive approach to satisfy everyone might be self-contradictory, and therefore, should be 'modified': Grabe and Stoller categorise processes activated when reading takes place into two categories: lower-level and higher-level processes. Lower level processes such as lexical access, syntactic parsing, semantic proposition formation and working memory activation are considered to be automatic linguistic processes whereas higher-level processes such as text model comprehension, situation model of reader interpretation and executive control processes relate more to the use of background knowledge and inferencing skills. Even though Grabe and Stoller maintain that reading comprehension is 'balancing and coordinating many (of these) abilities in a very complex and rapid set of routines' (ibid., 29), they also stress that automatic processes are 'carried out in a bottom-up manner with little interference from other processing levels or knowledge sources'. For example, fluent word recognition or initial syntactic parsing does not require interaction from context or background

information. When readers have problems at these levels, then structures are raised to the conscious level for the use of context and inferencing (ibid., 33). Grabe and Stoller, stressing the importance of different types of reading changing according to the reader purpose, note that the use of higher order skills might change according to the various purposes for reading; a reader might be using more top-down processing when skimming a text, for example.

2.3.2 Componential Models

Weir, et al. (2000) point out that process models aim at explaining the psycholinguistic process of reading according to temporal sequence (eye movement and computer on-line studies), and cast light especially on our understanding of lexical access and word decoding. These models usually account for lower level processes accurately but are quite vague about higher level processes (Rayner and Pollatsek; 1989, 471). Moreover, they are exclusively premised on 'careful reading' – normal, silent reading as in the careful reading of a newspaper article (ibid, 23.). As such, 'quick purposeful reading' – as in skimming a text to get the gist – is not adequately accounted for (Weir, et al, 2000).[25] Certain reading models that deliberately refer to the components involved in the reading process, whether they are bottom-up, top-down or interactive, are considered to emphasise higher level processes more thoroughly. Some of these are briefed below.

In Hoover and Tunmer's (1993) **simple view**, as they refer to their model, the components are 'word recognition' (or decoding: the ability to rapidly derive a

[25] See also Carr and Levy (1990, 3).

representation from the printed input that allows access to the appropriate entry in the mental lexicon) and 'linguistic comprehension' (the ability to take lexical information and derive sentence and discourse interpretations) both of which are equally important. This view asserts that word recognition should be accompanied by the full set of linguistic skills (such as determining the intended meaning of individual words, assigning syntactic structures to sentences, deriving meaning from sentences and building meaningful discourse) in order to comprehend language. Word recognition is assumed to be 'a guide' to linguistic skills. Reading comprehension involves the same ability as linguistic comprehension but one that relies on printed information arriving through the eye (ibid, 3-8). In Gough's (1992 in Alderson 2000) view, reading is also essentially divided into two components: decoding (word recognition) and comprehension. According to Gough, 'comprehension consists of parsing sentences, understanding sentences in discourse, building a discourse structure, and then integrating this understanding with what one already knows' (ibid., 12). Perfetti (1991, 33) also states that 'learning to read does not involve learning rules but is a matter of incrementing a store of graphemically accessible words (Restricted Verbal Efficiency Model)'. Therefore, according to the two-component approach, there is basically minimal difference between listening and reading and comprehension is not a reading but a centrally controlled linguistic skill.[26]

Carr and Levy (1990) emphasise the role of '**componential skills**' approach in the analysis of the reading process. This approach suggests that reading is the product of

[26] Urquhart and Weir (1998) caution that the simple view might pose difficulties for the evaluation of L2 reading in which the reading skill might develop well beyond and before the listening skill.

a complex but decomposable information-processing system. According to Carr and Levy, most problems in cognitive psychology are handled by reference to specialised processing mechanisms, each of which carries out one particular kind of mental operation. In the analysis of the reading process, too, these operations, their organisation, control and coordination, the flow of information among them and the parameters of the system in which they exist should be identified to account for individual and developmental differences. These operations are finite in number, theoretically distinct and empirically separable. Carr and Levy (1990) explain the existence and interaction of the components of their model through numerous case studies.

Among the process models cited in section 2.3.1, Coady's (1979) model involving three components (conceptual abilities, process strategies and background knowledge) and Bernhardt's (1991) model consisting of language, literacy and world knowledge components, Grabe's (1991) and Grabe and Stoller's (2002) taxonomic views are apparently componential in their approach to the reading process. The reader is referred to section 2.3.1 for brief comments on those models.

The last model that will be discussed here places exclusive importance on both propositional integration and discourse processing level. Kintsch and van Dijk (1978) in their **model of text comprehension and production** emphasise comprehension to the exclusion of word recognition, although they assume the latter must exist (Barnett, 1989, 27). The assumption is that the surface structure of a piece of discourse is interpreted as a set of propositions. Propositions (the meaning elements of a text, underlying semantic structures) become organised into a coherent whole (a

text base) in differential retention. The semantic structure of texts can be described both at the local microstructure level (structure of the individual propositions and their relations) and at a more global macrostructure (discourse level), that is, by micropropositions and macropropositions. The formation of a coherent (mental) semantic text base (a discourse topic) involves a cyclical process maintained through macrorules based on referential coherence (argument overlap), and if referential coherence is scarce, on inference. However, the formation of the text base is constrained by limitations of working memory or buffer capacity. Macro-operators reduce information in a text base to its gist, that is, the theoretical macrostructure. These operations are under the control of schema (involving schematic structures of discourse; superstructures), which is a theoretical formulation of a comprehender's goal. Macrorules are the semantic mapping rules that organise propositions into appropriate levels (Kintsch and van Dijk, 1978. See also van Dijk, 1977). Van Dijk and Kintsch (1983, in Grabe, 1999 and Kintsch, 1988) particularly emphasise three levels of comprehension representation: 1) verbatim representation which decays rapidly, 2) conceptual text-based representation that is generated through the process described above, and 3) the situation model that incorporates the reader's schemata and affective states; a deeper level at which the text loses its individuality and its information content. It is at the last level that not only comprehension but learning takes place (Kintsch, 1994). Kintsch (1988)[27] has later revised the model to integrate lower level processes; a **construction-integration model**, in which the initial processing is strictly bottom-up. In this model, a text base is constructed from the linguistic input in a construction process. The text base is integrated with the comprehender's knowledge base (an associative network the nodes of which are

[27] See Kintsch (1998) for a more detailed account of the theory.

concepts or propositions), while the text is integrated into a coherent whole through a spreading activation process (whose duty is to select the best interpretation through the control of inconsistencies and irrelevancies).

2.3.3 Skills, Strategies and Taxonomies

In reviewing reading theories, we have frequently mentioned reading skills, abilities or strategies. Although researchers have frequently attempted to identify reading skills as components of their models, these terms are not clearly distinguished. According to Urquhart and Weir (1998, 84),[28] the focus on skills is partly due to a need to break down the rather vague and undifferentiated concept of 'comprehension' into more accessible chunks. The components identified in models could be translated into 'skill' terms such as decoding, accessing lexicon, etc. The authors define a reading skill as 'a cognitive ability which a person is able to use when interacting with written texts'. They are seen as part of the generalised reading process (ibid, 88). 'Strategies' on the other hand are used in psychology to describe how an organism seeks to attain its goals. They are used for pragmatic reasons; they are ways of getting round difficulties encountered while reading; as such, their psychological validity does not need to be investigated. Pritchard (1990, 275) defines strategy as 'a deliberate action that readers take *voluntarily*[29] to develop an understanding of what they read'. Cohen (1998b, 11) also acknowledges that the element of *consciousness* is what distinguishes strategies from those processes that are not strategic. Paris et al. (1996), however, note that it is hard to differentiate reading strategies from other processes that might be called thinking, reasoning, etc.

---

[28] See also Carr and Levy (1990).
[29] Italics added.

The definition they provide also refers to skills as unconscious information-processing techniques, and strategies as actions selected deliberately to achieve particular goals. Strategies can become skills when automatised, and skills when used intentionally can work as strategies (ibid, 611; Ellis 1994). In reading research, strategies are usually identified by observing reader behaviour and associating certain aspects of it with 'good readers' or 'poor readers'. However, the distinction between a skill and a strategy is still unclear and the terms are used interchangeably as can be seen in some of the taxonomies given below. Urquhart and Weir (1998, 96-98) draw our attention to the distinction by stating the following:

1. strategies are reader-oriented, skills are text-oriented;
2. strategies represent conscious decisions taken by the reader, yet skills are deployed unconsciously (they are automatic);
3. strategies, unlike skills, represent a response to a problem.

Despite the lack of consensus in conceptions of skills and strategies and their labels, taxonomies grouping several reading skills and strategies are widely offered and used in the field probably due to their convenience for teaching and testing. Some of the frequently cited and recently developed taxonomies are given in Appendix 2.3. These taxonomies do not go without criticism: Urquhart and Weir (1998) state that in Davis and Lunzer's taxonomy categories overlap and that some of them are more inclusive than others. Grabe's categories might be considered too broad. Alderson (2000, 11) radically claims that these taxonomies are seductive since they appear to be theoretically justified means of isolating reading skills to be tested. However, they are rather 'armchair' productions than the result of empirical observations and

therefore they do not have empirical validity. This leads us to an issue bearing crucial importance for the arguments on the reading construct.

2.3.4 Is Reading Unitary or Componential?

As mentioned above in discussing the reading models, it is widespread practice in reading research to classify reading into a series of subskills and to construct test items to measure individual reading skills. Weir and Porter (1994, 3) point out that practitioners claim 'that sets of reading skill components provide useful frameworks on which to base course design, teaching and test and materials'. Grabe (1991, 382) also considers a reading component's perspective as an appropriate research direction leading to important insights into the reading process. However, the separate existence of skills has not been supported in all studies. Rost (1993, 80) summarises three emerging views as holistic general-factor theories, the multiple factor models and middle roaders. Weir and Porter (1994) refer to them as the 'unitary', the 'multi-divisible' and the 'bi-divisible' views. Several studies testing their empirical validity are given below.

Davis (1968) in order to measure distinct operations in his taxonomy used two forms of a 96-item test each comprising 8 subtests. By factor analysis he found that 4 factors clearly emerged and were consistent across two forms. He revised his taxonomy to involve 4 skills rather than 8.[30] Thus he argued that 'comprehension among mature readers is not a unitary mental skill or operation' (Davis, 1968, 542).

[30] See Appendix 2.3

Spearritt (1972) reanalysed Davis's data using maximum likelihood factor analytical procedures and found four factors: recalling word meanings; drawing inferences from the content; recognising a writer's purpose, attitude, tone; following the structure of a passage. The remaining correlating with these four factors, Spearitt found similar categories to Davis's, except for the last one: finding answers to questions answered explicitly or in paraphrase.

Guthrie and Kirsch (1987), using factor analysis, identified two 'negligibly' correlated factors; comprehension and locating information. Reading to comprehend, which involves reading carefully to understand the explicitly stated ideas, was clearly differentiated from reading to locate information, which requires selective sampling of the text.

Carver (1992, 358) reported that a principal component analysis of four standardised tests yielded two factors; accuracy level and rate level. He concluded that 'most standardised tests purporting to measure reading comprehension are also measuring individual differences in rate, that is, the ability to comprehend fast'.

Weir and Porter's (1994) The University of Reading data suggest that some students were able to cope with reading passages and questions at the global level but less successful on lower level microlinguistic items: cohesion markers, discourse markers, lexis and structural items. Researchers speculate that this might be due to the successful application of background knowledge to the text and/or transfer of higher level processing skills from L1 which compensate for deficiencies in lower

linguistic abilities. The authors also mention data from the ESP Centre in Alexandria where they tested reading comprehension at three levels:

a) reading a text quickly;

b) reading carefully to understand main ideas and important details;

c) a knowledge of more specifically linguistic contributory skills.

Point biserial correlation analysis revealed that levels (a) and (b) correlated more with their own subtests than with level (c) and vice versa. They note that similarly, there was a set of students who could cope with level (a) and (b) operations but failed at level (c) microlinguistic items. Their analysis confirmed differential performance on global as against specifically microlinguistic items.

Weir and Porter (1994, see also Urquhart and Weir, 1998) also found that items which focus on 'cohesion' or 'working out the meaning of words in context' - microlinguistic items - were out of place in the College English Tests (CET) administered to large groups in China. Weir et al.'s (2000) AERT project confirms this finding as well. They conclude that linguistic competence is not the same as performance ability in language skills and strategies in reading, and a measurement tool excluding these strategies will have a considerable negative washback in teaching.

To mention some previously discussed studies that have relevance for the present discussion, Perkins and Brutten (1992) categorise reading comprehension items as textually explicit, textually implicit and scriptially implicit (with a 94% agreement between the two raters) and find that these items behave differently. Buck et al.

(1997) identify 24 cognitive and linguistic attributes which account for 97% of the multiple choice reading test performance. These studies also reflect the assumption of separately identifiable reading skills.

However, these results are not confirmed by other studies. Lunzer et al. (1979) constructed four separate tests, each measuring eight distinct operations.[31] They administered the tests to 257 primary school children. Using factor analytic procedures, they concluded that subskills in comprehension do not exist.

Lee and Musumeci (1988), in an attempt to investigate the validity of the hierarchical reading descriptions in ACTFL guidelines, compared level of study, text types and reading skills. They found no evidence for hierarchy of text types, reading skills and the performances of test takers in relation to these as advised in the guidelines. They suggest that cognitively-based tasks might not differentiate levels of reading proficiency. Such descriptions should be linguistically oriented. However, they underline that the text itself might be influencing the assessment of reading performance more than the formulation of the question does.

More recently, Rost (1993) investigated younger native German speakers. He found one factor accounted for 77% of the total variance that he interpreted to be 'general reading competence'. He suggested that administering several subtests for the assessment of reading comprehension may be a waste of time without the gain of additional information beyond 'general reading comprehension' (ibid, 89).

[31] See Appendix 2.3

Hudson (1993), too, investigated the separability of subskills in grammar, reading comprehension and general reading tests (a multiple choice cloze) using IRT. His primary focus is to place skills and subskills in a top-down or bottom-up framework, however, his findings indicate that there is not a clear implicational scale among the items and skills investigated in this research. Hudson (1996) concludes that reading is not the sum of skills such as skimming, scanning, vocabulary identification and reading for main idea. Skills are difficult to define in practice and they are not ordered implicationally. Rather, they 'appear to cover wide bands of overlapping abilities', ranging 'from local text recognition and processing to broader text interpretation and use strategies' (ibid., 4).

It has also been previously mentioned that Schedl et al. (1996) could not find evidence for the unique factorial existence for the types of items categorised as 'reasoning' items in TOEFL specifications.

Just as the **quantitative studies** reviewed above suggest, there is also evidence for both unitary and componential approaches to reading from **qualitative studies**.

Alderson and Lukmani (1989) asked nine EFL teachers to determine what the items in a reading test are testing then categorise them as high, middle or low abilities. Following this, the teachers were asked to categorise the items into eight categories as described by the test writers. Alderson and Lukmani found considerable disagreement as to the level and content of the items and the test takers' performance was such that it was not possible to relate cognitive levels to their linguistic proficiency. Similarly, Alderson (1990a) presented the TEEP test to teachers to take

expert judgements as to what each item was measuring in terms of skills and asked them to categorise these skills into 'higher order' and 'lower order' skills.[32] In the second part (Alderson, 1990b), he analysed the introspections and retrospections from two students taking the test. Alderson claims that the study showed that teachers did not agree as to what each item tested and found a weak relationship between item statistics and what the item claimed to be testing. He also noted that test takers did not necessarily use the skills designated by the test writer in answering questions. Weir et al (1990) criticised Alderson's study on the grounds that no evidence was provided to indicate that the 'expert' judges had experience or had received any training prior to the task; reduction of skill taxonomy into high/low distinction may also reduce differentiation; Alderson provided no definition of high and low skills. Weir et al. pointed out that Alderson also assumes lower order skills were less difficult than higher order skills. However, it was not reasonable to expect a close correlation between cognitive levels and linguistic proficiency. Lumley (1993a) also criticised Alderson and Lukmani (1989)[33] and with a more careful research design and with a more linguistically oriented skill list, he found high agreement on the difficulty level of the items (validated by IRT analysis) and almost complete agreement in skills being tested by each item.[34]

Anderson et al. (1991) attempted to use various data sources to validate the reading construct of a multiple choice test. They processed data from candidates' retrospective think aloud protocols of their reading and test taking strategies, content analysis of the reading comprehension passages and questions, and candidates' test

---

[32] See Weir (1988).
[33] See also Matthews (1990).
[34] See also Bruten et al., (1991) and Weakley, (1993) cited in Urquhart and Weir, (1998).

performance. Firstly, items were classified into categories: understanding main ideas, understanding direct statements and drawing inferences. Secondly, they are classified according to Pearson and Johnson's (1978 in ibid) taxonomy of question and answer relationship. Finally, the researchers used Nevo's (1989) test-taking strategy checklist to classify verbal protocols. Chi-square analyses were carried out and results indicated that five categories were used differently depending on the type of questions used: stating failure to understand a portion of the text, paraphrasing, guessing, matching the stem with the text and making reference to time allocations.

The empirical evidence from the studies reviewed above has confirmed neither the multi-divisible nor the unitary hypothesis. However, considering that the primary implication of the multi-divisible reading approach for teaching and testing is that reading can be broken down into underlying skills and these skills can be differentially assessed, and a broad sampling across the skills is required in order to achieve construct validity in a reading test, further investigation of the issue of reading sub-skills is needed before one view is claimed and the other is refuted and diagnosis and measurement of reading are adjusted. Weir and Porter (1994, 14) warn against a wholesale adherence to either the unitary or the multi-divisible view in testing reading because of the blurred picture that each approach may give of the candidates' reading ability. Weir et al. (2000) also point out that the utilisation of test formats with a specifically linguistic focus of testing lower elements of reading (a possible consequence of the unitary approach) might disadvantage some candidates who are able to cope well with global comprehension items but fail at lower level elements of reading as Weir and Porter's (1994) data suggest. Weir and Porter (1994) also stress that it is widespread practice among test developers to focus on reading

skills components in test construction whether or not the sum of these components (answers to test items) equates fully with what the reader would normally take away from the test. The authors claim that 'whatever theoretical position the test developer takes, the need to construct individual test items will exert strong pressure to attempt to measure individual reading skill components and strategies, or combinations of them' (ibid., 3). Indeed, although taxonomic views are not strongly reflected in them, it is possible to see categorisation of test tasks according to text processing variables, reader purpose and task dimensions in certain reading assessment frameworks such as that of TOEFL 2000 (Enright et al., 2000).

2.3.5 An Expanded Model

Among the models that attempt to account for the reading process assuming a multi-divisible approach to the reading skill, a particular model is of central importance to the present study: Urquhart and Weir (1998). In order to account for a wider range of reading skills, including quick and selective reading behaviours such as skimming, search reading and scanning as well as careful reading, Urquhart and Weir (1998) propose a four dimensional reading taxonomy and an expanded reading model (See below). The researchers suggest that an overriding attention has been paid to careful reading in the theoretical literature, which is suggestive of the fact that 'expeditious' reading behaviours (selective reading) of both L1 and L2 readers are ignored. However, the academic reading needs of students have changed over the years with the influx of academic articles and information available on the Internet and students have to learn to read more quickly. Academic articles have also changed to include more information in abstracts and titles allowing the readers to access texts more

quickly to judge their relevance (Weir, 1999). Urquhart and Weir (1998) point out the following:

> 'We have theories of careful reading but very little on how readers process texts quickly and selectively, i.e. *expeditiously*, to extract important information in line with intended purpose(s). Given the value of these types of reading to the work forces of states in the northern hemisphere, let alone those of emerging nations, it is time more attention was paid to them in the professional and 'academic' literature.' (ibid, 101)

For the sake of convenience, Urquhart and Weir's taxonomy is repeated below:

Table 2.2 Matrix of reading types, Urquhart and Weir (1998)

	Global	**Local**
Expeditious	**A.** Skimming quickly to establish discourse topic and main ideas. Search reading to locate quickly and understand information relevant to predetermined needs.	**B.** Scanning to locate specific information; symbol or group of symbols; names, dates, figures or words.
Careful	**C.** Reading carefully to establish accurate comprehension of the explicitly stated main ideas the author wishes to convey; propositional inferencing.	**D.** Understanding syntactic structure of sentence and clause. Understanding lexical and/or grammatical cohesion. Understanding lexis/deducing meaning of lexical items from morphology and context.

The assumptions underlying this taxonomy may be summarised in a two-way distinction; 'process' and 'purposes' of reading:

I: The process of reading involves the use of different skills and strategies. Skills indicate the usually subconscious process of applying linguistic skills to extract main ideas and important details whereas strategies indicate the quick and usually conscious process of employing strategies for achieving the purposes of reading efficiently and quickly.

A. Careful reading and expeditious reading: The former is a slower process involving the use of probably different conscious reading skills (such as accessing mental lexicon, syntactic parser and thematic organiser) and the latter is a quicker process involving the use of reading strategies (as well as using careful reading skills when appropriate). In the expeditious mode, the reader does not usually attempt to understand every word in a passage but focuses on overall meaning.

B. Reading at the global and local level: Both careful and expeditious reading can be at the global and local levels. Global comprehension refers to the understanding of propositions beyond the level of microstructure, that is, any micropropositions in the macrostructure, including main ideas and important details. Local comprehension refers to the understanding of propositions at the level of microstructure, i.e., the meaning of lexical items, pronominal references, etc.

II: Purposes of reading: For different purposes of reading, the reader resorts to different skills and strategies and thus different processes are involved. The test should encompass these different skills and strategies as much as possible.

A. Expeditious reading: The reader's formal knowledge of the structure of the text and background knowledge can play an important role. Unlike careful reading, the linearity of the text is not necessarily followed. The reader is sampling the text, which can be words, topic sentences or important paragraphs, to extract information on a predetermined topic in search reading or to develop a macro structure of the whole text as in skimming. The process can be top-down when the reader is deciding

how to sample the text and which parts of the text to be sampled. It can also be bottom-up when the reader's attention is on the sampled parts of the text.

i. Skimming: Reading for the gist. The reader avoids details and tries to form a framework about what the text is about. The reading is selective and reader-driven with sections of the text either omitted or given very little attention. An attempt is made to build up a macrostructure (the gist) of the text. That is, propositions which the reader assumes to represent the macrostructure are committed to long term memory.

ii. Search reading: Locating information on predetermined topics. The reader does not have to establish a macropropositional structure for the whole text. The reader selects information to answer a set of questions or provide data as in completing assignments. It differs from skimming in that the search for information is guided by predetermined topics; so the reader does not necessarily have to establish a macropropositional structure for the whole text. While search reading, the reader keeps alert for words in the same or related semantic fields. The reader pays attention to titles, subtitles and other discourse clues and especially uses his/her formal knowledge of text structure to search for information on prespecified macropropositions. The parts that are deemed to be important will be read more carefully.

iii. Scanning: Reading selectively to achieve very specific goals, e.g. finding a number in a directory, looking for specific words/phrases, figures/percentages, names, dates of particular events or specific items in an index. Any part of the text

that does not contain the preselected symbol(s) (specific words, figures, names, etc.) is disregarded. During scanning, little or no syntactic processing is involved and only limited amount of lexical access is needed. The reader can scan with decoding alone. Coherence is not checked and macrostructure is not formed. In fact, the reader just checks whether the word or words being scanned fit the search description or not. If the search is successful, scanning will be over. There is even little need to read the sentence completely.

B. Careful reading: This type of reading is associated with reading to learn hence with the reading of text books or assigned articles. It is also the kind of reading favoured by many educationalists and psychologists to the exclusion of all other types. The reader attempts to handle the majority of information and to build up a macrostructure. The process is not selective. The reader adopts a submissive role and accepts the writer's organisation. He or she attempts to build up a macrostructure on the basis of the majority of the information in the text. In careful reading, the process can be sequentially bottom-up, from letters to words and from words to sentences and finally to texts. It can also be top-down, a process of confirming and correcting predictions by sampling the visual input. Most likely, the process is interactive involving both bottom-up and top-down reading by interactively using all sources of information and background knowledge to facilitate the inferencing of propositional meanings and the extraction of main ideas at the macroprositional level. Careful reading at the local level is more likely to be bottom-up, involving the use of skills at the micropropositional level such as inferring the meaning of lexical items and understanding the syntactic structure of sentences.

Urquhart and Weir (1998) report that there is empirical and theoretical evidence for the distinct nature of expeditious reading skills and strategies in that L2 readers find reading quickly and efficiently in the L2 particularly difficult, and perform differently in the expeditious reading sections of tests (Test for English Majors) compared to careful reading sections (Weir, 1983a; Shen et al., 1998 in ibid). They suggest that the reason why this difference has not surfaced in the literature is due to the fact that testing instruments failed to include items testing expeditious reading (skimming, search reading or scanning).[35] Paris et al. (1991) confirm that educational tests of reading do not reflect the notion of strategic reading. Students are required to read brief, disembodied paragraphs without titles, pictures, etc. unlike the ones they may encounter in content areas.

By definition, expeditious reading involves the conscious application of strategies (e.g. the alternating use of top-down and bottom-up strategies), the use of titles, initial summaries, expressions signalling the relative importance of propositions (e.g. 'the most important of all', 'above all', etc.) to form the macrostructure of certain parts of a text or to find answers to predetermined questions on the macrostructure. Therefore, the reader goes through the text quickly to select parts which are more important for closer scrutiny. Careful reading differs from expeditious reading in that readers at the same time deal with minor information in the text. There may be interactive application of several strategies during the reading process. In a test environment, time should be controlled strictly for expeditious reading components in that limited time for expeditious reading should be allocated in order to prevent careful, detailed reading (Urquhart and Weir, 1998, 130-133).

---

[35] Weir et al. (2000) give a review of major ESL tests.

2.3.6 Summary

Section 2.3 has discussed the nature of the reading construct and several reading theories and models both in L1 and L2 attempting to capture various aspects of the reading process. 'Process models' and 'componential models' have been discussed and several studies examining the existence and the types of reading subskills have been presented. Lastly, 'the expanded model' of reading on which the BUEPT reading test is operationalised has been detailed.

2.4 Conclusion

As discussed in section 2.2, a proper language test design should involve rigorous validation attempts. A validation study is by its nature a strenuous ongoing process and naturally is too broad to be captured with all its aspects in a single research task. However, the present study attempts to provide certain evidence for the construct validity of the reading module of the Boğaziçi University English Proficiency Test. In this attempt, Messick's (1989a) framework, which classifies construct validity evidence into six aspects, will be followed.

Therefore, the first research question is related with the content aspect of construct validity. Here, the concern is the specification of the boundaries and the structure of the construct to be assessed in relation to domain theory. The content aspect of the construct validity requires that the tasks selected for assessment should be both relevant to and representative of the domain. Thus the questions to be raised are:

1) How is the construct defined and reflected in the test?

2) Do the experts agree on the operations measured by the test items as specified by the test writers? It is hypothesised that the content analysis data by the language and testing experts will reveal that the items measure the operations specified by the test writers.

Secondly, to investigate the substantive aspect of construct validity, the operations used by test takers in answering the questions will be analysed. The third research question is as follows:

3) What are the operations utilised by the test takers to arrive at the correct answers? It is hypothesised that the test takers will use the operations specified in the test specifications to arrive at the correct answers.

The third aspect of the construct validity is the structural aspect. To investigate the structural aspect, the following question is posed:

4) What are the dimensions of the reading construct measured by the test? The first hypothesis concerning this research question is that the correlations between the scanning, search reading and careful reading parts of the test will correlate moderately.[36] The second hypothesis is that the items putatively testing different operations (scanning, skimming, search reading and careful reading) will load on different factors in the Principal Component Analysis.

In terms of the generalisability aspect, whether the same factorial picture emerges in the different versions of the text will be investigated. Therefore, the research question is the following:

---

[36] Skimming part cannot be included in the correlational analysis since it is represented in the test by one item.

5) Do the factor structures of the different versions of the test show similarities across versions? It is hypothesised that the items putatively testing different operations will load on different factors in the Principal Component Analysis in the same manner across four versions of the test, that is, similar component structures will be observed across four different versions of the test.

As for the external aspect of construct validity, the question below is raised:

6) What will be the relation between an established criterion measure and the test under investigation? It is hypothesised that the BUEPT reading test will correlate significantly with the IELTS reading module. It is also hypothesised that the content analysis of the two tests will yield evidence to support the meaningfulness of the correlation between the tests.

The consequential aspect of the construct validity will not be covered in the study for the reasons that will be discussed in the next chapter.

CHAPTER 3

METHODOLOGY

3.1 Introduction

The purpose of this chapter is to describe the methods used in the investigation of the research questions in the present study. In order to do this, each research question will be handled individually in succession and the details concerning the participants and the data collection methods (i.e. instrument, procedures and data analysis) will be described. The reader is reminded that the test under investigation in the present study is the new reading module of the Boğaziçi University English Proficiency Test (BUEPT). Five versions of this test were developed and administered between September 2000 and 2001 (September 2000, January 2001, June 2001, August 2001 and September 2001) before the test went through further changes reflecting the approaches of the testing office members who were appointed to the post after September 2001. As explained in section 1.2, September tests are administered to all students upon their entrance to the university. Students who score less than 60% in the test are required to attend to the prep year at the School of Foreign Languages for at least one semester. These students are grouped into three levels – beginner, intermediate and advanced levels – according to their scores in the placement test administered after their registration to the school. All advanced students and intermediate students who have scored 80% in achievement exams can take the test in January. The rest of the students take the June test on the completion of two-semester prep year. The ones who fail the June test attend a six-week remedial summer course and attempt the August test.

In the present study, when the data collection procedure required the submission of the test to other people, a short version of the September 2000 test (scanning, skimming, search reading I and careful reading I in Appendix 3.1) was used since this was the form the testing office could afford to release for research purposes. In the quantitative analyses, the August 2001 test was excluded due to the fact that this test is taken by a rather homogeneous, low scoring group of students who fail the June test.

As is mentioned above, the following sections will describe in detail the methods used in the investigation of each research question in the present study.

3.2 Research Question 1: How is the construct defined and reflected in the test?

This research question is related to the content aspect of construct validity, therefore, it involves the discussions on the evidence of content relevance and representativeness. These involve the specification of the nature and the boundaries of the construct domain, which is essentially done at the theoretical level. In order to ensure content relevance and representativeness, primarily, a recently developed reading framework was chosen to form the theoretical basis of test specifications. The Testing Office members of the School of Foreign Languages, including the researcher, worked on the reading framework by Urquhart and Weir (1998) to develop test specifications. Following this, in April 1999, they worked under the supervision of Professor Cyril Weir for the refinement of the test specifications and they were trained in methods of statistical analysis, text mapping and item development. After the completion of the training period, the Testing Office

members developed the first version of the new BUEPT reading test to be administered in September 2000 and performed the pilot testing. Five versions of the test were developed and administered as part of the Boğaziçi University Proficiency Test. The details of this a priori content validation are discussed in Chapter 4.

3.3 Research Question 2: Do the experts agree on the operations measured by the test items as specified by the test writers?

It has been mentioned above that content relevance and representativeness of assessment tasks involve specification of the nature and the boundaries of the domain. Besides this, content is also evaluated by the consensual professional judgement on the relevance and representativeness of the test items with respect to the domain (Messick 1989a, 36). No matter to what extent the test development process described in section 4.2 involved consensual procedures (i.e. text mapping and final revision of the tests by the teachers who were not involved in the test writing process), it is assumed that further confirmation by experts of the congruence between the test items and the test specifications would provide evidence for the content aspect of construct validity. Expert judgement was appealed to in order to confirm that the content analysis performed by them will reveal that the items measure the operations specified by the test writes in the manner as explained below.

3.3.1 Participants: Six experts participated as raters in the content analysis investigation. Four were teachers in the School of Foreign Languages at Boğaziçi University with extensive test development experience. These teachers had not been involved in the development of the BUEPT reading test. Two of the experts were

applied linguists with testing expertise at the Department of Foreign Language Education at the same university.

3.3.2 Instrument: The instrument used in this analysis is a simplified version of the test method facets (TMF) rating instrument and checklist given in Bachman et al. (1995). Test rubric characteristics, text characteristic and text span definitions are taken from Bachman et al. (1995) but in order to facilitate the identification of the operations used in answering questions, a list of operations that are specified in Urquhart and Weir (1998) and Khalifa (1997) was added to the item characteristic part. In the first and second parts (test rubric and text characteristics), the marking was done on a scale of five points. In the item characteristics part, the raters were asked to choose one or more operations and mark one text span for each item. Since the same content analysis scheme would also be used in the analysis of the sixth question, the scheme required the analysis of both the BUEPT and the IELTS reading tests.[1]

3.3.3 Procedure: The experts were given the content analysis scheme and a copy of the tests. They were explained the research aim and the rationale of the scheme. They were asked to follow the instructions carefully when responding to the scheme.

3.3.4 Data Analysis: The operations and the text spans that the experts marked in the BUEPT part of the content analysis scheme were summed to report the agreement among them and the results were compared to the test specifications.

---

[1] See Appendix 3.3.

3.4 Research Question 3: What are the operations utilised by the test takers to arrive at the correct answers?

In order to understand the underlying processes that take place during different types of reading as they are devised in the test, and thus provide evidence for the substantive aspect of construct validity, the third research question has been formulated. A number of researchers have noted the usefulness of verbal protocol data in the analysis of test taking processes and validation studies (i.e. Matsumoto, 1993; Green 1998). In line with their suggestions, the data were collected from 15 participants through immediate retrospective think aloud procedures to analyse whether the test takers use the operations specified in the test operations to arrive at the correct answers.

3.4.1 Participants: The subjects who participated in the study were chosen from the students of the School of Foreign Languages at Boğaziçi University. Initially, a large group of students that included participants from all levels of English proficiency and even certain students from language departments had been invited to participate in the study. The aim was to sample data from participants from a wide scale of English proficiency and compare their performances. However, after initial interviews with the students, it was seen that the students from low levels of proficiency were unable to provide adequate data, and the data collection procedure was quite frustrating for them. The students from the language departments also reported some background familiarity with the texts, which was not available for the students who did not start their studies at the departments. This might introduce a confounding effect which would not be eliminated unless the sampling was enlarged to include more

participants from various academic backgrounds. Therefore, the researcher decided to limit the group of participants only to 15 students who had just graduated from the School of Foreign Languages by passing the end of the year proficiency exam. The rest of the data was not taken into account in the analysis.

3.4.2 Instrument: The data were collected using the full version of the reading module of the September 2000 proficiency test consisting of the parts scanning (ten questions, SC1-10), skimming (one question, SR1), search reading (two texts, five questions on each, SR2-11), careful reading (two texts, five questions on each, CR1-10).

The operations (O) used by the test takers were classified based on the definitions of the operations and text spans specified in the content analysis scheme presented in section 4.3.[2] The text span in the scheme refers to the portion of the text to which the item relates. However, in the analysis of the verbal protocols, this was taken as 'the part of the text that the test taker processed in order to answer the test item'. For example, in cases where an item was originally designed to relate to a specific part of the text (TS2), yet the test taker had to process a larger span, the TS was specified as TS3 instead of TS2. Besides, a slight modification in the description of text spans was necessary. The definition of TS2 was that 'the item relates to a specific part of the passage, and requires only localised understanding of that part'. Since the extent of 'a specific part' was not clear in the definition and it could potentially range from one sentence to one paragraph, the researcher decided to distinguish between a very

[2] See also Appendix 3.3.

short span (one sentence or less than a sentence) and larger one (more than one sentence to one paragraph). The former was marked as TS2* and the latter, TS2.[3]

The analysis of the protocol data was extended beyond the operations and the text span by taking into consideration the test taking strategies (tts)[4] reported by the test takers during the protocols. The initial observations and tape recorded data revealed that certain test takers not only utilised reading operations in order to arrive at the answers but they employed several test taking strategies which might substantially change both the proceeding and the outcome of the test taking process. Therefore, together with the operations used and the text span processed, these strategies were analysed, too. The strategies differed from the operations in the sense that they did not aim at comprehension but they were used either when the comprehension failed or when the test takers wanted to arrive at the answers more quickly. For example, tts1 was aimed at locating the answer through the help of the order of the questions instead of using the prompts from the question and the text; tts2 involved matching the similar words in the text and the question to find the answer without sufficient comprehension; tts3 entailed the use of grammatical clues to extract the answer from the text, etc.[5] The use of these strategies was considered 'unfavourable' when the strategic process was not accompanied by a favourable reading comprehension process, the existence of which was checked by frequent questions by the researcher and demonstrated by the test takers through paraphrasing and translation. Otherwise, when the test takers used these as contributory strategies and understood both the question and the text sufficiently when providing an answer, the outcome was

[3] See Appendix 3.4 for the definitions.
[4] Lower case letters were used for the abbreviation of the test taking strategies (tts) to avoid confusion.
[5] See Appendix 3.4 for the definitions and examples.

considered to be 'favourable'.[6] Thus the analysis here was four-fold: the operations used (O), the text span processed (TS), the strategies (tts) employed by the text taker and the observations (OBS) on the outcome of the test taking process. The classification of the outcome of the test taking process as favourable and unfavourable depended on a judgement based on three questions:

Did the test taker

1) OBS1: understand the question? (Yes-Y, No-N, Partially-P)
2) OBS2: locate the part of the text that contained the answer correctly? (Yes-Y, No-N)
3) OBS3: understand the part of the text that contained (or assumed to contain) the answer correctly? (Yes-Y, No-N, Partially-P)

It was decided that any inadequacy (No or Partially) on the part of any of these three processes should indicate an 'unfavourable' comprehension process. Lastly, whether the test taker answered the question correctly or not (OBS4) was determined to give a four-way classification of observations: favourable comprehension/correct answer, favourable comprehension/incorrect answer, unfavourable comprehension/correct answer and unfavourable comprehension/incorrect answer. Finally, the questions that were not attempted or not answered were classified as 'unanswered' questions. The full form of the verbal protocol analysis scheme is given in Appendix 3.4.

3.4.3 Procedures: The data were collected through meetings with individual participants. Before the participants actually started to produce the data, they were explained and presented with an example of what they were supposed to do while they were describing their test taking process. They were asked to tell the researcher

[6] The definition of tts2 necessarily excludes detailed comprehension. Hence, tts2 does not apply here.

whether they had understood the question; which parts of the text they had read and the manner in which they had read the text; i.e. whether they had read the text line by line understanding every word or by skipping parts, etc. To facilitate their descriptions, they were given detailed descriptions of how a reader might process a text differently and they were asked to behave as they would normally do under test conditions. The participants were asked to report the reading and test taking processes they used when they were reading the text and answering the questions right after each question they answered. They were told that they might skip questions if ever they needed, however, they should explain why they did so. The exam time was suspended when the participant answered one question and following the think aloud protocol, the time was then restarted as the participant continued with the next question. The procedure was administered until a total elapsed testing time had passed. The researcher asked questions concerning the participant's test taking process during the think aloud when necessary. The verbal protocols produced by the participants were recorded for analysis and the researcher took observation notes during the procedure. Among the several techniques suggested for the collection of verbal protocol data[7], the immediate retrospective think aloud protocol technique with mediation through occasional questions was chosen for several reasons. First of all, it was seen at the beginning of the data collection phase, when the researcher worked with the beginner group, that it was quite challenging for the participants to answer questions at the same time verbalising their thought processes. It was also seen that the participants lost the train of their thoughts when they had to wait until the end of the test. Therefore, the best method seemed to be collecting the data right after each question was answered. The fifteen participants whose performances were

[7] See Green (1998).

analysed in this section of the study produced think aloud protocols in the manner described above.

3.4.4 Data Analysis: Initially, the researcher listened to the tape recording data and listed down the operations reported by the participants without doing any categorisation. As mentioned above, it was observed at this stage that certain subjects employed several test taking strategies to arrive at the answers and these could not be accounted for with the definitions given for the reading operations. Therefore, it was decided that it would be appropriate to note down these strategies and record them as they were reported by the subjects. The second listening was done to make the categorisation of the operations (O), text spans (TS), test taking strategies (tts) reported by the participants and the observations (OBS) according to the item type using the checklist given in Appendix 3.4. That is, the operations that a participant reported to use when answering a question were categorised in comparison with the list of operations that were designated by the test specifications as indicative of a certain type of reading (scanning, skimming, search reading, careful reading). The operations used in answering the questions were totalled to yield frequencies on an item basis. The part or parts of the text that individual test takers had to process to find the answer as well as test taking strategies were recorded and totalled on an item basis, too. Following these, the outcome of the reading process (OBS) was evaluated categorising the reading process as explained above.

The researcher reanalysed the data by listening to the recordings for the third time to verify the categorisation done in the second listening phase. There was a two-week gap between the last two listening sessions. The first and second categorisations done

in the second and third listening phases were compared to check intra-rater reliability. The dissimilarities between them were counted and compared to the total markings.

3.5 Research Question 4: What are the dimensions of the reading construct measured by the test?

This research question was formulated to assess the structural aspect of the construct validity. The structural aspect of the construct validity requires that the internal structure of the assessment should be consistent with the internal structure of the construct domain. Therefore, the investigation of this research question forms a substantial part in the present study since it is closely linked with the test development process. In order to investigate the congruence between the dimensions of the reading construct as reflected in the reading framework and in the test, two hypotheses were formed. The first hypothesis is that the correlations between the scanning, search reading and careful reading parts of the test will correlate moderately.[8] The second hypothesis is that the items putatively testing different operations (scanning, skimming, search reading and careful reading) will load on different factors in the Principal Component Analysis.

Therefore, after rigorous attempts to standardise the test development procedure, the next step in empirically developing the BUEPT test was to administer it and analyse the data statistically. It has been mentioned before that each test, before it is actually administered to the target group, should be piloted on a sample group to verify that

[8] Skimming part cannot be included in the correlational analysis since it is represented in the test by one question.

the items in it work desirably. Item analysis is a crucial part of a priori validation with which construct irrelevant variance in a test can be detected and reduced. Although item analysis relates to the content aspect of the construct validity, it is discussed here in order to give an intact representation of the test development process and statistical qualities of the tests.

Hence, for each test – September 2000, January 2001, June 2001 and September 2001- classical test analysis procedures (central tendency measures, reliability and item analysis) were employed both at the pilot and the actual test administration phases. The data from the actual administrations of the four versions of the test were further subjected to correlation and Principal Components Analysis in order to investigate the dimensions of the reading construct measured by the test. Before the details of the test versions are given, it will be more practical to define and describe the statistical procedures used in the study since these procedures were uniformly employed in the data analysis of all the tests.

Measures of central tendency: For the measures of central tendency and dispersion, mean, range and standard deviation estimates were used (Brown, 1996). These values and the Kolmogorov-Smirnof normality test were used to determine whether the data were normally distributed or not. In the Kolmogorov-Smirnof test, p values larger than 0.05 were taken as the indication of normally distributed data. The higher the p value, the closer the distribution to the normal distribution. In addition, skewness and kurtosis values, which indicate normality when they are equal to zero, were taken into consideration. The distribution is judged to be near normal if the skewness value is between –1.0 and +1.0. Kurtosis coefficients smaller than –1.0 are

considered platykurtic (flat distribution) whereas coefficients larger than 2.0 are considered to be leptakurtic (overly peaked distribution). For the estimate of internal consistency, Cronbach's Alpha (α), which is based on the average inter-item correlation, was used. SPSS 10.0 calculates alpha as equivalent to Kuder-Richardson 20 (KR20) coefficient for dichotomous data (SPSS Base 10.0 User's Guide, 1999). Kuder-Richardson formula (KR-21) was not used because the procedure is known to underestimate the reliability of the test basically because it assumes that all items have the same item difficulty (Brown and Hudson, 2002).

Item analysis criteria: Item analysis is done to evaluate the effectiveness of individual test items. Traditionally two procedures, item facility - IF (or item difficulty) and item discrimination - ID, are employed (Brown and Hudson, 2002). Item-total correlation patterns (CITC) and reliability estimates for individual items (alpha if item deleted - AIID) are also analysed.

Item facility (IF) is determined as the proportion of correct responses to total number of items. Item facility is inversely related to the actual difficulty of any given item; the higher the difficulty, the lower the proportion of correct responses in the whole group of test takers (Henning, 1987). IF values range from 0 to 1.00 and Henning (1987) and Alderson et al. (1995) suggest that items which are as near to a facility value of 0.5 as possible should be selected to have a widespread scores in a test. In terms of rejecting the item as too difficult or too easy, the suggested rule of thumb is to reject items with IF less than 0.40 or more than 0.70 (Brown and Hudson, 2002). However, this decision must be closely related with the purpose of the test, and as the test under scrutiny here is a reading test which is a part of a general proficiency test,

the limits for item rejection were set at 0.20 and 0.80 boundaries, following Green and Weir (1998).

Item discrimination index (ID) shows the degree to which an item discriminates between weak and strong examinees in the ability being tested (Henning, 1987). The groups of 'high-scorers' and 'low-scorers' are isolated as upper and lower third (sometimes as upper and lower 25%, 27%, or 33%). In order to calculate ID statistics, IF for the upper and lower groups are calculated separately (by dividing the number of examinees answering correctly in that group by the total number of examinees) and finally by subtracting the IF for the lower group from the IF for the upper group. Therefore, a discrimination index of 1.00 would be considered very good and 0.40 or above would be fairly high (Brown and Hudson, 2002). In the present study, item discrimination is analysed by both comparing the upper 33% and lower 33% groups and also by dividing them into six groups according to their total test performance (by analysing item discrimination patterns- IDPs). For the analysis of item discrimination patterns, groups are ranged from the lowest to the highest performing group and it is expected that the percentage of candidates answering a certain item correctly (IF) will increase from the lowest to the highest group systematically; weaker students responding to the item incorrectly and the good ones correctly. The groups on either side of the passing mark (60%) are kept narrower than the others since it is important to know which items, if any, are not discriminating around the pass/fail boundary. It is usually helpful to produce graphic representations of the way items perform across the six bands. If an item discriminates well between all the bands, we will see a line which moves from the bottom-left hand corner to the top-right-hand corner in the graph relatively similar to

the one in Figure 3.1. On the other hand, it is obvious from Figure 3.2 that the item SR1 is too difficult (the line does not reach the top-right-hand corner), and it does not discriminate well between the levels either, since firstly there is a dip at the point that corresponds to band 5 and secondly there is not much change in the slope of the line from band 3 to 6 except for the dip.

Figure 3.1: IDP graph for a favourable item

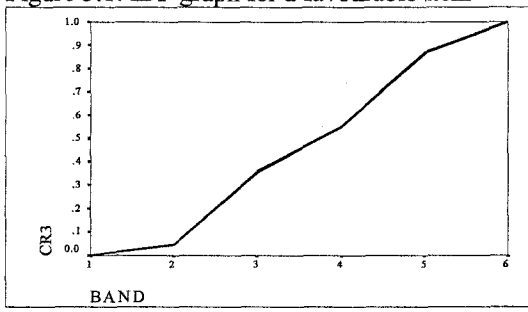
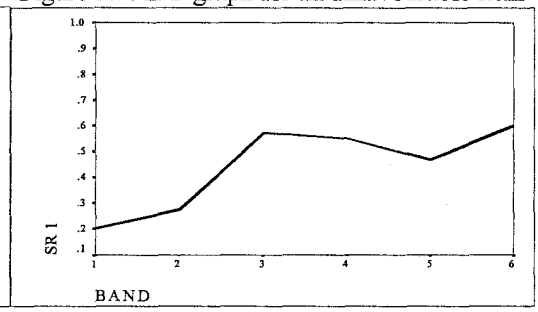


Figure 3.2: IDP graph for an unfavourable item



Item discriminability is also computed by looking at the correlation between an item and the total test/subtest score. This value is the correlation between the item and the score on the whole test/subtest minus the score on that item (corrected item-total correlation). A correlation of 0.20 and above is acceptable according to Green and Weir (1998). However, it should be born in mind that correlation is a function of sample size and ability range, and therefore may change with different samples (Henning, 1987).

Internal reliability estimates are additional data to evaluate the degree to which items fit together in a test, i.e., the test's homogeneity. Items that do not contribute to the test's overall reliability positively should be modified or rejected. This computation will tell us whether the test's internal reliability (alpha) would increase or decrease if the particular items were removed (alpha if item deleted). The last two calculations were done using Item-Total Statistics facility of SPSS 10.0.

Inter-correlations: In order to investigate the research question whether or not the four subtests (scanning, skimming, search reading, careful reading) in the reading test are testing different reading operations, the primary step was to correlate these subtests with each other using the Pearson Correlation Coefficient. The subtests are expected to show some degree of agreement since they are assumed to measure a certain aspect of reading ability; however, the correlations are expected to be fairly low. Green and Weir (1998) suggest that we might expect correlations of between 0.4 and 0.7 on different parts of the same test because of an underlying language ability which substantiates all language behaviour. If any two tests correlate very highly (e.g. 0.9), we might think that there is a high degree of relationship between the two measures and we might suspect that they are basically testing the same operation. Then we might argue whether it is indeed necessary to have both tests or not. When the subtests of a reading module predict each other to a high extent, we might also suspect that reading ability as it is measured by the test under investigation might be unidimensional - measuring some overall ability - rather than componential.

Principal Component Analysis: The second step of the analysis of operations tested by different subtests is the Principal Component Analysis (Henceforth: PCA). In the present study, PCA is conducted following Hatcher's (1994, 1-56) suggestions. Hatcher describes PCA as a variable reduction procedure in which a set of observed variables are reduced into a smaller set of artificial variables called 'principal components' that will account for most of the variance in the observed variables. In language testing, it is seen as 'a way of discovering factors that underlie language performance and of testing the relationship among them' (Hatch and Lazarton 1991,

489). The first component extracted in a PCA accounts for a maximal amount of total variance in the observed variables and the second component accounts for what is not accounted for by the first component and as such is uncorrelated with the first component. Resulting components will display varying degrees of correlation with the observed variables but will be uncorrelated with each other.

In a reading test for example, putatively different variables (e.g. skills) are expected to load on different components. If the variables load on the same component, we might assume that they function in a similar manner, and there is a strong possibility that there are no separate skills; they measure the same construct, undifferentiated reading ability. If, on the other hand, variables conceivably testing a certain skill load on a certain component while others load on a different component, we are led to think that reading ability is divisible as it is measured by that test (Green and Weir, 1998). In conducting PCA, the first step is to perform an initial extraction of the components. The number of the components in the initial extraction is equal to the number of variables being analysed. However, for the subsequent analysis only a few of them will be retained for the interpretation. To determine the number of components to be retained there are a few criteria to be taken into consideration:[9]

1. The eigenvalue-one criterion: An eigenvalue represents the amount of variance that is accounted for by a given component. This criterion suggests that any component with an eigenvalue greater than 1.00 should be retained. Since each

[9] In the present study, 'communalities' (percent of variance in a variable that is accounted for by the retained components) are also considered. Variables are expected to have a communality value of .3 or more.

variable contributes one unit of variance in the data set, any component that displays an eigenvalue that is more than 1.00 accounts for a greater amount of variance than itself. Components with eigenvalues less than one are viewed as trivial and are not retained.

2. The scree-test: A scree-plot displays eigenvalues against components. When there is a large break in the curve and it starts to flatten out, the components after the break are assumed to be unimportant and are not retained.
3. Proportion of variance accounted for: It is suggested that we may retain components that account for at least 5% of the total variance. Alternatively, researchers might retain enough components so that a cumulative percent of 70% is attained.[10]
4. Interpretability criteria: The basic question here is whether the retained components have substantive meaning and whether our interpretation of the components makes sense in terms of what is known about the constructs under investigation (Hatcher 1994, 22-26).

In the present study, the data were analysed by using PCA's first rule, namely, taking primarily the first criterion, eigenvalue-one rule, into consideration.[11] Then, the amount of variance and interpretability criteria were checked. After the number of components to be retained was decided, which generally corresponded to the number of components with eigenvalues over 1.00, the data were subjected to 'varimax

[10] However, Green and Weir (1998) warn us that when our data is of individual items on a 0/1 scale, the scale for any correlations is very restricted. Therefore, in our case too, we may expect to find lower cumulative percent of variance accounted for by the extracted components.

[11] Other extraction methods such as principal axis factoring were tried and did occasionally give better results. However, for the sake of consistency, the results of PCA with varimax rotation will be reported throughout the study.

rotation'.[12] The variables that have high loading on one particular component were determined. The loading is usually considered high when it is at least .40. Variables are expected not to have high loadings on more than one component. If the resulting component structure was not interpretable, the number of components to be extracted was controlled or variables were eliminated until an interpretable picture was attained. The ideal component structure would be with four components that account for four reading skills; scanning, skimming, search reading and careful reading.[13] In the interpretation of the components, subtest-factor correlations were also assessed. To determine whether the data are adequate for the PCA, Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was used. This measure compares the magnitudes of the observed correlation coefficients to the magnitudes of the partial correlation coefficients. A KMO below 0.50 is usually considered unacceptable. Another indicator of the strength of the relationship among variables, namely Bartlett's test of sphericity, was also used to check whether the observed significance level is small enough (SPSS Base 10.0 User's Guide,1999). SPSS version 10.0 was used in the PCA analysis of the data.

3.5.1 The September 2000 test – Pilot Version

3.5.1.1 Participants: The pilot version of the September 2000 test was trialed on 81 Turkish undergraduate EFL students studying at the pre-sessional year of English medium Koç University. Koç University prep students were chosen as an equivalent sample group after inspection of the curriculum and the materials of the institution

[12] Rotation is a linear transformation that is performed on the factor solution to make the interpretation easier. Varimax rotation produces uncorrelated components and maximises the variance of a column of the factor pattern matrix (Hatcher 1994, 28).
[13] In the cases where the skimming question is not included, the expected component number is 3.

and discussions with the language teachers. 100 subjects were randomly chosen among all students from three different levels (beginner, intermediate, advanced); therefore the group involved students with differing language abilities. However, 81 students could provide a complete set of data. Therefore, data from 19 students had to be eliminated from the analysis. Since these prep students were all registered in undergraduate programs, their academic background (in this case their high school specialisation) was considered unimportant.

3.5.1.2 Instrument: The pilot version of the September 2000 test included more items than the final version would so that the best items could be chosen and the items that did not perform well could be eliminated after statistical analyses. It is usually suggested that a pilot test include at least one third more items than needed (i.e. Green and Weir, 1998). However, the scanning part included only one extra item, as including more items might make scanning the text easier and it generally needs less effort to repair scanning items. Search and careful reading parts included three extra items.[14]

3.5.1.3 Procedures: The pilot version of the test was administered in May 2000, one month before the end of the school year. Koç University team decided that the scores they gained from the BUEPT test should be added to their within-year achievement average score so that the students would perform as they would do under normal exam conditions. 60% was the test's cut-off point for pass/fail distinction. The pilot version of the BUEPT September test was administered and scored by the Boğaziçi

[14] See Appendix 3.1.

University Testing Office team and the results were reported to the Koç University teachers.

3.5.1.4 Data Analysis: The data from the September 2000-pilot version of the test were analysed in terms of measures of central tendency, normality judgements and item analysis. In the light of the findings, the questions were evaluated and the ones with poor quality were either revised or dropped from the test to give the purged version to be administered in the September 2000 proficiency test. Factor analysis (PCA) could not be performed on the data since the data were not large enough.

3.5.2 The September 2000 Test

3.5.2.1 Participants: After the test was reduced to its purged version, it was administered to the group of incoming students in September 2000 as the reading part of the BUEPT. These students were mostly high school graduates who were placed in English-medium Boğaziçi University according to the scores they obtained in a nationwide university placement examination.[15] Among the group of all test takers, 341 were randomly selected for analysis.

3.5.2.2 Instrument: The test administered in September 2000 included a scanning subtest with ten questions, two search reading subtests, the first of which consisted of a multiple-choice skimming question and five search reading questions, and the second search reading subtest consisting of five search reading questions. The third

[15] A small group of postgraduate students also took the September BUEPT. However, they were excluded from the analysis.

subtest in the September 2000 test was the careful reading part with two reading texts with five questions on each.

3.5.2.3 Procedures: The administration of the September 2000 test was preceded by training workshops in which the staff of the school was introduced to the rationale and the structure of the new test. The new test administration procedures developed prior to the actual administration were explained to the staff, who normally work as test proctors. The test was administrated to approximately 870 incoming students on 8th September 2000. However, due to the time limitation, the data from 341 students could be typed in the SPSS program for analysis. It should also be noted that these students were the first 341 students on an alphabetically ordered name list. The researcher did not consider name as a confounding factor and thus considered the group as a randomly selected group of test takers.

3.5.2.4 Marking and Data Analysis: Following the administration of the test, the members of the testing office reviewed approximately 200 randomly selected answer sheets to check for alternatively correct answers that might appear in the data set. The keys were reviewed and a list of incorrect answers that frequently appeared was prepared. The next day, the teaching staff of the school who would work as the markers of the reading test was given a brief orientation session in which correct and incorrect answers were discussed. They were reminded that grammar and spelling mistakes would not be penalised. Each answer sheet was scored by two independent markers (first and second markers) and their marking was double-checked by a third marker (tripler). In the cases where there was a discrepancy between the first and second markers, the tripler resolved the discrepancy, discussing the case with a

testing office member if necessary. The scores from the marking session were typed in the SPSS program for statistical analysis. The procedures of measures of central tendency, normality judgements and item analysis were carried out. The data were further subjected to Principal Component Analysis and subtest-factor correlations were assessed for the investigation of the research question.

3.5.3 The January 2001 Test – Pilot Version

3.5.3.1 Participants: The January test was trialed on a mixed group of 86 students from the prep school of Koç University[16] and 66 students from the Boğaziçi University freshman students who were taking Advanced English courses. The reason why such a mixed group was formed was that due to the problems in scheduling, only beginner and intermediate level students from Koç University could participate in the pilot administration of the January test. Since the lack of advanced level students would have an effect on the results, a group of freshman students at Boğaziçi University were given the same test and the data from the two groups were mixed. However, it should be noted that the freshman students had almost completed a year's study in their departments and their level of English was presumably higher than that of prep year advanced students. The contribution of their high level of English might result in increased mean performance, however, since the lack of it would result in a more important imbalance, the testing office members decided to work with this mixed group. Therefore, the results from a total of 152 students were used for the analysis.

[16] See section 3.5.1.1 for the reason why Koç University students were chosen for the pilot study.

3.5.3.2 Instrument: The pilot version of the January test consisted of scanning, skimming, search reading and careful reading parts. There were 11 questions in the scanning part with one question being extra. Search reading part consisted of two texts, the first text having one skimming and six search reading questions and the second one six search reading questions. In the careful reading part, there were two texts with six accompanying questions. Thus, there was only one question extra in each part that could be eliminated from the test if its statistics was unfavourable. This was not very advantageous but the texts did not lend themselves for more questions as it may happen at times.

3.5.3.3 Procedures: The pilot version of the January 2001 test was administered in June 2000 right before the end of the academic year when the students would take the proficiency exam that would be given by their institution. The scores from this test were used to assign extra credit points to the Koç University students to encourage the students to perform as they would do under exam conditions. The pass/fail cut-off was set at 60%. The Boğaziçi University team scored the tests and reported the results to the Koç University team.

3.4.3.4 Data Analysis: The data from the January 2001 – pilot version were analysed using measures of central tendency, normality judgements and item analysis procedures. Evaluating the results, the questions were either eliminated or repaired to give the purged version to be administered in the January 2001 test. PCA could not be employed due to small sample size.

3.5.4 The January 2001 Test

3.5.4.1 Participants: The usual group that takes the January test is formed of advanced students and intermediate students who have scored 80% or above on average in the achievement exams. 650 Boğaziçi University prep school students who took the exam in January 2001 and provided complete data form the participant group of this set of data.

3.5.4.2 Instrument: In the light of the results from the pilot administration, the January 2001 test was revised and reduced to its purged version. The January 2001 test, just as the September 2000 test, was formed from four main parts. It had a scanning part with 10 questions, a skimming question, a search reading part with two texts and five search reading questions on each, and a careful reading part with two texts and five questions on each.

3.5.4.3 Procedures: The test was administered on 20th January 2001. However, there felt to be no need for a training workshop for the proctors since they were now familiar with the test and they were given written test administration procedures as usual.

3.5.4.4 Marking and Data Analysis: Marking and data analysis were carried out as explained in section 3.5.2.4.

3.5.5 The June 2001 Test – Pilot Version

After the administrations of the September and January tests, testing office members observed several problems related with the performance of the skimming item and the practicality of the test and decided to make certain changes with the test specifications. Before the June version of the BUEPT reading test is detailed, the changes the test specifications underwent will shortly be summarised. Firstly and most importantly, as it will be clear in the next chapter, the skimming item (SR1), both in the pilot and regular administrations of the September and January tests, had unfavourable item statistics (low ID and CITC values). The item had a negative impact on the reliability of the tests and the subtests. Moreover, almost the same percent of test takers could respond to the item correctly in the low- and high performing groups. It was also quite impractical to administer a part of one multiple-choice question taking five minutes. After repeated failure, it was obvious that the form in which the skimming skill was measured in the test was not the most feasible one. Thus, it was decided that it would be more beneficial to eliminate that part altogether until a better method was implemented. In the later versions of the test (June 2001 and September 2001) the skimming question (SR1) was not included.

Secondly, it was observed in the previous two administrations that a reading test taking almost 1 hour and 40 minutes, with the whole test (the BUEPT) adding up to four hours is too exhausting and the test takers might have been negatively affected by the fatigue they probably felt. It was also quite a challenging task to produce well performing tests with multiple texts when several versions had to be produced, trialed and administered in a relatively short time. It would be much preferable to

spend more time on fewer texts and produce tests with better-chosen texts and better-working items. Therefore, both search and careful reading parts are reduced to a single section with only one text. However, in order to reduce the risk of a possible reliability problem, the texts were lengthened and the number of questions was increased from five to seven for each section. Thus the reading test now had a scanning part of one text and ten questions as it had in the previous versions but a search reading part with one text and seven questions and a careful reading part with one text and seven questions.[17]

3.5.5.1 Participants: The pilot study for the June test was carried out with Koç University prep students in May 2001 as it was the case with the previous versions. Data were collected from randomly chosen 85 EFL students from differing levels of proficiency. However, the results from 71 students' performance could be processed due to incomplete data.

3.5.5.2 Instrument: The June pilot test had eleven scanning, nine search and nine careful reading questions so there were one scanning, two search and two careful reading questions extra.

3.5.5.3 Procedures: The conditions established for the September and January tests were similar for the June version.

3.5.5.4 Data Analysis: Data analysis was carried out as it was done with the previous tests. See section 3.5.1 and 3.5.3 for details.

---

[17] See Appendix 3.2 for test specifications.

3.5.6 The June 2001 Test

3.5.6.1 Participants: After the June test was piloted and reduced to its purged version, it was administered to the group of students who were graduating from Boğaziçi University, School of Foreign Languages prep year in June 2001. The ones who obtained a score of 60% (C) or above in the whole test (including the listening and writing sections) would pass the test and be eligible to register the departments to start their undergraduate studies. 1102 students, excluding the postgraduate students took the test and provided complete data.

3.5.6.2 Instrument: As explained above, the test had a scanning part with one text and ten questions, a search reading part with one text and seven questions and a careful reading part with one text and seven questions.

3.5.6.3 Procedures: The test was administered on 16th June 2001. The proctors were given written test administration procedures. No further training was needed.

3.5.6.4 Marking and Data Analysis: Marking and data analysis were carried out as explained in 3.5.2.4.

3.5.7 The September 2001 Test – Pilot Version

3.5.7.1 Participants: For the piloting of the September 2001 test, the testing office members of the Boğaziçi University could not make the necessary arrangements with Koç University due to a heavy schedule of both institutions. The pilot version of the

September 2001 was trialed on 75 freshman year students who were taking Advanced English course at the School of Foreign Languages at Boğaziçi University. These students were completing their second semester in their departments and were highly advanced in their level of English. If not the best match, these students could still be used for trialing purposes in the absence of a better group. However, since they formed a more homogeneous group, the reliability estimates might be negatively affected. This being taken into consideration, the test was given to these students in June 2001.

3.5.7.2 Instrument: The September pilot test consisted of eleven scanning questions, seven search reading questions and seven careful reading questions, which meant that there was no extra questions to eliminate. Once again, when preparing text-bound questions, test writers usually find themselves in an impasse of having to write several questions on main ideas when there are not so many of them. This was the case with the September 2001 test. Therefore, if there were malfunctioning items they would be repaired rather than eliminated.

3.5.7.3 Procedures: The test was administrated to the students taking two different sections of the Advanced English course offered by the same instructor. The instructor was previously a member of the prep school and she was familiar with the test and test administration procedures. The scores that the students obtained in the test counted 20% of their total course grade. The test was administered and corrected by the members of the testing office and the results were reported to the instructor.

3.5.7.4 Data Analysis: Data analysis was done as explained in section 3.5.1.4.

3.5.8 The September 2001 Test

3.5.8.1 Participants: After the test was revised in the light of the findings from the pilot administration, it was administrated to the incoming students in September 2001. When the postgraduate applicants were extracted from the data, there were 719 students who took the test and provided complete data.

3.5.8.2 Instrument: Similar to the June 2001 test, the September 2001 test had a scanning section with ten questions, search and careful reading sections with one text and seven questions in each.

3.5.8.3 Procedures: The test was administered on 8th September 2001. The procedure was carried out as explained in section 3.5.6.3.

3.5.8.4 Marking and Data Analysis: The marking and data analysis procedures were carried out as explained in section 3.5.2.4.

3.6 Research Question 5: Do the factor structures of different versions of the test show similarities?

For the investigation of the fifth research question, which is related with the generalisability aspect of the construct validity, the component matrices that the Principal Component Analysis yielded in the analysis of the data from the regular administrations of the tests were compared. It is hypothesised that similar component structures will be observed across four different versions of the test, giving

supporting evidence to the assumption that the reading construct as measured by these tests is generalisable across different versions. The reader is reminded that these tests are different versions of the BUEPT reading module and they were administered to different groups of test takers. However, all of them were developed based on the test specifications generated form the reading model discussed in section 2.3.5. There are quantitative differences between the first two (September 2000, January 2001) and the last two tests (June 2001 and September 2001), the last two being shorter versions with fewer texts and without the skimming part (See section 3.5.5); however, qualitatively, they reflect the same rationale in terms of test structure. Therefore, it is assumed that if the questions putatively measuring different operations (scanning, skimming, search reading, careful reading[18]) load on different factors in the same manner across different versions of the test, there is evidence for the generalisability of the construct.

3.7 Research Question 6: What will be the relation between an established criterion measure and the test under investigation?

In order to provide evidence for the external aspect of construct validity, which includes convergent and discriminant evidence from relationship of the assessment scores with criterion measures, the BUEPT reading module was compared with an established criterion measure; namely, the reading module of the IELTS (International English Language Testing System), hypothesising that the two tests would correlate significantly. The IELTS test was chosen among the two widely acknowledged international proficiency tests (the other being the TOEFL) to control

[18] Where the tests do not involve the skimming part (as in June 2001 and September 2001 tests), the component number is reduced to three.

the method variance because it is structurally more similar to the BUEPT, with longer texts and mostly short answer questions. Primarily, the comparison involved the investigation of the statistical correlation between the two tests. When evaluating the correlation, the tests were also compared content-wise to reveal the similarities and differences between them so that a more meaningful evaluation of the correlation could be done. It is assumed that the investigation of the correlation between two reading tests become more meaningful when the characteristics of the texts and items are well understood. Therefore, the hypothesis that the BUEPT and the IELTS reading tests will have a high correlation will be investigated through correlational statistics and this will be interpreted in the light of the content comparison. Each process is detailed below.

3.7.1 The Correlation between the BUEPT and the IELTS Reading Modules

3.7.1.1 Participants: 126 prep students attending the School of Foreign Languages at Boğaziçi University participated in the study. The students were chosen from three different levels of English proficiency (advanced, intermediate and beginner levels).

3.7.1.2 Instruments: For the investigation of the correlation between the two tests, the released version of the September 2000 test and a reading test from Cambridge IELTS 2 (2000, 83-94) were used.[19]

3.7.1.3 Procedures: The copies of the tests together with written instructions on how and when to administer the tests were given to the volunteering class teachers. They

[19] See Appendix 3.1.

were explicitly reminded to pay utmost attention to deliver the tests as was explained in the instructions applying the time limits strictly. They were also given verbal explanations as to the aim of the research. Most of the teachers agreed to give extra credit to the students in proportion to the scores they received in the tests. The teachers administered the tests in their classes within a week in two separate sittings. The tests were scored by the researcher and the results were reported to the class teachers.

3.7.1.4 Data Analysis: The two sets of scores from the BUEPT and the IELTS reading tests were correlated using two-tailed Pearson correlation coefficient procedure using SPSS 10.0.

3.7.2 Content analysis

3.7.2.1 Participants: The experts who participated in the study were introduced in section 3.3.1.

3.7.2.2 Instrument: In order to compare the content of these tests systematically, initially the researcher tried to obtain explicit definitions and test specifications for the IELTS reading test. She corresponded with the Cambridge ESOL Research and Validation Group. She was informed that the Cambridge tests were described depending on a range of features relating to texts and their accompanying items and they were developed as part of a local item banking system. It was also stated that the Cambridge approach to reading assessment was a rather holistic view of reading as a complex cognitive and linguistic ability, and that this was reflected in the way the

Cambridge group described item focus within their descriptive system (Email correspondence with Taylor and Underhill). Therefore, the group was unable to provide the information the researcher needed so she decided to include the analysis of the IELTS test in the content analysis scheme she developed for the analysis of the BUEPT reading test. The experts responded to the content analysis scheme that is discussed in section 3.3.2.[20]

3.7.2.3 Procedures: The experts were instructed as explained in section 3.3.3.

3.7.2.4 Data Analysis: The sum of the values that the experts marked on the scheme was calculated for the test rubric and text characteristics parts and also for the overall difficulty part, to give an overall score of comprehensibility and difficulty for the tests. For the item characteristics part, the marked operations and the text spans were summarised and the number of agreeing judges were given.

3.8 Conclusion

The study proposed above attempts to provide evidence for the construct validity of the BUEPT reading test in line with Messick's (1989a) framework. Although the study covers essential aspects of the framework, it leaves out the consequential aspect that deals with the value implications of score based interpretations and the intended and unintended consequences of the score interpretations. Messick (1996, 251) states that the short and long term consequences of score interpretation must be supportive of the general testing aims. On the other hand, consequences associated

[20] See also Appendix 3.3.

with testing are related with numerous factors 'in the context or setting and in the persons responding as well as in the content and form of the test'. He suggests that the positive or negative impact of the test on teaching and learning might be investigated by 'classroom observations or questionnaires documenting teacher and learner behaviour associated with the introduction of the test'. Such data were not available to the researcher because as explained above, after the administration of five versions of the test, the testing office members including the researcher left their positions and the test was basically changed to its previous format after September 2001. The reasons for this were related with institutional dynamics rather than the 'validity' of the test. Although such institutional consequences should definitely be considered in the introduction of new tests, it is assumed that they cannot be discussed within the concept of 'construct validity'. Therefore, the researcher will comment on the issue in the concluding chapters instead of bringing the issue in research focus.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

As discussed in the previous chapters, this study attempts to provide evidence for the construct validity of the BUEPT reading test within Messick's (1989a) framework, which conceptualises construct validity in six aspects as fundamental to all validation concerns. Six research questions were formulated to address five of these aspects; namely content, substantive, structural, generalisability and external aspects. In the present chapter, the research questions associated with these aspects will be investigated and the results from the analysis of the relevant data will be discussed.

4.2 Research Question 1: How is the construct defined and reflected in the test?

4.2.1 The construct theory and the test specifications

In Messick's (1989a) framework, the content aspect of construct validity includes evidence of content relevance and representativeness. The content relevance and representativeness of a test can be judged as a part of a broader set of construct related evidence supportive of score interpretations. Construct theory should specify the boundaries and facets of the behavioural domain of reference. The nature of the behavioural domain becomes important at the stage of test construction, where domain specifications serve as a blueprint for what kinds of items should be constructed for the inclusion in the test. The role of construct theory in developing

test specifications is to justify the implicit links between the knowledge, skills and other attributes measured by the test and the domain (ibid., 34-40). Therefore, the construct theory delimits the domain specifications, which in turn determine test specifications. Test specifications are important in determining item or task relevance and representativeness. It should also be remembered at this point that for Messick, two major threats to construct validity are construct –underrepresentation and construct-irrelevant test variance. Construct-underrepresentation occurs when the test is too narrow and fails to include important dimensions or facets of the construct. Construct-irrelevant test variance, on the other hand, stems from construct-irrelevant difficulty or construct-irrelevant easiness (See section 2.2.1). The former is taken care of by ensuring that the breadth of specifications for a test reflects the breadth of the construct invoked in score interpretation (i.e. test specifications are adequately based on the construct theory). The latter is usually investigated by item performance evidence (item analysis) (ibid., 34).

Taking content-related evidence as described above, the construct theory that specifies domain specifications becomes the heart of the discussion. Since the test specifications based on the construct theory will guide the test construction, the relevance and representativeness of the test tasks are closely linked with the relevance and representativeness of the test specifications, which are expected to be reflective of domain specifications. Therefore, in the discussion of content-related evidence, the construct theory, which is in this case an expanded reading model proposed by Urquhart and Weir (1998), and the test specifications developed based on the theoretical premises of that model should be explicitly referred to. The adequacy of the model and the adequacy of the extent to which the test and,

therefore, test specifications reflect the model are the crucial issues in such a discussion.

The reading model in question is extensively discussed in section 2.3. This model has been developed based on detailed analyses of the domain, that is, the analysis of the reading theories and the nature of the reading process. It assumes an interactive approach to reading in which the relationship between the reader and the text as well as the one between the reader and the task are acknowledged. Unlike the models that focus exclusively on careful reading at the global level, this model, the authors suggest, takes into account a wider range of reading needs of students that appeared in their needs analysis, observations and analysis of EFL curricula and teaching materials (Urquhart and Weir, 1998 and Weir et al., 2000). Within a componential framework, the model combines definitions from Kintsch and van Dijk (1978) and Just and Carpenter (1987) to adequately account for the reading behaviour both at the local and the global levels using expeditious and careful reading processes (Urquhart and Weir 1998, 100-108). Although it does not replace, and does not aim to replace these models, as it accounts for such important variables as reader purpose, comprehension focus, text coverage, rate of reading and relationship with underlying processes, it provides a satisfactorily combinatory framework for the testing of EAP reading.

However, it should be pointed out that although the framework apparently gives a clear delineation of the boundaries of the reading constructs to be assessed, the reflection of skill definitions to test specifications is not equivalently clear as manifest in Urquhart and Weir (1998, 299-305). Although in the matrix of reading

types, Urquhart and Weir (ibid., 123) draw a four-way distinction among the expeditious versus careful and global versus local reading types, they designate the focus for skimming, search and careful reading as being both global and local. That means, readers might shift from global to local focus or vice-versa depending on their perceived need along the reading process. For example, while search reading, readers might process certain parts of the text rapidly forming a general idea on what the text is about without much detailed comprehension. They, however, might shift to slow and detailed reading when they deem a particular part of the text relevant and important in relation to their predetermined purpose. In a test situation, this might mean that readers might use global and local expeditious reading skills as well as careful reading skill alternately in attempting to answer search and careful reading items despite the fact that the amount of global and local focus should presumably change according to item type. Although such an interactive description of reading behaviour is obviously more capable of reflecting real-life reading processes of test takers, it poses problems in terms of the componential categorisation of the reading skills. Particularly in terms of test development, it may not be very accurate to draw very solid boundaries between different reading skills when these could be used alternately. Therefore, it is suggested here that instead of a four-way clear cut categorisation of the reading skills, a classification that would reflect the overlapping nature of the skills might be more accurate and would remove the discrepancy between Urquhart and Weir's matrix and the test specifications they offer as exemplary.

Nevertheless, despite this inconsistency of the model, when the skills are allowed some interaction, a satisfactory formulation of test specifications (specification of

operations and performance conditions) based on this model is assumed to provide evidence for content relevance and representativeness for the test under investigation.

The test specifications

Test specifications for the BUEPT reading module are given in Appendix 3.2. Further details on the operations and performance conditions as they were embodied in the test are given below.

Exclusion of microlinguistic skills: It should be noted that the BUEPT test does not involve the assessment of microlinguistic skills, which correspond to the skills involved in 'careful reading at the local level' (See section 2.3.5). Although items tapping microlinguistic skills frequently appear in reading tests and they are found to correlate quite highly with more global reading items, there are certain concerns for their inclusion in a reading proficiency test. Based on suggestive evidence, Urquhart and Weir (1998, 134-140) state that microlinguistic skills function as 'contributory' skills in the performance of higher order reading operations but they may not constitute an adequate predictor of reading ability on their own. Where the aim is to assess whether a test taker has understood the main ideas and important details in a text, inclusion of discrete linguistic items may disadvantage some test takers since they may be able to comprehend over all ideas in a text irrespective of linguistic shortcomings. When the test specifications were being refined under the supervision of Weir, the testing office members decided to take this advice.

Selection of the texts: While the texts were selected, the operationalisations of each skill were carefully considered. For example, a text that involved several names, dates, figures, etc. for scanning and a text with clear rhetorical structure for search reading were chosen (See Appendix 3.2). Besides this, following Clapham's (1996) findings suggesting social sciences texts, when not too much subject specific, not disadvantaging students from any background, search and careful reading texts were chosen from humanities and social sciences fields. Scanning texts could involve non-specific scientific texts since in scanning the processing of text was very limited.

Format and the coverage of the items: Except for the skimming item that tapped on main idea generation, with very few exceptions, the items were written in short-answer format to ensure a direct testing of comprehension excluding test-taking strategies such as guessing.[1] In line with the test specifications, the scanning items were usually 'wh-' questions that asked for a specific date, name, figure, etc. There was a considerable overlap between the wording of the items and the text and the items could be answered within the scope of a sentence. Search reading items, on the other hand, did not involve any wording overlap in that sense; thus they were formulated by total paraphrase of the text. Search reading items also had reference to subtitles or the first sentences of the paragraphs to facilitate expeditious reading. Their scope did not extend a few sentences or a paragraph so that they could be answered without processing the whole text. As for the careful reading items, attempts were made to make sure that all the main ideas in the text were covered by the items and a test taker having answered them would arrive at the total understanding of the text.

---

[1] See Hughes (1989) and Weir (1993) for the advantages of short-answer question format.

Text length: The length of the scanning and search reading texts was adjusted so that the test takers could not read the whole text in the given time but they were rather encouraged to read selectively. The careful reading texts were shorter and could be processed in the given time limit.

Time control: The timing for expeditious reading skills was strictly controlled so that the test takers would read expeditiously rather than slowly and carefully.

Instructions: Clear instructions were given at the beginning of each section and these were read out loud in the actual test administrations to ensure that the test takers understood what they were expected to do.

4.2.2 Item writing through text mapping

In the previous section, the importance of delineation of the boundaries of the construct to be assessed and the linking function of test specifications have been discussed in relation to construct validation. It has been stressed that tests should be valid operationalisations of what we believe to be important elements of the construct in question. It is further suggested in this section that procedures concerning the item writing stage of test development should also be systematised not only to guard against threats to validity but also to allow replicability across various applications (Urquhart and Weir, 1998). Therefore, it is suggested that item writing procedure should closely reflect the operations and performance conditions designated in test specifications. Below is the discussion of a crucial procedure of test development which is expected to standardise item writing process by

minimising (more correctly, by eliminating) the possible subjectivity of a single test constructor that might potentially harm the validity of the items and, therefore, of the test; text mapping.

Jafarpur's (2003) study clearly shows that a test constructor designs questions on the basis of his/her understanding of the passage and the expected response becomes dependent on *this* interpretation, and differentially constructed tests lead, in turn, to differential performances. Sarig (1989), too, discusses that there might be various interpretations of a text depending on the variations in readers' content schemata and 'what a teacher (or a test constructor)[2] takes "to be" the meaning of a text is not necessarily the one and only possible "correct" meaning of that given text (Mann 1983, 80 in ibid). Discrepant interpretations are not restricted to test designer-test taker situation but expert readers may disagree among themselves, too (i.e. Afflerbach, 1990). Therefore, when learners come up with their own interpretations not matching the tester's, the deviations may not necessarily be due to incorrect text processing but merely to different but unexpected reading of the text. At this point, an important question arises. When the meaning is somewhat relative, how can testers meet both the standards of validity and fairness when developing reading tests in which test takers are expected to construct a certain meaning? When the argument is pushed to the extreme it may even be claimed that 'no criterion answer can be fair, since it necessarily reflects a biased interpretation' (Sarig 1989, 81). However, Alderson and Short (1981, in ibid.) claim that while a text may have multiple interpretations, the text itself imposes a limit to these infinite interpretations and 'some sort of consensus' among proficient readers should reflect 'what a given text

[2] Parentheses added.

actually means' (ibid., 81). In order to ensure that a test is a fair one, Sarig proposes that it should be based upon an interpretation validated by a consensus. She suggests 'The Meaning Consensus Criterion Answer' (MCCA) procedure in which model comprehension products of several readers from diverse content-domains are compared and the main ideas on which the majority of the readers agree are chosen to form the scoring key of a reading test based on the text in question.

Suggestions on text mapping are not limited to Sarig's. Since the interest in text analysis rose in the second half of the 1970's, many applications of teaching text structure in reading instruction have found their ways, this being supported especially by the research findings suggesting that remembering the author's text structure facilitates recall and good readers can follow the text structure better (Pearson and Fielding in Barr, et al 1991). Several techniques with different names such as networking, graphic representations, hierarchical summaries, flow charting, etc. are all based on the same logic of analysing macrostructures of text organisation. It is claimed that the use of these techniques facilitates understanding of meaning relations within a text and results in better comprehension and retention (i.e. Armbruster and Anderson, 1987, Hegarty, et al., 1991). Grant's (1993) SCROL strategy and Peterson et al.'s (2000) text mapping strategies are two recent examples of such techniques. One application of the concept mapping idea to computer environment is Carlson and Larralde's (1995).

In the field of EFL testing, Khalifa (1997) adapts Sarig's (1989) MCCA principles to validate the content relevance of the items in a battery of reading tests. She asks a large group of content-domain and language experts to mind map the texts used in

the reading tests by concentrating on main ideas. In the first round, the experts look quickly at the texts and without referring back to them they write down what they assume to be the main ideas. In the second round, they read the texts more carefully and take further notes. Two different reading styles are supposed to reflect different operations readers perform; expeditious versus careful reading. Khalifa (ibid.) calculates the ratio of consensus that the experts reached on the main idea each item is based on. She uses this information as the content related evidence for the test items and suggests this procedure as an important qualitative criterion in item analysis.

Urquhart and Weir (1998) and Weir et al. (2000) suggest text mapping as a step of a priori test validation. To determine the aptness of texts to be used for a test population (whether it satisfies performance conditions determined in test specifications, and whether intended operations could be tested with a particular text), and to decide the important ideas to be tested, texts should be 'text mapped' individually and then consensus should be established before item writing. What seems to be important in this procedure is that text mapping procedure should replicate the actual testing conditions; scanning text mapping should be done under scanning test conditions, etc. Urquhart and Weir (1998) maintain that

> 'This is a crucial first step in trying to ensure the validity of our tests. We would be concerned that the answers to the questions we then wrote revealed the important information in the text that could be extracted by the particular type of reading being assessed. An ability to answer the items should indicate

that the candidate has understood the passage in terms of successful performance of the specified operation(s).' (ibid., 276)

In line with the above suggestions, in the present study text mapping was employed as a technique to establish a consensus in both text selection and in the decisions to select important ideas to be tested. All the five test writers of the BUEPT test including the researcher were involved in the mapping procedure. Three types of text mapping were used. The following procedures were common to all.

Five texts of the nature determined in the test specifications for each skill were chosen individually (the texts were never read carefully before the mapping phase). Initially, each text was read and mapped by the test writers individually. They used as much time as is given for each section in the test specifications (i.e. 10 minutes for scanning) to read the texts since this would be what test takers would do.[3] When the time was over, they wrote down whatever they remembered to form individual text maps. They did not take notes while they read the texts since they were trying to avoid writing down too many peripheral details. If one only transfers important information from working memory to long-term storage, then writing down what ever remembered after the reading session might be a way of extracting the important pieces of information in the text. Secondly, the points that were present (recognised and remembered) in at least four of the five text maps were recorded to form a consensus text map. This was done for all the five texts consecutively. Among the five consensus maps (therefore, texts), the best was chosen for the test. Others, if

[3] Urquhart and Weir (1998) correctly state that test writers usually use considerably long periods of time to do the reading and reach deeper understanding of texts but expect test takers to reach the same level in a much shorter time. Therefore, it is important to allow for text mapping as much time as test takers would use.

they were satisfactory, were kept for future tests and the problematic ones were discarded. Each test writer had a copy of the consensus map and wrote questions on the points in it individually. They were advised not to refer to the text frequently and use the information on the map to write questions. So, there would be five different questions written on each point in the consensus map. It might seem a burdensome endeavour to write several questions on a point. However, as Jafarpur (2003) points out, each test writer formulates a question in line with his/her understanding of the passage. Besides, his/her style of item writing might differ significantly due to his/her language style and creativity. Therefore, the BUEPT test writers decided that even if it might take longer, each member of the group should work on the questions individually and come up with her alternatives. Next, the group decided on the best items to be included in the test among the groups of five alternative questions. Sometimes, a sixth version was written pulling in strengths and creativity of the alternative questions together. If there were too many questions, the group decided which ones to exclude. Finally, they went back to the texts to check for possible inaccuracies, overlaps and alternative answers so that they could guarantee the uniqueness of answers. At this stage, they edited the texts when necessary.

In particular, for scanning text mapping, five texts of informative nature with several names, numbers, dates, etc. were chosen individually. For the individual text maps, the test writers went through the texts to locate names, dates, figures, etc. in 8 minutes for each. The aim was to find out the most repeated, prominent elements of the text. In the last two minutes they put down the information they deemed important. Consensus maps were formed with the inclusion of the points that appeared in at least four of the five individual maps. Comparing the consensus maps,

the best text was chosen. Any text that was not suitable for the purpose was eliminated. Others were filed for future use. Each test writer wrote questions on the points in the consensus map of the chosen text. The questions were compared and the best ones were chosen for the inclusion in the test. The questions in this part were limited to 11 or 12 items for the pilot versions.

Text mapping for search reading part was done in two sections. The first mapping was for the skimming – the main idea question and the second one was for search reading questions. Firstly, texts with features given in test specifications were chosen. The test writers skimmed each text for 3-4 minutes individually to get the gist of the texts, focusing on titles, subtitles and the first and last sentences of the paragraphs. They recorded the points they deemed important in 2-3 minutes without referring to the texts. By comparing individually taken notes, they formed the main idea of the text as precisely as possible. As the content and form of the main idea question was predetermined (multiple choice), they formulated the distractors and wrote the question together for this item. In the second part, test takers took another 10 minutes to read the text quickly to extract the main ideas and formed a more detailed text map. They did not refer to the text when they formed their text maps. Later, as in the manner explained above, they formed a consensus map for each text, determined the text to be worked on and wrote the questions. When the questions were being finalised, however, they re-read the texts to check the uniqueness of answers. They edited the texts if necessary.

For careful reading, the time allowed for reading each text was 20 minutes. Test writers read the texts carefully by paying attention to major arguments. As careful

reading is an incremental process and readers cannot form the macrostructure until they read the whole text, the test writers did not take notes while they were reading just as in the previous mapping sessions. They used five minutes to write down main ideas without referring to the texts after they finished reading. The procedure was completed as it is explained for the search reading.

It should be pointed out here that as the part of the item writing stage, the forms of the tests prepared following the procedure described above were taken by a native speaker editor who checked for the possible language inaccuracies in the texts and the items. The tests were then taken by two (on occasion three) experienced language teachers who were not involved in test writing process. This was necessary since the test writers after a prolonged work on the tests may become blind to them and may not be aware of certain inaccuracies. These teachers took the tests under exam conditions and judged them in terms of difficulty level, freedom of ambiguity, the appropriateness of keyed answers, etc. Necessary adjustments were done according to their suggestions before the tests were piloted.

4.2.3 Item analysis

As mentioned in section 4.2.1, construct irrelevant test variance can be taken care of by item performance analysis, which establishes the item facility and discriminability values and the homogeneity of the items in the test. As part of the test development process, piloting the tests and evaluating the performance of the items to either revise or exclude them from the test was of crucial importance in minimising if not eliminating the construct irrelevant variance. Therefore, all the tests reported in this

study were trialed on a sample group and item statistics were evaluated before the tests were used in actual administration. For the sake of convenience, the item analysis results are given in section 4.5 together with factor analysis results.

4.2.4 Conclusion

As discussed above, the content aspect of the construct validity of the BUEPT reading test was taken care of by the operationalisation of the Urquhart and Weir's (1998) reading framework through the test specifications. This particular framework was found useful because it is one of the few frameworks that can function as a blue-print in the EAP testing. Despite the overlapping categories of reading skills, explicit definition of reading purposes and processes in the framework helped bridge the gap between reading theories and testing practice especially at the stage of the formation of the test specifications. Besides, the text mapping technique was used to standardise item writing process. Item analysis that will be detailed in section 4.5 facilitated the reduction of construct irrelevant variance. Therefore, all the necessary a priori validation requirements were satisfied to ensure the content aspect of construct validity.

4.3 Research Question 2: Do the experts agree on the operations measured by the test items as specified by the test writers?

For the analysis of the second research question, the part of the content analysis scheme[4] that relates to the BUEPT reading test is used. The scheme requires the

[4] See Appendix 3.2 and section 3.3 for the definitions and the explanations.

analysis of a test in terms of 'test rubric characteristics' involving test organisation (TO), time allocation (TA) and instructions (I); 'text characteristics' denoting several linguistic and rhetorical properties of the texts used in a test; 'item characteristics' specifying the reading operations (O) and the part(s) of a text processed (TS) in answering a test item, and 'overall difficulty of the questions' part tapping on, as the title suggests, the difficulty level of the questions as perceived by the evaluators. As it stands, the points of consideration in the content analysis scheme overlap extensively with the operations and performance conditions identified in the test specifications.[5] Therefore, it is assumed that judgemental evidence on the congruence between the test specifications and the evaluations of the test characteristics in the content analysis scheme by the experts will provide support for the content relevance and representativeness of the test. The analysis of the test by the help of the content scheme requires the experts (henceforth: the raters) to evaluate several properties of the test by rating these using a five-point scale. The lower ratings suggest facilitative characteristics that would improve the 'comprehensibility' of the test (See Bachman et al., 1995) whereas higher ratings would denote such characteristics that would raise the difficulty level. In line with the level of 'comprehensibility' one would like to implement in a test, it is assumed that these characteristics should similarly be reflected in the test specifications. Therefore, the scores from the six raters on the test rubric and text characteristics were averaged, the operations (O) and the text spans (TS) they marked were listed with the number of raters that opted for the alternatives given in parentheses and these are compared to the test specifications (frequencies). Where the raters commented on the nature (NT) and the rhetorical organisation (RO) of the texts, their

[5] See Appendix 3.2 for the test specifications.

choices are again listed with their frequencies. Additional comments from the raters are also taken into consideration. Tables 4.1 and 4.2 show the results of the content analysis of the BUEPT reading test.

Table 4.1: Test and text characteristics – BUEPT (scores averaged over 5)

Test Rubric Characteristics	BUEPT		
TO	1.6		
TA	2.3		
I	1.2		
Text Characteristics	**BUEPT-SC**	**BUEPT-SR**	**BUEPT-CR**
NT	magazine article (6)	magazine article (1) research/journal article (5) textbook article (1)	Magazine article (4) Research/journal article (4)
RO	narration (1) description (1) information (4) argumentation (1)	information (3) discursive (1) argumentation (4)	Information (1) Discursive (4) Argumentation (2)
GR	3	3.7	3.5
VOC	2.5	3.3	4
COH	1.6	2.8	2.2
RO'	1.8	2.3	1.8
DC	1.6	2.7	2.2
DNI	1.8	2.2	3.2
TI	1.8	3.5	3
TS	3.3	3.5	4.2
CS	1.3	1.8	1.5
OD	2.3	3.7	3.7
Av. comprehensibility score	**2.1**	**2.95**	**2.93**
Overall diff. of the questions	scanning: 1.7	skimming (SR1): 3.2 search reading: 4	careful reading: 2.7

Table 4.2: Operations and text spans (with frequencies) – BUEPT

BUEPT	Operations (O)	Text Span (TS)
scanning		
sc1	1(5), 2(3), 3(5)	2(6)
sc2	1(5), 2(3), 3(5)	2(6)
sc3	1(6), 2(3), 3(4)	2(6)
sc4	1(5), 2(3), 3(5)	2(6)
sc5	1(5), 2(3), 3(6)	2(6)
sc6	1(6), 2(3), 3(4)	2(6)
sc7	1(6), 2(3), 3(4)	2(6)
sc8	1(5), 2(3), 3(5)	2(6)
sc9	1(4), 2(3), 3(6)	2(6)
sc10	1(6), 2(3), 3(4)	2(6)
sum:	**1(53), 2(30), 3(50)**	**2 (60)**
search reading		
sr1	1(1), 2(1), 4(3), 5(3), 6(3), 7(4), 9(4)	1(1), 2(1), 4(4)
sr2	1(1), 2(3), 4(4), 5(3), 6(3), 8(5), 11(1), 12(2)	2(6)
sr3	1(1), 4(4), 5(3), 6(1), 8(5), 11(1),12(2)	1(1), 2(2), 3(3)
sr4	1(1), 2(2), 4(4), 5(1), 6(1), 8(3), 11(3), 12(2)	1(1), 2(2), 3(3)
sr5	1(1), 2(2), 4(4), 6(2), 8(4), 10(1), 12(1)	2(6)
sr6	1(1), 2(3), 4(3), 5(3), 6(1), 8(3), 10(2), 12(2)	1(1), 2(4), 3(1)
sum:	**1(6), 2(11), 4(22), 5(13), 6(11), 7(4), 8(20), 9(4), 10(3), 11(5), 12(9)**	**1(4), 2(21), 3(7), 4(4)**
careful reading		
cr1	2(1), 4(4), 8(4), 10(2), 11(2), 12(1), 13(1)	2(4), 3(2)
cr2	1(1), 2(1), 4(4), 8(4), 9(1), 10(2), 11(2), 12(2), 13(2)	2(1), 3(5)
cr3	2(1), 4(3), 8(3), 9(1), 10(3), 11(2), 12(2), 13(2)	2(1), 3(5)
cr4	2(1), 4(3), 5(1), 8(3), 9(1), 10(3), 11(3), 12(3)	2(1), 3(5)
cr5	1(1), 2(1), 4(4), 8(4), 9(1), 10(2), 11(2), 12(2), 13(1), 15(1)	2(1), 3(5)
sum:	**1(2), 2(5), 4(18), 5(1), 8(18), 9(4), 10(12), 11(11), 12(10), 13(6), 15(1)**	**2(8), 3(22)**

4.3.1 Results

Test rubric characteristics

In Table 4.1, test rubric characteristics show that the test organisation (TO: 1.6) and instructions (I: 1.2) have been evaluated as quite clear. Time allocation score (TA: 2.3) is a little higher than these showing that time allowed for the completion of certain sections might have been perceived as insufficient. In fact, two of the raters explicitly commented that the time allocated to the scanning section was not enough.

Text and item characteristics

Scanning: Starting to analyse the test characteristics by the scanning section, the nature of the text (NT) is unanimously (6 raters) identified as magazine article, and the majority of the raters (4) considered the type of rhetorical organisation (RO) informative.[6] Grammar (GR: 3) and vocabulary (VOC: 2.5) scores show that the complexity of the text is perceived as average. The low scores of cohesion (COH: 1.6) and rhetorical organisation (RO°: 1.8) show that the text has an explicit structure. The information in the text is sufficiently contextualised (DC: 1.6) and diffused (DNI: 1.8) and it is highly concrete (TI: 1.8). However, the topic is judged to be somewhat specific (TS: 3.3) although it is culture-free (CS: 1.3). The overall difficulty of the text (OD: 2.3) is found not to be very high. These evaluations give an average 'comprehensibility' score of 2.1 for the scanning text. The raters evaluated the difficulty of the scanning items as quite easy (Overall diff.: 1.7). As for

[6] Note that for NT and RO, the raters could mark more than one option.

the operations, the raters showed a high agreement in the operations (O) they used in answering scanning questions. Table 4.2 shows that they used O1, O2 and O3.[7] In sum, O1 is marked 53 times, O2 is marked 30 and O3 is marked 50 times. They marked the text span (TS) as 2 unanimously.

Search reading: The search reading text is mostly seen as a research/journal article (5) involving information (3) and argumentation (4). In terms of grammar (GR: 3.7) and vocabulary (VOC: 3.3), the text is designated as complex rather above the average. Coherence (COH: 2.8) and rhetorical organisation (RO°: 2.3) scores show that it has less than a very explicit structure. The information in the text is contextualised (DC: 2.7) at medium level and it is rather diffused (DNI: 2.2). Abstractness (TI: 3.5) and topic specificity (TS: 3.5), on the other hand, are considered moderately above the average. It is perceived as culturally not too specific (CS: 1.8) but the overall difficulty of the passage is above the average (OD: 3.7). These give an average 'comprehensibility' score of 2.95 for the text. The difficulty of the questions is 3.2 for the skimming question and 4 for the search reading question indicating that the raters found the search reading questions quite difficult. The raters reported the use of a wide range of operations for the search reading questions. As a whole, the operations with the number of instances each operation is marked in parentheses are given in Table 4.2. The most frequently used operations are 2(11), 4(22), 5(13), 6(11), 8(20) and 12(9). Besides this, except for two questions (SR2, SR5), there was not a very clear picture as to what text span should be processed to answer the search reading questions. However, the most frequently marked text span is 2(21).

[7] Note that the raters can identify more than one operation for each item.

Careful reading: The careful reading text is observed to be a magazine (4) or a research/journal article (4) of a basically discursive (4) and argumentative (2) nature. Grammar (GR: 3.5) and especially vocabulary (VOC: 4) are complex, that is, above average; however, the text is seen somewhat explicitly organised (COH: 2.2, RO°: 1.8). The information in it is perceived to be compact and abstract (DNI: 3.2, TI: 3) at average level but the topic is seen to be highly specific (TS: 4.2). Cultural specificity is rather low (CS: 1.5). The passage as a whole is difficult above average (OD: 3.7). The average 'comprehensibility' score for the passage is 2.93. Careful reading questions are rated as moderately difficult (Overall diff.: 2.7). It is interesting that although the raters reported the use of several operations in their responses to the careful reading items and they reported them in different combinations, there can be observed a consistent pattern. The most frequently used operations used are listed here with the frequencies in parentheses: 4(18), 8(18), 10(12), 11(11) and 12(10). In terms of text span (TS), the first question is analysed to be based on TS2 by four raters and TS3 by two raters but the rest of the questions are judged to be processed in the span of TS3 by the majority of the raters (5).

4.3.2 Discussion

Test rubric characteristics

The analyses the raters made yielded desirable results in terms of test rubric characteristics of the BUEPT reading test. One point to be mentioned on the insufficiency of the test time allocated for the scanning part is that such a comment was also received from the teacher-test takers at the construction stage of the test but the observations of the testing office members at the pilot stage of the test confirmed

that most student-test takers were able to finish the test on time. Therefore, no adjustments were seen necessary.

Text and item characteristics

Scanning: In the test specifications, the scanning text is designated as accessible non-specialist academic/ semi-academic journal article of informative/ descriptive nature. The vocabulary should involve no technical jargon. Rhetorical organisation should be explicit. The text should involve contextualised factual knowledge and require no background knowledge on the part of the reader. As such, the raters' scores of those characteristics of the text are in congruence with the test specifications: medium level linguistic difficulty, explicit structure, high level of contextualisation and concreteness and medium level topic specificity. With 2.1 average comprehensibility score and 1.7 overall difficulty of the questions, scanning part is designed and perceived to be the simplest part of the test. As for the operations, the first three operations in the scheme correspond to the purpose and the operationalisations of scanning in the test specifications.[8]

Since the questions asked for a specific detail (figure, date, etc.) and the wording of the questions and the text overlapped extensively, the raters reported frequent use of O1 and O3. Besides, as the questions did not appear in the order of the information in the text, they needed to go back and forth in the text to locate the answer (O2). They reported that they could find the answer in a specific part of the passage (TS2). Since the focus of scanning questions is local, that is expected as well. On the whole, it can

[8] See Appendix 3.3 for the definitions of the operations.

be said that the scanning part has been operationalised in the test in complete congruence with the test specifications.

Search reading: The picture in search reading analysis is seemingly not as clear as it is in scanning. In terms of the nature of the text and the type of rhetorical organisation, the search reading text, on which one skimming and five search reading questions are based, is perceived to be a research/journal article with information and argumentation in it. These are in line with the test specifications. The linguistic properties of the text (GR: 3.7, VOC: 3.3) are above average, which again might be expected in an academic text. However, organisation characteristics (COH: 2.8, RO°: 2.3) and especially the structure of the information in the text (DC: 2.7, DNI: 2.2) could be more facilitating since speeded expeditious reading of a text requires a very explicit text and information structure. Especially, the topic is categorised as highly abstract (TI: 3.5) and specific (TS: 3.5). The search reading text received the same OD (3.7) score with and even a slightly higher 'comprehensibility' score than the careful reading text. This was unexpected and did not match the test specifications since such qualities as implicit text structure and abstractness may impede quick accession to the main ideas in the text; thus, the appropriateness of the text for search reading purposes is challenged. Moreover, skimming (3.2) and search reading questions (4) are found to be the most difficult questions of the test. The reasons for the difficulty were determined as the lack of correct answers and the blurred focus of certain questions by one rater.

As for the operations used in reaching the correct answer, for the skimming question the operations O4-O9 are in line with the purpose and operationalisations determined in the test specifications.[9]

And the text span TS4 was the expected part to be processed in arriving at an answer for the skimming question. However, one rater claimed that the item had no relationship to the passage, and another marked TS2 because she claimed that the question could be answered by reading the conclusion only.

The other search reading questions could be answered by using several operations from O1 to O12, predominantly by O2, O4-O6, O8 and O12. No matter how the results seem to be complicated, they are not totally incompatible with the test specifications since search reading starts as expeditious reading, and after the part of the text that contains the answer has been determined through the help of several textual clues, such as the formal organisation of the text and the organisation of the information in it, the reader reverts to careful reading for the detailed understanding of a specific part of the passage. The answer to the questions should not be found with the mere matching of the words in the text as in scanning but the reader should be alert to words and phrases within the same semantic field of the key words in the question. O8 is typical of search reading and is used in 20 cases. However, the existence of 17 cases of typically careful reading operations of O10-12 show that the experts had to read more extensively and in detail to answer the questions. O1 reported 6 times, however, cannot be accounted for within the specifications for search reading. One rater consistently marked O1, O4 and O8 for search reading

[9] See Appendix 3.3 for the definitions of the operations.

questions. It is possible that she misjudged O1 since the use of O4 and O8 are consistent with search reading. As for O2, it should also be noted that by its nature, search reading requires a rapid inspection of the text for the relevant words, phrases and concepts in the same semantic field with the key words in the question. Even if search reading is characterised by occasional careful reading and it is a slower and more linear process than scanning, it necessitates a faster inspection of the text. Therefore, taking the process of 'going back and forth in the text' as optional for search reading, it can be said that O2 is a favourable operation for search reading as well.

As for the text spans, the problems are that one rater judged the text span for four search reading questions (SR3, 4, 6) as 'unclear' (TS1), and the text span for SR3 and SR4 was seen as TS3 by three raters although the expected response was TS2 for SR2-6. Here the problem might arise from the definitions of the text spans. Looking back at the definitions of text spans in the content analysis scheme,[10] it may be claimed that the distinction between TS2 and TS3 is not very clear because the term 'specific part' of a text may be understood as either a single sentence or a paragraph. Therefore, whether a sentence is related to another sentence or whether several paragraphs are linked when 'the test taker relates one part of the passage to several others' is not clear. In the future use of the scheme, with a slight modification, the text span that relates to one sentence or less than a sentence is marked as TS2* and one paragraph, as TS2.[11]

[10] See Appendix 3.3 for the definitions and explanations.
[11] See section 3.4.

The comments on the problematic aspects lead us to another discussion that should be raised here. All these results should be evaluated with the precaution that three items (SR1, SR4 and SR6) in the search reading test are found to be statistically deficient in the item performance analysis, which is discussed in section 4.5. Unfortunately, only this version of the BUEPT reading test could be released by the testing office. It is the researcher's contention that had she been able to work with a statistically better version, the results could have been much clearer. It should also be commented here that in the experience of the testing office members who developed the reading tests under investigation, search reading was the most challenging subtest to write since it required extreme accuracy and care in the selection of the texts and the wording of the items. The texts had to satisfy the clearly determined qualities; any obscurity on the part of the text organisation and the topic would risk the level of effectiveness with which the text can be processed expeditiously. In addition, the questions should make reference to either subtitles or the first sentences of the paragraphs to facilitate the search reading of the test taker, and the questions would be based on the main ideas in the text. These requirements made the formulation of the questions extremely difficult, resulting in items with blurred focus at times. However, as it can be judged from the statistical analyses, the test writers managed to cope with this problem more efficiently after the first version, which is used as the instrument here.

Careful reading: The careful reading text is judged as either a magazine article (4) or a research/journal article (4) with discursive (4) and argumentative (2) nature. Careful reading texts are supposed to be academic in nature, therefore, it is important that the majority of the raters perceived it a research/journal article. However, it

could also be considered an appropriate article for a science magazine, as two raters noted. The above average scores for grammar and vocabulary (GR: 3.5, VOC: 4), rather compact nature of the information in the text (DNI: 3.2), higher levels of abstractness (TI: 3) and specificity (TS: 4) of the topic are in line with the test specifications that suggest a careful reading text should be propositionally more demanding involving more abstract argumentation. The text is rated as having an explicit organisation (COH: 2.2, RO°: 1.8), though a careful reading text could afford more implicitness. The operations used in answering the questions are varied again ranging from O1 to O15. Predominantly used operations are O2, O4, O8, O10-12. Typically, the operations between O10-15 are considered definitive of careful reading, the last three (O13-15) indicating decoding of microlinguistic elements at the local level.[12]

O8 also involves careful reading after the information is searched expeditiously in the text. However, O1, O2 and O4 were not expected in careful reading. In the two instances reported of O1 and five instances of O2, the test takers might have made use of the names mentioned in the questions to locate where they should be reading for the answer. The high frequency of O4 is interesting and the only explanation that could be brought to this is that in order to complete the sentence-completion type of questions in the careful reading part, the test takers had to focus on the wording of the question and the text carefully to fill in the blanks with the correct phrases. Similarly, O13, though a complementary skill in careful reading which can be used when the linguistic properties of the text had to be decoded, might have served the same purpose. As for the text spans, the first question is judged by the majority of the

[12] See Appendix 3.3 for the definitions of the operations.

raters (4) to be a rather local processing (TS2). Otherwise, most of the raters (5) estimated that larger portions of the text had to be processed for the successful completion of CR2-5 items. This is again in line with the careful reading specifications that require the understanding of the main ideas of the text to form a macrostructure. It can be said that with the large proportions of the text covered by the test items, the majority of the information in the text is processed.

All in all, the content analysis by the language experts gave important support for the successful operationalisations of the test specifications in the test except for a few problems with the search reading test. Therefore, the hypothesis that the content analysis by the language and testing experts will reveal that the items measure the operations specified by the test writers is essentially confirmed. It is also seen that the content analysis scheme based on Bachman et al. (1995) and developed by the additions from the definitions of Urquhart and Weir (1998) and Khalifa (1997) is a useful instrument to gain systematic information on a test. It is possible to reveal both the strenghts and weaknesses of a test and the congruence and incongruence between the test and the specifications with the use of this instrument.

4.4 Research Question 3: What are the operations utilised by the test takers to arrive at the correct answers?

The third research question is formulated to investigate the substantive aspect of construct validity. In order to do this, immediate recall protocol data were collected and analysed as explained in section 3.4. Intra-rater reliability was investigated through the comparison of the categorisation of the operations (O), text spans (TS),

test taking strategies (tts) and observations (OBS) done in the second and third listening of the protocol data. The operations and text spans used by the test takers were totalled and compared to the test specifications. Test taking processes and observations were also analysed to investigate construct-irrelevant variance. Besides, the operations that were used by the test takers who could process the texts and the questions favourably were compared in relation to search and careful reading tests. Lastly, some illustrative definitions that the test takers provided for their reading processes were given.

4.4.1 Results

4.4.1.1 Intra-rater reliability

In order to determine the intra-rater reliability, the dissimilarities in the markings between the first and second categorisations were counted and compared to the total number of markings in the second classification. For example, if the test taker S1 was determined to have used O1, O2 and O4 for SC1 in the first categorisation but O2, O3, O4 and O5 in the second, the discrepancy was calculated to be 3 in 4 for that item. For the scanning part, there was no discrepancy between the first and second categorisations. For the search reading part, there were a total of 600 operations (O) marked in the second classification and 71 marks did not appear either in the first or in the second categorisation giving 12% discrepancy. For careful reading, 542 operations were marked and 41 marks did not appear either in the first or second categorisation giving 8% discrepancy. Among 452 text spans (TS) marked in the second classification, 20 cases did not match with the first one (4% discrepancy) and among 193 test taking strategies (tts) marked, there were only 4 cases of discrepancy

(2%). No discrepancy in the observations (OBS) between the first and second classifications was observed. Roughly, this comparison suggested that intra-rater reliability was around 0.9, which was quite satisfactory.

4.4.1.2 The operations and the text spans

Table 4.1 summarises the findings of this analysis. The numbers in parentheses give the frequencies of the instances noted. The information collapsed in this table will be discussed in two separate parts. In the first part, the operations used and the text spans processed by the test takers when answering questions will be discussed in relation to the test specifications. In the second part, the test taking strategies and the observations will be discussed to pin down any construct-irrelevant variance.

Table 4.3: Verbal protocol analysis results

	Operations (O)	Text Span (TS)	Test Taking Strategies (tts)	f/c*	f/inc	unf/c	unf/inc	unans
Scanning								
SC1	1(15), 2(15), 3(15)	2*(15)					1	
SC2	1(15), 2(15), 3(15)	2*(15)					2	
SC3	1(15), 2(15), 3(15)	2*(15)	5(1)				2	
SC4	1(15), 2(15), 3(15)	2*(15)					1	
SC5	1(15), 2(15), 3(15), 8(1)	2*(15)	1(1), 5(1)				1	
SC6	1(15), 2(15), 3(14), 4(1), 8(1)	2*(15)	1(1), 5(1)				2	
SC7	1(15), 2(15), 3(15)	2*(15)					1	
SC8	1(15), 2(15), 3(15)	2*(15)	1(1)				3	
SC9	1(15), 2(15), 3(15)	2*(11)	1(3), 5(2)					4
SC10	1(15), 2(15), 3(15)	2*(15)	1(2), 6(1)				2	
Search reading								
SR1 (SK)	2(6), 3(4), 5(6), 6(15), 7(15), 9(10), 13(1)	3(1), 4(14)	6(7)	8	4	0	3	
SR2	2(15), 3(12), 4(9), 8(12), 10(3), 11(2), 12(2), 13(2)	2*(3), 2(11), 3(1)	1(3), 2(3), 5(1)	11	0	2	3	
SR3	2(14), 3(9), 4(14), 8(7), 10(1), 11(5), 12(2)	2*(3), 2(6), 3(5)	1(6), 2(2), 5(4), 6(1)	3	2	0	9	1
SR4	2(14), 3(8), 4(13), 5(1), 6(1), 8(3), 10(4), 11(5), 14(1)	2*(2), 2(3), 3(7)	1(6), 2(1), 4(2), 5(7)	2	2	1	5	5
SR5	2(10), 3(9), 4(9), 6(14), 8(7), 10(6), 12(1)	2*(3), 2(7), 3(4)	2(2), 5(2)	7	0	2	5	1
SR6	2(15), 3(13), 4(13), 8(8), 10(6), 12(1), 13(1)	2*(1), 2(11), 3(3)	1(4), 2(2), 4(1), 6(1)	5	1	1	8	
SR7	2(15), 3(15), 4(10), 6(2), 8(11), 10(3), 11(1), 12(3), 14(3)	2(14), 3(1)	1(1), 2(7), 3(1), 4(1), 5(2)	6	1	3	5	
SR8	2(14), 3(10), 4(13), 6(2), 8(8), 10(5), 12(1), 15(1)	2*(3), 2(6), 3(5)	1(5), 2(6), 5(5)	5	0	5	3	2
SR9	1(9), 2(12), 3(12), 4(2), 6(3), 8(12), 10(1), 12(1)	2*(5), 2(9)	1(1), 5(1)	14	0	0	0	1
SR10	2(14), 3(7), 4(11), 6(7), 8(9), 10(2)	2*(2), 2(7), 3(3)	1(2), 2(1), 5(4)	8	0	0	4	3
SR11	1(2), 2(11), 3(9), 4(7), 6(11), 8(8), 10(3), 12(1), 15(1)	2*(2), 2(10), 3(3)	1(1), 2(5), 4(1)	9	0	6	0	
Careful reading								
CR1	1(5), 2(7), 3(7), 4(5), 8(3), 10(12), 12(1), 13(2), 14(1)	2*(1), 2(5), 3(9)	1(2), 4(1), 6(1)	10	1	1	3	
CR2	1(10), 2(12), 3(10), 4(6), 5(1), 6(1), 8(9), 10(7), 12(4), 13(2), 14(1)	2(9), 3(6)	1(1), 2(3), 4(2), 5(1)	9	2	1	3	
CR3	1(13), 2(10), 3(8), 4(6), 5(2), 6(1), 8(7), 10(6), 12(4), 13(2)	2*(4), 2(9), 3(2)	1(1), 2(2), 4(3), 5(1)	6	1	5	3	
CR4	1(14), 2(8), 3(6), 4(1), 5(1), 8(5), 9(1), 10(6), 11(2), 12(5), 13(1)	2*(2), 2(3), 3(10)	1(2), 2(3), 5(2), 6(2)	3	1	6	5	
CR5	1(14), 2(11), 3(8), 4(6), 8(8), 10(9), 12(4), 13(3), 14(2)	2*(1), 2(4), 3(10)	2(4), 6(3)	8	2	2	3	
CR6	1(2), 2(15), 3(14), 4(7), 8(8), 10(6), 13(5)	2*(2), 2(9), 3(4)	1(1), 2(6), 3(1)	8	0	5	2	
CR7	2(13), 3(15), 4(2), 8(8), 10(7), 12(2), 13(8), 14(1)	2*(1), 2(13), 3(1)	1(1), 2(3), 4(1)	10	0	2	3	
CR8	2(15), 3(15), 5(3), 8(6), 10(6), 11(1), 12(1), 13(1)	2*(2), 2(9), 3(4)	2(5), 3(1), 4(2)	7	1	4	3	
CR9	2(14), 3(14), 4(4), 8(6), 10(6), 11(3), 12(5), 13(1)	2*(2), 2(8), 3(5)	1(4), 2(5), 3(1), 4(4), 5(2), 6(2)	5	0	1	9	
CR10	2(13), 3(13), 4(1), 8(7), 10(3), 11(5), 12(3)	2(11), 3(4)	1(2), 2(3), 4(1), 6(1)	12	0	2	1	

* **f/c**: favourable comprehension/correct answer **f/inc:** favourable comprehension/incorrect answer **unf/c:** unfavourable comprehension/correct answer
unf/inc: unfavourable comprehension/incorrect answer **unans:** unanswered

Scanning: An inspection of Table 4.3 shows us that the subjects uniformly used O1, O2 and O3 in arriving at the answers to the scanning questions. This means that they looked for figures, dates, names, etc rapidly by going back and forth in the text and they could answer questions by matching the exact key words or phrases in the question and the text. They read a sentence or sometimes less than a sentence to answer scanning questions (TS2*). The emerging picture is very clear and completely in line with the test specifications.

Skimming: SR1, the multiple choice skimming question requiring the extraction of the main idea, prompted several operations including rapid inspection, exact key word matching, but basically reading the title, subtitles, first and last sentences of paragraphs (O6: 15) and reading the introduction and conclusion carefully (O7: 15). Six test takers explicitly reported the importance of introduction and conclusion and section beginnings and endings demonstrating their knowledge of text structure (O5). 10 test takers mentioned that they had to form a general idea pulling together what they had read in order to answer this question (O9). Except for one, all the subjects inspected the whole text though skipping large portions especially mid-paragraphs (TS4). These operations are also congruent with the test specifications.

However, for search and careful reading subtests, such a clear conclusion could not be reached. Depending on the characteristics of individual questions, the test takers employed several reading skills. Table 4.4 shows the sum of the operations and the text span used for each subtest.

Table 4.4: The sums of operations and text spans by the subtests

Oper.	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10	O11	O12	O13	O14	O15
SC	150	150	149	1	0	0	0	2	0	0	0	0	0	0	0
SR	11	140	108	101	7	55	15	85	10	34	13	12	4	3	2
CR	58	118	110	38	7	2	0	67	1	68	11	29	25	5	0

Text S.	TS1	TS2*	TS2	TS3	TS4	TS5
SC	0	146	0	0	0	0
SR	0	24	84	32	14	0
CR	0	15	80	55	0	0

Search reading: To remember, search reading process is defined in the test specifications as rapid, selective reading process performed by keeping alert for words in the same or related semantic field with the topic of search, using formal/textual knowledge, titles and subtitles and paragraph beginnings and endings to locate the relevant information. Once the relevant information is located, the process is likely to involve careful reading. Therefore, the reader is likely to begin the process by focusing on the key words/concepts in the question and searching for them in the text (O2-O4). The reader might make use of the subtitles, and read first and last sentences of the paragraphs to inspect the relevance of the content (O6). Finally, he or she reads carefully to confirm the answer after deciding the location (O8). O8 is more likely to involve a shorter text span (e.g. one or two -short-sentences) compared to O10-O12, which require detailed understanding of main ideas (e.g. within one full paragraph). The sums for SR in Table 4.4 show that the test takers did indeed use search reading operations (O2: 140, O3: 108, O4: 101, O6: 55 and O:8 85) but there are also 59 cases of careful reading operations (O10-O12) recorded in SR row. This should indicate either that once the search reading process failed, the test takers reverted to careful reading process or that they preferred to read larger spans of text so as to ensure comprehension. Note that majority of text spans is marked as TS2 (84) and there are 32 cases in which the test takers reported the reading of more than one paragraph (TS3).

Careful reading: In the test specifications, careful reading is defined as a process of establishing a macrostructure for the text by reading it thoroughly in order to comprehend main ideas and supporting information. What is surprising with the sums of the operations given in the CR row of Table 4.4 is that the test takers employed expeditious reading strategies also in careful reading subtest. A typically scanning operation O1, which involves looking for figures, names, etc. was used to locate the names of the researchers mentioned in the questions in 58 cases. Similarly, O2 and O3 were again used to locate the key words/concepts given in the questions (118, 110 cases, respectively). 67 cases of O8 show that certain test takers attempted to answer questions reading a short span of text. However, the existence of 108 instances of careful reading operations (O10-O12) shows that the test takers had to revert to careful reading of larger spans of the text almost twice as much as they did in search reading (59 instances compared to 108). They made use of contributory linguistic skills (O13-O15) in careful reading more than they did in search reading (30 cases compared to 9). There are also slightly more instances of TS3 in which the test takers had to read more than one paragraph in careful reading in comparison to search reading (55 cases compared to 32).

In sum, it is evident that in the careful reading test, certain test takers started the reading process by expeditious reading (O1-O4 and O8) and continued with careful reading operations (O10-O12) when necessary. Considering that the test takers had to complete the test in a given test time, it was legitimate for them to try to arrive at the answer as quickly and economically as possible. But to what extent the nature of the search reading items allowed expeditious reading on one hand, and on the other, to what extent careful reading items required careful processing of larger text spans

are also a matter of construct validity and necessitates closer analysis of the nature of the items. This will be investigated through the analysis of the test taking strategies used by the subjects and the observations on their comprehension process.

4.4.1.3 The test taking strategies and the observations

Table 4.5 below shows the frequencies and the percentages in relation to the observations made on the outcome of the test taking process. Table 4.6 summarises the frequencies of the test taking strategies used in each subtest.

Table 4.5: The outcome of the test taking process

	Scanning	Search reading	Careful reading	Search and Careful reading
No. of cases	150	165	150	315
favourable/correct	131 (87.3%)	79 (47.9%)	78 (52%)	157 (49.5%)
favourable/incorrect	0	10 (6.1%)	8 (5.3%)	18 (5.7%)
unfavourable/correct	0	20 (12.1%)	29 (19.3%)	49 (15.5%)
unfavourable/incorrect	15 (10%)	44 (26.7%)	35 (23.3%)	79 (25%)
unanswered	4 (2.7%)	12 (7.3%)	0	12 (4.1%)

Table 4.6: Test taking strategies

	SC	SR	CR
tts1	8	29	14
tts2	0	29	34
tts3	0	1	2
tts4	0	5	14
tts5	5	26	6
tts6	1	2	6

It is seen that the majority of the processes that enabled the test takers to arrive at the correct answers are favourable in all of the subtests (favourable/correct). There are a certain number of cases in which the test takers failed to show sufficient comprehension and could not answer the questions (unfavourable/incorrect), as it would be expected in any test situation. What is of interest here are the cases in which the test takers processed the question and the text favourably but could not answer the question correctly (favourable/incorrect) and the cases in which the test

takers did not exhibit sufficient comprehension but could get the answer right (unfavourable/correct). When the outcome is not favourable despite favourable comprehension, it might be assumed that the test involves 'construct-irrelevant difficulty' and when the opposite is true, it might be suspected that the test involves 'construct-irrelevant easiness', both of which are major threats for construct validity. Although the ratio of such cases is low in the data, they will be discussed in relation to the subtests and the items below.

Scanning: It is seen that most scanning questions could be answered correctly by favourable processing and there are no odd cases in this section. Nevertheless, it is worth noting here that despite the relative easiness of the scanning test for the subject group, there is a 10% incorrect answers, which simply resulted from the fact that some test takers did not even read the sentence in which the answer was located. This was considered 'unfavourable' processing. There are four instances where SC9 could not be answered (See Table 4.3). The reason for this was that the test takers could not locate the key word 'West'[13] most probably due to its physical location in the text.

Search reading: In this section, there are 10 cases (6.1%) of **favourable comprehension/incorrect answer**. The distribution of these to questions is as follows:

SR1: (4 cases: subjects S2, S4, S9, S10) All these subjects read the first paragraph and the conclusion carefully, inspected the paragraphs by reading the beginnings and endings and paid attention to the subtitles. They demonstrated their comprehension

[13] See the test and the questions in Appendix 3.1

by translating or summarising the parts they read. S4 tried to guess the answer before reading the text (tts6). However, while marking the correct option that gives the main idea of the passage, they decided that between the options *c* and *d*, *d* gives a more general idea and it is more logical therefore *d* should be the correct answer. The problem here seems to be that the subjects used an inferential reading comprehension process where literal understanding was tested. They were mislead by their own logical assumptions rather than the insufficiency of their comprehension.

SR3: (2 cases: subjects S3, S6) S3 understood the question and the text, found the answer but the answer 'to push beyond the natural growth stages' did not mean much to her and she wrote 'something more than a neutral position' as the answer. S6 found the answer but brought in his interpretation that it is 'play and creative activities' that should be pushed so he wrote 'play and creative activities should be made more genuinely demanding'. Since the idea that states the correct answer was not supported with further explanation in the text, it seems that it was to a certain extent hard to interpret that part of the text for these test takers.

SR4: (2 cases: subjects S3, S6) Here the problem was that both of the subjects found the answer 'wide range of symbols' too general. Instead, they wrote 'mathematical symbols' as the answer, again using their judgement. 'Mathematical symbols' appeared to be a strong distractor in the text.

SR6: (1 case: Subject 14) She combined two pieces of information to form an answer that seemed logical to her. She assumed if the child was asked to make further responses, his capacity would improve and she wrote 'making further responses' as the answer.

SR7: (1 case: Subject 3) S3 provided partially correct and logical answer to this question. Instead of 'concealed from family and community', she wrote 'concealed

from community and don't come to light'. Since the information is partial and partial credits were not assigned, her answer was deemed incorrect.

Shortly, what seems generally to be the problem in these cases was the way the subjects brought in their interpretation to the question and the text. However, checking the very low performance of SR3 (See Table 4.3), it should be pointed out that the answer to the third question was not easily understandable even for the test takers who could understand the paragraph in which it was located. This may be both due to the grammatical structure of the sentence with an expletive subject (i.e. It is necessary to push beyond the natural growth stages), or to the semantic obscurity of the concept 'natural growth stages'. It is known that unless such concepts are explicitly discussed in the text, the lack of background knowledge on the part of the test taker might be a disadvantage. It has also been observed that SR4 received too few correct responses in general. Most of the test takers, even the ones who read frequently carefully, could not locate the answer correctly. Some test takers explicitly commented that the question was too general and there were no clear key words to help them in their search.

Careful reading: There are 8 cases (5.3%) of **favourable comprehension/incorrect answer** in the careful reading part. The cases are as follows:

CR1: (1 case: Subject S12) S12 understood the question and read extensively but eventually decided that what was discussed in the passage was a matter of 'intelligence'. What S12 had to do in order to give the correct answer was to go back

to the question and the text again to determine the specific information required in the question. Her behaviour could be attributed to mere carelessness.

CR2: (2 cases: Subject S1, S12) S1 could find and write the answer but doubted its accuracy. He crossed it out to go back to it later on but did not have time to do that. S12 again found the answer 'inhibit their aggression' and translated it but thought that it was not an ability and, therefore, she should find another answer.

CR3: (1 case: Subject S6) Although S6 read extensively and understood the details of the experiment, he was mislead by the word 'through' in the question and thought it should be followed something like 'somebody's perspective'. He read the text with this information in mind and wrote 'dominant's perspective'.

CR4: (1 case: Subject S10) S10 was one of the few test takers who made comments on the hypothetical nature of the question (...might have been proven if only..) and read extensively. After she read the correct part of the text and understood the relation, she went back to the beginning and got confused. She skipped the question for later attempt (tts5) but did not go back to the correct part in the second trial and could not provide the correct answer.

CR5: (2 cases: subjects S2, S5) S2 tried to guess the answer (tts6) before she read (it must be something like 'influence') and read the text with this in mind. Although she filled in the second blank correctly, she could not find the word for the first blank and could not be given credit. S5 uttered the correct response 'empathy' depending on what she had read to that point before she read the text for CR5 but after she found the correct answer, she did not check the structure of the question again and wrote 'appears to be a response of empathy', which yielded an incorrect answer.

CR8: (1 case: Subject S2) S2 interestingly said that the answer for the question seemed to be 'to pull his observations together', which was right. However, she

added that she was not satisfied with the answer because 'thinking capacity' should also be mentioned. Depending on her own conclusion, she wrote 'thinking and observation capacity', an answer which was not specific enough for that particular question.

Once again, for these readers there were no major problems of comprehension but either through carelessness or the use of inappropriate reading strategy, they provided faulty answers. In sum, except for SR3, the data did not reveal evidence of construct-irrelevant difficulty.

The next analysis concerns the **unfavourable comprehension** and **correct answering** of the questions. The cases in which the test takers could answer the questions without substantial understanding of the text might be indicative of construct-irrelevant easiness and must be analysed with care.

Search reading: In the search reading subtest, there were 20 such instances (12.1%). The distribution is as follows:

SR2: (2 cases: subjects S5, S13) Both S5 and S13 located the key word 'mass media' in the second paragraph, rapidly read the sentences having partial understanding of the text. Both paid attention to 'type of' 'kind of' synonymity and decided that the answer should be located in that part of the paragraph (tts2). S13, who could locate the answer in his second attempt (tts1), additionally said that 'generate' in the question and 'produce' in the text matched, too.

SR4: (1 case: Subject S4) S4 went back and forth among last three paragraphs on the first page in order to find something equivalent to 'advances humankind is likely to make'. She thought these might be innovations, technological developments, etc. She had a general idea what the text was about but could not summarise it sufficiently. She reported that she could eventually match 'handling' and 'cope with' and wrote the answer (tts2).

SR5: (2 cases: subjects S3, S4) Both S3 and S4 understood the question and by the help of the subtitle 'pre-linguistic stage', they decided that the answer should be located in that section of the text. However, they processed the first paragraph only to find the answer. Although they basically understood the discussion in the paragraph, they did not pay attention to the 'recorded material' part in the question, thus answering it prematurely. However, since the discussion was introduced in that paragraph and the correct answer was also explicitly stated there, the question could be answered with partial correspondence with the text and the question. Hence, instead of expeditiously reading three paragraphs, the question could be answered within the limit of one paragraph. Whether this is a weakness of the question and whether such a processing of the text is unfavourable is debatable. However, it is a fact that this was not taken into consideration in the design of the item.

SR6: (1 case: Subject S8) S8 reported that he had difficulty in answering the question because the question was too general. He reported that he found 'one-way presentation devices' and read quickly till the end of the paragraph roughly understanding that such devices were not considered adequate. But since the paragraph was too long, he speeded up and rolled his eyes over the lines to catch 'cognitive development'. He kept on until he saw 'active involvement' which appeared in the question, too. Since there was 'cognitive development' also in that

sentence, he wrote it as the answer without being sure. He said he picked that phrase because it was repeated and it seemed to be the subject matter in that paragraph (tts4). It sounded logical, too.

SR7: (3 cases: subjects S4, S5, S12) All the test takers had problems with the vocabulary in the question but located the paragraph to be read by mere matching 'Islamic culture' in the text and the question. The question had two blanks to be filled so they looked for an answer with two words combined with 'and' (tts2). There were three such structures in the paragraph (public and private, good and evil, family and community). By minimum understanding, these subjects could detect the right answer. S4 also used tts1 and tts5 in locating the answer.

SR8: (5 cases: subjects S2, S5, S9, S11, S13) These test takers searched for 'cultural differences' and 'at home and outside' that appeared in the question. When they could not find the exact or near matches they had difficulty in locating the relevant part of the text and utilised tts1 and tts5. They again looked for an answer of two parts since the question had two blanks combined with 'and'. When they could match 'lead to' in the question with 'subject to' in the text followed by phrases 'transformation and acculturation', they wrote the answer.

SR11: (6 cases: subjects S2, S4, S5, S7, S9, S12) In completing SR11, these test takers mainly looked for the phrase 'historical evidence' in the text. When they could not find the phrase, some looked for dates and some made use of the title. But what enabled them to arrive at the answer was that it was the last question and hence should be located in the last parts of the text (tts1) and the mere match of the word 'go through' in the question and the text (tts2). None could explain what was discussed in the text and most had problems with the word 'acknowledge' in the question.

For the search reading part, the expeditious reading of these test takers consisted of scanning for the exact matches, searching for the synonyms or related words with the key words in the question and matching the part of the text that contained the answer with the question to extract the answer with either with no or partial understanding or with the comprehension of a single sentence. What their process lacked according to the test specifications is that they did not check the paragraphs content-wise, did not form a general idea what they may be about to decide on their relevancy but used the order of the questions to narrow down the text span to be processed, and either read too little or did not read at all to confirm their answer. Therefore, in certain cases, their process involved only matching of words. Among several test taking strategies, they made use of tts1, tts2, tts4 and tts5 frequently.

Careful reading: 29 cases, 19.3% of the responses, in the careful reading section classified as unfavourable/correct are given below:

CR1: (1case: subject S7) S7 reported that he looked for the word 'capability' in the question or its synonyms 'ability, capacity', etc. in the paragraphs where the name of the monkey, Binti, mentioned and as he found the words, he underlined them. He said: 'I checked for the other words in the question, too. If I also found these words where a 'capability' was mentioned, I knew that the answer was there (tts2). I read that sentence once again.'

CR2: (1 case: subject S7) S7 looked for the name of the researcher, Stammbach, and 'high ranking monkeys' by scanning the text. When he found other key words in the question repeated within the same part of the text (tts2), he narrowed down the text span to read only that sentence carefully.

CR3: (5 cases: subjects S1, S4, S11, S12, S13) Most of these test takers reported that they did not understand the question when they first read it. They especially did not figure out what type of an answer they were required to give after the word 'through' in the question. They scanned for 'Kummer and Cord' to locate the paragraph to be read but after a brief glance to that paragraph, they decided that it was about the set up of the experiment including details, which they did not need to read since they were looking for some kind of a result, which might be given towards the end of the paragraph (tts4). The phrase 'social rules' was also important for them so they read quickly until they found 'social conventions'. The concluding words 'Here then…' were also helpful in signalling them that the result of the experiment was given in that sentence. They read the last sentence of that paragraph carefully to understand that the answer was there.

CR 4: (6 cases: subjects S4, S5, S6, S8, S11, S13) The test takers again scanned for Miller, the researcher's name, to locate which paragraph to read. The problem with this question was that its hypothetical nature ('…the existence of empathy might have been proven if…') could not be detected by these test takers (except for S8, who declared that he did not understand why the sentence was hypothetical). They simply thought that in that experiment one group of monkeys got the shock and the other saw it. With this in mind, they read the first and second paragraphs of Miller's experiment carefully to find the implicitly stated relation of 'the actor' and 'the receiver' there. Here, the process was indeed careful reading and extended over two paragraphs. However, the students did not pay attention to the fact that the question was based on a weakness of the experiment and they did not read until the paragraph in which it was discussed. This was unforeseen by the test writers.

CR5: (2 cases: subjects: S4, S9) Both S4 and S9 correctly understood that the first blank had to be related with 'empathy' depending on what they had previously read (tts6). They scanned for Masserman, the researcher, and skipped the first paragraph because they understood that it related the details of the experiment. They confirmed their guess by quickly reading the second paragraph of Masserman's experiment (merely understanding that the subject discussed in that part of the text was 'emphathy'). So far, the process was a successful search reading process (not careful reading though) but for the second part of the question, they rapidly inspected the text to match the words 'conscious' and 'of others' in the question with 'aware' and 'others' in the text and filled in the question correctly by reading one sentence carefully (tts2).

CR6: (5 cases: subjects S5, S7, S9, S13, S14) All these test takers exhibited similar behaviours in their attempt to answer CR6. They read the question, understood it, determined the key words and looked for them in the text: They found 'the preschool child' given in the question and looked for what this child is 'unable to differentiate between'. They determined that the question is negative and read the text very rapidly until they found 'but … does not make a clear separation between…' without clear understanding of the argument. They decided that the answer was there because this sentence matched the question (tts2). None could tell what that sentence meant.

CR7: (2 cases: subjects S1, S7) S1 and S7 read the question and decided that they should find the words 'mathematics and physics' in the text. They expected to find the exact matches since they thought these words could not be paraphrased. Besides, since the question asked something that a child could not grasp, they should look for some negative expression. They found 'mathematics and physics' by scanning the text. The phrase 'because of this lack' within the same sentence showed that the

answer was there. They jumped to the beginning of the paragraph to discover that some concept of 'reversibility' was discussed there and decided that it should be the answer (tts4). They did not confirm their answer by reading from the beginning to the part in which they found 'mathematics and physics'.

CR8: (4 cases: subjects S1, S3, S11, S13) What these test takers did to arrive at the answer without detailed processing of the text was to find the key word 'pinball' in the text and search that paragraph for one 'capacity' that an 'older child' has. Most thought that the question was related with the result of the experiment, therefore, did not pay much attention to the details by skipping a large part until they found what they considered to be a concluding statement (the last sentence) where they also found the word 'ability' (tts2). The answer, 'ability to pull his observations together', however, did not make sense to them.

CR9: (1 case: subject S9) S9 reported that he could not find many key words in this question and had difficulty in locating the answer. He limited the text span to be processed by the help of the order of the questions (tts1 and tts5) and scanned for the word 'possibility' given in the question to write what follows it as the answer. He only considered that the text was giving a negative fact and he had to write some positive structure for the answer.

CR10: (2 cases: subjects S7, S13) CR10 exemplifies an ability that a child at 'the third stage' will have and asks what this ability may be. Both S7 and S13 looked for the example in the text but since it was not explicitly given, they decided that any mention of an ability where they could find the phrase 'the third stage' would correspond to the example given in the question. They did so to find the answer (tts2).

In short, the above examples illustrate that certain test takers could arrive at the correct answers without favourable processing of the question, the text or both in the careful reading part. The analysis of the unfavourable comprehension/correct answering process revealed that some test takers employed expeditious reading skills (scanning and search reading) to locate the answer, followed only briefly by careful reading. They had partial understanding of the parts they read and were unable to form macrostructure of the text. Among the test taking strategies, they used tts1, tts5, tts6 to facilitate their search and especially tts2 and tts4 to extract the answer.

Another analysis that could be informative at this stage was to investigate the operations utilised by the test takers who exhibited favourable comprehension. Both the favourable/correct and the favourable/incorrect processes were analysed and categorised into three groups (SR, CR, SR+CR) to see whether these test takers indeed used search reading operations to answer search reading items (SR), and similarly, careful reading operations, for careful reading items (CR). [14] The test takers might have used a combination of these operations starting with search reading to locate the answer and the processing of a short text span (e.g. one sentence) but followed by extensive reading (e.g. the whole paragraph) resulting in the comprehension of the main idea in that part of the text (SR+CR). Table 4.7 shows the distribution of the operations to the items and the sums.

[14] The scanning test is excluded from this analysis for the reason that it was observed that the operations specified in the test specifications were uniformly utilised by the test takers.

Table 4.7: Processes in favourable comprehension by item

No: 88	SR1	SR2	SR3	SR4	SR5	SR6	SR7	SR8	SR9	SR10	SR11	Sum
SR	12	5	2	1	3	1	2	1	14	8	6	55
CR	0	2	3	3	0	0	1	1	0	0	0	10
SR+CR	0	4	0	0	4	5	4	3	0	0	3	23
No: 85	**CR1**	**CR2**	**CR3**	**CR4**	**CR5**	**CR6**	**CR7**	**CR8**	**CR9**	**CR10**	**Sum**	
SR	1	2	2	0	0	3	0	1	0	1	10	
CR	9	5	4	4	7	3	1	4	5	11	53	
SR+CR	1	4	1	0	3	2	9	2	0	0	22	

Among 88 attempts to answer the search reading questions, 55 (62.5%) were mere search reading operations, but in 33 (37.5%) cases the test takers had to read carefully (CR or SR+CR). For example, items such as SR2-8 required some careful reading for the majority of the text takers. In the careful reading part, the questions could be answered by mere search reading operations just in 10 cases (11.8%) but in the rest of the instances, the test takers had to read carefully even if they used search reading operations before careful reading in 22 (25.9%) cases. Then, it can be assumed that the majority of the processes the successful 'comprehenders' utilised were in congruence with the test specifications but in a significant number of cases in the search reading part, they had to shift to careful reading processes.

As the last note of this section, some test takers' definitions of the various reading processes they reported to have used will be given to provide support for the assumption that readers adjust their reading process in line with their predetermined purposes, which basically underlies the reading framework the test under investigation was founded on. Whether the definitions that the test takers provided match the test specifications or not is also of importance.

Scanning: I just look for key words. When I find them, I check a few words that come before them. If they match with the question, I don't read more (S1). I just check for the key words as if they were pictures. I don't read and understand them

(S2). I quickly look for what ever easy to find; a number, name, for example. It is enough to read only one or two words when I locate the key words. I don't understand the text (S3). It is easy to find words with capital letters or numbers in the text quickly. So I only look for them in the text (S4). I read all the questions first and underline the key words such as names and numbers. Then, I try to find them in the text without reading anything else. Sometimes it is enough to read a few words to answer the question (S6).

Skimming: First of all, I read all the options to understand what the text might be about, and decided what to pay attention to in the text. For example, media looked important to me, also whether the child is under five or not. Then I read the title and the first paragraph carefully. I read the conclusion, too. I could not be sure of the answer so I quickly read the beginnings of the paragraphs. I read the first sentences. If they were not clear, I read one more. Then I quickly went over the lines, sometimes understanding a little. On the second page, I read the subtitle and the first paragraph carefully but I skipped the rest because it was about visual stimulation. I could understand this because 'visual stimulation' was repeated in all the paragraphs. I think I just looked at some words very quickly. That was enough to have a very slight idea about that part (S1). The first lines of the paragraphs give us an idea. I read them. But for the rest, I just browsed[15] but I understood what it was about. I don't read but just passing along the lines, I see some words and they give me an idea. For example, if I see 'experience' 'television' 'visual stimulation', I understand that part is related with television and its effects, etc. When I see words such as 'first, second, lastly, therefore, thus', etc., I read the sentence because usually important information is given after these expressions. I read the conclusion carefully because

[15] 'browse' is used as the equivalent of 'göz gezdirmek' in Turkish.

it usually summarises main points (S2). I read the introduction and conclusion and the subtitles carefully but I browsed the rest because I can guess what the text is about when I read a few words. Subtitles are important, too (S3).

Search reading: Reading the question, I can tell whether I need to read carefully or quickly. If the text and the question look similar, I don't lose time reading everything. I read very quickly for whatever I remember from the question. When similar words appear, I understand I am getting closer to the answer. Then, I slow down and pay attention to what is said there in more detail (S7). I read the question and try to understand it. Sometimes, I don't understand the question clearly because I haven't read the text yet. Still, I keep it in mind and read the first sentence of each paragraph. If the paragraph looks relevant, I browse it quickly to see whether some of the key words appear in it or not. I don't usually understand details when I read quickly but I can roughly tell what the paragraph is about. If I scan for a specific word, I don't understand anything. What I do is something in between. I have to understand at least a little. If I see something similar to the question, I slow down and read more carefully (S9). When I read quickly, I can have a general idea what the text is about but if you ask me to summarise it, I cannot. Usually important words give me some idea. For example, I saw 'perception' and I thought it was related with 'mental development' so I slowed down and read that part carefully. Or, if I look for an answer to a question that asks what should be done, I expect to find words such as 'necessity', 'suggestion' or 'should, must', etc. I read the sentences carefully if I see these (S10). I don't try to understand everything when I read quickly. My eyes sweep along the lines. I understand as much as I can and don't worry about the rest. I usually have an idea about the text. Of course, if I do it very quickly, as in scanning,

I don't understand much (S11). There is scanning, which is very fast, and slow and careful reading in which I try to understand everything, think about the words I don't know, etc. This is something in between (S14). When I scan, I don't read much. I just try to find the word I am looking for. Sometimes, when I read quickly I understand only a few words and they gave me an idea. Sometimes, I try to understand more but do not stop when I don't. This is slower. And when I want to understand in detail, I read slowly and re-read the parts that I don't understand. I think there are four types of reading for me (S15).

Careful reading: This is what I do when I think the answer is there; read every word, try to understand what it says. If I don't understand, I go back a little and read again, try to find the meanings of the words if I don't know them, etc (S6). When I read carefully, I try to keep what I read in mind. So the more I read the more information accumulates. If I get lost at some point, I go back and read again. I try to understand what a word means when I don't know it. Or, if a sentence is too long for me to understand, I try to divide it (which one is the verb, which part is a relative clause, etc.) In search reading you don't have time for that (S8). When I think some part of the text is important, or if I cannot understand it by fast reading, I slow down and read every word and I try to visualise what is said in the text. For example, I imagine a cage and monkeys in it and food tubes hanging down, etc (S15).

4.4.2 Discussion

The categorisation of the operations and the text spans showed that the scanning items and the skimming question (SR1) triggered reading operations congruent with

the test specifications. For the search reading items, it could be said that the test takers employed search reading operations as designated in the test specifications but in certain cases, they had to revert to careful reading operations and read larger text spans when their search was unsuccessful. For the careful reading items, certain test takers attacked the items by quickly locating the answer and reading carefully only briefly, but 108 cases of careful reading operations in this section show that they had to revert to reading more slowly and carefully to find the answer almost twice as much as they did in the search reading part. There are also more instances in careful reading test in which the test takers had to read larger spans of the text. Therefore, it could be said that in the majority of the cases, the test takers utilised the operations specified in the test specifications.

The second group of data that consists of the observations on the outcomes of the test taking process and the analysis of test taking strategies showed that the majority of the test takers read and understood the questions and the text favourably and could arrive at the correct answers. These were not discussed extensively since it could be concluded that their reading process reflected the operations specified in the test specifications, reserving the fact that some search reading questions required careful reading operations.[16] By the same token, the test takers who could neither exhibit sufficient comprehension nor answer the questions correctly merely exhibited a reading behaviour similar to the group of test takers who read unfavourably but answered the questions correctly (unfavourable/correct). However, the data from the rest of the groups were analysed to detect construct-irrelevant variance. The data from the group of test takers who could read but could not answer correctly

[16] See Table 4.7

(favourable/incorrect) revealed that SR3 and SR4 are problematic but there is no construct-irrelevant difficulty otherwise. On the other hand, the processes that the test takers who did not read sufficiently but could extract the correct answer (unfavourable/correct) showed that for the search and careful reading tests, there is a possibility of construct-irrelevant easiness. However, it should also be noted that among 315 possible responses given to search and careful reading questions, the total unfavourable/correct responses form only 15.5%. On the other hand, in 25% of the cases, the same processing resulted in incorrect answers. There have been observed several factors that have facilitated test-wiseness and might be influential in construct-irrelevant easiness. As Allan (1992) suggests, test-wiseness is an important source of test construct invalidity since scores of some learners may be influenced by skills which are not the focus of the test and these should be investigated. Some factors that could influence the test performance observed in the present study are discussed below.

Firstly, the wording overlap between the question and the text either in the form of exact matches or synonyms facilitates expeditious reading and enables the test takers to narrow down the span to be read very quickly. It is common sense that under test taking conditions, no matter whether they are pressed for time or not, test takers usually want to find the answer quickly. They are prompted by the questions and their primary aim is to answer them rather than fully grasp the subject matter in the text as they would do, for example, in reading a text for later retention. Therefore, it is important to control the extent to which the items trigger expeditious reading by word match and whether such reading is favourable in a particular case or not. By definition, search reading items were designed not to include exact matches between

the text and the question to prevent mere scanning whereas there was no such restriction in careful reading items. Search reading items should also make reference to first and/or last sentences of the paragraphs so that they could be located by the inspection of them. In certain cases, such restrictions worked against the expectations and restricted the possibility of expeditious reading. For example, in SR3, the paraphrasing of 'genuinely demanding intellectual tasks' as 'challenging problems' resulted in a very general question (What should be done to promote a child's mental development so that he can handle challenging problems?) which cannot be easily located due to its blurred focus. Similarly, 'the advances humankind is likely to make' in SR4 and 'two-way presentation equipment' in SR6 could not be easily associated with the text. On the other hand, when the wording of the question matched the text closely, the process turned into mere scanning followed by a brief careful reading. For example, in SR9 (Immigrant have arrived in France for several reasons; ____ and ____ being the main ones.), simply finding the words 'France', 'immigrant' and 'reason' in the text would suffice to locate the answer. Similarly, in careful reading items, in which the wording of the question might overlap with the text, expeditious reading strategies could easily be used to narrow down the text span. For example, in CR6, after finding the phrase 'pre-school', which was given in the question, matching the synonymous phrases 'unable to differentiate' and 'does not make a clear separation' both in the text and the question enabled five test takers to answer this question without sufficient comprehension. Such possibilities should be considered carefully before tests are actually administered and gathering verbal protocol data seems a very informative procedure in that respect.

Secondly, on occasion, the order of the questions helped the test takers to determine the text spans to be processed unfavourably especially in the search reading part. Instead of using textual and content clues to read expeditiously, some test takers narrowed down the text span by the help of the order of the questions, then either scanned on word basis or read carefully to find the answer. Therefore, when the explicit organisation and argument structure in a text is guaranteed, and when the test takers are not expected to build a macrostructure of the text by incremental reading, the items could be given in a mixed order in search reading tests.

Besides, careful reading questions, again by definition, are formed on main ideas. The text mapping procedure also restricts the arguments on which the items would be written to the ones that only appear in the consensus map (See section 4.2.2). It is usually the case that the arguments that are most explicitly stated in the text appear in the consensus maps. Therefore, the items are usually based on explicitly stated arguments, which are by nature less challenging than the ones that are implicitly stated (See Perkins and Brutten, 1988 and Anderson et al., 1991 in section 2.2.3.1). For example, in the first careful reading text, the result of Kummer and Cord's experiment was explicitly marked with the phrase 'Here, then, is an intriguing example of how inhibition plays a crucial role…', which is usually the case in academic articles. Kummer and Cord's study involves a main idea and is to be used in the test. Since the item could not be based on any minor detail in the study, there was one thing to be asked and it was 'inhibition'. However, this explicitness enabled five test takers to arrive at the answer without substantial understanding of the study. Likewise, certain test takers were able to extract the answer easily without sufficient comprehension when there is one main concept discussed in a paragraph (See also

tts4 in Appendix 3.4). For example, in the second careful reading text, the ability of 'pulling observations together' is discussed in relation to a pinball machine example. The related question (CR8) based on this idea requires the test taker to name that ability. When six test takers found 'an ability' mentioned in that paragraph, they did not need to read more to write the correct answer. This is not to claim that such explicit arguments should not be the focus of comprehension items. On the contrary, puzzle-like questions tapping on minor details should always be avoided. Nevertheless, it is clear that when there are not any distractors in a text part on which the question is based, the amount of comprehension decreases and test-wiseness is set free. Just as a multiple choice item involves distractors in the form of options, a short answer question could have distractors in the text, too. Therefore, maximum care should be given to base careful reading questions on arguments that have, for example, alternative explanations, comparisons, etc. and a careful reading text should include implicit arguments as well as explicit ones.

One major issue to be discussed in relation to the point given above is the format of the items. It is designated in the test specifications that except for the skimming item, the preferred item format for the test is short-answer questions, which usually require the extraction of information printed in the text. Hence, test takers are not asked to formulate a response themselves but they should see it in the text. It is the researcher's observation and experience that with such texts as in the first careful reading test and with only short-answer questions, it is usually hard to design items that require the test taker to relate several parts of the text to form a macrostructure as well as items that tap on implicit arguments naturally because they are not stated in the text. The test writer is usually bound to write items on a specific part of the

text and usually on explicitly stated arguments if he or she works with short answer questions. Therefore, despite all their weaknesses discussed fervently by testing specialists, multiple choice items can occasionally be utilised to form items that are based on implicit arguments and items that are related with several parts of the text. It is also the researcher's contention that reading comprehension tests with discrete items (whether short answer or multiple choice) are generally limited measures of the ability no matter how carefully they are designed and they are less than authentic by nature.[17] Thus, although psychometric concerns abound in the discussions of integrative tests, such as summary writing (i.e. Cohen, 1994), as the pendulum swings back (i.e. Enright et al., 2000), it would be desirable to investigate the possibilities of including such tasks in the reading tests.

It should also be noted here that 'Wh-' and 'sentence-completion' type of short answer questions might be initiating different processes, basically because mutilated sentences are harder to understand since they include incomplete information. While filling a sentence, grammatical structure should also be paid attention to. It is observed that the test takers in this study went back and forth between the text and the question more when they were answering such items. As it is discussed in section 2.2.3.1, test methods and item formats affect the interaction of the reader with the text (i.e. Riley and Lee, 1996). There needs to be further consideration on this point, too.

Another observation that would help the improvement of the reading test is that since the search reading subtest has proved to be most difficult, it should be administered

[17] See also Riley and Lee (1996)

as the last part. It is the usual practise to order the test parts from the easiest to most difficult. When the search reading test, the most challenging one, was administered after such an easy test as scanning, the test takers might have formed faulty expectation as to the difficulty of the test and experienced disorientation.

After the discussion of possible improvements to the test, the last comments should be made on the test takers' definitions of their reading process. The data revealed that different speeds of reading and, accordingly, different levels of comprehension exist for these readers, and they adjust their reading type according to their perceived needs. It was also evident in the protocol data that the difficulty of the text affected the reading speed. When the test takers were not able to get enough information to locate the question, they slowed down to understand more. This finding is in line with the assumption that different types of reading exist as designated in the test specifications. It is also in line with the definitions of the skills given in Urquhart and Weir (1998), integrating Just and Carpenter's (1987) and Carver's (1992, 1997) assumptions.

All in all, it can be said that verbal protocol data give positive support to the confirmation of the hypothesis that the test takers will use the operations specified in the test specifications to arrive at the correct answers. The test specifications are operationalised in the scanning and skimming parts successfully. They are operationalised in the search and careful reading subtests with considerable success though there is room for improvement.

4.5 Research Question 4: What are the dimensions of the reading construct measured by the test?

The fourth research question concerns the structural aspect of the construct validity. In order to investigate the congruence between the dimensions of the reading construct as reflected in the reading framework and in the test, two hypotheses were formed. The first hypothesis is that the correlations between the scanning, search reading and careful reading parts of the test will correlate moderately.[18] The second hypothesis is that the items putatively testing different operations (scanning, skimming, search reading and careful reading) will load on different factors in the Principal Component Analysis. In the present section, these two hypotheses will be examined in relation to four versions of the BUEPT reading test. However, for each version of the test, the central tendency measures, item analysis and subsequent item revisions will be presented before the correlation and factor analysis results. Therefore, for each version of the BUEPT reading test – September 2000, January 2001, June 2001 and September 2001 – firstly, descriptive statistics (central tendency measures), item analysis and item revision results at the pilot stage will be presented. Secondly, the results from the formal administrations of the test will be given starting with central tendency measures and item analysis, followed by correlation statistics and factor analysis.

[18] Skimming part cannot be included in the correlational analysis since it is represented in the test by one question.

4.5.1 The September 2000 Test – Pilot Version

4.5.1.1 The September 2000 Test – Pilot Version: Descriptive Statistics

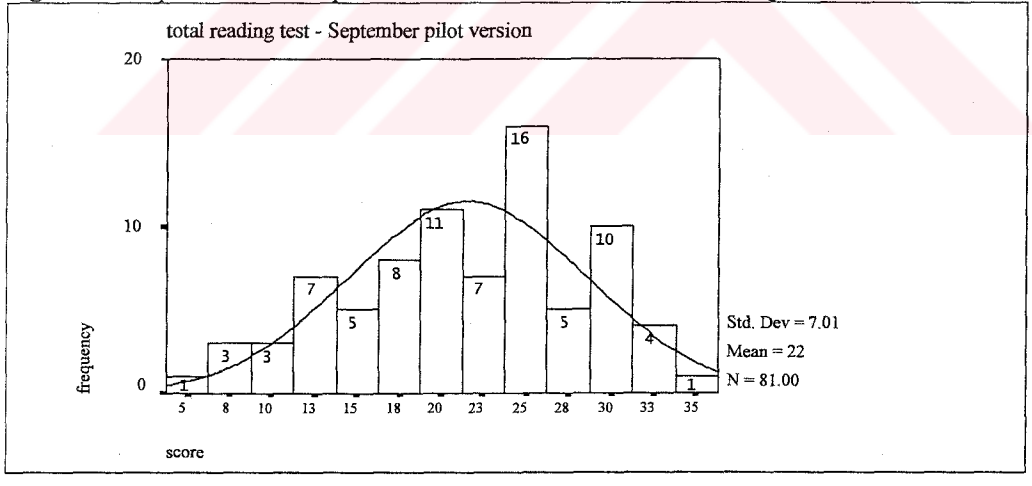
Table 4.8 shows the score distribution data for the pilot version of the September 2000 reading test.

Table 4.8: September 2000 – pilot version: Descriptive statistics of total reading scores

N	Item n	Range	Min	Max	Mean	SE	Std.	Skewness	Kurtosis	Alpha
81	38	28	6	34	21.51 (56.60%)	.78	7.01	-.253	-.726	0.87

The mean (21.51=56.60%) for the pilot version of the September 2000 test is only slightly lower than the cut-off (60%). The distribution is normal (See Figure 4.1).[19] Alpha coefficient of reliability (.87) is found quite satisfactory for the whole reading test.

Figure 4.1: September 2000 – pilot version: Distribution of total reading scores



As for the individual subtests of the reading test, in the descriptive statistics in Table 4.9, it is seen that the mean is quite high for the scanning part (80.54%) and the

[19] See Appendix 4.2 for the normality tests and plots.

distribution of the scores is skewed negatively whereas for the search and careful reading tests, the distribution is normal, the mean being higher for careful reading (51.77%) than search reading (41.43%).[20]

Table 4.9: September 2000 – pilot version: Descriptive statistics of subtests

Substest	N	Item n	Range	Min	Max	Mean	SE	Std.	Skewness	Kurtosis	Alpha
Scanning	81	11	8	3	11	8.86 (80.54%)	.22	2.01	-.935	-.019	0.68
Search R.	81	14	13	0	13	5.80 (41.43%)	.36	3.22	.201	-.712	0.74
Careful R.	81	13	12	0	12	6.73 (51.77%)	.34	3.07	-.346	-.501	0.75

It is also observed that all the subtests have lower reliability coefficients than the overall alpha (0.87), obviously because the number of items in each subtest is less than the whole test (scanning: 0.68, search reading: 0.74, careful reading: 0.75). Yet, the reliability estimates in each part can be considered within acceptable limits especially if it is considered that there is room for improvement by either deleting the items that do not contribute positively to the test and the subtests, or repairing weak items in the light of the responses given by the subjects.

4.5.1.2 The September 2000 Test – Pilot Version: Item Analysis

Table 4.10 below gives the item analysis results of the pilot version of the September 2000 test. The values that are not within the established limits are marked in boldface.

[20] See Appendix 4.3 for the score distribution graphs and Appendix 4.4 for the normality tests and plots of the subtests of September 2000.

Table 4.10: September 2000 – pilot version: Item analysis statistics

Item	IF	ID	CITC	CITC Subtest	AIID*	AIID* Subtest	Item	IF	ID	CITC	CITC subtest	AIID*	AIID* subtest
SC1	**1.00**	**0.00**	**0.00**	**0.00**	**0.8734**	**0.6889**	**SR9**	0.42	**0.27**	0.23	**0.16**	0.8728	**0.7493**
SC2	**0.88**	0.43	0.39	0.49	0.8696	0.6370	**SR10**	0.49	0.61	0.38	0.37	0.8695	0.7270
SC3	**0.86**	**0.29**	0.27	0.25	0.8714	0.6727	**SR11**	0.40	0.76	0.53	0.51	0.8662	0.7113
SC4	**0.90**	**0.21**	0.22	**0.12**	0.8721	**0.6897**	**SR12**	0.49	0.83	0.47	0.38	0.8674	0.7260
SC5	**0.89**	**0.39**	0.29	0.31	0.8712	0.6651	**SR13**	0.37	0.68	0.45	0.38	0.8681	0.7262
SC6	**0.85**	**0.32**	0.36	0.46	0.8701	0.6396	**SR14**	0.59	0.74	0.48	0.43	0.8672	0.7199
SC7	0.79	0.46	0.33	0.39	0.8705	0.6507	**CR1**	0.64	0.75	0.50	0.48	0.8669	0.7220
SC8	0.69	0.42	0.36	0.34	0.8699	0.6603	**CR2**	0.61	0.54	0.33	0.28	0.8706	0.7453
SC9	0.74	**0.33**	0.20	0.36	**0.8729**	0.6549	**CR3**	0.43	0.72	0.60	0.55	0.8646	0.7137
SC10	0.47	0.61	0.41	0.32	0.8688	0.6661	**CR4**	0.28	**0.04**	**0.02**	**0.03**	**0.8766**	**0.7698**
SC11	0.77	0.50	0.45	0.47	0.8682	0.6333	**CR5**	0.67	0.56	0.45	0.48	0.8679	0.7222
SR1	0.44	**0.21**	**0.12**	**0.06**	**0.8752**	**0.7596**	**CR6**	0.64	0.75	0.53	0.47	0.8661	0.7231
SR2	0.67	0.64	0.48	0.40	0.8674	0.7237	**CR7**	0.22	0.42	0.34	0.39	0.8702	0.7331
SR3	0.40	0.57	0.39	0.44	0.8693	0.7197	**CR8**	0.77	0.63	0.45	0.51	0.8681	0.7211
SR4	**0.11**	**0.25**	0.36	0.45	0.8701	0.7245	**CR9**	0.57	0.81	0.46	0.44	0.8677	0.7267
SR5	0.35	**0.38**	0.22	0.23	0.8728	0.7416	**CR10**	0.49	0.65	0.38	0.33	0.8695	0.7393
SR6	0.41	0.76	0.52	0.49	0.8663	0.7136	**CR11**	0.22	0.50	0.42	0.30	0.8688	0.7426
SR7	0.28	0.63	0.49	0.47	0.8672	0.7166	**CR12**	0.56	0.62	0.40	0.43	0.8690	0.7285
SR8	0.38	0.54	0.32	0.28	0.8707	0.7371	**CR13**	0.63	**0.28**	**0.11**	**0.17**	**0.8751**	**0.7567**

SC: scanning **SR**: search reading **CR**: careful reading **IF**: item facility **ID**: item discrimination **CITC**: corrected item-total correlation **AIID:** alpha if item deleted ***Alpha; overall:** 0.8728 **SC:** 0.6820 **SR:** 0.7432 **CR:** 0.7502

The items that have item facility (IF) values not within the limit of 0.20 - 0.80 are SC1-6 and SR 4. Item discrimination values (ID) for items SC1, SC3-6, SC9; SR1, SR4- 5, SR9; CR4 and CR13 are lower than the acceptable limit 0.40. Corrected item-total correlation (CITC) was found to be lower than 0.20 for items SC1, SR1, CR4 and CR13. Alpha if item deleted (AIID) is checked by comparing the overall alpha of the test (0.8728) with the alpha of each item. If alpha is lower than 0.8728 when a particular item is deleted, then that item is contributing positively to the test's internal reliability and it will be kept in the test. If the reverse is true, i.e. alpha is higher when the item is deleted, then the item is likely to be removed or repaired. In that respect, SC1, SC9, SR1, SR4, CR4 and CR13 need consideration. It is also suggested that where the various parts of a test might be testing different dimensions, separate correlation and reliability analyses on each part should also be carried out (Green and Weir, 1998) since reliability coefficients are generally calculated based on the assumption that tests are homogeneous. Anastasi and Urbina (1997, 97) underline that 'the more homogeneous the domain, the higher the interitem consistency'. Thus it will be possible to see the individual contribution of items to the particular part of the test (subtest) as well as the whole test if both the subtest and the whole test are taken into consideration. An item may seem to be decreasing the internal reliability of the whole test when the test is multidimensional. However, if an item works in accordance with the items it is grouped with in a subtest, it deserves a second consideration. The columns titled 'CITC-subtest' and 'AIID-subtest' give us CITC and AIID values of an item in the subtest in which it appears (i.e. scanning, search reading, etc.). Here again, we see that SC1, SC4, SR1, SR9, CR4 and CR13 do not contribute to the specific subtests positively.

The last item statistics that are analysed in this study are item discrimination patterns (IDPs). The details of the procedure are as follows:

The possible highest mark in the test is 38 (38 items, 1 point each) and pass/fail cut-off is set at 60%, which is 22.8 ≈ 23. The test takers are assigned to six groups in such a way that the groups on either side of the pass mark (23) are narrower than the others since it is important to know which items are not discriminating around the pass/fail boundary. Table 4.11 shows the distribution of the scores and the number of candidates in each group, i.e. band.

Table 4.11: September 2000 – pilot version: Distribution of total reading score by band

band	range of scores	total score mean	no of test takers	pass/fail	percent	cumulative percent
1	0 – 10	7.60	5	fail	6.2%	6.2%
2	11 – 18	14.82	22	fail	27.2%	33.3%
3	19 – 22	20.29	14	fail	17.3%	50.6%
4	23 – 26	24.35	20	pass	24.7%	75.3%
5	27 – 31	29.60	15	pass	18.5%	93.8%
6	32 – 38	32.60	5	pass	6.2%	100%

It can be seen in the table that the numbers in the bands are reasonably well distributed and almost 50% of the test takers passed. While the band increases from 1 to 6, as expected, the total mean score increases, too. For example, the mean for band1 is 7.60 whereas it is 32.60 for band6.

The next step is to identify the overall item facility values for each item in the test in each of these bands. We expect to find gradually increasing values from the lowest in band1, close to 0%, and the highest in band6, close to 100%. Table 4.12 shows item discrimination patterns of the items in the pilot version of the September 2000 test.

Table 4.12: September 2000 – pilot version: Item discrimination patterns by band

band	SC1	SC2	SC3	SC4	SC5	SC6	SC7	SC8	SC9	SC10	SC11	SR1	SR2
1	1.00	0.40	0.80	0.80	0.60	0.40	0.40	0.40	0.20	0.00	0.20	0.20	0.20
2	1.00	0.77	0.68	0.77	0.77	0.77	0.73	0.45	0.68	0.18	0.64	0.27	0.45
3	1.00	0.93	0.93	0.86	1.00	0.86	0.71	0.71	0.79	0.50	0.57	0.57	0.50
4	1.00	0.95	0.95	1.00	0.95	0.90	0.85	0.75	0.85	0.65	0.95	0.55	0.80
5	1.00	1.00	1.00	1.00	0.93	1.00	0.93	0.93	0.80	0.80	1.00	0.47	1.00
6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.40	1.00	0.60	1.00

band	SR3	SR4	SR5	SR6	SR7	SR8	SR9	SR10	SR11	SR12	SR13	SR14	CR1
1	0.00	0.00	0.00	0.00	0.00	0.00	0.40	0.20	0.00	0.00	0.20	0.00	0.20
2	0.27	0.00	0.18	0.18	0.00	0.18	0.27	0.23	0.14	0.23	0.09	0.23	0.36
3	0.29	0.00	0.50	0.14	0.29	0.36	0.29	0.43	0.21	0.57	0.29	0.71	0.57
4	0.35	0.15	0.45	0.50	0.30	0.60	0.40	0.65	0.45	0.50	0.35	0.75	0.80
5	0.67	0.20	0.40	0.80	0.60	0.53	0.60	0.73	0.87	0.80	0.73	0.93	0.93
6	1.00	0.60	0.40	1.00	0.80	0.40	1.00	0.80	0.80	1.00	1.00	0.80	1.00

band	CR2	CR3	CR4	CR5	CR6	CR7	CR8	CR9	CR10	CR11	CR12	CR13
1	0.20	0.00	0.00	0.20	0.20	0.00	0.20	0.00	0.00	0.00	0.00	0.20
2	0.45	0.05	0.27	0.55	0.32	0.09	0.68	0.41	0.36	0.00	0.45	0.64
3	0.50	0.36	0.29	0.43	0.64	0.14	0.64	0.43	0.36	0.14	0.43	0.64
4	0.70	0.55	0.40	0.75	0.75	0.20	0.90	0.65	0.55	0.30	0.55	0.65
5	0.80	0.87	0.27	1.00	1.00	0.40	1.00	0.93	0.73	0.40	0.87	0.67
6	1.00	1.00	0.20	1.00	1.00	0.80	0.80	0.80	1.00	0.80	1.00	0.80

In Table 4.12, it is clearly seen that SC1 does not discriminate between any levels. All test takers could answer the item correctly. Therefore, there is no use in keeping this item in the test. SC2 is an easy item, but it can differentiate between band1-5. Besides, the facility percentages from band 1 to band 5 are incremental. SC3-5 have low discrimination values if not zero since almost all the test takers in the last three bands could answer the item correctly and it cannot be claimed that the distribution in the first three bands shows us how the test takers in these bands perform differentially. SC9 can differentiate between band1 and band 2 only. The rest of the scanning items have acceptable item discrimination patterns (IDPs). SR1 is a problematic item that cannot differentiate in bands3-6, e.g. more test takers in band 3 can get the item right than do those in band 5. SR4 is a difficult item and it does not discriminate among low levels. SR5, on the other hand, cannot discriminate among higher levels. The rest of the SR items have relatively good IDPs. CR1-3 exhibit the examples of desirable IDPs. CR4 could be answered by very few test takers from all bands. Thus, it has a very low discrimination power. CR13 received correct

responses in similar percentages from the test takers in band2-5. The rest of the CR items have relatively satisfactory IDPs. The band score graphs for the items in the September 2000 pilot version of the test are given in Appendix 4.5.

4.5.1.3 The September 2000 Test – Pilot version: Evaluation of the Items

As mentioned above, item analysis provides us with the information necessary to evaluate the appropriateness of items that constitute a test. It is important to know how each item behaves in terms of difficulty, its ability to discriminate weak test takers from good ones, its compatibility with the whole test to form a test suitable for our purposes. With the information pooled through item facility values, item discrimination indices and patterns, correlation and reliability analyses, we have a solid statistical basis to decide whether or not an item is functioning as we expect it to function. However, these are not the only considerations for the inclusion or exclusion of an item. Test writers also consider the characteristics of the intended group of test takers, the overall difficulty level they want to establish in the test i.e. whether they want to form a difficult or an easy test, etc. Test writers should also analyse the responses designated as incorrect to see whether there are any unintended distractors either in the text or in the question rubric. Therefore, information from statistical analyses is usually coupled with subjective judgements with reference to the useability of an item. The following part gives the details of item evaluation on the basis of each subtest in the reading test we are analysing.

Scanning: On the whole, the scanning part with a mean of 80.54% was found to be too easy. Item facility values (IF) of SC items are too high for the first six items and

item discrimination indices (ID) for these items are too low. Especially, SC1 with IF of 1.00 and ID, 0.00 does not work at all. When the actual test was checked, it was seen that the answer for SC1 is in the first paragraph of the text and very easy to locate. Therefore, this question was removed from the test. As mentioned above, in general, there is a tendency for SC items to have high IF and low ID for the first part of the test (SC1-6), but the second part is less problematic (SC7-11). It seems that the test takers used their time less sparingly for the first part and when more than enough time is spent on scanning items, the likelihood of answering an item correctly is too high. It is only in the second part that the test takers had to do quick, selective reading as suggested in the test specifications.[21] Therefore, the problem with scanning items might be related to timing rather than the nature of the questions themselves. Since the nature of the scanning items is clearly designated in the test specifications, all possible scanning items would resemble the ones in the test. What makes the difference seems to be the time spent to scan the text to answer the questions. Therefore, the only change made in the test was the exclusion of SC1 from the test.

Search reading: The problematic items in the search reading part were SR1, 4, 5 and 9. SR1 is a multiple choice-skimming item testing the ability to read selectively to establish discourse topic and main ideas.[22] The test taker is asked to choose the alternative that best expresses the main idea of the text, focusing on salient parts such as the introduction, conclusion, titles and subtitles, etc. The time for this task is limited to five minutes; therefore the test taker is not allowed to read the whole text but use text organisation features and read selectively to arrive at a general idea

[21] See Appendix 3.2 for the test specifications
[22] See Appendix 3.1 for the tests.

about the text. Although the IF value for the item was not too low (0.44), its CITC, AIID, IDP values showed that the item was not functioning as expected and it could not discriminate well between the bands. On the other hand, SR1 is a multiple-choice question and there is probably a guessing effect included. However, since there is only one question testing skimming ability in the test, it had to be retained but the alternatives were revised and tightened. The problem with SR4 was that there were some alternative answers. IF, ID and IDP showed that when the answer in the key was taken as the only correct response, the question became too difficult and only 60% of the top group could answer the question correctly. Since the question was formulated on a point all the test writers included in their text maps,[23] and the item discrimination pattern exhibited an incremental structure, the question was repaired by the addition of extra information and retained in the test. SR5 and 9 were eliminated from the test. SR8 could also be excluded from the tests since five questions in each section in addition to the first skimming question were adequate. The reason why SR8 was chosen was its erratic item discrimination pattern. Looking at the IDP of the item in Table 4.12, it can be seen that fewer test takers in bands 5-6 could respond correctly to the item than did those in band3.

With all these changes in the search reading part, it was obvious, however that the subtest would still be a difficult section since its mean was only 41.43%. Nevertheless, this was taken into consideration in relation with the other parts of the reading test and the whole proficiency test (the BUEPT), and no further changes were made in the search reading subtest.

---

[23] See section 3.2.2.

Careful reading: In this section of the test, two items CR4 and 13 were problematic and they were excluded from the test. There was one more extra item. Because CR5 and CR6 were based on the same main idea one of them could be eliminated. CR6 had better ID and IDP so CR5 was eliminated.

No further quantitative analysis was carried out on the pilot version of the September 2000 test since the data was not large enough for more sophisticated techniques such as factor analysis and at this stage of test validation it was more important to focus on item quality. However, as it would be obvious in the later parts of the study, qualitative analyses of such data as expert opinion and introspection would have been helpful in determining item quality. Unfortunately, testing schedule of the institution did not allow further investigation at that moment.

4.5.2 The September 2000 Test

After the analyses explained in the previous section were completed and the test was reduced to its purged version, it was administered to the group of incoming students as part of the BUEPT. The statistical analyses done on the data are given below.

4.5.2.1 The September 2000 Test: Descriptive Statistics

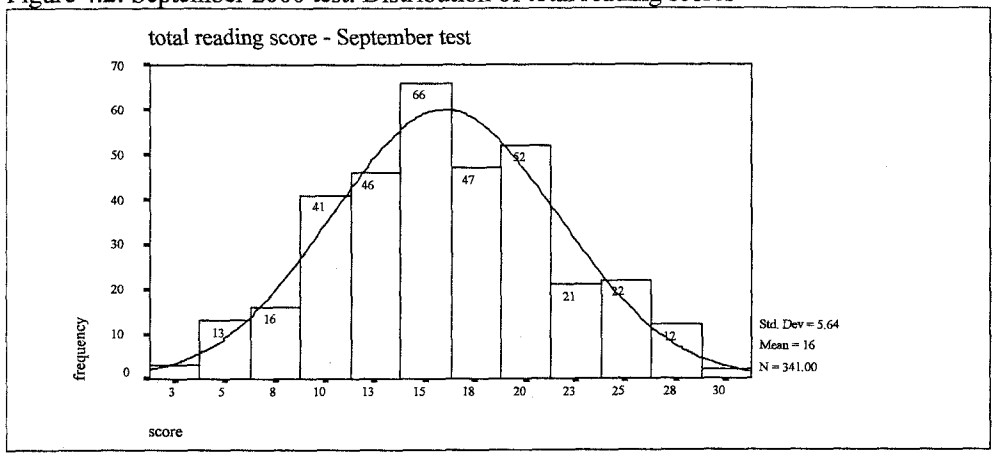
Table 4.13 shows descriptive statistics of the reading module of the BUEPT administered in September 2000.

Table 4.13: September 2000 test: Descriptive statistics of total reading scores

N	Item n.	Range	Min.	Max.	Mean	SE	Std.	Skewness	Kurtosis	Alpha
341	31	27.00	2.00	29.00	15.95 (51.45%)	0.31	5.65	0.076	-0.410	0.84

The mean for the overall reading module is 15.95 (51.45%), lower than the cut-off 60% and the mean attained in the pilot version (21.51 = 56.60%). The distribution of the scores is normal (See Figure 4.2).[24]

Figure 4.2: September 2000 test: Distribution of total reading scores



The calculated alpha value (0.84) is found satisfactory. However, the statistics from the individual subtests (See Table 4.14) show that the scanning part has a very high mean (7.26 = 72.6%) with a negatively skewed distribution of scores.[25] However, this value is slightly lower than the one in the pilot version (8.86 = 80.54%). The mean of careful reading is lower than expected (4.87 = 48.7%) and lower than the pilot version mean, too (6.73 = 51.77%). Search reading mean is the lowest of all (3.82 = 34.73%) as in the pilot version (5.80 = 41.43%).

Table 4.14: September 2000 test: Descriptive statistics of subtests

Subtests	N	Item n	Range	Min	Max.	Mean	SE	Std.	Skewness	Kurtosis	Alpha
Scanning	341	10	10	0	10	7.26 (72.6%)	0.12	2.16	-0.674	0.146	0.71
Search R.	341	11	10	0	10	3.82 (34.73%)	0.12	2.26	0.517	-0.314	0.63
Careful R.	341	10	10	0	10	4.87 (48.7%)	0.13	2.45	0.075	-0.696	0.71

[24] See Appendix 4.6 for the normality tests and graphs.
[25] See Appendix 4.7 for the score distribution graphs by subtest and Appendix 4.8 for the normality tests and graphs.

4.5.2.2 The September 2000 Test: Item Analysis

The next step was the analysis of individual items calculating item facility values (IF), item discrimination indices (ID), corrected item-total correlation (CITC- both for overall reading test and the subtest) and internal consistency analysis (alpha if item deleted: AIID). Table 4.15 gives the item analysis values for each item in the reading test. The values that are not within the acceptable limits are given in boldface.

Table 4.15: September 2000 test: Item analysis statistics

Item	IF	ID	CITC	CITC Subtest	AIID*	AIID* subtest	Item	IF	ID	CITC	CITC subtest	AIID*	AIID* subtest
SC1	**0.89**	**0.19**	**0.19**	**0.16**	0.8367	**0.7137**	**SR7**	0.50	0.44	0.28	0.28	0.8353	0.6049
SC2	**0.85**	0.41	0.34	0.30	0.8333	0.6968	**SR8**	0.36	0.61	0.35	0.35	0.8327	0.5880
SC3	**0.87**	**0.39**	0.31	0.30	0.8341	0.6971	**SR9**	0.55	0.59	0.26	0.21	0.8360	0.6190
SC4	**0.89**	**0.32**	0.24	0.25	0.8357	0.7027	**SR10**	0.30	0.65	0.44	0.44	0.8297	0.5685
SC5	**0.85**	**0.37**	0.30	0.40	0.8342	0.6823	**SR11**	0.50	0.81	0.46	0.43	0.8289	0.5686
SC6	0.68	0.55	0.37	0.40	0.8318	0.6808	**CR1**	0.80	0.44	0.29	0.27	0.8344	0.7072
SC7	0.65	0.73	0.40	0.47	0.8310	0.6666	**CR2**	0.48	0.51	0.29	0.27	0.8348	0.7100
SC8	0.63	0.59	0.37	0.49	0.8319	0.6627	**CR3**	0.42	0.78	0.45	0.46	0.8290	0.6760
SC9	0.33	0.58	0.39	0.38	0.8314	0.6853	**CR4**	0.52	0.83	0.47	0.44	0.8284	0.6806
SC10	0.61	0.66	0.43	0.48	0.8299	0.6646	**CR5**	**0.15**	0.44	0.40	0.37	0.8317	0.6948
SR1	0.36	**0.12**	**-0.02**	**-0.02**	**0.8450**	**0.6654**	**CR6**	0.72	0.69	0.42	0.40	0.8303	0.6871
SR2	0.57	0.58	0.33	0.28	0.8334	0.6033	**CR7**	0.60	0.66	0.39	0.37	0.8315	0.6924
SR3	0.24	0.63	0.46	0.40	0.8294	0.5790	**CR8**	0.52	0.69	0.39	0.37	0.8314	0.6923
SR4	**0.07**	**0.19**	0.27	0.24	0.8351	0.6144	**CR9**	0.22	0.36	0.28	0.31	0.8347	0.7022
SR5	0.32	0.63	0.42	0.35	0.8305	0.5893	**CR10**	0.45	0.84	0.50	0.46	0.8274	0.6767
SR6	**0.08**	**0.18**	0.23	0.20	0.8359	0.6184							

SC: scanning **SR**: search reading **CR**: careful reading **IF**: item facility **ID**: item discrimination **CITC**: corrected item-total correlation **AIID:** alpha if item deleted ***Alpha**; **overall:** 0.8373 **SC:** 0.7087 **SR:** 0.6259 **CR:** 0.7143

Table 4.15 shows that the first five scanning items are problematic; they have too high IF and too low ID except for SC2. However, except for SC1 they do not have a negative effect either on the overall or subtest reliability. SC9 has relatively low IF among other scanning items. In the search reading part, skimming item (SR1) has a low ID, negative correlation both with the whole test and the subtest, and if it is deleted the reliability coefficients of both the whole test and the subtest increase markedly. SR4 and 6 had low IF and ID values, too. In the careful reading part, there is only one item with a low IF value and it is CR5.

As the last check of the items' performance, item discrimination patterns (IDPs) were analysed. The possible highest mark in the test is 31 points (31 items, 1 point each) and pass/fail cut-off is $18.6 \approx 19$ (60%). Table 4.16 gives us the general distribution of scores by band.

Table 4.16: September 2000 test: Distribution of total score by band

band	range of scores	total score mean	no of test takers	pass/fail	percent	cumulative percent
1	2-8	6.06	32	fail	9.4%	9.4%
2	9-15	12.33	127	fail	37.2%	46.6%
3	16-18	16.93	73	fail	21.4%	68.0%
4	19-21	19.92	52	pass	15.2%	83.3%
5	22-25	23.25	36	pass	10.6%	93.8%
6	26-29	27.09	21	pass	6.2%	100.0%

32% of the test takers have been assigned to passing groups and 68% to failing groups with a pile of 37.2% amassing in band 2. Item discrimination patterns in Table 4.17 below and band score graphs in Appendix 4.9 show that the scores in the first five scanning items are too easy for the sample population.

Table 4.17: September 2000: Item discrimination patterns by band

band	SC1	SC2	SC3	SC4	SC5	SC6	SC7	SC8	SC9	SC10	SR1	SR2	SR3
1	0.72	0.50	0.59	0.63	0.44	0.25	0.13	0.22	0.03	0.19	0.31	0.22	0.00
2	0.87	0.79	0.83	0.85	0.80	0.54	0.53	0.46	0.15	0.43	0.34	0.44	0.06
3	0.90	0.92	0.95	0.97	0.96	0.75	0.71	0.75	0.38	0.73	0.36	0.58	0.21
4	0.98	0.98	0.94	0.92	0.94	0.85	0.85	0.83	0.44	0.79	0.44	0.71	0.38
5	0.97	1.00	1.00	1.00	0.94	0.94	0.92	0.86	0.67	0.89	0.36	0.89	0.58
6	1.00	1.00	1.00	0.95	1.00	1.00	0.95	1.00	0.81	1.00	0.38	0.90	0.81

band	SR4	SR5	SR6	SR7	SR8	SR9	SR10	SR11	CR1	CR2	CR3	CR4	CR5
1	0.00	0.03	0.00	0.28	0.09	0.19	0.03	0.06	0.44	0.25	0.06	0.03	0.00
2	0.02	0.13	0.02	0.34	0.20	0.47	0.15	0.33	0.73	0.33	0.26	0.39	0.05
3	0.03	0.29	0.11	0.55	0.38	0.59	0.25	0.51	0.89	0.49	0.34	0.47	0.07
4	0.10	0.48	0.10	0.63	0.44	0.63	0.42	0.73	0.90	0.62	0.56	0.75	0.19
5	0.19	0.69	0.06	0.75	0.64	0.69	0.61	0.83	0.92	0.69	0.92	0.92	0.42
6	0.38	0.81	0.43	0.90	0.86	0.95	0.95	0.95	0.95	0.95	0.95	1.00	0.67

band	CR6	CR7	CR8	CR9	CR10
1	0.13	0.16	0.13	0.00	0.00
2	0.63	0.47	0.36	0.12	0.24
3	0.84	0.64	0.58	0.22	0.53
4	0.85	0.81	0.65	0.35	0.65
5	0.97	0.83	0.83	0.39	0.86
6	1.00	1.00	1.00	0.57	0.95

Besides, SC3-5 do not discriminate between failing (band3) and passing (band4) test takers well. The rest of the SC items have favourable IDPs. For SR1, the percent of correct responses are approximately the same in all the bands, the highest being band4 rather than band6. Too few of the test takers in the highest bands could answer SR4 and SR6 correctly. The remaining SR items have satisfactory IDPs. As for CR items, although CR5 discriminates well in higher bands (band 5and 6), we see that it does not do so in lower bands (1 and 4). Only 67% of the test takers in band 6 could answer this item correctly. CR9 is also a difficult item since only 57% of the test takers in band6 could answer it correctly. However, it has a better IDP than CR5 (with a lower ID = 0.36, though). There were no problems observed with other CR items.

4.5.2.3 The September 2000 Test: Evaluation of the Items

It was interesting to see that the first scanning item was the most problematic one among the others just as in the pilot version. However, this time the location of the item should not pose any problems. Again, just as in the pilot version IF values for the first half of the scanning test (SC1-5) were much higher than the second half. The assumption on how the students used their time concerning scanning items seems to have been confirmed with this finding. It seems that the test takers used their time more comfortably for the first part, and scanning items can easily be answered correctly when too much time is spent on them. It is only in the second part that the test takers might have to do quick, selective reading as assumed. It seems that the ones who could do scanning quickly enough could get more items correct and slow readers might have been unable to answer the questions in the later parts of the test. Unfortunately, since the scoring system did not permit differentiation between incorrect responses and unattempted questions, this assumption has to remain speculative with the data in hand. Besides, due to the apparent easiness of the scanning test, assigning 0.5 point to SC items rather than 1 point in calculating the overall scores of the test takers was legitimate. However, this distinction is not reflected in the present study to maintain a regularity in the calculations.

As mentioned above, SR1 was the most problematic item in the test with its negative correlation with the rest of the test and the subtest, and negative effect on the overall and the subtest alpha. Its peculiarity might be merely due to its being a multiple-choice item, more specifically, to the possible effect of the guessing factor. SR1 is a skimming task

which requires test takers to choose the alternative that best expresses the main idea of a four-page text in five minutes. Test takers might simply have chosen an alternative without actually skimming the text. This might well be due to the seeming impossibility of the task or the lack of the ability to skim a text in order to get the gist in a short time. A second explanation might be that there might be an unintended factor in the design of the item, especially in the alternatives that led to the unusual performance of SR1. When the distribution of the responses was checked, it was seen that the alternative *d* was marked by 31% of the test takers as the correct answer whereas the actual correct answer *c* was marked by 36% (a: 11%, b: 14%, c: 36%,d: 31%, e: 8%). It seems that the alternative *d* was too strong a distractor, possibly because the difference between *c* and *d* was too subtle to discern in a short time. However, before we have further data on how test takers respond to this question especially in terms of the strategies and skills they used, this discussion remains inconclusive.

In the search reading part, SR4 and SR6 are the other problematic items besides SR1. SR4 had low IF and ID values in the pilot version, too. Although the question was rewritten, its performance did not improve. The extra information added to the question apparently had a negative impact on its clarity.[26] As for SR6, the item had 0.28 IF and 0.63 ID in the pilot version and was not considered a problematic item. However, in the actual test administration, IF decreased to 0.08 and so did ID to 0.18. One explanation for that could be related to its being the last item in the first search reading section. It is possible that the test takers who could not use their time efficiently could not respond to

[26] See Appendix 3.1 for the test items.

the item. As mentioned before, since the items could only be marked as correct (1) or incorrect (0), this claim cannot be supported by statistical evidence either.

CR5's IF of 0.22 in the pilot version decreased to 0.15 in the September administration although its ID slightly increased from 0.42 to 0.44. One major problem observed with this item is that it has two blanks and both have to be filled out correctly. Otherwise, since no partial credits could be assigned the response is considered incorrect. It was observed that the incorrect responses for the first blank were more than the ones for the second one. However, since the test writers contended that both pieces of information were necessary for the formation of a complete main idea, the scoring method was not changed. 'End of test effect' may also be valid for this item since it is the last question of the section.

To assign the test takers fairer scores and to increase the performance of the test, each scanning item was assigned 0.5 point, as mentioned before. Besides, the scores from SR1 and SR4 were not included in the actual score calculation of the students who took the test since there were deficiencies in the way the items were formulated. Otherwise, there was a reasonable balance between the difficult and easy items and item statistics yielded favourable values.

One last point that has to be mentioned here is that, in general, approximately 60% of the students are expected to fail the September exam depending on the statistics from previous years. General observations by the faculty staff indicated that there might be test dependent as well as test independent reasons for this. It is expected that roughly

one third of the incoming student population taking the BUEPT in September has zero or very basic level of English. Since the BUEPT is a proficiency test rather than a placement test, it aims at discriminating performance at a high academic level. Therefore, the presence of a relatively large group of non-speakers may have an effect on the results, too. It has also been observed that a number of students may prefer to attend the prep school before they register in their departments. Therefore, not all students taking the test in September may be striving to pass the test. At this point, it should be emphasised that it is important to pool in data from various versions of the test administered at different testing seasons before the discussions on the validity of the BUEPT reading test is concluded. Thus, the other three versions of the BUEPT reading test to be discussed in later sections of this chapter will be informative in that respect.

4.5.2.4 The September 2000 Test: Inter-correlations and PCA

The next step in the statistical analysis of the September 2000 test data was firstly to investigate the degree to which the subtests correlated and, secondly, to identify the component structures of the tests in order to assess differential performance of the subtests so that it can be evaluated whether or not the present data lend support to congruent operationalisation in the test of the multicomponential nature of the reading skill as embodied in the Urquhart and Weir's (1998) framework. In doing this, the information attained from the item analysis was frequently resorted to in order to determine possible random factors due to defective item characteristics.

The September 2000 Test: Inter-correlations: Table 4.18 shows the inter-correlations of the subtests in the September 2000 reading test.

Table 4.18: September 2000 test: Subtest inter-correlations

Subtests	**Search reading**	**Careful reading**
Scanning	.430*	.481*
Search reading	-	.610*

*: Correlation is significant at the 0.05 level (2- tailed).

All the correlations in the table are moderate (within .4 and .7 limit) and significant at .05 level. However, the scanning test has lower correlation with the other two subtests (SR: .430, CR: .481) than search reading has with careful reading (.610). The correlation between the search and careful reading tests is moderately high suggesting a stronger link between the two compared to the one between the scanning test and the rest. Since all the correlations are within the given limits, it can be concluded that the overlap between the tests is not large enough to claim that these tests measure exactly the same construct. There is support for the differential performance of the tests.

The September 2000 Test: Principal Components Analysis: The September 2000 test has 31-item data[27] and the whole set was submitted to PCA with varimax rotation without constraining the number of components to be extracted. KMO measure of sampling adequacy was .843 and Bartlett's test of sphericity was significant at .000 level, both of which were quite satisfactory. No communialities below .30 were observed.[28] 10 components with eigenvalues higher than 1.00 were extracted and these accounted for 54% variance in the data. Rotated component matrix is given in Table 4.19 in which the highest loading of the items on the components are marked in bold.

[27] See Appendix 3.1 and 3.2 for the description of the subtests and the distribution of questions.
[28] See Appendix 4.10 for details.

Table 4.19: Rotated component matrix: September 2000 test - whole set

	Component									
	1	2	3	4	5	6	7	8	9	10
SC1	,064	,041	,184	,007	-,003	,016	,021	**,651**	-,041	,206
SC2	**,413**	,222	,031	-,061	-,055	,038	,137	,285	-,115	-,020
SC3	,234	,207	**,277**	-,062	,132	,126	,185	-,011	-,636	-,016
SC4	,239	,258	-,048	-,110	-,074	,196	,397	-,086	-,192	**,420**
SC5	-,085	**,621**	-,054	-,060	,060	,214	,107	,242	,122	-,157
SC6	,246	**,449**	,219	-,140	-,122	-,020	,195	,190	,051	-,273
SC7	,273	**,653**	-,117	,142	,004	,118	,019	-,134	-,022	,158
SC8	,106	**,705**	,060	,188	-,032	,034	,066	-,167	-,161	,033
SC9	,093	**,527**	,184	,106	,306	-,250	-,090	,151	,230	,118
SC10	,032	**,612**	,309	-,013	,201	-,031	,016	,140	-,022	-,089
SR1	,007	-,044	,077	,029	-,052	-,027	-,045	,080	,017	**,722**
SR2	,139	,014	**,708**	,028	,058	,076	,041	,034	-,072	,072
SR3	,135	,167	**,517**	,279	,190	-,138	,164	,200	,054	,064
SR4	-,004	-,036	,179	**,686**	-,062	,162	,202	,062	,110	-,067
SR5	,206	,217	,055	**,656**	,230	,117	-,060	-,065	,012	,109
SR6	,192	,015	-,102	,060	,151	,205	-,039	**,567**	,135	-,164
SR7	,060	,097	-,002	,098	**,704**	,071	,014	,170	-,137	-,028
SR8	,200	,045	,175	,028	**,566**	,070	,118	-,136	,242	-,110
SR9	,280	,081	,015	-,189	,303	,140	,193	-,261	**,366**	-,116
SR10	**,580**	-,004	,220	,002	,329	-,042	,070	,049	-,026	-,013
SR11	**,531**	,070	,104	,012	,294	,272	-,011	,081	-,023	,178
CR1	,005	-,002	,149	,048	,312	**,697**	,050	,039	,017	,086
CR2	,023	,080	,073	,157	,115	,020	**,815**	-,008	,058	-,049
CR3	**,511**	,017	,192	,320	-,067	-,119	,351	,156	,111	-,011
CR4	**,664**	,111	,116	,136	,047	,091	-,034	,028	,077	-,008
CR5	**,486**	,157	-,100	,385	,098	-,128	,170	,120	-,085	-,138
CR6	**,479**	,133	,359	-,002	-,065	,280	-,172	-,056	,233	,028
CR7	,232	,105	**,384**	,248	,057	,221	-,008	-,027	-,087	-,343
CR8	,182	,112	,153	,111	,096	,145	,317	,134	**,536**	,024
CR9	,157	,143	-,026	,219	-,137	**,563**	,022	,193	,041	-,136
CR10	,240	,291	**,403**	,156	,015	,260	,034	-,043	,335	-,049

The matrix displayed a structure hard to interpret with 7 weak components with less than three variables loading on them. In order to have a clearer picture, factor loadings were saved as variables and correlated against total subtest scores. In Table 4.20 below, the first column gives the components (factors),[29] the second, eigenvalues and the third, percent of variance accounted by each component. We see in the third column that only the first two components account for more than 5% of variance. The next four columns

[29] The terms 'component' and 'factor' are used synonymously.

give the Pearson correlation coefficients of subtest scores (total scanning, search and careful reading, skimming) against factors (F1, F2, etc.).

Table 4.20: September 2000 test – whole set: Subtest – Factor correlations

Factors	Eigen value	% of variance (cum: 54.04%)	scanning	search reading	careful reading	skimming (SR1)
F1	5.597	18.054	.298*	**.475***	**.561***	.007
F2	1.761	5.682	**.865***	.131*	.216*	-.044
F3	1.379	4.448	.202*	**.396***	**.339***	.077
F4	1.326	4.279	.019	.279*	**.329***	.029
F5	1.208	3.895	.089	**.570***	.084	-.052
F6	1.202	3.879	.060	.165*	**.368***	-.027
F7	1.119	3.609	.174*	.119*	**.305***	-.045
F8	1.087	3.505	.203*	.103	.095	.080
F9	1.054	3.400	-.114*	.120*	.250*	.017
F10	1.020	3.289	.047	.158*	-.111*	**.722***

Pearson Correlation: * *correlation is significant at 0.01level (2-tailed) / * correlation is significant at 0.05 level (2-tailed)

When the cut-off for meaningful correlation is taken .3, it is seen that scanning correlates with F2, search reading with F1, F3 and F5 and careful reading with F1, F3, F4, F6 and F7. The skimming question correlates with F10. The reader should be reminded that singling out one particular question as a subtest may not be meaningful and also violates the criterion that every factor should have at least three variables loading on it. However, since skimming skill is tested by only one question (SR1) in this test, and since the item behaves peculiarly, the researcher decided to treat it as a subtest in the calculations. This item usually correlates with one factor that does not correlate with other items. F8 and F9 do not correlate with any subtests within the set limit.

In terms of the interpretations of the factors, it is seen that F1 is a factor that accounts for a combination of search and careful reading (r= .475 and .561 respectively).[30] F2 has a

[30] The correlation of the scanning part with F1 is very close to .30 (.298). If this were considered high enough, F1 would be considered as a general factor accounting for an underlying reading ability. However, since there is a significant difference between *r* of the scanning and the other tests' the researcher preferred not to include the scanning test in the interpretation.

clearly high correlation with scanning (.865). F3 is another factor that has correlation with both search and careful reading (r= .396 and .339 respectively). F4, F6, F7 all seem to account for the variables only in careful reading (r= .329, .368, .305 respectively). F5 has a relatively high correlation with search reading (.570). F8 and F9 are not interpretable and F10 has a high correlation with the skimming question (.722). Shortly, it can roughly be said that in the correlation pattern discussed above there are two factors that account for an operation that involves search and careful reading simultaneously (F1 and F3), one factor for scanning (F2), one factor for search reading (F5), one factor for skimming (.722) and three factors for careful reading (F4, F6 and F7). There are two factors that do not correlate at all with any subtests within the set limit. Both the rotated component and the correlation matrix are hard to interpret and suggestive of random factors that are not accounted by the reading operations solely. Then the next step would be to eliminate random factors as far as possible. Reduction of random factors relating to item characteristics was done depending on the item analysis results and thus eliminating the items that proved to be problematic in that stage. The reader is referred to section 4.5.2.2 (item analysis for the September 2000 test version) for the performance of the items in the test. Table 4.21 shows the rotated component matrix of the purged form of September 2000 test data excluding the items SC1-4, SR1, 4, 6 and CR5 (N=22).

Table 4.21: Rotated component matrix: September 2000 test – purged set

	Component						
	1	2	3	4	5	6	7
SC6	.163	**.490**	.242	.114	.030	-.272	.248
SC7	.160	.120	**.778**	.040	.092	.041	.097
SC8	.045	.302	**.740**	.055	.046	.008	-.029
SC9	.068	**.589**	.263	.064	-.071	.225	.073
SC10	-.016	**.715**	.246	.011	.112	.142	-.001
SR2	**.416**	.350	-.138	.078	.131	.114	-.294
SR3	.322	**.484**	-.025	.347	-.036	.156	-.196
SR5	.282	.039	.338	.193	.199	**.346**	-.123
SR7	.099	.129	.071	.044	.033	**.713**	.036
SR8	.185	.273	-.164	.136	.100	**.451**	.402
SR9	.168	.032	.055	.146	.046	.077	**.779**
SR10	**.674**	.055	.076	.160	-.105	.261	.085
SR11	**.583**	.030	.132	.009	.186	.283	.060
CR1	.069	-.041	.010	-.018	**.637**	.358	.178
CR2	-.130	.094	.070	**.791**	.047	.073	.156
CR3	.470	.079	.104	**.609**	.053	-.106	-.092
CR4	**.639**	.097	.163	.093	.138	-.061	.166
CR6	**.527**	.242	-.006	-.079	.371	-.147	.204
CR7	.196	.382	-.121	.138	**.453**	-.028	.027
CR8	.197	.093	.013	**.518**	.213	.112	.117
CR9	.046	-.039	.212	.165	**.687**	-.036	-.134
CR10	.265	.380	.111	.190	**.414**	-.021	.136

PCA extracted 7 factors in the purged data. KMO measure of sampling adequacy was .855 and Bartlett's test of sphericity was significant at .000 level. No communalities below .30 were observed.[31] 7 components with eigenvalues higher than 1.00 accounted for 54% of variance. The highest factor loading of each variable is given in boldface in Table 4.21. Although less complex compared to the previous structure, the rotated component structure of September 2000 test excluding problematic items did not yield an easy to interpret picture. To facilitate interpretation, the factor loadings were correlated against the total subtest scores. In Table 4.22 below, it is seen that the first factor accounts for 21.79% of all the variance in the data set – a considerably large amount – the second, 6.97%, and all the others, approximately 5%. Scanning correlates

with F2 and F3 (.661 and .679 respectively), search reading with F1, F2 and F6 (.628, .315 and .560 respectively), and careful reading with F1, F4 and F5 (.478, .518 and .592 respectively).

Table 4.22: September 2000 test – purged set: Subtest – Factor correlations

Factors	Eigen value	% of variance (cum: 53.7%)	scanning II (SC 6-10)	search r. (-SR1,4,6)	careful r. (-CR5)
F1	4.794	21.790	.106*	**.628***	**.478***
F2	1.533	6.967	**.661***	**.315***	.276*
F3	1.182	5.373	**.679***	.079	.113*
F4	1.165	5.296	.090	.249*	**.518***
F5	1.105	5.025	.086	.133*	**.592***
F6	1.030	4.683	.048	**.560***	.021
F7	1.002	4.554	.124*	.185*	.158*

Pearson Correlation: * *correlation is significant at 0.01level (2-tailed) / * correlation is significant at 0.05 level (2-tailed)

The subtests' correlations with more than one factor might be because of the fact that the subtests themselves have composite structures. However, each component can still be assigned a certain meaning. The first component F1 seems to account for an operation which is a combination of search and careful reading. F2 on the other hand is a factor that accounts for a combination of scanning and search reading. While at first sight these combinations may not seem meaningful, when the reading taxonomy that underlies the skills classification and operational definitions of the BUEPT reading test is considered, it is seen that F1 might be considered to account for 'global' level operations and F2 for 'expeditious' reading skills. F1 suggests a link between search and careful reading in both of which readers should identify main ideas in the text (comprehension at a global level), and between scanning and search reading, in both of which readers should read expeditiously and selectively. The reader is referred to Appendix 3.2 for detailed definitions of operations and skills.[32] On the other hand, F3 correlates with scanning

[31] See Appendix 4.11 for details.

[32] For brief reference: scanning (reading expeditiously at the local level), search reading (reading expeditiously at the global level), careful reading (reading carefully at the global level).

alone, F4 and F5 correlate basically with careful reading and F6 is the factor that seems to explain search reading. F7 does not correlate meaningfully with any of the subtests.

The purged version of September 2000 test yielded the picture above. Although it is not always very likely to achieve the ideal distribution of variables to components as expected in such complex analyses as PCA, all possible random factors should be identified and eliminated as far as possible before final conclusions are made. One way of attaining a better matrix would be to constrain the number of factors to four (or three if SR1 is excluded), and interpret the resulting matrix. In fact the researcher of the study gradually dropped each uninterpretable component by reducing the number of factors to be extracted one by one and worked on several matrices with fewer components, which eventually produced more interpretable data. However, each time a component is dropped, the cumulative variance accounted for by the retained components is reduced. By four components, it drops to 32%. Even though very high percent of variance (70, 80% as Hatcher 1994 suggests) may not be expected to be accounted for with dichotomous data (See Green and Weir, 1998), such low accountable variance as 32% is not very desirable. Several analyses of the data mentioned above suggested that other than item characteristics, the subtests themselves might have a composite nature, a factor that might affect the results. Therefore, the subtests were submitted to PCA analysis individually and the results were not surprising.

Table 4.23 below shows that subtests in themselves are componential.[33]

---

[33] See Appendix 4.12 for details.

Table 4.23: Rotated component matrix: September 2000 test - subtests

	Component		
	1	2	3
SC1	.118	.020	**.840**
SC2	.160	**.462**	.414
SC3	.051	**.705**	.190
SC4	.063	**.668**	-.087
SC5	**.635**	.038	.049
SC6	**.525**	.165	.221
SC7	**.554**	.412	-.207
SC8	**.600**	.398	-.261
SC9	**.666**	-.108	.170
SC10	**.683**	.113	.098

	Component		
	1	2	3
SR1	-.133	-.108	**.780**
SR2	.261	.332	**.348**
SR3	.263	**.524**	.355
SR4	-.098	**.821**	-.129
SR5	.175	**.674**	-.028
SR6	.223	**.263**	.063
SR7	**.488**	.107	.035
SR8	**.629**	.155	-.163
SR9	**.591**	-.063	-.269
SR10	**.637**	.179	.227
SR11	**.596**	.154	.333

	Component	
	1	2
CR1	**.632**	-.116
CR2	-.046	**.657**
CR3	.190	**.730**
CR4	**.471**	.400
CR5	.116	**.655**
CR6	**.681**	.099
CR7	**.475**	.270
CR8	**.433**	.318
CR9	**.508**	.101
CR10	**.611**	.247

As was suggested in the item analysis part, apparently, the test takers behaved differently in the first and second part of the scanning test. An explanation posed for this observation was that test takers might have used time less sparingly for the first few items and did more slow and careful reading with higher ratios of correct answers. In the second part, they had to do fast, selective reading with fewer correct answers. Differential performance of the test takers in the first and second part of the scanning test is supported by the component structure of the test, too. In the scanning matrix in Table 4.23 above, F1 accounts for SC5-10, F2 for SC2-4 and F3 for SC1. Rotated component matrix of the search reading test displays three factors; first component loading clearly on the second-text search reading items (SR7-11), the second component on the first-text search reading items (SR2-6), the third component basically accounts for the skimming (SR1) element. SR2 loads on the third component as well, but consider that its loading on the second component is almost the same. PCA analysis of the careful reading test produced two components; the first component accounts for the second-text careful reading items (CR6-11), the second component for the first-text careful reading items (CR1-5). Among the first-text items, only CR1 does not load on F2. CR4 has

meaningful loading both on the first and second component, which in fact shows that it is a componentially heterogenous, therefore, problematic item.

These findings suggest a strong text effect for search and careful reading tests and it became clear that the scanning test should also be treated as bi-componential for the reason stated above. Thus, it appeared reasonable to reduce some of these factors that might have been influential in the 10-component matrix output. Differential item properties together with text difference factor might be producing a more complex effect and might be yielding factors that cannot be accounted for only by the interpretations already brought in. Therefore, a new data set was formed with the inclusion of SC5-10 (scanning II), SR1 (skimming), SR7-11 (search reading II) and CR6-10 (careful reading II). SR1 was retained in the analysis despite the fact that it was a problematic item because it was the only item testing skimming skill. The researcher also preferred to include the second search reading test since the first one had certain defective items and the second careful reading text seemed more homogeneous. This set was designated as 'half-set I' data and was submitted to PCA analysis with varimax rotation.[34] As seen in Table 4.24 below PCA extracted four components with eigenvalues over 1.00. KMO measure of sample adequacy was .824 and Bartlett's test of sphericity was significant at .000 level.

[34] See Appendix 4.13 for details.

Table 4.24: Rotated component matrix: September 2000 test – half-set I

	Component			
	1	2	3	4
SC5	**.594**	.142	-,013	-,276
SC6	**,495**	,243	.076	-,184
SC7	**,656**	,159	,030	,138
SC8	**,728**	,104	-,051	,087
SC9	**,544**	,004	,334	,067
SC10	**,642**	,102	,206	-,003
SR1	-,018	-,009	-,066	**,833**
SR7	,185	-,131	**,607**	,025
SR8	,025	,178	**,662**	-,145
SR9	,022	,230	**,438**	-,198
SR10	,071	,261	**,589**	,166
SR11	,102	,393	**,448**	,356
CR6	,082	**,593**	,181	,054
CR7	,119	**,522**	,144	-,005
CR8	,092	**,439**	,277	-,194
CR9	,113	**,623**	-,159	-,006
CR10	,250	**,622**	,180	,011

Table 4.25: September Test – half-set I: Subtest – Factor correlations

Factors	**Eigen value**	**% of variance (cum: 43.9%)**	**scanning II (SC 5-10)**	**search r. II (SR 7-11)**	**careful r. II (CR 6-10)**	**skimming (SR1)**
F1	3.704	21.785	**.960***	.134*	.216*	-.018
F2	1.544	9.085	.196*	**.303***	**.905***	-.009
F3	1.120	6.591	.159*	**.895***	.217*	-.066
F4	1.088	6.400	-.026	.066	-.050	**.833***

Pearson Correlation: * *correlation is significant at 0.01level (2-tailed) / * correlation is significant at 0.05 level (2-tailed)

Four factors accounted for 43.86% of total variance. There is a clear distribution of subtests to components; F1 accounting for scanning subtest, F2 for the careful reading, F3 for search reading and F4 for the single item skimming (SR1). Table 4.25 above shows that the first component (F1) accounts for a large percent of variance (21.79%), the second one for 9.09%, the third and fourth for approximately 6.50%. The correlation of F1 with the scanning II is .960, that of F2 with careful reading II is .905. F3 correlates with search reading II with r= .895 and F4 with SR1 with r= .833. F2 also correlates with search reading II at r= .303 suggesting a link between search and careful reading. However, that correlation is much lower than the correlation between F2 and careful reading (.905).

To see whether a similar structure appears with the remaining items of search and careful reading items, a second half set (half-set II) was formed with the inclusion of SC5-10 (scanning II), SR1 (skimming), SR2-6 (search reading I) and CR1-5 (careful reading I) (See Table 4.26 below).

Table 4.26: Rotated component matrix: September 2000 test – half -set II

	Component				
	1	2	3	4	5
SC5	**.650**	-.034	.091	-.186	.256
SC6	**.500**	.312	-.111	-.049	.235
SC7	**.658**	.115	.268	.012	-.241
SC8	**.700**	.115	.207	.047	-.241
SC9	**.515**	.179	-.117	.371	.062
SC10	**.627**	.104	-.091	.286	.211
SR1	-.096	-.164	.019	**.613**	-.357
SR2	.058	.130	.160	**.546**	.195
SR3	.159	.379	.078	**.521**	.225
SR4	-.090	.393	**.587**	-.023	.044
SR5	.197	.258	**.654**	.141	-.110
SR6	.036	.078	.120	.104	**.739**
CR1	.074	-.131	**.621**	.100	.318
CR2	.072	**.590**	.052	-.099	.016
CR3	.043	**.740**	.060	.196	-.018
CR4	.149	**.465**	.167	.217	.151
CR5	.166	**.605**	.089	.050	.013

The half-set II data were submitted to PCA. The output was different but still worthy of consideration. For half-set II, KMO measure of sample adequacy was .771 and Bartlett's test of sphericity was significant at .000 level. Four factors accounted for 49.25% of total variance.[35] PCA extracted five components with eigenvalues over 1.00. All scanning II items loaded on F1 and except for CR1, careful reading I items loaded on F2. However, the case was not so clear for search reading I items. SR2 and 3 loaded on F4 together with the skimming item (SR1). SR4 and 5 loaded on F3 and SR6 was the only item loading on F5. The only subtest that could be accounted for by one factor was the scanning subtest. Obviously, certain search and careful reading items shared some

properties (either search reading items functioning as careful reading items or vice-versa) needless to mention the confounding effect of the problematic item characteristics of certain items in this set. The subtest-factor correlations in Table 4.27 also show that scanning strongly correlates with F1 (.953). F2 correlates with careful reading I (.779) and search reading I (.404), similarly F3 correlates with search reading I (.422) and careful reading I (.311). F4 seems to account for skimming (.613). F5 seems to have a weak correlation with search reading I.

Table 4.27: September test – half-set II: Subtest – Factor correlations

Factors	Eigen value	% of variance (cum: 49.5%)	scanning II (SC 5-10)	search r. I (SR 2-6)	careful r. I (CR 1-5)	skimming (SR1)
F1	3.581	21.066	**.953***	.227*	.162*	-.096
F2	1.559	9.170	.217*	**.404***	**.779***	-.164*
F3	1.111	6.534	.062	**.422***	**.311***	.019
F4	1.089	6.408	.145*	**.414***	.158*	**.613***
F5	1.073	6.315	.058	.294*	.153*	-.357*

Pearson Correlation: * *correlation is significant at 0.01level (2-tailed) / * correlation is significant at 0.05 level (2-tailed)

After certain evidence supporting text factor, 10-component output was re-analysed correlating the sub-sections of the subtests against the factors extracted by PCA. Therefore, scanning I (SC1-5), scanning II (SC6-10), skimming (SR1), search reading I (SR2-6), search reading II (SR7-11), careful reading I (CR1-5), careful reading II (CR6-10) were correlated against the factors hoping that these parts would correlate with different factors to differing levels to explain the existence of the factors.[36] However, except for scanning II (.932 compared to .865) and search reading II (.720 compared to .570) and careful reading I (.568 compared to .561) there was not significant increase in the correlation of the sub-sections to factors compared to the correlation of subtests to factors. Nevertheless, it should notice that among all the parts in the data, scanning II,

[35] See Appendix 4.14 for details.
[36] See Appendix 4.15 for the correlations.

search reading II and careful reading I were the parts that had fewer problems so they could be identified better as factors in this matrix. The same procedure was applied to the purged data in which problematic items were excluded.[37] This time the data did not include scanning I (SC1-5) since these items were found to be problematic. Moreover, SR1, 4 and 6 and CR5 were also excluded since were found to be defective items, too. Among all the parts included in the analysis, only careful reading II had higher coefficient (.681) as compared to total careful reading (.592). Since no clear results could be achieved, these data were disregarded.

4.5.2.5 The September 2000 Test: Discussion

Shortly, the procedures and the findings detailed above, especially half-set I data, suggest that the subtests of the BUEPT reading test can be identified as factors in the data. The reason why initial extractions did not produce clear pictures might be several: defective item properties, composite nature of some items, or interrelatedness of certain operations that were used in answering the questions as well as text difference could be the factors that affected the item performance and item correlations. In that respect, it has been observed through item analysis that certain items such as scanning have proved to be very easy and have low discriminating power. On the other hand, certain items such as SR4, SR6 and CR5 received too few correct responses resulting in undesirable item performance. Deleting these weakly performing items resulted in simpler factor structures. On the other hand, the subtests themselves included items of differing level of

[37] See Appendix 4.15.

difficulty as well. In fact, a neat balance between the difficult and easy items is desirable in proficiency tests; however, such difference also suggests that difficult items may require different reading processes as compared to easy items. For example, easy careful reading items might be answered by using mere search reading operations and vice-versa. In certain cases, it might be the case that all reading operations were used interchangeably to facilitate test taking process. This is also noted down by Shih (1992) that readers adjust their reading process according to the reading purpose and it is possible that they use more than one operation to arrive at an answer when taking a test. It should also be pointed out that by definition, search and careful reading skills overlap considerably in that search reading entails careful reading when the answer is located. Both require test takers to identify the main ideas in the text and comprehend at the global level. This is also clearly reflected in the inter-correlations. The fact that the search and careful reading tests included two different texts with five items based on each must have also introduced a text effect to the data to further complicate the emerging factor structure. It has been noted in the literature that the items based on a text may exhibit local item dependence in reading tests (Yong-Won 2000, 2004). Unfortunately, the statistical procedures that might shed light on such effects require considerable expertise on the part of the researcher, which the researcher of this study did not possess. On the other hand, the qualitative data gathered through verbal protocols can be integrated into the discussion to unfold several points made above.[38] For example, it has been observed that under test taking conditions, test takers usually wished to answer test items as quickly as possible and they usually started with reading

[38] See section 4.4.

quickly to locate the answer. Where the answer was not easily accessible, they had to read more carefully and in some cases, larger text spans. This happened both with some careful and search reading items. A minority of the test takers used certain test taking strategies that were not explained by any reading operations to find the answer successfully.

However, almost all the test takers who participated in the verbal protocol study reported the use of reading skills as they are explained in the test specifications and the framework the specifications were based on. The circumstances under which they performed those operations, however, changed according to the characteristics of the items and of the text as well as their preferred test taking style. In that respect, the verbal protocol data suggested that item and text characteristics have been controlled to a large extent though there is room for improvement with the tests used for that part of the study. To remove emerging ambiguities in the quantitative data, it can be suggested that if evidence on one test is replicated on another, the arguments are better grounded and the conclusions become more robust if supported by repeated evidence. Therefore, the statistical analysis of three more tests will be discussed before final conclusions on the nature of the reading process will be made.

4.5.3 The January 2001 Test – Pilot Version

The January 2001 test was piloted following the procedures explained in 3.5.4. The details are as follows.

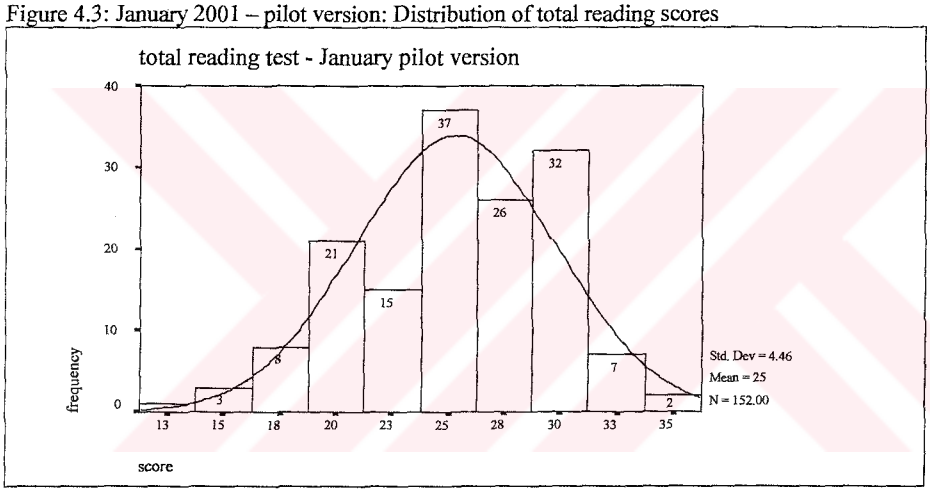
4.5.3.1 The January 2001 Test – Pilot Version: Descriptive Statistics

Table 4.28 below provides the score distribution data for the pilot version of the January 2001 test. As seen in the table, the mean is over the cut-off (70.27%) and Figure 4.3 below shows that the distribution is slightly skewed.[39] Alpha coefficient (0.71) for the total reading test is satisfactory.

Table 4.28: January 2001 test – pilot version: Descriptive statistics of the total reading scores

N	Item n.	Range	Min.	Max.	Mean	SE	Std.	Skewness	Kurtosis	Alpha
152	36	22	13	35	25.3 (70.27%)	0.36	4.45	-.372	-.449	0.71

Figure 4.3: January 2001 – pilot version: Distribution of total reading scores



As for the individual subtests, it is seen in Table 4.29 that the scanning test has the highest mean (86.36%) with a negatively skewed distribution.[40] The scanning test also

Table 4.29: January 2001 test – pilot version: Descriptive statistics of the subtests

Subtests	N	Item n	Range	Min	Max.	Mean	SE	Std.	Skewness	Kurtosis	Alpha
Scanning	152	11	5	6	11	9.5 (86.36%)	0.11	1.34	-.893	.406	0.41
Search R.	152	13	11	1	12	8.18 (62.92%)	0.18	2.26	-.567	-.118	0.52
Careful R.	152	12	12	0	12	7.56 (63%)	0.19	2.37	-.579	.113	0.63

[39] See Appendix 4.16 for the normality tests and plots.
[40] See Appendices 4.17 and 4.18 for the score distribution graphs and normality tests and plots for the subtests.

has the lowest alpha (0.41). The search reading test has a mean of 62.92%. The score distribution in the search reading test is negatively skewed, too. Alpha coefficient of 0.52 is lower than expected. The careful reading test is similar to the search reading test except for a slightly higher alpha (0.63). In general, the pilot version of the January test yielded higher mean scores with lower alpha coefficients compared to the September version. The item analysis data below will shed light on the nature of the items in the test, and thus, the possible reasons for this.

4.5.3.2 The January 2001 Test – Pilot Version: Item Analysis

Item analysis for the January pilot test was carried out as explained in section 3.5. However, for the calculation of the item discrimination (ID) index, 'the upper and lower 33%' percent rule could not be used with the January pilot test data since the scores were considerably high and there were no test takers scoring in the lower 33%. So 'the lower' group was enlarged to include the scores in the lower 50% (0-18) and the 'upper group' was reduced to include approximately the upper 20% (29-36). If this hadn't been done, ID indices would have been augmented unrealistically. Table 4.30 below presents the item analysis results of the pilot version of the January 2001 test. The values that are not within acceptable limits are marked in boldface.

Table 4.30: January 2001 test – pilot version: Item analysis statistics

Item	IF	ID	CITC	CITC Subtest	AIID*	AIID* Subtest	Item	IF	ID	CITC	CITC subtest	AIID*	AIID* subtest
SC1	**0.94**	**0.06**	**0.0955**	**-0.0117**	0.7101	**0.4280**	**SR8**	0.71	**0.30**	**0.1178**	**0.0321**	**0.7112**	**0.5414**
SC2	0.72	**0.08**	**0.0696**	0.2089	**0.7141**	0.3602	**SR9**	0.32	**0.25**	**0.0427**	**0.0938**	**0. 7163**	**0.5280**
SC3	0.62	**0.25**	**0.1811**	**0.1730**	0.7073	0.3827	**SR10**	0.49	0.70	0.3041	0.3114	0.6984	0.4719
SC4	**0.96**	**0.08**	**0.1222**	**0.0913**	0.7092	0.4019	**SR11**	**0.81**	**0.33**	0.2195	0.2106	0.7045	0.5004
SC5	**0.81**	0.53	0.2432	**0.1863**	0.7031	0.3706	**SR12**	0.66	0.43	0.2190	0.1605	0.7045	0.5121
SC6	**0.94**	**0.08**	**0. 1592**	0.3019	0.7079	0.3463	**SR13**	0.21	**0.25**	**0.1309**	**0.1563**	0.7097	0.5118
SC7	**0.94**	**0.15**	**0.1210**	**0.1406**	0.7092	0.3895	**CR1**	0.72	**0.35**	**0.1462**	0.2047	0.7093	0.6207
SC8	**0.99**	**0.00**	**0.1121**	**0.1373**	0.7099	0.3982	**CR2**	0.78	0.45	0.2650	0. 3134	0.7016	0.6007
SC9	**0.89**	**0.25**	0.2728	0.2246	0.7024	0.3582	**CR3**	0.54	0.50	0.2664	0.2118	0.7012	0.6211
SC10	**0.97**	**0.17**	**0.1892**	**0.1979**	0.7076	0.3813	**CR4**	**0.84**	**0.20**	**0.1845**	0.2543	0. 7064	0.6115
SC11	0.77	**0.21**	**0.1330**	**0.1180**	0.7097	0.4040	**CR5**	0.75	0.55	0.2390	0.2705	0.7032	0.6083
SR1	0.64	**-0.05**	**0.0223**	**-0.0425**	**0.7211**	**0.5612**	**CR6**	**0.16**	**0.08**	**0.0842**	**0.1651**	**0.7118**	0.6251
SR2	0.78	0.45	0.2838	0.3251	0.7005	0.4740	**CR7**	0.41	0.85	0.4011	0.4025	0.6912	0.5803
SR3	0.62	0.85	0.4035	0.3733	0.6912	0.4556	**CR8**	0.60	0.63	0.2343	0.2229	0.7035	0.6185
SR4	0.74	0.70	0.3867	0.3288	0.6935	0.4715	**CR9**	0.50	0.72	0.2894	0.2948	0.6965	0.6037
SR5	0.68	0.85	0.3502	0.2766	0.6954	0.4827	**CR10**	**0.84**	0.60	0.2781	0.2762	0.7014	0.6082
SR6	0.77	0.68	**0.1803**	**0.1267**	0.7068	0.5185	**CR11**	0.73	0.62	0.4853	0.4606	0.6866	0.5712
SR7	0.76	**0.35**	**0.1609**	0.2690	0.7081	0.4863	**CR12**	0.70	0.63	0.2658	0.2310	0.7013	0.6159

SC: scanning **SR**: search reading **CR**: careful reading **IF**: item facility **ID**: item discrimination **CITC**: corrected item-total correlation
AIID: alpha if item deleted ***Alpha**; **overall:** 0.7107 **SC:** 0.4076 **SR:** 0.5227 **CR:** 0.6282

It is clearly seen in the table that in general the scanning items have high IF values whereas their ID power is low as was the case with the September test. Moreover, most of the items have low correlation with the total test and the subtest. SC1 has negative impact on the reliability of the subtest, and SC2 on the whole test. As for the skimming item (SR1), it has unfavourable item statistics as the SR1 in the September test except for the relatively good IF of 0.64. Similarly, five more items in the search reading test needed consideration (SR6-9 and SR13) as well as CR1, CR4 and CR6 in the careful reading test. After the item discrimination patterns (IDPs) of these items were analysed, these items were either repaired or eliminated from the test.

For the analysis of IDPs, six groups were formed as shown in Table 4.31.

Table 4.31: January 2001 test – pilot version: Distribution of total score by band

band	range of scores	total score mean	no of test takers	pass/fail	percent	cumulative percent
1	0-11	0	0	fail	0%	0%
2	12-17	16.2	10	fail	6.6%	6.6%
3	18-21	19.9	23	fail	15.1%	21.7%
4	22-24	23.8	40	pass	26.3%	48%
5	25-29	27.5	52	pass	34.2%	82.2%
6	30-36	31.2	27	pass	17.8%	100.0%

The possible highest mark in the test was 36 and the pass/fail cut-off was set at 21.6≈ 22 (60%). There were no test takers assigned to the first band. Only 21.7% of the test takers were assigned to the failing groups. The majority were in the passing groups, the largest group being in the fifth band. The item discrimination patterns based on this grouping are given in Table 4.32 below.

Table 4.32: January 2001 test – pilot version: Item discrimination patterns by band

band	SC1	SC2	SC3	SC4	SC5	SC6	SC7	SC8	SC9	SC10	SC11	SR1
2	0.90	0.70	0.50	0.90	0.50	0.90	0.90	1.00	0.70	0.90	0.70	0.80
3	0.91	0.70	0.35	0.91	0.65	0.87	0.83	0.91	0.78	0.87	0.52	0.48
4	0.90	0.60	0.55	0.95	0.80	0.90	0.95	1.00	0.83	0.98	0.78	0.63
5	0.96	0.75	0.69	0.98	0.87	0.98	0.98	1.00	0.96	1.00	0.81	0.63
6	1.00	0.89	0.85	1.00	0.96	1.00	0.96	1.00	1.00	1.00	0.93	0.78
band	**SR2**	**SR3**	**SR4**	**SR5**	**SR6**	**SR7**	**SR8**	**SR9**	**SR10**	**SR11**	**SR12**	**SR13**
2	0.40	0.00	0.30	0.10	0.20	0.50	0.60	0.00	0.00	0.60	0.50	0.10
3	0.61	0.35	0.48	0.52	0.74	0.65	0.52	0.26	0.30	0.74	0.39	0.09
4	0.73	0.57	0.68	0.63	0.80	0.73	0.70	0.30	0.40	0.75	0.60	0.20
5	0.87	0.71	0.88	0.77	0.79	0.77	0.81	0.42	0.58	0.83	0.73	0.23
6	0.96	0.96	0.96	0.93	0.93	0.96	0.74	0.30	0.78	1.00	0.93	0.33
band	**CR1**	**CR2**	**CR3**	**CR4**	**CR5**	**CR6**	**CR7**	**CR8**	**CR9**	**CR10**	**CR11**	**CR12**
2	0.40	0.40	0.30	0.70	0.20	0.10	0.00	0.20	0.00	0.50	0.40	0.30
3	0.70	0.57	0.26	0.61	0.65	0.13	0.22	0.39	0.43	0.70	0.17	0.57
4	0.57	0.78	0.48	0.83	0.73	0.10	0.26	0.57	0.38	0.80	0.73	0.65
5	0.79	0.87	0.62	0.90	0.81	0.17	0.44	0.67	0.62	0.96	0.90	0.77
6	0.93	0.93	0.81	0.96	0.89	0.30	0.89	0.81	0.70	0.93	1.00	0.93

Table 4.32 makes it clear that the scanning items have low discrimination values since they could be answered correctly by the majority of the test takers in all the bands. Among all the scanning items, SC3 is the only one behaving differently since in lower bands it received the fewest correct responses. SR1 once again presents a problematic pattern especially because it was answered correctly by approximately the same number of test takers in band1 and band6. Among the search reading items, SR6 does not discriminate well between bands 3-6. Similarly, too many test takers could answer SR7 and SR8 in lower bands. On the other hand, SR9 and SR13 are problematic because only the minority of the test takers in each band could answer these questions. In the careful reading test, there were four problematic items (CR1, CR4, CR6 and CR10). CR1 was an easy item and did not discriminate well between passing and failing groups. Although the percent of correct answers for CR4 is relatively incremental, it was an easy item for

lower bands. CR6 was answered incorrectly by the majority of the test takers in all the bands. CR10 received too many correct answers in the lower bands.

4.5.3.3 The January 2001 Test – Pilot Version: Evaluation of the Items

In general, the relative easiness of the test for the subject group evident in the statistical analyses given above was the major factor to consider. Hence, except for a few cases, the problematic items were relatively easy for the lower bands, which reduced their discrimination power. The apparent reason for that was the high performance of the 66 test takers who were about to finish their freshman year in Advanced English classes. Looking at Table 4.31, it is seen that there were relatively fewer test takers in the lower bands, a fact that influences the calculation of the percentages of the correct answers. This was taken into consideration in the final decisions and the hardest items; SC3, SR6, SR13, CR6 and CR7, were dropped from the test. SR9 was re-written. For SR1, the multiple-choice skimming question, the test takers' responses were checked to detect strong distractors. Since none of the distractors seemed to be too strong, no other major changes were done.

4.5.4. The January 2001 Test

After the analyses explained in the previous section were completed, the test was reduced to its purged version. It was administered as part of the proficiency exam to the group of advanced students and the intermediate students who achieved an average of

80% or above in the achievement tests in the School of Foreign Languages of Boğaziçi University. The statistical analyses done on the data are given below.

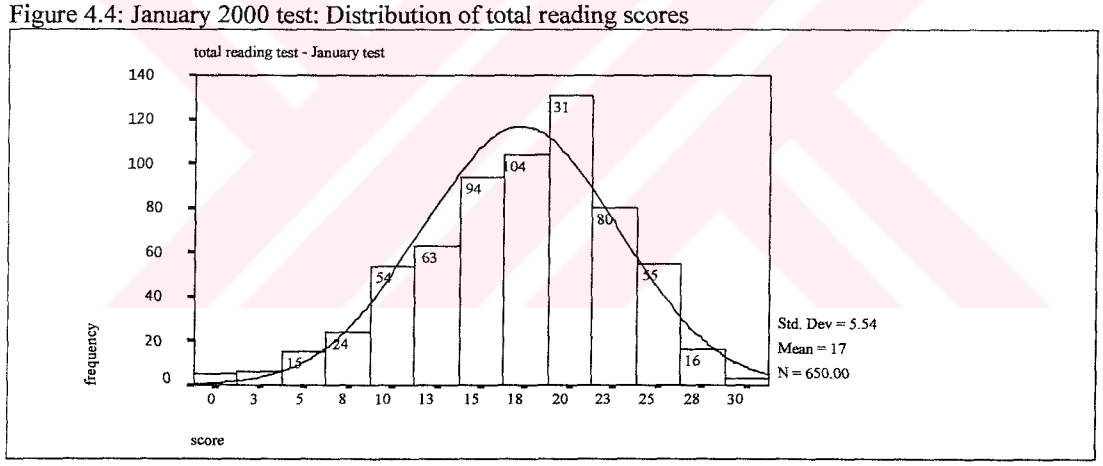
4.5.4.1 The January 2001 Test: Descriptive Statistics

The data from the actual administration of the January 2001 test revealed a mean score of 55.39% (See Table 4.33) with a slightly peaked normal distribution (See Figure 4.4).[41] The alpha coefficient of reliability (0.82) is quite satisfactory.

Table 4.33: January 2001 test: Descriptive statistics of total reading scores

N	Item n.	Range	Min	Max	Mean	SE	Std.	Skewness	Kurtosis	Alpha
650	31	30	0	30	17.17 (55.39%)	0.22	5.54	-0.474	-0.044	0.82

Figure 4.4: January 2000 test: Distribution of total reading scores



The distribution of the scores and the alpha coefficients of the subtests are given in Table 4.34 below.

[41] See Appendix 4.20 for the normality tests and graphs.

Table 4.34: January 2001 test: Descriptive statistics of the subtests

Substest	N	Item n	Range	Min	Max	Mean	SE	Std.	Skewness	Kurtosis	Alpha
Scanning	650	10	10	0	10	6.79 (67.9%)	0.08	2.10	-0.717	0.559	0.67
Search R.	650	11	11	0	11	4.60 (41.8%)	0.09	2.38	0.223	0.191	0.64
Careful R.	650	10	10	0	10	5.78 (57.8%)	0.09	2.48	0.096	-0.426	0.69

It is seen in Table 4.34 that the mean of the scanning test is 67.9% and the alpha coefficient is 0.67. The distribution of the scores in the scanning test is negatively skewed. The search reading test has the lowest mean, 41.8% with a negatively skewed distribution. The search reading test has the lowest alpha (0.64). The careful reading test has a mean of 57.8%, and an alpha of 0.69. The score distribution in the careful reading test is near normal.[42] As expected, the scores of central tendency and reliability were improved compared to the pilot version. As such, they were within acceptable limits, sparing the negative skewedness of the scanning test.

4.5.4.2 The January Test: Item Analysis

To evaluate the performance of the items in the January test, item analysis was performed as explained in section 3.5. Table 4.35 below provides the item analysis statistics for the test. The values that are not within acceptable limits are given in boldface.

[42] See Appendix 4.21 for the score distribution graphs and 4.22 for the normality tests and graphs.

Table 4.35: January 2001 test: Item analysis statistics

Item	IF	ID	CITC	CITC Subtest	AIID*	AIID* subtest	Item	IF	ID	CITC	CITC subtest	AIID*	AIID* subtest
SC1	**0.95**	**0.19**	0.24	0.25	0.8146	0.6627	**SR7**	0.66	0.44	0.25	0.25	0.8143	0.6280
SC2	0.48	**0.38**	0.22	0.21	0.8156	**0.6749**	**SR8**	0.34	0.54	0.38	0.39	0.8094	0.6008
SC3	0.66	**0.39**	0.21	**0.18**	0.8156	**0.6794**	**SR9**	0.61	0.81	0.33	0.29	0.8110	0.6203
SC4	0.42	0.56	0.34	0.31	0.8108	0.6530	**SR10**	0.43	0.53	**0.10**	**0.13**	**0.8201**	**0.6535**
SC5	**0.88**	0.41	0.34	0.39	0.8116	0.6401	**SR11**	**0.06**	**0.13**	**0.18**	0.21	0.8156	0.6348
SC6	**0.87**	0.44	0.33	0.43	0.8116	0.6329	**CR1**	0.50	0.63	0.35	0.35	0.8104	0.6664
SC7	**0.88**	0.43	0.38	0.51	0.8106	0.6223	**CR2**	0.69	0.75	0.47	0.46	0.8059	0.6481
SC8	0.45	0.58	0.37	0.39	0.8098	0.6342	**CR3**	0.39	0.59	0.35	0.36	0.8104	0.6651
SC9	0.80	0.57	0.37	0.48	0.8102	0.6182	**CR4**	0.68	0.66	0.37	0.37	0.8095	0.6646
SC10	0.42	0.65	0.36	0.35	0.8100	0.6436	**CR5**	0.45	0.69	0.41	0.40	0.8079	0.6575
SR1	0.66	**0.33**	**0.17**	**0.10**	**0.8171**	**0.6558**	**CR6**	0.56	**0.37**	**0.15**	**0.18**	**0.8184**	**0.6977**
SR2	0.40	0.60	0.38	0.37	0.8094	0.6045	**CR7**	0.70	0.64	0.38	0.40	0.8092	0.6594
SR3	0.28	0.54	0.38	0.38	0.8096	0.6032	**CR8**	0.52	0.60	0.35	0.34	0.8104	0.6688
SR4	0.41	0.71	0.44	0.42	0.8069	0.5918	**CR9**	0.58	0.50	0.25	0.27	0.8143	0.6811
SR5	0.46	0.69	0.45	0.42	0.8065	0.5921	**CR10**	0.72	0.60	0.34	0.38	0.8108	0.6616
SR6	0.30	0.54	0.33	0.32	0.8111	0.6146							

SC: scanning **SR**: search reading **CR**: careful reading **IF**: item facility **ID**: item discrimination **CITC**: corrected item-total correlation **AIID:** alpha if item deleted ***Alpha; overall:** 0.8166 **SC:** 0.6704 **SR:** 0.6414 **CR:** 0.6903

As was the case in the previous tests, the scanning test included some items with high IF values. Surprisingly, not all these items were the first few items, and among the first items, there were two with relatively low IF values (SC2 and SC4). SC2 and SC3 had negative impact on the reliability of the scanning test. SR1, the skimming item, was problematic as before. SR10 also appeared having low correlation with the reading test in general (CITC) and the search reading test in particular (CITC Subtest) and it had negative impact on the reliability of both (AIID and AIID Subtest). This was not expected since SR10 (SR11 in the pilot version) was considered to be an easy item with only high IF and low ID. SR11 (SR12 in the pilot version) also had acceptable values in the pilot test data but in the actual test data, it appeared with very low IF, ID and CITC. There was only one item with unfavourable values in the careful reading test, and this was CR6. CR6 was numbered as CR8 in the pilot data and did not appear problematic then.

Item discrimination patterns were extracted by grouping the test takers into six bands as given in Table 4.36. 43.8% of the test takers were assigned to the passing groups.

Table 4.36: January 2001 test: Distribution of total score by band

band	range of scores	total score mean	no of test takers	pass/fail	percent	cumulative percent
1	0-18	5.7	50	fail	7.7%	7.7%
2	9-15	12.4	170	fail	26.2%	33.8%
3	16-18	17.1	145	fail	22.3%	56.2%
4	19-21	20	131	pass	20.2%	76.3%
5	22-25	23.1	126	pass	19.4%	95.7%
6	26-31	27.1	28	pass	4.3%	100.0%

The item discrimination patterns are given in Table 4.37 below.

Table 4.37: January 2001 test: Item discrimination patterns by band

band	SC1	SC2	SC3	SC4	SC5	SC6	SC7	SC8	SC9	SC10	SR1	SR2	SR3
1	0.68	0.14	0.34	0.10	0.40	0.40	0.44	0.10	0.28	0.04	0.34	0.06	0.00
2	0.95	037	0.56	0.24	0.86	0.81	0.82	0.22	0.72	0.22	0.60	0.15	0.09
3	0.96	0.46	0.60	0.34	0.88	0.92	0.92	0.39	0.81	0.39	0.64	0.36	0.21
4	0.98	0.55	0.77	0.47	0.95	0.95	0.97	0.59	0.88	0.50	0.79	0.51	0.31
5	0.98	0.63	0.79	0.71	0.98	0.94	0.98	0.70	0.96	0.69	0.71	0.69	0.56
6	1.00	0.82	0.93	0.89	1.00	0.96	1.00	0.89	1.00	0.86	0.86	0.79	0.82

band	SR4	SR5	SR6	SR7	SR8	SR9	SR10	SR11	CR1	CR2	CR3	CR4	CR5
1	0.06	0.04	0.00	0.32	0.08	0.16	0.30	0.02	0.10	0.14	0.02	0.14	0.04
2	0.14	0.18	0.11	0.52	0.10	0.45	0.38	0.01	0.28	0.46	0.18	0.51	0.19
3	0.36	0.39	0.33	0.68	0.28	0.60	0.36	0.04	0.50	0.73	0.39	0.76	0.47
4	0.56	0.66	0.33	0.70	0.42	0.73	0.44	0.05	0.59	0.87	0.44	0.79	0.57
5	0.71	0.74	0.53	0.83	0.63	0.83	0.60	0.13	0.77	0.94	0.67	0.85	0.75
6	0.93	1.00	0.75	0.93	0.86	0.89	0.61	0.29	0.82	1.00	0.86	1.00	0.86

band	CR6	CR7	CR8	CR9	CR10
1	0.24	0.24	0.12	0.12	0.20
2	0.48	0.46	0.30	0.46	0.56
3	0.59	0.77	0.54	0.64	0.80
4	0.63	0.89	0.62	0.65	0.79
5	0.63	0.84	0.75	0.71	0.90
6	0.75	1.00	0.89	0.89	0.93

In general, there were no significant problems with the item discrimination patterns of the January test. SC1 and SC5-7 seemed unable to differentiate band1 and the rest. SR6 did not differentiate between the band3 and band4 (failing and passing test takers). SR10 was answered correctly by relatively few number of test takers in higher bands, and SR11 received too few correct responses from the test takers in all the bands. Otherwise, the distribution of scores from the lower bands to the higher ones was incremental and a satisfactory number of the test takers in the higher bands answered the items correctly.

4.5.4.3 The January 2001 Test: Evaluation of the Items

Depending on the analysis done above, the problematic scanning items were not dropped from the test but as it was done before, all the scanning items were given 0.5 point credit. SR1 was also included in the score calculations since it had relatively high IF. The reason why it had low correlation with the rest of the items in the test might be due

to the fact that it was a multiple choice item and presumably behaved differently in comparison to the other items. SR10 and CR6 were also scored. However, SR11 was not included in the score calculations due to its dramatically low IF and ID. Thus, the final calculation of the scores to be assigned to the actual students who took this test to pass the prep year was done over 30 items.

4.5.4.4 The January 2001 Test: Inter-correlations and PCA

The January 2001 Test: Inter-correlations: Table 4.38 shows the inter-correlations of the subtests in the January 2001 reading test.

Table 4.38: January 2001 test: Subtest inter-correlations

	Search reading	Careful reading
Scanning	.444*	.395*
Search reading	-	.502*

*: Correlation is significant at the 0.05 level (2- tailed).

It can be seen in the table that all correlations are moderate and significant at 0.05 level. However, the scanning test has slightly lower correlation with the careful reading test than it has with the search reading test (search reading: .444, careful reading: .395). The correlation between the search and careful reading tests is .502. It can be asserted that the overlap between the tests is not large enough to claim that they test exactly the same reading skills.

The January 2001 Test: Principal Component Analysis: The January 2001 test has 31-item data and the whole test was submitted to PCA with varimax rotation without constraining the number of components to be extracted. KMO measure of sampling adequacy was .861 and Bartlett's test of sphericity was significant at .000 level. No

communalities below .30 were observed.[43] 9 components with eigenvalues higher than 1.00 were extracted and these accounted for 48.73% of variance in the data. Rotated component matrix is given in Table 4.44 in which the highest loadings of the items on the components are marked in bold.

Table 4.39: Rotated component matrix: January 2001test - whole set

	Component								
	1	2	3	4	5	6	7	8	9
SC1	.111	-.007	.074	**.693**	-.153	.159	.058	.065	.102
SC2	.126	.082	.088	**.342**	.241	-.067	-.495	.144	-.167
SC3	.155	.185	.002	.289	.128	-.210	.042	-.443	**.317**
SC4	.268	.203	.114	**.403**	.154	-.229	.131	-.169	-.258
SC5	-.039	.132	.301	**.567**	.172	.040	-.047	-.030	.173
SC6	.006	.107	**.756**	.111	-.013	.018	.011	.136	.124
SC7	.012	.102	**.795**	.143	.040	.071	.015	-.007	.108
SC8	.277	.022	**.352**	.266	.233	.066	-.102	-.134	-.179
SC9	.148	.001	**.782**	.056	.025	.113	-.002	-.022	-.027
SC10	.313	.189	**.474**	-.065	.072	-.140	.066	-.188	-.181
SR1	.068	.084	.104	.106	.072	.018	-.016	.013	**.789**
SR2	**.493**	.213	.080	.013	.106	-.062	-.044	.139	.034
SR3	**.507**	.079	.024	.165	.174	-.002	.130	.258	-.132
SR4	**.638**	.159	.060	.110	.028	.004	.127	-.055	.021
SR5	**.605**	.137	.098	.039	.053	.202	.009	.148	.087
SR6	**.578**	.093	.068	-.056	.080	.071	-.245	-.039	.034
SR7	.064	.066	-.005	.067	**.748**	-.002	.054	-.050	.137
SR8	.282	.087	.110	-.043	**.645**	.148	.018	.132	-.067
SR9	.315	.023	.095	.174	.146	.376	**.334**	-.153	-.045
SR10	-.006	.057	.046	.090	.143	-.079	**.760**	.133	-.076
SR11	.275	.065	-.029	.070	.085	-.154	.059	**.591**	.083
CR1	.306	**.518**	-.026	.071	-.091	-.042	-.075	-.177	.033
CR2	.372	**.492**	.060	.039	.016	.174	-.043	-.082	.087
CR3	.215	**.493**	.070	-.024	-.042	.060	.040	.315	.067
CR4	-.058	**.600**	.089	.283	.134	-.034	-.116	.247	-.102
CR5	.186	**.528**	.081	.100	.004	.054	.198	.237	.039
CR6	.015	.065	.064	-.042	.037	**.633**	-.057	.009	.072
CR7	.135	.388	.095	-.023	.178	**.434**	.092	-.103	.108
CR8	.053	**.512**	.183	-.115	.227	.071	.051	-.026	-.031
CR9	.053	.223	.007	.264	-.024	**.527**	-.027	.000	-.181
CR10	.097	**.541**	.009	.075	.031	.195	-.030	-.158	.044

Table 4.39 displays a rather meaningful distribution except for the components with less than three variables loading on them (F5, 7, 8 and 9). The first five scanning items load on F4 (SC3 loads more heavily on F9 but it also has a positive loading on F4 both being below .400 though.), and the last five scanning items load on F3. The skimming item

[43] See Appendix 4.23 for details.

(SR1) loads on F9, the first-text search reading items (SR2-6) load on F1 neatly but the second-text search reading items (SR7-11) load on three different factors (F5, F7 and F8). All the first-text careful reading items and two second-text careful reading items (CR8 and CR10) load on F2 but three second-text items (CR6-7 and CR9) load on F6. It can be said that the first and the second part of scanning, the skimming (SR1), the first-text search reading and the first-text careful reading tests are identifiable as components in this data. In order to verify this observation, subtest-factor correlation matrix of the 9-component extraction was analysed as it was done with the previous data. Table 4.40 presents the subtest-factor correlations in the data.

Table 4.40: January 2001 test – whole set: Subtest – Factor correlations

Factors	Eigen value	% of variance (cum: 48.8%)	total scanning	total search r.	total careful r.	skimming (SR1)
F1	5.094	16.433	**.304***	**.741***	.267*	.068
F2	1.941	6.260	.212*	.208	**.844***	.084*
F3	1.438	4.639	**.692***	.135*	.123*	.104*
F4	1.237	3.991	**.503***	.141*	.121*	.106*
F5	1.161	3.747	.214*	**.448***	.090*	.072
F6	1.156	3.728	-.075	.122*	**.403***	.018
F7	1.057	3.409	-.084*	.240*	.008	-.016
F8	1.024	3.304	-.163*	.165*	.054	.013
F9	1.009	3.254	-.044	.166*	.025	**.789***

In Table 4.40 above, it can be seen that F1 correlates mostly with search reading (.741) and to a degree with scanning (.304). F2 seems to be the factor that defines careful reading since its correlation is very high with this subtest (.844). F3 and F4 seem to account for the aspects of scanning (r= .692 and .503 respectively). F5 again correlates with search reading however, less strongly than F1 does. F6, as is the case with search reading, accounts for the second and less prominent aspect of careful reading (.403). F7 and F8 do not correlate with any subtests. F9 is the component that correlates only with the skimming item at a quite high level (r= .789). The data here show that the subtests

have primary and secondary correlations and there are two factors (F7 and F8) that cannot be accounted for by total subtest scores.

The next step was to identify the problematic items and eliminate them from the data to see whether PCA yielded a neater distribution of the items to components.

The item analysis of the January 2001 test revealed certain problems with the items: SC1 had a very high IF and low ID. SC2 had a slightly low ID and affected the subtest reliability negatively. SC3 had a slightly low ID and low subtest correlation. SC5-7 had IF values over .80 but were not defective otherwise. SR1 had a low ID, did not correlate with the whole test and the subtest sufficiently and affected the test and subtest reliability negatively. SR10's correlation with the whole test and the individual subtest was low and the item affected the reliability of both the whole test and the subtest negatively. SR11 had a very low IF and ID and its correlation with the whole test was too low. Among the careful reading items, CR6 had a low ID and low correlation both with the whole test and subtest. It also affected the subtest and test reliability negatively. When these items were extracted from the test and the data was analysed by PCA with varimax rotation, fewer components were yielded. PCA analysis of the purged data of January 2001 test (N of items: 24) yielded five components with eigenvalues over 1.00 accounting for 42.26% of total variance (See Table 4.46). KMO measure of sampling adequacy was .867 and Bartlett's test of sphericity was significant at .000 level. No communalities under .30 were observed.[44] Table 4.41 gives the component matrix of this analysis.

[44] See Appendix 4.24 for details.

Table 4.41: Rotated component matrix: January 2001 test – purged set

	Component				
	1	2	3	4	5
SC4	.279	.123	-.091	.317	**.520**
SC5	-.138	**.383**	.155	.339	.370
SC6	.035	**.765**	.122	-.047	.067
SC7	.010	**.810**	.126	.061	.077
SC8	.195	**.403**	.049	.317	.105
SC9	.143	**.782**	.054	.068	-.050
SC10	.370	**.425**	-.007	.057	.135
SR2	**.566**	.052	.089	.033	.203
SR3	**.538**	.027	.061	.216	.099
SR4	**.616**	.076	.159	.116	.051
SR5	**.570**	.115	.319	.098	-.135
SR6	**.563**	.055	.095	.075	.023
SR7	.066	-.043	.057	**.680**	.129
SR8	.306	.060	.153	**.562**	-.017
SR9	.194	.144	.329	**.368**	-.234
CR1	.279	-.002	**.444**	-.109	.184
CR2	.333	.070	**.526**	.033	.136
CR3	.272	.042	**.443**	-.118	.213
CR4	.064	.075	.317	.017	**.662**
CR5	.272	.061	.365	-.042	**.429**
CR7	.085	.104	**.616**	.189	-.105
CR8	.144	.125	**.390**	.082	.203
CR9	-.075	.070	.521	.190	-.050
CR10	.056	.021	.572	.054	.161

In the matrix, all the scanning items except for SC4 load on F2. The first text-search reading items (SR2-6) load on F1 and the second-text search reading items (SR7-9) load on F4. All careful reading items except for CR4 and CR5 load on F3. CR4 and CR5 load on F5 together with SC4 but they also have moderate positive loading on F3, too. Subtest-factor correlations given in Table 4.42 also support this distribution.

Table 4.42: January 2001 test – purged set: Subtest – Factor correlations

Factors	**Eigen value**	**% of variance (cum:42.3%)**	**scanning II (SC4-10)**	**search r. (-SR1, 10, 11)**	**careful r. (-CR6)**
F1	4.773	19.889	.257*	**.778***	.295*
F2	1.895	7.895	**.820***	.112*	.116*
F3	1.325	5.520	.076	.291*	**.857***
F4	1.090	4.542	.295*	**.486***	.058
F5	1.059	4.414	**.304***	.024	**.377***

Pearson Correlation: * *correlation is significant at 0.01level (2-tailed) / * correlation is significant at 0.05 level (2-tailed)

In this analysis, the scanning test excludes SC1-3, the search reading test excludes SR1, SR10-11 and the careful reading test does not involve CR6. Table 4.42 shows that F1

correlates with search reading with r= .778. F2 seems to account for scanning since their correlation is quite high; .820. F3 accounts for careful reading with a high correlation of .857. F4 is correlated with search reading, too, however at a lower level; .486. F5 moderately correlates with scanning and careful reading, a combination hard to explain.

On the whole, the PCA analysis of the purged version of the January test provides a considerably neat distribution of items to the components. Among the search reading items, it is possible to see a differential text effect. However, this becomes harder to observe in the careful reading items. For example, while in the 9-component matrix all the first-text careful reading items piled together loading on one factor, in the purged version CR4 and CR5 loaded on different factors. Therefore, although the five-component matrix of the purged data would suffice to discuss the componential nature of the January test, it would still be informative to assess the internal structures of the subtests.

The subtests of the January 2001 test were individually subjected to PCA to give the component matrices in Table 4.43. [45]

[45] See Appendix 4.25 for details.

Table 4.43: Rotated component matrices: January 2001 test – subtests

	Component		
	1	2	3
SC1	.124	.002	**.690**
SC2	.003	**.385**	.276
SC3	-.034	.210	**.472**
SC4	-.045	**.675**	.261
SC5	.308	.146	**.612**
SC6	**.808**	.000	.145
SC7	**.812**	.093	.192
SC8	.256	**.547**	.176
SC9	**.757**	.274	-.025
SC1	.324	**.689**	-.249

	Component		
	1	2	3
SR1	.189	**.213**	-.561
SR2	**.581**	.121	.024
SR3	**.490**	.176	.469
SR4	**.603**	.188	.118
SR5	**.623**	.171	.094
SR6	**.673**	-.050	-.142
SR7	.131	**.621**	-.230
SR8	.382	**.471**	.066
SR9	.238	**.489**	-.048
SR10	-.228	**.665**	.363
SR11	.247	.068	**.564**

	Component	
	1	2
CR1	**.549**	.099
CR2	**.555**	.310
CR3	**.627**	-.028
CR4	**.589**	.033
CR5	**.605**	.092
CR6	-.110	**.790**
CR7	.349	**.527**
CR8	**.462**	.208
CR9	.169	**.548**
CR10	**.498**	.254

PCA extracted three components in the scanning test, the first of which (F1) seems to account for the second part of the scanning test (SCII: SC6-10). Although SC8 and SC10 loaded on F2, their loading on the first component is positive, too. The second component (F2) has loading both from the items of the first and the second parts of the test. The items in the first part (SC I: SC1-5) basically load on the third component (F3). For the search reading test, the distribution of the items to the components is neater. F1 accounts for mostly the first-text search reading items (SRI: SR2-6) and F2, for the second-text items (SRII: SR7-11). S11 with its very low IF and ID behaves differently and loads on a separate factor. The skimming item (SR1), however, does not constitute a separate factor as it did in the 9-component structure matrix. The first-text careful reading items (CRI: CR1-5) load neatly on F1. However, two second-text careful reading items (CR8 and CR10) also load on F1. They also have positive loading on F2, the factor the other second-text careful reading items load on (CRII: CR6-10).

Even if the picture was roughly as explained, the subtest-factor correlations were checked for further verification. The factor scores were correlated against the subsections of the tests and the correlations are given in Table 4.49.

Table 4.44: Subsection-factor correlations: January 2001test – subtests

Scanning	SC I	SC II	Search R.	S R I	S R II	Careful R.	CR I	CR II
F1	-.009	**.806***	F1	**.931***	.253*	F1	**.935***	**.470***
F2	**.623***	**.525***	F2	.192*	**.911***	F2	.159*	**.819***
F3	**.644***	.175*	F3	.171*	.177*	-	-	-

Pearson Correlation: * correlation is significant at 0.05 level (2-tailed)

Among the correlations of the subsections of the scanning test and the factors, the highest correlation is between F1 and SCII (.806). It can comfortably be claimed that SC II emerges as a factor in the data. The subsection – factor correlations of the search reading parts confirms that the two parts are identifiable as factors (SRI with r= .931 and SRII with r= .911), and this finding suggests that there might be a text difference effect in the data. For the careful reading test, PCA extracted two factors the first of which correlates very highly with CRI (CR1-5) with r= .935 and moderately with CRII (CR6-10) with r= .470. The second factor identified correlates strongly with CRII (.891). Here, too, it can be claimed that there is a bi-componential distribution, which suggests a possible text effect.

The next step in the analysis was to separate the subsections of the test and assess the factor structure of the data in which only one search and one careful text were included together with the second part of the scanning test. The aim in doing this was to verify the subtest-factor distribution when a possible additional text effect was removed. Therefore, the first split set is formed by the inclusion of the second part of the scanning

text (SCII: SC6-10), skimming question (SR1), the first-text search reading items (SRI: SR2-6) and first-text careful reading items (CRI: CR1-5).

Table 4.45 below shows that SCII, SRI and CRI load on different factors (F2, F1 and F3 respectively) and emerge as separate skills in the data.[46] The skimming question (SR1) does not load on the same factor as SRI but it has some loading on F3 with careful reading items.

Table 4.45: Rotated component matrix: January 2001 test – half-set I

	Component		
	1	2	3
SC6	-.026	**.754**	.206
SC7	-.005	**.809**	.170
SC8	.421	**.426**	-.085
SC9	.147	**.793**	.004
SC10	.335	**.448**	.059
SR1	-.085	.163	**.346**
SR2	**.493**	.059	.260
SR3	**.601**	.051	.044
SR4	**.649**	.075	.127
SR5	**.628**	.095	.184
SR6	**.545**	.043	.133
CR1	.270	-.036	**.488**
CR2	.399	.060	**.470**
CR3	.226	.017	**.556**
CR4	.047	.108	**.649**
CR5	.200	.071	**.622**

It is also noteworthy that SR1's loading on F3 is below .40. This item in fact does not load on very strongly with any of the factors although it does not appear to form a separate factor itself either. The correlations of the subtest with the factors are given in Table 4.46.

[46] See Appendix 4.26 for details.

Table 4.46: January 2001 test – half-set I: Subtest – Factor correlations

Factors	Eigen value	% of variance (cum: 41.1%)	SCII (SC 6-10)	SRI (SR 2-6)	CRI (CR1-5)	skimming (SR1)
F1	3.648	22.800	**.306***	**.913***	**.364***	-.085*
F2	1.795	11.200	**.916***	.102*	.069	.163*
F3	1.139	7.116	.082*	.238*	**.890***	**.346***

These correlations provide support for the above interpretation. F1 correlates with search reading I with r= .913. F1 also has a moderate correlation with CRI and SCI (r= .364 and .306 respectively) but these are insignificant compared to .913. F2 accounts for SCII with a correlation of .916 and F3, for CRI with r=.890. SR1's correlation to F3 is only moderate (.346).

To assess the structures emerging with the other texts in the search and careful reading tests, the half-set II, which was formed including SCII (SC6-10), skimming (SR1), SRII (SR7-11) and CRII (CR6-10), was also analysed. PCA extracted five factors for this data.[47] The distribution of the items to the components is given in Table 4.47.

Table 4.47: Rotated component matrix: January 2001 test – half-set II

	Component				
	1	2	3	4	5
SC6	**.772**	.015	-.040	.072	.133
SC7	**.811**	.098	-.018	.035	.069
SC8	**.427**	.192	.347	-.203	-.329
SC9	**.794**	.100	.016	.018	-.014
SC10	**.481**	.101	.253	.036	-.055
SR1	.119	.142	.133	-.054	**.832**
SR7	.029	.100	**.570**	.144	.067
SR8	.122	.257	**.585**	.167	-.136
SR9	.147	.409	.136	**.432**	-.204
SR10	.038	-.078	.131	**.844**	.034
SR11	-.005	-.117	**.598**	-.052	.115
CR6	.034	**.526**	-.269	.152	.188
CR7	.094	**.639**	.102	.091	.108
CR8	.153	**.396**	.249	.002	.148
CR9	.058	**.567**	-.042	-.040	-.216
CR10	.043	**.558**	.193	-.216	.045

[47] See Appendix 4.27 for details.

It is seen in the matrix that the scanning items load on F1 and careful reading items on F2 neatly. SR1 again surfaces as a separate factor in this data (F5). However, search reading items load on two different factors; F3 and F4. It is difficult to explain the reason why SR9 and SR10 should form a separate factor. Especially SR11 and, to a certain extent, SR10 were deemed to be problematic in the item analysis data. SR9, on the other hand, had favourable item properties. The distribution in Table 4.43 was not repeated here.

4.5.4.5 The January 2001 Test: Discussion

Up to this point, the factor structures of two tests, September 2000 and January 2001 are discussed. In general, it has been observed that initial extractions yielded relatively more complex structures than it was expected. However, it has also been observed that problematic items might be adding random factors to the data resulting in hard to interpret factor-subtest distributions. The removal of the problematic items facilitated better distributions. Moreover, the analysis of the internal structures of the subtests showed that the subtests themselves are at least bi-componential, either because of time related differential performance as in the scanning test or differential text effect as in search and careful reading tests. The analysis of the purged and spilt half sets have supported that scanning, skimming, search reading and careful reading can be identified as loading on different factors therefore as separate skills in the data. These claims are supported by repeated evidence from both of the tests.

In the following part, the June 2001 and September 2001 reading tests are going to be analysed to assess whether these tests exhibit similar structures in terms of factor-subtest distributions as well. These tests have reduced forms in terms of the texts and questions.[48] Therefore, it is expected that their PCA will yield simpler component structures.

4.5.5 The June 2001 Test – Pilot Version

In the light of the analyses done on the September and the January versions, several changes were made concerning the structure of the BUEPT reading test to improve its practicality. The changes were effective from the June 2001 test and they are explained in section 3.5.5 in detail. The following tests were again written in strong dependence on the test specifications given in Appendix 3.2.

4.5.5.1 The June 2001 Test – Pilot Version: Descriptive Statistics

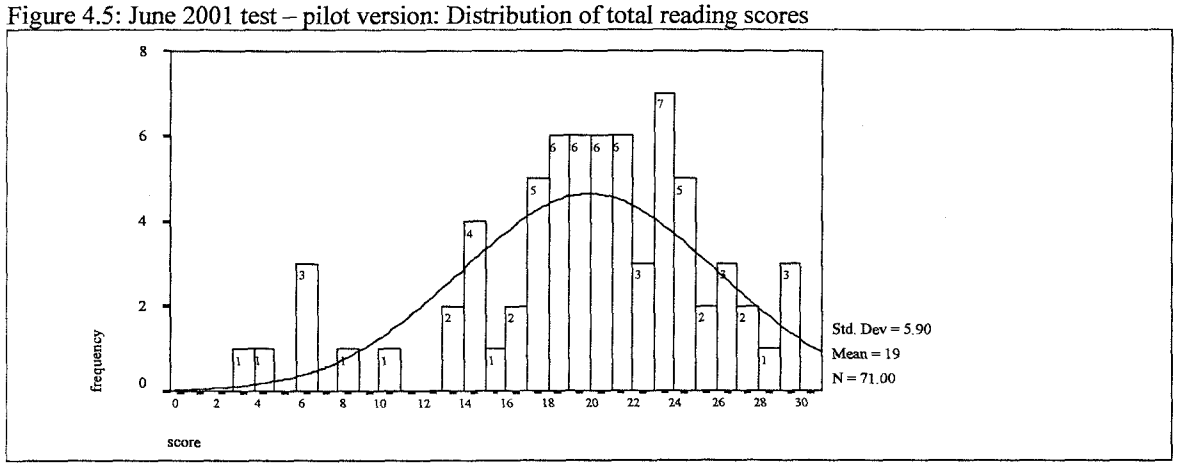
The June-version pilot test had 11 scanning, 9 search and 9 careful reading items so there were 1 scanning, 2 search and 2 careful reading items extra. Table 4.53 shows that the mean is higher than the cut-off (60% = 17.40). The distribution is normal (See Figure 4.5 below).[49] Alpha coefficient (0.87) is satisfactory for the whole reading test.

Table 4.48: June 2001 test – pilot version: Descriptive statistics of total reading scores

N	Item n.	Range	Min.	Max.	Mean	SE	Std.	Skewness	Kurtosis	Alpha
71	29	26.00	3.00	29.00	19.28 (66.48%)	0.70	5.91	-0.848	0.724	0.87

[48] See section 3.5.5 for the explanations.
[49] See Appendix 4.28 for the normality tests and graphs.

Figure 4.5: June 2001 test – pilot version: Distribution of total reading scores



As for the individual subtests (See Table 4.49 below), the mean for the scanning part is again too high (80.8%), and the distribution is highly skewed.[50] The mean for the search and careful reading parts are almost the same and close to the cut-off point; 57.6% and 57.9% respectively. The score distribution in these tests are normal. The reliability coefficients for the subtests are lower than the overall alpha (0.87) and the scanning part has the highest alpha (0.78), obviously because of the fewer number of the items included in each calculation.[51]

Table 4.49: June 2001 test – pilot version: Descriptive statistics of subtests

Substest	N	Item n	Range	Min	Max	Mean	SE	Std	Skewness	Kurtosis	Alpha
Scanning	71	11	8	3	11	8.89 (80.8%)	0.27	2.30	-1.300	0.886	0.78
Search R.	71	9	9	0	9	5.18 (57.6%)	0.27	2.28	-0.375	-0.421	0.70
Careful R.	71	9	9	0	9	5.21 (57.9%)	0.30	2.52	-0.218	-0.972	0.75

[50] See Appendix 4.29 for the score distribution graphs and Appendix 4.30 for the normality tests and graphs of the subtests.

[51] For example, the alpha would be 0.82 if the search and careful reading parts were combined.

4.5.5.2 The June 2001 Test – Pilot Version: Item Analysis

The calculations for the item analysis were done following the method explained in section 3.5 and the results are given in Table 4.50 below. The values that are not within the acceptable limits are given in boldface.

Table 4.50: June 2001 test – pilot version: Item analysis statistics

Item	IF	ID	CITC	CITC subtest	AIID*	AIID* subtest	Item	IF	ID	CITC	CITC Subtest	AIID*	AIID* subtest
SC1	**0.96**	0.43	0.47	0.48	0.8655	0.7621	**SR5**	0.41	0.52	0.41	0.45	0.8646	0.6565
SC2	**0.89**	0.43	0.44	0.31	0.8645	0.7713	**SR6**	0.68	0.46	0.20	**0.18**	**0.8703**	**0.7103**
SC3	**0.97**	**0.14**	0.27	0.34	0.8680	0.7720	**SR7**	0.58	0.77	0.46	0.46	0.8631	0.6547
SC4	**0.97**	**0.29**	0.39	0.38	0.8669	0.7700	**SR8**	0.75	0.72	0.46	0.48	0.8633	0.6537
SC5	0.75	**0.39**	0.20	**0.10**	**0.8698**	**0.8024**	**SR9**	**0.89**	0.55	0.42	0.41	0.8648	0.6731
SC6	0.59	0.63	0.40	0.42	0.8648	0.7640	**CR1**	0.68	0.51	0.31	0.28	0.8672	**0.7554**
SC7	0.65	0.73	0.34	0.44	0.8666	0.7605	**CR2**	0.63	0.63	0.34	0.33	0.8664	0.7476
SC8	**0.82**	0.50	0.45	0.51	0.8637	0.7489	**CR3**	0.52	0.56	0.39	0.39	0.8651	0.7383
SC9	**0.83**	0.81	0.56	0.73	0.8613	0.7221	**CR4**	0.42	0.45	0.39	0.49	0.8652	0.7217
SC10	0.79	0.93	0.57	0.65	0.8606	0.7302	**CR5**	0.51	0.68	0.49	0.46	0.8621	0.7265
SC11	0.68	0.84	0.54	0.63	0.8608	0.7307	**CR6**	0.61	0.66	0.56	0.56	0.8603	0.7096
SR1	0.47	**0.36**	0.22	0.23	**0.8701**	**0.7022**	**CR7**	**0.83**	0.81	0.52	0.41	0.8622	0.7359
SR2	0.66	0.63	0.38	0.39	0.8653	0.6703	**CR8**	0.55	0.54	0.40	0.35	0.8649	0.7452
SR3	0.34	0.52	0.51	0.34	0.8617	0.6792	**CR9**	0.46	0.61	0.48	0.63	0.8624	0.6969
SR4	0.42	0.57	0.39	0.46	0.8652	0.6547							

SC: scanning **SR**: search reading **CR**: careful reading **IF**: item facility **ID**: item discrimination **CITC**: corrected item-total correlation **AIID:** alpha if item deleted ***Alpha; overall:** 0.8688 **SC:** 0.7763 **SR:** 0.6990 **CR:** 0.7541

As was previously the case, the scanning items have high IF values in general, and in particular IFs of SC1-4 and SC8-9 are above the limit. ID values of SC3-5 are especially low. SC5 does not correlate with its subtest sufficiently and has negative effect on the alpha of both the whole and the subtest. In the search reading part, SR1, SR6 and SR9 are problematic. SR1 has a low ID and has negative impact on the alpha of both the whole and the subtest. SR6 has low correlation with its subtest and similar to SR6, its contribution to the alpha is negative. SR9 has a high IF value but otherwise, it performs well. Among the careful reading questions CR1 and CR7 are problematic. When CR1 is dropped, the alpha of the careful reading part increases slightly and CR7 is a relatively easy item with an IF of 0.83.

The next step in the item analysis was to extract the discrimination patterns (IDPs) of the items by assigning the test takers into six performance groups (bands) and compare the item facility values across these bands. Table 4.51 gives the details of the band score distribution. It is immediately seen that when 17 is taken as the lowest passing score (60% = 17.4) the majority (77.4%) of the test takers are assigned to passing bands and a small group of 22.6% is in the failing bands. This would obviously cause problems in the analysis since percentage calculations might be artificially high or low when there are too few test takers assigned to a band.

Table 4.51: June 2001 test – pilot version: Distribution of total reading score by band

band	range of scores	total score mean	no of test takers	pass/fail	percent	cumulative percent
1	0-8	5.50	6	fail	8.5%	8.5%
2	9-13	12.00	3	fail	4.2%	12.7%
3	14-16	14.71	7	fail	9.9%	22.5%
4	17-18	17.55	11	pass	15.5%	38.0%
5	19-23	20.96	28	pass	39.4%	77.5%
6	24-29	26.06	29	pass	22.5%	100.0%

Looking at the item discrimination patterns in Table 4.52 below and band score graphs in Appendix 4.31, it can be seen that many items have erratic patterns resulting from the unbalanced distribution of the scores to the bands. Therefore, IDPs were taken into consideration as secondary information.

Table 4.52: June 2001 test – pilot version: Item discrimination patterns by band

band	SC1	SC2	SC3	SC4	SC5	SC6	SC7	SC8	SC9	SC10	SC11	SR1	SR2
1	0.50	0.50	0.83	0.67	0.50	0.00	0.00	0.33	0.17	0.00	0.00	0.00	0.17
2	1.00	0.67	1.00	1.00	0.33	0.67	0.33	0.33	0.33	0.67	0.33	0.33	0.33
3	1.00	0.57	0.86	1.00	0.71	0.29	0.57	0.57	0.71	0.43	0.29	0.43	0.43
4	1.00	0.91	1.00	1.00	0.73	0.36	0.82	0.91	0.91	0.91	0.73	0.64	0.73
5	1.00	1.00	1.00	1.00	0.79	0.71	0.68	0.89	0.93	0.93	0.75	0.43	0.71
6	1.00	1.00	1.00	1.00	0.88	0.88	0.81	1.00	1.00	0.94	1.00	0.63	0.88

band	SR3	SR4	SR5	SR6	SR7	SR8	SR9	CR1	CR2	CR3	CR4	CR5	CR6
1	0.00	0.00	0.00	0.33	0.00	0.00	0.33	0.33	0.17	0.17	0.17	0.00	0.17
2	0.00	0.33	0.33	0.67	0.00	1.00	1.00	0.67	0.00	0.00	0.00	0.00	0.00
3	0.14	0.14	0.14	0.71	0.60	0.57	0.86	0.71	0.57	0.29	0.29	0.29	0.29
4	0.00	0.27	0.36	0.55	0.30	0.73	0.82	0.36	0.55	0.27	0.09	0.36	0.45
5	0.36	0.46	0.36	0.68	0.70	0.82	0.96	0.68	0.75	0.64	0.43	0.57	0.68
6	0.81	0.75	0.81	0.88	0.90	0.94	1.00	1.00	0.81	0.81	0.88	0.88	1.00

band	CR7	CR8	CR9
1	0.17	0.00	0.00
2	0.33	0.33	0.00
3	0.86	0.29	0.14
4	0.82	0.55	0.45
5	0.96	0.61	0.46
6	0.94	0.81	0.88

4.5.5.3 The June Test – Pilot Version: Evaluation of the Items

The test writers took into consideration the statistical findings given above together with the analysis of the responses given by the test takers. They incorporated their subjective judgements as to the value of the items in the test considering the information contained in them, etc. As a result, SC5, SR1, SR6, CR1 and CR4 were eliminated from the test.

4.5.6 The June 2001 Test

After the analyses explained in the previous section were completed and the test was reduced to its purged version, it was administered to the group of students who were

graduating from Boğaziçi University, School of Foreign Languages prep year in June 2001. The data from this group excluding postgraduate students' were subjected to the analyses described in section 3.5.

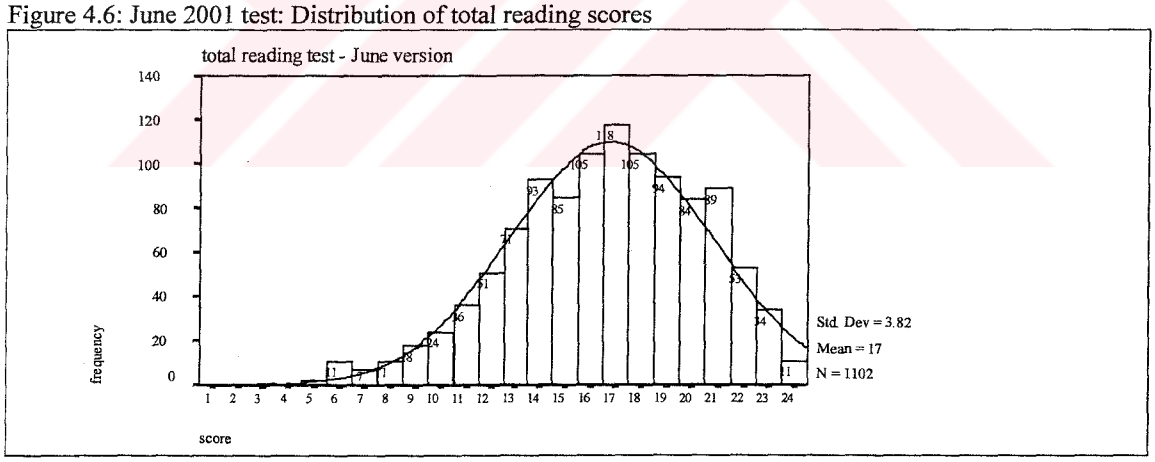
4.5.6.1 The June 2001 Test: Descriptive Statistics

Table 4.53 shows the descriptive statistics of the reading test administered in June. The overall mean of the reading test is slightly higher than that of the pilot version and above the cut-off point (60% = 14.4), however the alpha is lower (0.77 as opposed to 0.87). The distribution of scores is slightly negatively skewed (see Figure 4.6).[52]

Table 4.53: June 2001 test: Descriptive statistics of total reading scores

N	Item n	Range	Min	Max	Mean	SE	Std	Skewness	Kurtosis	Alpha
1102	24	19	5	24	16.56 (69%)	0.12	3.82	-0.401	-0.201	0.77

Figure 4.6: June 2001 test: Distribution of total reading scores



The score distribution in the subtests is as in Table 4.54.

[52] See Appendix 4.32 for the normality tests and graphs.

Table 4.54: June 2001 test: Descriptive statistics of subtests

Subtest	N	Item n	Range	Min	Max	Mean	SE	Std	Skewness	Kurtosis	Alpha
Scanning	1102	10	7	3	10	8.94 (89.4%)	0.04	1.37	-1.471	1.874	0.57
Search R.	1102	7	7	0	7	3.43 (49%)	0.05	1.88	0.082	-0.864	0.66
Careful R.	1102	7	7	0	7	4.20 (60%)	0.05	1.68	-0.277	-0.576	0.58

The scanning test has the highest mean (89.4%), and the score distributions in the test has a sharp negative skeweness.[53] The search reading test had the lowest (49%) mean. The score distribution in the test is slightly positively skewed. The careful reading test gives an ideal mean of 60%. The distribution of the scores in the careful reading test is slightly negatively skewed. It can be observed that there is a dramatic decrease in the alpha coefficients as compared to the ones obtained in the pilot version. The reason might be that fewer items were included in the subtests of the actual version of the June test.[54] Here the alpha coefficient of the scanning test is 0.57 (compared to 0.78), that of search reading is 0.66 (compared to 0.70), and the careful reading test reveals 0.58 (compared to 0.75).

[53] See Appendix 4.33 for the score distribution graphs and Appendix 4.34 for the normality tests and graphs of the subtests.

[54] For example, the alpha of combined search and careful reading subtests would be 0.75.

4.5.6.2 The June 2001 Test: Item Analysis

Item analysis for the June test was performed as explained in section 3.5. The results are given in Table 4.55. The values that are not within the acceptable limits are given in boldface.

Table 4.55: June 2001 test: Item analysis statistics

Item	IF	ID	CITC	CITC subtest	AIID*	AIID* subtest	Item	IF	ID	CITC	CITC subtest	AIID*	AIID* subtest
SC1	**0.98**	**0.12**	**0.08**	**0.10**	**0.7685**	**0.5750**	**SR3**	0.70	0.80	0.39	0.37	0.7538	0.6244
SC2	**0.85**	**0.35**	**0.19**	**0.19**	0.7661	0.5694	**SR4**	0.35	0.59	0.42	0.40	0.7519	0.6151
SC3	**0.98**	**0.12**	**0.06**	**0.10**	**0.7690**	**0.5756**	**SR5**	0.20	**0.38**	0.35	0.37	0.7569	0.6276
SC4	**0.98**	**0.10**	**0.11**	**0.16**	**0.7680**	0.5699	**SR6**	0.45	0.76	0.48	0.46	0.7467	0.5969
SC5	**0.87**	0.60	0.26	0.30	0.7622	0.5342	**SR7**	0.57	0.78	0.45	0.43	0.7492	0.6069
SC6	0.76	0.71	0.30	0.27	0.7600	0.5476	**CR1**	0.28	**0.03**	0.29	0.28	0.7610	0.5480
SC7	**0.88**	0.53	0.25	0.29	0.7627	0.5392	**CR2**	0.78	0.32	0.33	0.32	0.7584	0.5380
SC8	**0.88**	0.54	0.27	0.38	0.7619	0.5113	**CR3**	0.70	0.64	0.33	0.31	0.7580	0.5409
SC9	**0.87**	0.54	0.27	0.31	0.7618	0.5328	**CR4**	0.57	0.81	0.39	0.37	0.7541	0.5167
SC10	**0.88**	0.81	0.33	0.42	0.7590	0.4985	**CR5**	0.36	0.56	0.33	0.27	0.7581	0.5538
SR1	0.42	0.60	0.35	0.35	0.7571	0.6310	**CR6**	0.66	0.67	0.28	0.31	0.7618	0.5401
SR2	0.73	0.55	0.23	0.21	0.7647	**0.6680**	**CR7**	**0.85**	0.64	0.27	0.24	0.7620	0.5625

SC: scanning **SR**: search reading **CR**: careful reading **IF**: item facility **ID**: item discrimination **CITC**: corrected item-total correlation **AIID:** alpha if item deleted ***Alpha; overall:** 0.7676 **SC:** 0.5733 **SR:** 0.6607 **CR:** 0.5813

It can be seen in Table 4.55 that the scanning items persist to be easy; except for SC6, all IF values are above 0.80. The first four scanning items are the most problematic ones since they have unfavourable ID and CITC values. SC1 and SC3 have negative effects on both the overall and the scanning subtest alpha. The overall alpha increases when SC4 is dropped.

In the search reading part, SR2 decreases the subtest alpha and SR5 has slightly low ID. CR1 in the careful reading section has a very low ID, almost zero, showing its inability to discriminate among the low- and high-scoring test takers. CR7 is a relatively easy item but does not have any other unacceptable performance.

As the last check of items' discriminability characteristics, item discrimination patterns were analysed. Table 4.56 shows how the test takers were assigned to the groups and the range of scores in each band.

Table 4.56: June 2001 test: Distribution of total reading score by band

band	range of scores	total score mean	no of test takers	pass/fail	percent	cumulative percent
1	0-6	5.85	13	fail	1.2 %	1.2%
2	7-11	9.74	96	fail	8.7%	9.9%
3	12-13	12.58	122	fail	11.1%	21.0%
4	14-15	14.48	178	pass	16.2%	37.2%
5	16-19	17.45	421	pass	38.3%	75.5%
6	20-24	21.26	270	pass	24.5%	100.0%

Although the majority of the test takers (79%) were assigned to passing groups (e.g. the bands 4-6), since the data was large enough, there were sufficient number of candidates in each band. Table 4.57 gives the item discrimination patterns in six bands and Appendix 4.35 presents the band score graphs.

Table 4.57: June 2001 test: Item discrimination patterns by band

band	SC1	SC2	SC3	SC4	SC5	SC6	SC7	SC8	SC9	SC10	SR1	SR2	SR3
1	0.77	0.38	1.00	0.85	0.38	0.00	0.23	0.31	0.38	0.23	0.08	0.23	0.00
2	0.96	0.66	0.92	0.96	0.55	0.43	0.57	0.59	0.56	0.55	0.14	0.43	0.22
3	0.98	0.70	0.98	0.98	0.80	0.63	0.83	0.80	0.82	0.76	0.16	0.58	0.40
4	0.97	0.81	0.99	0.99	0.85	0.65	0.89	0.87	0.83	0.87	0.25	0.65	0.66
5	0.99	0.90	0.98	0.99	0.92	0.84	0.94	0.93	0.91	0.93	0.41	0.78	0.77
6	0.99	0.94	0.99	1.00	0.97	0.95	0.96	0.97	0.99	0.99	0.79	0.90	0.94

band	SR4	SR5	SR6	SR7	CR1	CR2	CR3	CR4	CR5	CR6	CR7
1	0.00	0.00	0.00	0.08	0.00	0.23	0.23	0.00	0.00	0.23	0.23
2	0.03	0.00	0.03	0.11	0.06	0.47	0.35	0.18	0.04	0.32	0.60
3	0.11	0.04	0.12	0.20	0.08	0.53	0.43	0.26	0.13	0.51	0.74
4	0.13	0.06	0.22	0.40	0.15	0.66	0.62	0.46	0.20	0.51	0.80
5	0.36	0.16	0.50	0.65	0.27	0.86	0.76	0.61	0.37	0.71	0.89
6	0.74	0.52	0.85	0.91	0.56	0.96	0.93	0.88	0.67	0.89	0.98

It is seen in Table 4.57 that SC1, SC3 and SC4 do not have favourable IDPs. SR5 cannot discriminate between the first five bands and only 52% of the candidates in band6 could respond to the item correctly. The same problem exists in CR1 and CR5, too.

4.5.6.3 The June Test: Evaluation of the Items

As compared to previous tests (i.e. the September and January tests), the June results suggest a higher performance on the part of the test takers. Except for the search reading mean, both the overall and the subtest mean scores are higher than the ones in the previous versions. Concerning the scanning part especially, not only the first half but the whole test could be completed with significant success. All the scanning items have mean values (IF) above 0.80. This is an expected result since this group of students had extended training in the skills tested in the exam. Otherwise, no serious problems were observed with the items. There is a proper balance among the difficult and easy items in the rest of the test; search reading items ranging from 0.20 to 0.73 and careful reading items, from 0.28 to 0.85 in terms of item facility.

Scanning items were assigned 0.5 point in the actual scoring of the test as it was in the previous versions and no further adjustments were done.

4.5.6.4 The June 2001 Test: Inter-correlations and PCA

The June 2001 Test: Inter-correlations: Table 4.58 shows the correlations between the subtests in the June 2001 reading test.

Table 4.58: June 2001 test: Subtest inter-correlations

	Search reading	Careful reading
Scanning	.363*	.275*
Search reading	-	.509*

*: Correlation is significant at the 0.05 level (2- tailed).

It can be seen that there is moderate correlation between the search and careful reading tests (.509) but the correlations between the scanning and search reading, and scanning and careful reading tests are quite low (.363 and .275 respectively). The lowest correlation of all is between the scanning and careful reading tests as it was the case in the previous tests, too. It can be concluded that the three subtests do not overlap to a degree that might lead us to assume that these tests measure the same skill. On the contrary, there is evidence that they test different skills. However, it must be noted that the overlap between search and careful reading is larger than that between scanning and the others. Scanning has an extraordinarily low correlation with careful reading.

The June 2001 Test: Principal Component Analysis: The June 2001 reading test has 24-item data (10 scanning, 7 search and 7 careful reading items) and the whole test was submitted to PCA analysis without constraining the number of components to be extracted. Eigenvalue-one criterion is applied in this data, too. KMO measure of sampling adequacy was .852 and Bartlett's test of sphericity was significant at

.000 level. Except for two items (CR3 and CR5), no communalities below .30 were observed.[55] 6 components with eigenvalues above 1.00 were extracted and these accounted for 41.39%of total variance in the data. The rotated component matrix is given in Table 4.59 in which the highest loadings of the items to factors are given in boldface.

Table 4.59: Rotated component matrix: June 2001 test – whole set

	Component					
	1	2	3	4	5	6
SC1	,106	,037	-,110	,105	,101	**,799**
SC2	-,079	,023	,281	**,639**	-,048	,210
SC3	**,139**	,104	-,080	,003	-,699	-,034
SC4	,048	,102	-,272	**,575**	,028	-,034
SC5	,218	**,436**	-,071	,109	-,186	,294
SC6	,107	,246	,193	**,442**	,002	,040
SC7	,061	**,525**	,068	,101	,112	-,066
SC8	,016	**,660**	,095	,071	-,049	,081
SC9	,126	**,609**	,093	-,074	-,158	-,149
SC10	,071	**,688**	,040	,169	,089	,065
SR1	**,590**	,119	-,144	,019	,208	-,009
SR2	,297	,128	-,220	**,373**	,333	-,382
SR3	**,408**	,087	,164	,338	-,118	-,187
SR4	**,584**	,036	,144	,056	,065	-,083
SR5	**,583**	,015	,040	,011	-,122	,149
SR6	**,584**	,105	,216	,142	-,141	,069
SR7	**,531**	,039	,229	,262	-,200	-,018
CR1	,312	,077	,232	-,034	**,478**	,035
CR2	,337	-,037	**,393**	,057	,103	,162
CR3	**,398**	,097	,226	-,048	,207	,114
CR4	,322	,060	**,516**	,061	,020	-,160
CR5	**,465**	,132	,122	-,095	,069	-,017
CR6	,102	,050	**,603**	,092	,096	-,099
CR7	,095	,181	**,518**	-,027	,036	,025

In Table 4.59, except for SR2, all the search reading items load on F1, scanning items in the second part of the test load on F2 and most of the careful reading items load on F3. Three scanning items (SC2, SC4 and SC6) load on F4. There is only one item loading on F5 and F6 each (CR1 and SC1 respectively). Subtest-factor correlations for the data are given in Table 4.60.

[55] See Appendix 4.36 for details.

Table 4.60: June 2001 test – whole set: Subtest – Factor correlations

Factors	Eigen value	% of variance (cum: 41.4%)	SC I (SC 1-5)	SC II (SC 6-10)	Scanning	Search R.	Careful R.
F1	3.853	16.056	.142*	.130*	.161*	**.892***	**.553***
F2	1.665	6.937	**.311***	**.878***	**.809***	.133*	.146*
F3	1.189	4.952	.034	.171*	.146*	.113*	**.689***
F4	1.103	4.595	**.574***	.261*	**.451***	.299*	.004
F5	1.081	4.505	-.271*	-.002	-.120*	.005*	.270*
F6	1.043	4.344	**.467***	-.006	.199*	-.114*	.006

According to the correlations, F1 seems to account for search reading because it correlates with the subtest with a high correlation of .892. However, F1 also correlates with careful reading with a moderate correlation of .553. F2 correlates with scanning (.809) but more specifically with its second part (SC II: SC5-10), since its correlation with SC II is .878 yet with the first part of the scanning test (SC I: SC1-5), it is only .311. F3 has a correlation of .689 with careful reading. F4 correlates with scanning by r= .451 and with SC I, by r= .574. It also has a correlation with search reading just below the limit (.299). F5 does not correlate with any of the subtests within the limit of .30 and F6 has moderate correlation with SC I (.467). It can be observed that SC I loads on several factors, not one more dominant than the other. This is a phenomenon that was observed in the previous tests as well. The first few items of the scanning test always seem to be problematic. SC II, on the other hand, clearly emerges as a factor in the data as was the case with the previous tests. Search reading loads on one factor heavily (F1) but it has a moderate loading on another factor (F4) too. Careful reading, on the other hand, clearly loads on two factors equally (F1 and F3). With this finding at hand, it can be concluded that F1 is a factor that combines search and careful reading subskills (reading at the global level), F2 accounts for SC II, and F3, for careful reading only. F4 and F6 account for the rest of the problematic section of scanning; SCI. F5 is not interpretable.

As the next step of the analysis, the items that were indicated as defective in the item analysis were extracted from the data and the factor structure analysis was repeated. Item analysis of the June test is discussed in section 4.5.6.2 in detail. According to this, SC1-4, SR2, 5 and CR1, CR7 were excluded from the data and PCA was run to give the distribution of the purged data in Table 4.61.[56]

Table 4.61: Rotated component matrix:
June 2001 test – purged set I

	Component		
	1	2	3
SC5	.380	**.446**	-.195
SC6	.292	**.346**	.046
SC7	.020	**.535**	.132
SC8	-.029	**.691**	.111
SC9	.071	**.574**	.129
SC10	.148	**.691**	-.014
SR1	**.556**	.032	.082
SR3	**.503**	.139	.172
SR4	**.522**	.005	.333
SR6	**.669**	.118	.174
SR7	**.620**	.088	.217
CR2	.197	.006	**.545**
CR3	.203	.075	**.487**
CR4	.227	.075	**.575**
CR5	.264	.086	**.392**
CR6	-.062	.116	**.671**

As seen in Table 4.61, the three subtests load on three different factors, scanning on F2, search reading on F1 and careful reading on F3 and their correlation with the matching factors given in Table 4.62 below are quite high: .921, .904, .909 respectively.

Table 4.62: June 2001 test – purged set I: Subtest – Factor correlations

Factors	**Eigen value**	**% of variance (cum: 36.7%)**	**SC II (SC 5-10)**	**Search R. (-SR 2,5)**	**Careful R. (-CR 1,7)**
F1	3.323	20.767	.184*	**.904***	.281*
F2	1.539	9.619	**.921***	.120*	.124*
F3	1.015	6.343	.131*	**.307***	**.909***

[56] See Appendix 4.37 for details.

However, note that F3 also correlates with search reading at a moderate r= .307 level, which is quite expectable, since these two skills correlate as subtests with each other more than they do with scanning.

On the whole, this finding is a substantial support for the differential performance of the subtests and therefore for differential existence of reading subskills as they are measured in the reading test of the BUEPT. Nevertheless, it would still be feasible to analyse the internal structures of the subtests both to verify the assumptions that scanning items performed differentially according to their place in the test, and that unless there were certain problematic items included, data from single-text search and careful reading tests would be uni-componential. Therefore, the individual subtests of the June test were subjected to PCA to give the component structures in Table 4.63.[57]

Table 4.63: Rotated component matrices: June 2001 test – subtests

	Component		
	1	2	3
SC1	-.225	**.557**	.433
SC2	.104	**.660**	-.129
SC3	.039	-.191	**.775**
SC4	.103	**.456**	.020
SC5	.327	.200	**.530**
SC6	.348	**.389**	.009
SC7	**.571**	.093	-.076
SC8	**.628**	.146	.093
SC9	**.657**	-.123	.117
SC1	**.628**	.224	.150

	Comp.
	1
SR1	.549
SR2	,344
SR3	,569
SR4	,611
SR5	,576
SR6	,685
SR7	,650

	Component	
	1	2
SR1	**.549**	.214
SR2	.344	**.818**
SR3	**.569**	.219
SR4	**.611**	.008
SR5	**.576**	-.303
SR6	**.685**	-.309
SR7	**.650**	-.219

	Comp.
	1
CR1	.514
CR2	.557
CR3	.543
CR4	.620
CR5	.495
CR6	.548
CR7	.447

	Component	
	1	2
CR1	**.514**	-.209
CR2	**.557**	-.244
CR3	**.543**	-.376
CR4	**.620**	.126
CR5	**.495**	-.289
CR6	**.548**	.380
CR7	.447	**.678**

As expected, scanning appeared multi-componential with a rather uniform second part.[58] The items in the first part of the scanning test load on two factors; F2 and F3 supporting the subtest-factor correlation results given in Table 4.60. For search and

[57] See Appendix 4.38 for details.
[58] Since SC6 has almost equal loading both on F1 and F2, it can still be considered as a part of SC II.

careful reading, PCA extracted single components. These tests seem to be unidimensional in themselves. This is also an important finding since it supports the assumption that the multicomponential nature of the subtests in September 2000 and January 2001 might be due to differential text effect as well. However, since the communalities for SR2 and CR7 were too low, it was worth performing the analysis again by increasing the number of factors to be extracted to 2, and checking whether there are any items loading on a second factor. In bi-componential factor matrices of the search and careful reading tests, SR2 and CR7, which were also designated as problematic items by the item analysis, loaded on a second factor.[59] When these items were removed from the data together with the first four scanning items to form purged set II, the PCA gave the expected factor-item distribution in Table 4.64, once again lending support for the multicomponential nature of the test. [60]

Table 4.64: Rotated component matrix:
June 2001 test – purged set II

	Component		
	1	2	3
SC5	.286	**.486**	-.136
SC6	.170	**.382**	.149
SC7	.030	**.522**	.110
SC8	.012	**.668**	.050
SC9	.050	**.578**	.107
SC10	.089	**.699**	.025
SR1	**.560**	.057	.077
SR3	**.446**	.184	.187
SR4	**.503**	.040	.329
SR5	**.657**	.014	.023
SR6	**.638**	.160	.192
SR7	**.586**	.137	.206
CR1	.107	.039	**.537**
CR2	.198	.016	**.525**
CR3	.231	.074	**.450**
CR4	.197	.093	**.576**
CR5	.319	.085	**.326**
CR6	-.043	.100	**.645**

[59] See Table 4.63 and Appendix 4.38 for details.
[60] See Appendix 4.39 for details.

Therefore, 6-component initial structure matrix was reduced to 3-component one in which each factor accounted for a different reading skill. The subtest-factor correlations of the purged set II are given Table 4.65 in which it can be seen that F1 accounts primarily for search reading (r= .895) and secondarily for careful reading (r= .366); F2 accounts for scanning II (r= .922) and F3 for careful reading (r= .884).

Table 4.65: June 2001 test – purged set II: Subtest – Factor correlations

Factors	Eigen value	% of variance (cum: 33.7%)	SC II (SC 6-10)	Search R. (-SR 2)	Careful R. (-CR 7)
F1	3.707	19.510	.128*	**.895***	**.366***
F2	1.585	8.340	**.922***	.186*	.121*
F3	1.111	5.847	.166*	.240*	**.884***

*: Correlation is significant at the 0.05 level (2- tailed).

4.5.6.5 The June 2001 Test: Discussion

As seen in the above discussion, the reduction of external factors such as defective item characteristics and text effect led to more interpretable solutions supporting the assumptions raised in the study. The June 2001 test displayed a neater component structure congruent with multicomponential operationalisation of the reading skill in the test. However, it is preferable to have repeated evidence for the findings attained. Therefore, the September 2001 test, which has the same test structure as the June test, will also be discussed to assess whether a second test with a similar reduced format will yield the same characteristics or not.

4.5.7 The September 2001 Test – Pilot Version

As explained in section 3.5.7, in the absence of an equivalent group, the pilot testing of the September 2001 test was done on the freshman university students who had almost completed a year's study in their departments. Despite the fact that such pilot testing would not be as accurately informative as the one done on a truly equivalent

sample, especially in terms of alternative answers and item difficulty, the findings would still be suggestive. Therefore, the following analyses were made on the data from the freshman group.

4.5.7.1 The September 2001 Test – Pilot Version: Descriptive Statistics

The September test pilot version had only one scanning item extra. Otherwise, there were seven questions in the search and careful reading tests each. The extra items written for these sections were found defective by the test reviewers (See section 3.2.2) and were not included in the test. Therefore, there were 25 items tested in the September 2001 pilot version. The descriptive statistics are given in Table 4.66 below.

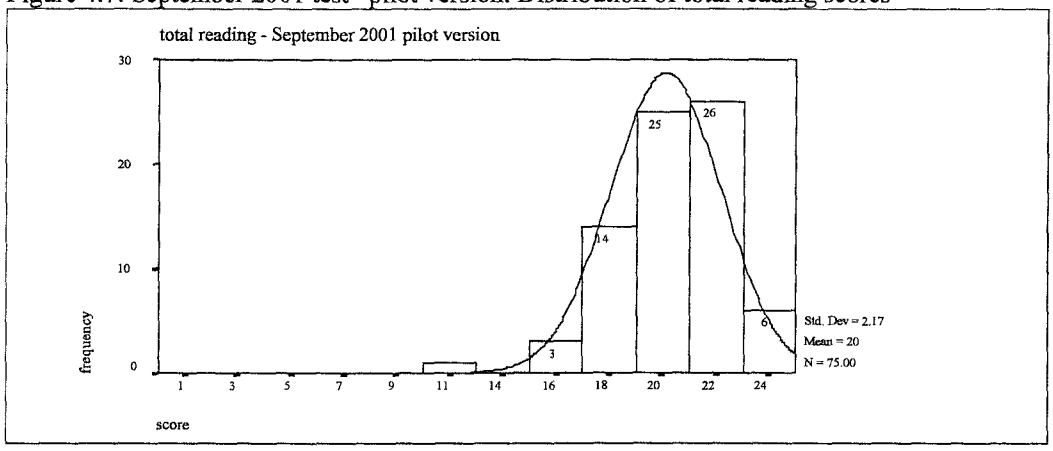
Table 4.66: September 2001 test – pilot version: Descriptive statistics of total reading scores

N	Item n	Range	Min	Max	Mean	SE	Std	Skewness	Kurtosis	Alpha
75	25	12	12	24	19.9 (82.9%)	0.25	2.17	-.810	1.477	0.30

As seen in the table, the mean is very high and the distribution of the scores is negatively skewed as expected (See Figure 4.7)[61] The coefficient of reliability for the test is below the acceptable limit (0.30). However, since the sample group was a rather homogeneous group with a higher level of English language proficiency than the group the test was geared at, this was again an expected result.

[61] See Appendix 4.40 for the normality tests and graphs.

Figure 4.7: September 2001 test– pilot version: Distribution of total reading scores



The distribution of the scores in the subtests of the September 2001 pilot test was as in Table 4.67 below.

Table 4.67: September 2001 test – pilot version: Descriptive statistics of subtests

Substest	N	Item n	Range	Min	Max	Mean	SE	Std	Skewness	Kurtosis	Alpha
Scanning	75	11	5	6	10	10.4 (94.6%)	0.13	1.08	-1.987	3.580	0.59
Search R.	75	7	5	2	7	4.6 (65.7%)	1.17	1.49	-.013	-.837	0.44
Careful R.	75	7	5	2	7	4.8 (68.6%)	0.18	1.52	-.246	-1.001	0.45

The mean for the scanning test was again too high; the test takers in the sample group were able to complete the scanning test with almost complete success. The distribution for the scanning test was overly peaked.[62] The alpha however, was higher than that of the whole test (0.59). For the search reading test, the mean was slightly higher than the cut-off (65.7%) and the distribution was near normal. The alpha of the search reading test was again low (0.44) but higher than the alpha of the whole test. The mean of the careful reading was 68.6%, higher than the cut-off and the distribution of the scores was rather flat. The alpha coefficient for the careful reading test (0.45) was similar to that of the search reading, again lower than expected. It seemed that the test on the whole for this group was easy. Especially, the

[62] See Appendix 4.41 for the score distribution graphs and Appendix 4.42 for the normality tests and graphs of the subtests.

extreme easiness of the scanning test contributed negatively to the reliability of the test.

4.5.7.2 The September 2001 Test – Pilot Version: Item Analysis

Despite the problems observed above, for further information on the test performance, item analysis was made on the data. Table 4.68 below provides the findings. The values that are not within the acceptable limits are given in boldface.

Table 4.68: September 2001 test – pilot version: Item analysis statistics

Item	IF	ID	CITC	CITC subtest	AIID*	AIID* subtest	Item	IF	ID	CITC	CITC subtest	AIID*	AIID* subtest
SC1	**0.93**	**0.00**	**-0.08**	0.24	0.3209	0.5657	SR3	**0.81**	0.61	0.20	0.21	0.2567	0.3952
SC2	**0.99**	**0.00**	**0.10**	**0.16**	0.2959	0.5805	SR4	**0.84**	1.00	0.38	0.25	0.2103	0.3800
SC3	**0.95**	**0.33**	**0.19**	0.30	0.2762	0.5521	SR5	0.79	**0.27**	**0.09**	0.21	0.2889	0.3968
SC4	**0.97**	**0.33**	**0.19**	**0.07**	0.2836	**0.5941**	SR6	0.51	0.65	**-0.01**	0.29	**0.3236**	0.3473
SC5	**0.93**	0.67	0.25	0.30	0.2627	0.5503	SR7	0.41	**0.14**	**-0.10**	**0.18**	**0.3531**	0.4088
SC6	**0.92**	**0.00**	**0.14**	**0.02**	0.2821	**0.6315**	CR1	**0.85**	**0.21**	**-0.08**	**0.14**	**0.3309**	0.4450
SC7	**0.95**	**-0.06**	**-0.03**	0.30	**0.3117**	0.5521	CR2	**0.85**	**0.33**	0.31	0.25	0.2327	0.4026
SC8	**0.97**	**0.00**	**-0.08**	**0.07**	0.3141	**0.5941**	CR3	0.67	0.49	**-0.05**	0.24	**0.3200**	0.4039
SC9	**0.92**	**0.33**	**0.14**	0.55	**0.2821**	0.4736	CR4	0.47	0.49	**0.15**	0.38	0.2682	0.3177
SC10	**0.96**	**0.00**	**0.03**	0.45	0.3023	0.5216	CR5	0.57	0.49	**0.13**	**0.18**	0.2771	0.4346
SC11	**0.93**	**0.27**	**0.10**	0.36	0.2907	0.5340	CR6	0.63	**0.38**	**-0.05**	**0.01**	**0.3346**	**0.5164**
SR1	0.52	0.43	**0.04**	**0.18**	**0.3050**	0.4102	CR7	0.75	0.88	**0.15**	0.32	0.2707	0.3620
SR2	0.80	**0.21**	**-0.06**	**0.10**	**0.3294**	**0.4415**							

SC: scanning **SR**: search reading **CR**: careful reading **IF**: item facility **ID**: item discrimination **CITC**: corrected item-total correlation **AIID:** alpha if item deleted ***Alpha**; **overall:** 0.3027 **SC:** 0.5857 **SR:** 0.4357 **CR:** 0.4541

Table 4.68 shows that all IF values are above 0.50, and especially in the scanning test, they were above 0.90. Since there were no test takers in the lowest 33% percent, the ID values were calculated comparing the groups performing at the lower 50% and the upper 20% of the scores. There were several problems with the item performance in terms of ID values, too. This was taken to be due to the high-performing homogeneous nature of the sample group. CITC and AIID problems should also be stemming from the same problem.

The band score distribution in Table 4.69 below also shows that there were no test takers in the lowest two bands (bands 1 and 2) and the majority was piled at band5.

Table 4.69: September 2001 test – pilot version: Distribution of total reading score by band

band	range of scores	total score mean	no of candidates	pass/fail	percent	cumulative percent
3	12-14	12	1	fail	1.3	1.3
4	15-17	16.4	8	pass	10.7	12
5	18-21	19.7	49	pass	65.3	77.3
6	22-25	22.5	17	pass	22.7	100

The distribution of item discrimination scores to bands is given in Table 4.70.[63] This analysis suggested that besides the scanning items, SR2, SR5, SR7, CR4 need further revision.

Table 4.70: September 2001 test – pilot version: Item discrimination patterns by band

band	SC1	SC2	SC3	SC4	SC5	SC6	SC7	SC8	SC9	SC10	SC11	SR1	SR2
3	1.00	1.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00
4	1.00	0.88	1.00	1.00	0.88	0.63	1.00	1.00	0.75	0.90	0.75	0.50	0.75
5	0.90	1.00	0.94	0.98	0.94	0.94	0.94	0.96	0.92	1.00	0.96	0.45	0.78
6	1.00	1.00	1.00	1.00	1.00	1.00	0.94	1.00	1.00	1.00	0.94	0.76	0.88
band	SR3	SR4	SR5	SR6	SR7	CR1	CR2	CR3	CR4	CR5	CR6	CR7	
3	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	
4	0.63	0.25	0.63	0.25	0.50	0.60	0.50	0.63	0.13	0.25	0.63	0.38	
5	0.82	0.90	0.76	0.51	0.39	0.90	0.88	0.63	0.39	0.55	0.61	0.78	
6	0.94	1.00	0.94	0.65	0.47	0.90	1.00	0.82	0.82	0.82	0.71	0.88	

4.5.7.3 The September 2001 Test – Pilot Version: Evaluation of the items

In the absence of dependable statistical findings, the testing office members made a qualitative analysis on the responses given by the test takers in the sample group. They listed the responses given by the low and high performing groups to each item in the test (excluding the scanning test), and analysed them to see whether any adjustments to the tests were necessary. Where there were alternatively correct answers, the items were repaired by either tightening the wording of the question or

[63] See Appendix 4.43 for band score graphs.

editing the text. Where there was obscurity in the main ideas discussed in the texts, explanatory information was added. As for the scanning part, SC8 was eliminated from the test but no other adjustments were done on this part.

4.5.8 The September 2001 Test

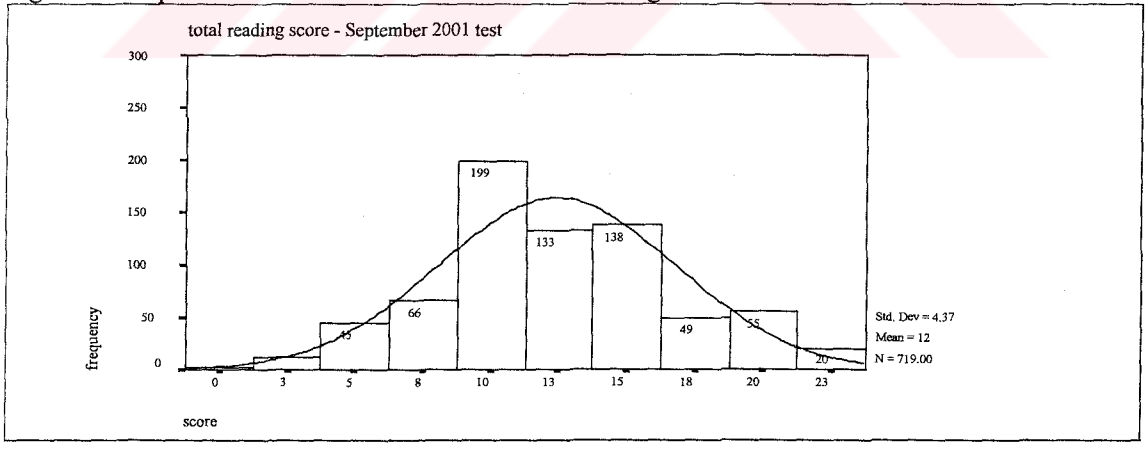
4.5.8.1 The September 2001 Test: Descriptive Statistics

After the necessary adjustments were made, the September 2001 test was administered to the incoming students. As seen in Table 4.71 below, the mean was 51.63% with near normal distribution (See Figure 4.8 below).[64] The alpha coefficient of 0.81 was quite satisfactory.

Table 4.71: September 2001 test: Descriptive statistics of total reading scores

N	Item n	Range	Min	Max	Mean	SE	Std	Skewness	Kurtosis	Alpha
719	24	23	0	1	12.39 (51.63%)	0.16	4.36	.166	-.181	0.81

Figure 4.8: September 2000 test: Distribution of total reading scores



The distribution of the scores in the subtests are given in Table 4.72.

[64] See Appendix 4.44 for the normality tests and graphs.

Table 4.72: September 2001 test: Descriptive statistics of subtests

Substests	N	Item n	Range	Min	Max	Mean	SE	Std	Skewness	Kurtosis	Alpha
Scanning	719	10	10	0	10	8.08 (80.8%)	0.08	2.13	-1.191	.937	0.75
Search R.	719	7	7	0	7	1.89 (27%)	0.08	2.11	.954	-.290	0.81
Careful R.	719	7	7	0	7	2.42 (34.6%)	0.06	1.61	.593	-.243	0.51

The mean of the scanning test was again high (80.8%) and the distribution was negatively peaked.[65] The alpha of 0.75 was satisfactory. The search reading mean (27%), on the other hand, was very low and the score distribution was positively skewed. However, search reading alpha was quite high (0.81). In the careful reading test, the mean was again low (34.6%) and the distribution of the scores was positively skewed. The alpha of the careful reading test was the lowest; 0.51.

4.5.8.2 The September 2001 Test: Item Analysis

The item analysis results of the September 2001 test are given in Table 4.73. The values that are not within acceptable limits are given in boldface.

[65] See Appendix 4.45 for the score distribution graphs and Appendix 4.46 for the normality tests and graphs of the subtests.

Table 4.73: September 2001 test: Item analysis statistics

Item	IF	ID	CITC	CITC Subtest	AIID*	AIID* subtest	Item	IF	ID	CITC	CITC subtest	AIID*	AIID* subtest
SC1	**0.84**	**0.36**	0.26	0.37	0.8055	0.7350	**SR3**	0.34	0.79	0.48	0.57	0.7949	0.7866
SC2	**0.89**	**0.23**	0.22	0.24	0.8068	**0.7499**	**SR4**	0.33	0.87	0.54	0.63	0.7914	0.7736
SC3	0.80	0.46	0.26	0.40	0.8058	0.7306	**SR5**	0.31	0.86	0.53	0.60	0.7922	0.7793
SC4	**0.95**	**0.20**	0.22	0.30	0.8069	0.7441	**SR6**	0.22	0.70	0.49	0.58	0.7952	0.7845
SC5	**0.83**	**0.18**	0.31	0.38	0.8038	0.7338	**SR7**	**0.15**	0.50	0.38	0.41	0.8009	0.8101
SC6	0.66	0.56	0.35	0.41	0.8021	0.7304	**CR1**	0.36	0.67	0.40	0.38	0.7992	0.4037
SC7	**0.88**	0.40	0.35	0.49	0.8024	0.7204	**CR2**	0.39	0.50	0.27	0.24	0.8064	0.4692
SC8	0.74	0.64	0.36	0.50	0.8011	0.7141	**CR3**	0.38	0.66	0.37	0.27	0.8007	0.4544
SC9	0.64	0.71	0.37	0.53	0.8008	0.7085	**CR4**	**0.12**	**0.25**	**0.17**	0.21	**0.8088**	0.4840
SC10	**0.84**	0.48	0.34	0.49	0.8023	0.7182	**CR5**	0.24	**0.20**	**0.09**	**0.12**	**0.8141**	**0.5176**
SR1	0.20	0.59	0.41	0.49	0.7991	0.7985	**CR6**	0.56	0.47	0.21	**0.18**	**0.8099**	0.4985
SR2	0.34	0.84	0.49	0.57	0.7942	0.7868	**CR7**	0.37	0.72	0.41	0.33	0.7985	0.4272

SC: scanning **SR**: search reading **CR**: careful reading **IF**: item facility **ID**: item discrimination **CITC**: corrected item-total correlation **AIID:** alpha if item deleted ***Alpha**; **overall:** 0.8086 **SC:** 0.7495 **SR:** 0.8136 **CR:**0.5058

As seen in Table 4.73, several scanning items had too high IF and too low ID values. In the search reading part, only SR7 had low IF. The remaining items were unproblematic. CR4 in the careful reading test had too low IF and ID, and it had low correlation with the test in general. It had a slight negative impact on the test reliability as well. CR5 also had low ID. Its correlation with the test in general and the careful reading subtest in particular was low too. It had a negative impact both on the subtest and the whole test reliability. CR6 also had low correlation with the subtest and a slight negative impact on test reliability. Otherwise, no problems were observed with the items.

The band score analysis in Table 4.74 shows that 63.6% of the test takers were assigned to failing groups. The majority was in bands 2 and 3. Except for slightly more test takers in band5 as compared to band4, the distribution of the test takers to the bands was neat.

Table 4.74: September 2001 test: Distribution of total reading score by band

band	range of scores	total score mean	no of test takers	pass/fail	percent	cumulative percent
1	1-6	4.6	59	fail	8.2	8.2
2	7-11	9.5	265	fail	36.9	45.1
3	12-13	12.6	133	fail	18.5	63.6
4	14-15	14.5	96	pass	13.4	76.9
5	16-19	17.2	116	pass	16.1	93
6	20-23	21.2	50	pass	7	100

The item discrimination patterns extracted based on this grouping are given in Table 4.75 below.

Table 4.75: September 2001 test: Item discrimination patterns by band

band	SC1	SC2	SC3	SC4	SC5	SC6	SC7	SC8	SC9	SC10	SR1	SR2	SR3
1	.46	.59	.36	.78	.42	.17	.31	.15	.14	.36	.02	.03	.03
2	.79	.85	.75	.92	.76	.52	.86	.64	.49	.78	.06	.13	.12
3	.90	.92	.91	.98	.94	.77	.96	.89	.81	.94	.14	.27	.27
4	.94	.96	.92	1.00	.92	.89	.99	.91	.80	.96	.21	.41	.49
5	.95	.97	.86	.99	.94	.80	.97	.86	.79	.96	.45	.74	.72
6	.98	1.00	.98	1.00	.96	.98	1.00	.98	.96	.98	.76	.94	.90

band	SR4	SR5	SR6	SR7	CR1	CR2	CR3	CR4	CR5	CR6	CR7
1	.00	.00	.02	.02	.05	.10	.08	.02	.22	.24	.03
2	.09	.09	.05	.05	.18	.24	.20	.08	.18	.47	.22
3	.23	.21	.11	.08	.30	.43	.38	.11	.17	.58	.26
4	.46	.42	.23	.13	.43	.57	.49	.09	.25	.59	.46
5	.80	.73	.52	.32	.68	.57	.65	.21	.30	.76	.71
6	.94	.94	.90	.70	.90	.72	.90	.36	.58	.84	.96

It is seen in Table 4.75 above that the scanning items had no discrimination power for higher groups; the majority of the test takers in band3 and above could answer the items correctly. The item patterns concerning the search reading questions were neatly incremental and did not suggest extreme difficulty for higher groups. Among the careful reading items, CR4 and CR5 showed weak discrimination and too few test takers could answer the items correctly even in higher bands.[66]

4.5.8.3 The September 2001 Test: Evaluation of the items

In general, the September 2001 test appeared to be a difficult test for the group of test takers who actually took the test as part of the proficiency test for registering for Boğaziçi University. However, as explained in section 4.5.2, September tests are normally taken by incoming students that include a considerable number of beginner level students. Therefore, the low means especially of the search and careful reading test were not unexpected. When the scores from the reading test were considered in relation with the other parts of the proficiency test (i.e. listening, writing), the

[66] See Appendix 4.47 for the band score graphs.

performance of the BUEPT was deemed to be favourable. However, pooling in the information from all the analyses explained above, it was decided that CR4 should be excluded from the final score calculations. The scanning items were assigned 0.5 points as determined before.

4.5.8.4 The September 2001 Test: Inter-correlations and PCA

The September Test: Inter-correlations: Table 4.76 below shows the inter-correlations of the subtests in the September 2001 reading test. It can be seen that correlations between scanning and the other two tests are lower (with search reading: .249, with careful reading: .231) than the correlation between search and careful reading (.541). In short, these correlations suggest that the three subtests of the September 2001 test do not test exactly the same skill. However, what scanning seems to be testing is 'more different' than what the other two tests are testing since .249 and .231 are too low correlations to suggest any similarity.

Table 4.76: September 2001 test: subtest inter-correlations

	Search reading	**Careful reading**
Scanning	.249*	.231*
Search reading	-	.541*

*: Correlation is significant at the 0.05 level (2- tailed).

The September Test: Principal Component Analysis: The September 2001 test had 24-item data and the whole set was submitted to PCA with varimax rotation without constraining the number of components to be extracted. KMO measure of sampling adequacy was .870 and Bartlett's test of sphericity was significant at .000 level. Except for CR3, no communalities below .30 were observed. Five components with eigenvalues over 1.00 were extracted and these accounted for 44.96% of

variance in the data.[67] Rotated component matrix is given in Table 4.77 in which the highest loadings of the items on the components are marked in bold.

Table 4.77: Rotated component matrix: September 2001 test – whole set

	Component				
	1	2	3	4	5
SC1	,062	,040	**.744**	,029	,107
SC2	,055	-,005	,253	,112	**,538**
SC3	-,041	,279	**,520**	,178	,098
SC4	,106	,168	,215	-,186	**,501**
SC5	,161	,126	**,619**	-,082	,132
SC6	,113	,388	,179	,043	**,460**
SC7	,047	,362	**,417**	,080	,332
SC8	,097	**,813**	,084	,028	,019
SC9	,081	**,685**	,219	,030	,131
SC10	,048	**,818**	,054	,060	,003
SR1	**,616**	-,020	,001	,105	,027
SR2	**,693**	,061	,006	,082	,080
SR3	**,679**	,050	,087	,092	-,031
SR4	**,728**	,086	,131	,120	-,068
SR5	**,711**	,057	,104	,102	,029
SR6	**,708**	,013	,068	,031	,069
SR7	**,539**	,053	-,022	-,029	,202
CR1	,392	,031	-,020	**,463**	,146
CR2	,077	,137	,111	**,616**	-,021
CR3	**,343**	,223	,074	,340	-,079
CR4	,134	-,076	-,161	**,502**	,196
CR5	,108	-,084	-,347	,234	**,459**
CR6	,122	,001	,311	**,443**	-,271
CR7	**,498**	,083	-,087	,267	,154

It can be seen in Table 4.77 that scanning items are dispersed over three components (F2, F3 and F5). However, all the search reading items load on F1 neatly and most of the careful reading items load on F4 (excluding CR5 and 7). Subtest-factor correlations in Table 4.78 below show that F1 accounts for search reading with almost a perfect correlation (r=.975). It also correlates with careful reading with r= .486.

[67] See Appendix 4.48 for details.

Table 4.78: September 2001 test – whole set: Subtest – Factor correlations

Factors	Eigen value	% of variance (cum: 44.96%)	SC I (SC 1-5)	SC II (SC 6-10)	Scanning	Search R.	Careful R.
F1	4.746	19.774	.107*	.116*	.130*	**.975***	**.486***
F2	2.621	10.921	.206*	**.889***	**.703***	.064	.105*
F3	1.238	5.158	**.831***	.264*	**.575***	.082*	-.005
F4	1.118	4.657	.044	.066	.066	.110*	**.807***
F5	1.068	4.449	**.407***	.276*	**.381***	.044	.137*

F2 explains scanning, more specifically the second part of the test (SCII: SC6-10) since its correlation with SCII is .889. F3, on the other hand correlates with SCI (SC1-5) with r= .831. F4 accounts for careful reading with a correlation of .807. F5 is the second factor that correlates with SCI. Therefore, it can be concluded that SCII, search reading and careful reading and another factor that relates to a combination of search and careful reading (reading at the global level) appear as factors in the data. SCI is again not as homogeneous as it would be expected.

The next step was to eliminate the problematic items from the data to see whether PCA yielded a neater distribution of the items to components. The item analysis had revealed that most scanning items have high item facility and low item discrimination values.[68] Among search reading items, SR7, and three careful reading; CR4, CR5 and CR6, are problematic. With some subjectivity, the researcher decided to eliminate the first five scanning items rather than all the problematic items because it would not be sensible to make an analysis with too few items. Besides, the second part of the scanning test showed some homogeneity. These items did not have any further unfavourable statistics either. SR7 was also retained in the set since its only problem was low IF. Among three problematic careful reading items, CR5 exhibited

[68] See section 4.5.8.2

the worst item quality and it was extracted from the data. Therefore, the purged version included SC6-10, SR1-7 and CR1-4 and CR6-7 (N=18).

PCA with varimax rotation of the purged data of September 2001 test yielded three components with eigenvalues over 1.00 accounting for 42.6% of total variance. KMO measure of sampling adequacy was .881 and Bartlett's test of sphericity was significant at .000 level. Two low communalities were observed; .274 and .271 for CR3 and CR4 respectively.[69] Table 4.79 shows the rotated component structure of the purged set of the test in which all the search items load on the first component (F1) and SCII items, on the second component (F2). Careful reading items load on the third component (F3) however, CR1 and CR7 load on F2 together with search reading items more heavily than they do on F3.

Table 4.79: Rotated component matrix:
September 2001 test – purged set

	Component		
	1	2	3
SC6	.163	**.548**	-.013
SC7	.111	**.588**	-.023
SC8	.039	**.771**	.120
SC9	.076	**.735**	.020
SC10	-.010	**.755**	.136
SR1	**.616**	-.012	.104
SR2	**.693**	.084	.095
SR3	**.673**	.076	.100
SR4	**.710**	.106	.162
SR5	**.702**	.102	.135
SR6	**.703**	.057	.066
SR7	**.556**	.101	-.037
CR1	**.421**	.059	.394
CR2	.063	.142	**.623**
CR3	.322	.202	**.360**
CR4	.160	-.116	**.481**
CR6	.037	.046	**.608**
CR7	**.505**	.067	.264

[69] See Appendix 4.49 for details.

Subtest-factor correlations of the purged data were analysed and the results are given in Table 4.80 in which it is seen that F1 accounts basically for search reading with a correlation of .969 and secondarily for careful reading (.476). F2 has almost a perfect correlation with SCII and F3 accounts for careful reading with a correlation of .833.

Table 4.80: September 2001 test – purged set: Subtest – Factor correlations

Factors	Eigen value	% of variance (cum: 42.6%)	SC II (SC 6-10)	Search R.	Careful R. (-CR5)
F1	4.422	24.569	.113*	**.969***	**.467***
F2	2.174	12.076	**.981***	.108*	.141*
F3	1.072	5.958	.068	.136*	**.833***

It is observed that there is a neat distribution of the items to the components in the purged version of the September 2001 data. Once again, there is substantial support for the differential performance of the subtests.

The analysis of the internal structures of the individual subtests would be informative in assessing whether previous assumptions about scanning items and text effect are supported or not. More specifically, scanning items perform differentially according to their place in the test, and unless there are certain problematic items, data from single-text search and careful reading tests would be unidimensional.

Table 4.81 below shows the component structures of the subtests of the September 2001 test.[70]

[70] See Appendix 4.50 for details.

Table 4.81: Component matrices: September 2001 test – subtests

	Component		
	1	2	3
SC1	-.005	**.776**	.130
SC2	-.026	.100	**.701**
SC3	.244	**.703**	-.081
SC4	.139	.060	**.662**
SC5	.072	**.517**	.370
SC6	.375	.142	**.508**
SC7	.338	**.541**	.192
SC8	**.812**	.082	.132
SC9	**.675**	.283	.128
SC10	**.824**	.093	.046

	Comp.
	1
SR1	.631
SR2	.701
SR3	.702
SR4	.761
SR5	.736
SR6	.715
SR7	.547

	Comp.	
	1	2
CR1	**.680**	.068
CR2	**.487**	-.209
CR3	**.561**	-.267
CR4	**.428**	.223
CR5	.275	**.783**
CR6	**.387**	-.486
CR7	**.630**	.132

As expected, scanning is multi-componential with items dispersed over three factors. Search reading is clearly unidimensional and careful reading, excluding one item (CR5) is unidimensional, too. Therefore, September 2001 subtests also give support to the assumptions discussed above.

4.5.8.5 The September 2001 Test: Discussion

The correlation analysis between the subtests of the September 2001 test showed that the skills in the test do not overlap extensively. On the contrary, especially the correlation between the scanning and the search reading test, as well as the careful reading test, is low enough to consider scanning as separate from the others. The moderate correlation between the search and the careful reading tests was also indicative of differential characteristics of these tests. Besides, with the September 2001 test data, it is once again confirmed that when the tests were refined from confounding factors such as defective item characteristics and multiple text effect, interpretable component matrices were attained. In the analysis of the purged version of the test, the items from different subtests loaded on related components, supporting that test takers show different performances on the subtests.

4.5.9 Summary of the Findings and Discussion

In section 4.5 of the study, the question whether or not the subtests of the BUEPT reading test are testing different reading operations has been investigated. Four versions of the test have been analysed through correlation and Principal Component Analysis. Pearson product-moment correlation analysis was used to assess the overlap between the subtests. Principal Component Analysis (PCA) was used to assess the internal structure of the tests, i.e. whether the items putatively testing a particular reading operation load on the same factor as opposed to the others that are deemed to test different skills. Firstly, the whole data from the test administration were submitted to PCA. Then, the items that were designated as problematic through item analysis were eliminated from the data to form the purged versions, which were analysed by the same factor analysis method. The internal component structures of the subtests were also analysed for further check for random factors. The subtests that included two texts were analysed in split halves to see the differential effects of the subsections on the component structure.

In the September 2000 test data, the inter-correlations were at moderate level, with search and careful reading correlating more highly. Initially extracted 10-component matrix was reduced to 7 components by the exclusion of the items designated as defective through the item analysis. The analysis of the individual subtests suggested that they were not unidimensional in themselves. The scanning items performed differentially according to their position in the test. The first four items did not load on a specific component but they were dispersed over several components. The items in the second half of the scanning test formed a component. For search and careful

reading, the items from different texts loaded on different components suggesting a text effect. The skimming question appeared as different from the other items in the test and loaded on and correlated with a specific component. In several subtest-factor correlation matrices, search reading and careful reading correlated with the same component to differing levels suggesting a link between these two types of reading operations. Half-set I, which was the set formed with non-problematic set of subsections, yielded a four-component ideal matrix with SCII, skimming, SRII and CRII loading on separate components. In half-set II, SCII and CRI could be identified as factors.

In the January 2001 test, the subtest inter-correlations were again moderate. Initially 9 components were extracted. SCI and SCII, skimming, SRI and CRI were identifiable as factors. The purged data revealed an interpretable 5-component data. The analysis of the individual tests revealed non-unidimensional subtests in which different sections of the subtests were easily identifiable. The subtest data of the January test also suggested differential performance of the scanning items and text effect. There was observable link between search and careful reading in correlation matrices in the January data as well. Half-set data revealed that different subskills were identifiable as separate factors.

Structurally reduced forms of the BUEPT reading test, the June and September 2001 versions, revealed less complex structures. In these tests, the difference between scanning and the other two tests, search and careful reading became more prominent by lowered correlations of the scanning with the other tests. The purged sets revealed 3-component matrices in which SCII, the search and careful reading subtests loaded

on separate components. The first few scanning items behaved in the peculiar manner as they did in the previous tests. The subtest-factor correlation analysis revealed a link between the search and careful reading subtests in this set of data, too.

On the whole, it can be said that there is repeated quantitative evidence supporting the assumption that the subtests are testing different reading operations since when the random factors were reduced, the subtests could be identified as loading on separate components in all four data sets. The statistical analyses were also indicative of successful operationalisation of the test specifications in the tests. Two observations emerged during data analysis. One is the assumption that test takers perform differentially on the first and second half of the scanning test possibly due to unequal time allocation for the two parts. This assumption was also supported by repeated evidence. The September 2000 and January 2001 data also suggested that the text itself was a strong test facet and different texts might lead to differences in statistical analysis. In the inter-correlation analyses, scanning was found to correlate only weakly with the other subtests. On the other hand, search and careful reading correlated moderately. In several subtest-factor correlation matrices they correlated with the same factor suggesting a stronger link between the two compared to the other combinations.

There are several points to be made based on these findings. Firstly, the statistical analysis of the four tests suggested that the scanning skill, as it is operationalised in the tests, is very different from the other two skills, namely search and careful reading skills obviously because of the fact that scanning, as delineated in the test specifications, is the only skill that has merely local focus. Therefore, it is quite

expected that scanning would appear less related to search and careful reading skills, both of which have global focus as well. As such, this finding is in congruence with the premises of the reading model on which the test is based. Besides, the performance in the scanning test is mostly related with speed rather than comprehension. It is widely acknowledged that word recognition skill is essential to reading comprehension. As designated in the test specifications, the scanning test basically taps on that ability. However, the findings in this study suggest that scanning, as it stands, only differentiates at very low levels of proficiency and has little measurement value in a highly demanding proficiency test as the BUEPT. Therefore, its value for future use should be questioned.

Secondly, search reading emerges as a highly demanding skill in the BUEPT. By definition, it requires fast process of the text and considerable comprehension at the same time. Test takers also need to resort to their textual formal schemata in order to facilitate their search reading. In essence, successful search reading combines skimming and careful reading, both of which have global focus. The fact that the search and careful reading tests in this study correlate highly significantly as opposed to scanning is also supportive of the global versus local distinction in reading framework. However, as discussed in section 4.2.1, the combinatory nature of the search reading should be reflected more explicitly in the model since search reading in Urquhart and Weir's (1998) reading matrix does not overlap with careful reading skill.

Thirdly, it has become clear that items based on different texts might behave differently. Therefore, although Urquhart and Weir (1998) suggest the opposite, it is

the researcher's view that the analysis of the differential performance of the items that operationalise different skills would be free of confounding text effect if the items in a test were based on the *same* text.

In general, as stated above the statistical analyses gave substantial support to the confirmation of the hypotheses discussed in this section. Search reading and careful reading correlated moderately reflecting the shared global focus. Scanning as the only skill with local focus correlated with search and careful reading skills only weakly. When the tests were reduced to the simpler forms, the items putatively testing different operations loaded on different factors, yielding supportive evidence for the operationalisability of Urquhart and Weir's (1998) reading framework.

4.6 Research Question 5: Do the factor structures of the different versions of the test show similarities across versions?

For the investigation of the generalisability aspect of the construct validity of the BUEPT reading test, it was hypothesised that the items putatively testing different operations would load on different factors in the Principal Component Analysis in the same manner across four versions of the test, that is, similar component structures would be observed across four different versions of the test. Among the four tests analysed in this study, the September 2000 and the January 2001 tests had similar structures: they were composed of the scanning, skimming, double-text search reading and double-text careful reading sections. PCA analyses made on these tests yielded similar structures. On the other hand, the June 2001 and the September 2001 tests had reduced components: They were composed of the scanning, single-text

search reading and single-text careful reading sections. PCA analyses made on these tests also yielded similar structures. Comparing both groups of test, it could be claimed that when the random factors were reduced, the subtests could be identified as loading on separate components in all four tests. Therefore, this is substantial evidence supporting the successful operationalisation of the reading construct as determined in the framework by Urquhart and Weir (1998) across test versions. It can be claimed that the reading construct underlying these tests is generalisable across different versions.

Since the statistical findings supporting the generalisability hypothesis are presented extensively in section 4.5, they will not be repeated here to avoid redundancy. The reader is referred to section 4.5 for details.

4.7 Research Question 6: What will be the relation between the criterion measure and the test under investigation?

The external aspect of the construct validity of the BUEPT reading test is investigated by the correlational evidence from a criterion reading test; the IELTS. However, the extent to which these tests are similar with respect to the elements of the reading construct assessed in each of them is questioned before the two tests are statistically correlated. Therefore, in this section of the study, first, the content comparison of the tests will be discussed. Then, the correlation between them will be investigated. The content analysis is expected to shed light on the meaningfulness of this correlation.

4.7.1 Content Comparison of the IELTS and the BUEPT Reading Tests

4.7.1.1 Results

The content analysis scheme explained in the previous sections is used in this analysis too. In section 3.3, the scheme was introduced and the analysis for the BUEPT test was given in section 4.3. The full version of the scheme was presented in Appendix 3.2. The results from both tests are given in Tables 4.82, 4.83 and 4.84 below. For the sake of convenience, the results from the BUEPT reading test-content analysis are given here, too. But the reader is referred to section 4.3 for the detailed analysis of the BUEPT test. In the following section, the results from the IELTS reading module will be analysed.

Table 4.82:Test and text characteristics – IELTS and BUEPT (scores averaged over 5)

Test Rubric Characteristics	**IELTS**	**BUEPT**
TO	2	1.6
TA	1.7	2.3
I	2.7	1.2

Text Characteristics	**IELTS 1**	**IELTS 2**	**IELTS 3**	**BUEPT-SC**	**BUEPT-SR**	**BUEPT-CR**
NT	newspaper article (4) magazine article (4) research/journal article (2)	newspaper article (3) magazine article (4) research/journal article (3)	newspaper article (1) magazine article (5) research/journal article (2) textbook article (1)	magazine article (6)	magazine article (1) research/journal article (5) textbook article (1)	magazine article (4) research/journal article (4)
RO	description (1) narration (1) information(4)	information (2) comparison and contrast (1) discursive (3) argumentation (1)	description (1) information (3) argumentation (4)	narration (1) description (1) information (4) argumentation (1)	information (3) discursive (1) argumentation (4)	information (1) discursive (4) argumentation (2)
GR	3.8	3.8	3.3	3	3.7	3.5
VOC	2.8	3.2	3.8	2.5	3.3	4
COH	2.3	2.3	2.5	1.6	2.8	2.2
RO°	2	2.3	2.7	1.8	2.3	1.8
DC	2	2.2	2.5	1.6	2.7	2.2
DNI	2.2	2.3	4	1.8	2.2	3.2
TI	1.2	2.5	2.8	1.8	3.5	3
TS	1.8	2.3	4.2	3.3	3.5	4.2
CS	3.7	2	1.4	1.3	1.8	1.5
OD	2.8	2.7	3.3	2.3	3.7	3.7
Av. comprehensibility score	**2.46**	**2.56**	**3.05**	**2.1**	**2.95**	**2.93**
Overall diff. of the questions	**questions 1-6: 3.5** **questions 7-9: 1.5** **questions 10-13: 3.3**	**questions 14-17: 2.3** **questions 18-21: 3** **questions 22-25: 3.2** **question 26: 1.5**	**questions 27-32: 2.5** **questions 33-40: 4**	**scanning: 1.7**	**skimming: 3.2** **search reading: 4**	**careful reading: 2.7**

Table 4.83: Operations and text spans with frequencies in parentheses– IELTS

IELTS	Operations (O)	Text Span (TS)
Passage 1		
q1	2(4), 4(4), 5(1), 9(3), 11(1), 12(4)	1(1), 2(4), 3(1)
q2	2(5), 4(6), 8(2), 10(1), 12(2)	2(4), 3(2)
q3	1(1), 2(6),	2(4), 3(2)
q4	2(4), 3(1), 4(1), 5(1), 9(1), 10(1), 12(2)	2(1), 3(3), 5(1), ?(1)
q5	2(4), 4(4), 5(1), 11(4)	2(5), 3(1)
q6	1(2), 2(5), 3(1), 4(1), 8(1), 12(1)	2(1), 3(2),4(2), ?(1)
q7	2(4), 3(3), 8(3), 10(3)	2(5), 3(1),
q8	1(1), 2(1), 3(5), 4(2)	2(6)
q9	2(2), 3(3), 4(3), 8(3), 12(1)	2(6)
q10	2(1), 4(1), 5(1), 6(1), 8(2), 9(4), 10(2), 11(1), 12(1), 13(1), 14(1)	3(1), 4(4), 5(1), ?(1)*
q11	2(1), 4(1), 5(1), 8(3), 9(4), 10(2), 11(2), 12(1), 13(1), 14(1)	3(1), 4(4), 5(1), ?(1)*
q12	2(1), 4(1), 5(1), 8(3), 9(4), 10(2), 11(2), 12(1), 13(1), 14(1)	3(1), 4(4), 5(1), ?(1)*
q13	2(1), 4(1), 5(1), 8(3), 9(4), 10(2), 11(2), 12(1), 13(1), 14(1)	3(1), 4(4), 5(1), ?(1)*
Passage2		
q14	2(2), 3(4), 4(3), 8(2), 10(3), 14(1)	2(6)
q15	2(2), 3(1), 5(2), 9(6)	4(6)
q16	3(2), 4(4), 8(4), 11(3), 12(1)	2(4), 3(1), 4(1)
q17	2(2), 4(4), 5(1), 8(2), 9(2), 11(3), 12(3)	2(3), 3(3)
q18	2(2), 3(1), 4(3), 5(3), 9(1), 10(1), 11(2), 12(1), ?(1)	1(1), 2(3), 3(1), 4(1)
q19	2(3), 3(3), 8(1), 10(3), 12(3)	2(5), 3(1)
q20	2(2), 4(3), 8(1), 11(3), 12(2), ?(1)	2(4), 3(2)
q21	2(2), 4(2), 9(3), 11(3), 12(2), ?(1)	1(1), 3(2), 4(3)
q22	2(1), 4(4), 5(3), 8(3), 9(1), 11(3), 12(1)	2(3), 3(2), ?(1)
q23	2(1), 4(4), 5(2), 8(3), 9(1), 11(2), 12(1)	2(3), 3(2), ?(1)
q24	2(1), 4(4), 5(2), 8(3), 9(1), 11(3), 12(1)	2(3), 3(2), ?(1)
q25	2(1), 4(3), 5(2), 8(3), 9(1), 11(2), 12(1), 15(1)	2(3), 3(2), ?(1)
q26	5(1), 9(6)	4(5), ?(1)
Passage3		
q27	1(4), 2(2), 4(2), 8(4), 10(1)	2(6)
q28	1(4), 2(4), 3(2), 8(3), 10(1)	2(6)
q29	1(3), 2(5), 3(5), 8(1), 10(1)	2(6)
q30	2(5), 3(2), 5(1), 8(3), 11(1)	2(5), 3(1)
q31	1(2), 2(5), 4(3), 8(4), 10(1)	2(6)
q32	1(2), 2(5), 4(3), 5(3), 8(3)	2(6)
q33	5(1), 2(3), 8(3), 9(4), 10(1), 11(1), ?(1)	2(1), 3(4), ?(1)
q34	5(1), 2(3), 8(4), 9(3), 11(1), ?(1)	2(2), 3(3), ?(1)
q35	1(3), 2(3), 8(3), 9(1), 11(2), ?(1)	2(1), 3(4), ?(1)
q36	1(1), 2(2), 8(4), 9(3), 11(1), ?(1)	2(2), 3(3), ?(1)
q37	1(1), 2(4), 8(4), 9(2), 11(1), ?(1)	2(3), 3(1), 5(1), ?(1)
q38	1(1), 2(4), 8(5), 9(3), ?(1)	2(2), 3(3), ?(1)
q39	1(1), 2(3), 8(5), 9(2), ?(1)	2(1), 3(4), ?(1)
q40	1(1), 2(4), 7(1), 8(4), 9(2), 11(2), ?(1)	2(2), 3(3), ?(1)

*One rater marked both TS4 and TS5 for these items.

Table 4.84: Operations and text spans with frequencies in parentheses– BUEPT (repeated)

BUEPT	Operations (O)	Text Span (TS)
scanning		
sc1	1(5), 2(3), 3(5)	2(6)
sc2	1(5), 2(3), 3(5)	2(6)
sc3	1(6), 2(3), 3(4)	2(6)
sc4	1(5), 2(3), 3(5)	2(6)
sc5	1(5), 2(3), 3(6)	2(6)
sc6	1(6), 2(3), 3(4)	2(6)
sc7	1(6), 2(3), 3(4)	2(6)
sc8	1(5), 2(3), 3(5)	2(6)
sc9	1(4), 2(3), 3(6)	2(6)
sc10	1(6), 2(3), 3(4)	2(6)
sum:	**1(53), 2(30), 3(50)**	**2 (60)**
search reading		
sr1	1(1), 2(1), 4(3), 5(3), 6(3), 7(4), 9(4)	1(1), 2(1), 4(4)
sr2	1(1), 2(3), 4(4), 5(3), 6(3), 8(5), 11(1), 12(2)	2(6)
sr3	1(1), 4(4), 5(3), 6(1), 8(5), 11(1),12(2)	1(1), 2(2), 3(3)
sr4	1(1), 2(2), 4(4), 5(1), 6(1), 8(3), 11(3), 12(2)	1(1), 2(2), 3(3)
sr5	1(1), 2(2), 4(4), 6(2), 8(4), 10(1), 12(1)	2(6)
sr6	1(1), 2(3), 4(3), 5(3), 6(1), 8(3), 10(2), 12(2)	1(1), 2(4), 3(1)
sum:	**1(6), 2(11), 4(22), 5(13), 6(11), 7(4), 8(20), 9(4), 10(3), 11(5), 12(9)**	**1(4), 2(21), 3(7), 4(4)**
careful reading		
cr1	2(1), 4(4), 8(4), 10(2), 11(2), 12(1), 13(1)	2(4), 3(2)
cr2	1(1), 2(1), 4(4), 8(4), 9(1), 10(2), 11(2), 12(2), 13(2)	2(1), 3(5)
cr3	2(1), 4(3), 8(3), 9(1), 10(3), 11(2), 12(2), 13(2)	2(1), 3(5)
cr4	2(1), 4(3), 5(1), 8(3), 9(1), 10(3), 11(3), 12(3)	2(1), 3(5)
cr5	1(1), 2(1), 4(4), 8(4), 9(1), 10(2), 11(2), 12(2), 13(1), 15(1)	2(1), 3(5)
sum:	**1(2), 2(5), 4(18), 5(1), 8(18), 9(4), 10(12), 11(11), 12(10), 13(6), 15(1)**	**2(8), 3(22)**

Test rubric characteristics

It is seen in Table 4.82 that the IELTS reading test is judged to be organised clearly (TO: 2) and the time allocation (TA: 1.7) for the sections is found to be sufficient. However, in terms of the instructions (I: 2.7), the raters decided that the IELTS test is less than very clear.

Text and item characteristics

IELTS 1: The first passage in the IELTS reading test is either a newspaper or a magazine article of informative nature (NT) for most of the raters. The difficulty level of grammar in the text is above average (GR: 3.8) and that of vocabulary (VOC: 2.8) is average. In terms of coherence (COH: 2.3) and explicitness of the

rhetorical organisation (RO°: 2), the text is judged to be quite explicit. The information in the text (DC: 2) is contextualised, relatively diffused (DNI: 2.2) and concrete (1.2). The topic is not subject specific (TS: 1.8) but has rather high cultural specificity (CS: 3.8). The overall difficulty of the text (OD: 2.8) is designated as being at the medium level. The average of all these characteristics gives a comprehensibility score of 2.46. There are three sets of questions formulated on the first passage of the IELTS test; the first group received the highest difficulty rating (questions 1-6: 3.5), the second group of questions (questions 7-9: 1.5) is quite easy and the rating for the last group (questions 10-13: 3.3) is again slightly above average.

As for the operations used in arriving at an answer to the questions (q), the raters reported the use of a variety of operations (See Table 4.83). The raters attempted q1, q5, q7, q9 with expeditious reading and careful reading operations.[71] According to the operations that the raters reported, they mostly scanned to answer q2, q3 and q8. It is not possible to determine a dominant operation for q4 and q6. Another problematic group of questions is q10-13 where the raters had to fill in a cloze paragraph. All sorts of operations are noted for the completion of these items, the highest frequency being that of O8 and O9.

The dominant text span for q1, q2, q3, q5, q7, q8 and q9 is TS2. For q4 and q6, there was not a clear text span marked. One rater commented that TS could not be determined for these questions (indicated by a question mark). The four raters chose TS4 for q10-13. One rater marked TS5 for those questions indicating that

[71] See Appendix 3.3 for the definition of the operations.

information outside the text needed to be incorporated in the answers and another rater again marked that the text span is not clear.

IELTS 2: The second IELTS text is perceived as an article that can appear in a newspaper (3), magazine (4) or in a research journal (3). It is reported to be of informative (2) and discursive (3) nature. Its level of grammar (GR: 3.8) and vocabulary (VOC: 3.2) is above average and it has moderately explicit organisation (COH: 2.3 and RO°: 2.3). The textual information is rather contextualised (DC: 2.2) and diffused (DNI: 2.3) and it is moderately concrete (TI: 2.5) and non-specific (TS: 2.3) and culture free (CS: 2). The overall difficulty level (OD) is 2.8 and all these characteristics give a comprehensibility score of 2.56. Four groups of questions received the difficulty ratings of 2.3, 3, 3.2 and 1.5, respectively.

The questions on the second passage of the IELTS test also received a variety of responses in terms of the operations. q14 and q16 seem to be a combination of search and careful reading. q15 apparently required a summarisation of a certain part of the text. q17 mostly needed careful reading. The operations reported for q18 are inconclusive. q19 seems to combine scanning and careful reading skills. q20-21 mostly required careful reading skills. q22-25 are again seem to prompt a combination of search and careful reading skills. q26 is perceived as a skimming item.

The text spans for certain questions in the IELTS2 could be identified with more ease than others. q14 is seen as a local question (TS2) and q15 requires the understanding of the entire passage (TS4). Four raters identified q16 as requiring TS2 but others

determined that the question might require the processing of larger spans (TS3, TS4). There was not an agreement on the text span for q17 since half the raters (3) determined that the question required TS2 and whereas the others concluded that it required TS3. For q18, three raters identified the text span as TS2 but one rater determined that the question has no relation to the text (TS1), another one, that it required the processing of TS3, and yet another, that it needs the integration of the information in the whole passage. Obviously, the text span for q18 is as inconclusive as its operations. For the majority of the raters, text span for q19 and q20 is TS2. q21 received mixed responses; one rater determining that it had no relationship with the passage (TS1), two raters concluding that it needed the process of TS3, and three raters judging that the entire passage had to be processed (TS4) for the completion of q21. For q22-25, three raters opted for TS2, and two for TS3. One rater again could not specify a particular text span for the question. q26 is identified as requiring the process of the entire passage (TS4).

IELTS 3: Most of the raters judged the third reading passage in the IELTS test to be a magazine article (5) involving information (3) and argumentation (4). Grammar (GR: 3.3) and vocabulary (VOC: 3.8) are above average and it has moderately less explicit organisation (COH: 2.5, RO°: 2.7). The information in the text is somewhat contextualised (DC: 2.5) but it is highly compact (DNI: 4). It is moderately abstract (TI: 2.8) but highly specific in terms of topic (TS: 4.2). Its culture specificity is quite low (CS: 1.4). The overall difficulty of the passage is judged to be 3.3, above average, and the average of the text characteristics (comprehensibility score) is 3.05. Two groups of questions based on this passage are of differing levels of difficulty:

the first group (questions 27-32) is moderately difficult (2.5) but the second group (questions 33-40) is highly difficult (4).

As for the operations used for the answering of the questions in this part of the test, there seems to be a clear difference between the first group of questions (questions 27-32) and the second one (questions 33-40). The first group seems to involve mostly expeditious reading skills (scanning and search reading) which may involve some careful reading after the answer has been located (O8). The second group of questions also involves scanning and search reading skills (O2, O8) but O9 is reported frequently as well, showing that the raters had to refer to the macrostructure of the text they formed in their minds in order to answer this set of questions. One rater was consistently unable to determine the operations for these questions (indicated by question mark).

The distinction between the two sets of questions is also clear in the text spans. The text for the first set of questions is determined to be locally processed (TS2) almost unanimously. For the second set, TS3 seems to be the dominant operation for the questions except for q37. For q37, three raters marked TS2, one rater marked TS3 and one rater commented that the processing of the question needs information outside the text (TS5). One rater was again unable to determine the text span for the questions.

4.7.1.2 Discussion of the results from the content analysis of the IELTS test

To sum up the results of the content analysis, the main sections of the scheme will be referred to briefly and the emerging picture will be evaluated before the content comparison presented in the next section. The analysis of the text characteristics by the raters revealed that the IELTS test reading passages are rather journalistic articles with some research focus. The rhetorical structure of the passages is organised such that the texts involve an increasing level of argumentation, the first text being generally informative, the second of discursive nature and the third involving argumentation for most of the raters. The comprehensibility scores also increase from the first text to the last. The types of the questions as well as their difficulty levels are varied. But each text has groups of relatively easy and difficult questions, the most difficult set of questions being based on the last text.

The analysis of the operations throughout the test did not reveal any sort of grouping according to skills used in responding to the questions. Rather, several types of questions tapping different skills were written on each text. Therefore, it was not possible to provide a summary of the operations for the text parts. However, the raters mostly reached a consensus on the majority of the questions. The analysis of the items on which there was no clear consensus is given below.

The first group of questions (q1-6) requires the test takers to judge if the statements in the questions agree or disagree with the writer's claims or not mentioned in the text at all, a type of question peculiar to the IELTS. The test taker can respond to the question by stating 'yes', 'no' or 'not given'. It is the 'not given' questions (q4 and

q6) that caused the difficulty for the raters most obviously because neither the statement in the question nor the answer exists in the text. So they went back and forth in the text (O2) and read differentially to find the answer. The text span for these questions could not be identified either. Three raters also reported that it was difficult for them to differentiate between a 'no' and 'not given' answer, and as a side observation, some of the raters incorrectly answered these questions. q7-9 are multiple choice items and could be answered by matching the item and the text and reading the text carefully afterwards (O2, O3, O4, O8). q10-13 require filling in the blanks in a cloze test and prompted a varied set of operations. However, the majority of the raters (4) reported that they formed a summary in their minds (O9), which, in the present classification, refers to the skimming skill, i.e., quickly going through the text in order to establish a general sense of the text. The raters might have taken O9 as reading carefully and summarising a part of the text, which is actually what the cloze test requires them to do. The fact that TS4 was marked as the dominant text span also supports that the raters were summarising a large text part in answering these questions. One rater who could not identify the text span for these items (indicated by the question mark) commented that identification of the text span does not apply well to cloze items. In the second text, q18 and q21 were the most problematic items to classify. These are again 'yes/no/not given' types of items. Marking the text span for the items q22-25 was also problematic. These items require the test taker to go back to the text again to locate specific information in the text. Several paragraphs need to be checked, although the answer lies in a specific part of the paragraph. Thus, TS2 (3) and TS3 (4) are marked for these items. One rater did not specify the text span for these items at all. For the same rater, items q33-40 were

problematic, too. These items also require the test takers to go back to several different parts of the text.

Otherwise, the majority of the items apparently require a combination of expeditious and careful reading skills for the raters. Two comments can be made at this point. Firstly, considering that the IELTS passages are relatively short and could be read by the expert raters in a short time, the raters might well have read the texts before they attempted the questions. If this is the case, in order to locate the answers, they might first have tried to go back to the part of the text where they remembered the answer was located and read that part of the text carefully to extract the answer. In this case, the nature of the expeditious reading changes substantially in the analysis of the raters since the definitions of the operations in the present classification do not suggest a priori reading. Secondly, the questions in the IELTS test do not always follow the information structure in the text; the test takers may need to go back to the beginning of the text after they have answered several questions from the other parts. This means that they might have had to go back to the parts they read before. This might have facilitated initial expeditious reading followed by careful reading. Under these circumstances, it is hard to designate the extent to which the raters processed the text expeditiously. One illustrative example to this may be q26, an item marked as a skimming item that required forming a summary of the main ideas/text topic in mind. As defined in the present analysis, this requires a quick processing of the text to arrive at a general understanding of the main ideas in the text without a detailed understanding of the whole text span. Nevertheless, if this is a question asked after several careful reading questions as is the case with q26, it is not a skimming question any more but an item that requires summarisation of the passage after

detailed processing. What it all boils down to is that the IELTS items may all be careful reading items rather than anything else.

4.7.1.3 Are the IELTS and the BUEPT reading tests comparable?

The results from the content analysis of the IELTS test are given above and the BUEPT reading test is analysed in section 4.3. Final comments on the comparability of these tests will be given below.

Test rubric characteristics

The analysis of the test rubric characteristics shows that both tests are well organised but some sections of the BUEPT might need too speedy processing and the instructions of the IELTS are not as clearly given as they are in the BUEPT.

Text and item characteristics

The nature of the texts used in two tests is quite similar. The IELTS texts are not strictly academic, yet they cover a range of rhetorical styles including argumentative texts as the BUEPT does. The texts are arranged according to their difficulty in the IELTS. In the BUEPT, however, the texts are selected according to their appropriacy to the skills being tested. Still, from scanning to careful reading, the texts are expected to be of increasing difficulty in the BUEPT as well. The questions in each subsection of the BUEPT are defined according to the componential skills approach. Therefore, the questions in each subtest represent a skill and are expected to be processed accordingly. The tests consist generally of short-answer questions with occasional multiple-choice items especially in the skimming part. The questions in

the IELTS, on the other hand, are apparently designed taking into consideration textual features (content, difficulty), item type (multiple-choice, cloze, etc.), item difficulty, etc., and the approach underlying the test does not rest on a componential, multi-divisible conceptualisation of reading.[72] Besides, a wide range of item types is used in the IELTS.

As such, there are similarities but also substantial differences between the tests. Whether the IELTS forms a true criterion for the BUEPT is challenged by the content analysis to a certain extent. However, since these two tests are designed to measure the construct of academic reading ability no matter to what extent the underlying conceptualisations of the reading construct differ, if not a full statistical overlap, some degree of correlation might be expected between them,.

4.7.2 The correlation between the IELTS and the BUEPT

Table 4.85 below shows that the means of the two tests. The means are quite comparable although the IELTS has a slightly higher mean.

Table: 4.85 The means of the IELTS and the BUEPT

	Mean	STD. deviation	N
IELTS	57.48	15.35	126
BUEPT	54.47	17.79	126

When the two tests are correlated in their full forms, a coefficient of .483 is found (See Table 4.86). The correlation is significant at 0.01 level and can be considered

Table: 4.86 The IELTS and BUEPT correlations

	BUEPT (SC+SR+CR)	BUEPT (SR+CR)
IELTS	.483*	.723*

*Pearson correlation significant at the 0.01 level (2-tailed)

[72] See also section 3.7.2.2.

moderate. However, as a further analysis, the researcher considered the results of the item-operation analysis in the content scheme discussed above, and taking into consideration the fact that the analysis of the IELTS yielded very few mere scanning operations,[73] she decided to repeat the correlation after eliminating the scanning section from the BUEPT reading test. The correlation between the IELTS and the search and careful reading sections of the BUEPT (SR+CR) yielded a much higher correlation coefficient of .723 significant at 0.01 level (See Table 4.86). This is considered to be sufficient support for the criterion related construct evidence for the BUEPT reading test.

In sum, the content analysis of the two tests provided valuable information as to the characteristics of the tests and helped the analysis of the correlation between them. It is usual practice in the language testing field to correlate newly designed tests with the standardised criterion measures to confirm their utility. Nevertheless, unless an analysis of the nature of them confirms construct congruence between them, such a comparison is not very informative. From that respect, by the help of the content analysis, the researcher has been able to approximate the operations tested in the tests to have a higher correlation, which is otherwise not very supportive of the criterion relatedness, and therefore, she has been able to confirm the hypothesis that the BUEPT and the IELTS reading modules will correlate significantly. However, it should still be noted that the tests compared are not exactly similar measures, which is also reflected in the only moderately high correlation (.723) between the IELTS and the BUEPT reading tests.

---

[73] Scanning operations, when they appear in the analysis, are generally complemented with search and/or careful reading operations (See section 4.7.1.1 and Table 4.83).

4.8 The Academic Implications of the Findings

The present chapter has given the results of the research attempting to implement a reading framework in the test design and the validation of the operationalisation of that framework in the tests. The findings presented and discussed in this chapter have several suggestions for the reading and testing research.

Firstly, and most importantly, the findings in combination give substantial support for the successful operationalisation of the reading framework on which the test under investigation is based on, thus suggesting the existence of different types of reading for different types of purposes.

The assumption that reading might involve subskills has been disputed in several studies and arguments. For example, Rost (1993) argued for a 'general reading competence' and Hudson (1996) claims that it is difficult to define skills in practice because they largely overlap.[74] On the other hand, reading research has shed light on several aspects of the reading process that suggest differential reading behaviour in different circumstances and with different purposes. For example, Kintsch and van Dijk (1978) and Kintsch and Yarbrough (1982) illustrate that there are two levels of comprehension: macro-processes – understanding at the global level, and micro-processes – understanding at the local level. Different levels of processing of a text place different loads on memory and result in different products. Just and Carpenter (1987) also underline that unlike at lower levels, at higher levels of comprehension, the reader must construct a representation of the text in relation to the situation the

[74] See section 2.3.4 for the details of the studies.

text is referring to. Guthrie and Kirsch (1987) show that reading to comprehend (reading carefully to understand the explicitly stated ideas) and reading to locate information (selective sampling of the text) are clearly differentiated. Carver (1992) identifies 'accuracy' and 'rate' as factors. Bernhardt (1991) in her literacy model, describes the flexible reading process as the deployment of various process strategies for different purposes. As a reader's literacy develops, he or she becomes more efficient in deciding how to approach a text, what to read carefully and how to monitor the process of reading to achieve the purposes of reading. Similarly, Grabe and Stoller (2002) also differentiate several types of purposes for reading and point out that the readers usually make initial decisions as to what to read and how to read in certain settings.

Taking all these into account, it would not be wrong to assume that different types of reading, or reading skills could be assessed separately when these were operationalised in the test items and the texts carefully. In the present study, the PCA analysis of the four versions of the BUEPT reading test clearly showed that scanning is a distinguishable skill (supporting Guthrie and Kirsch, 1987), and although they overlap, search reading and careful reading can be identified in the data as well.

This was also evident in the data from the expert analysis of the September version of the test. Unlike the experts in Alderson and Lukmani (1989) and Alderson (1990a)[75], the experts in this study showed considerable agreement as to what items in the test measure. The major strength of the analysis made here was that a content analysis scheme adopted from Bachman et al. (1995) was used to systematically

[75] See section 2.3.4.

gather information about both the items and the texts. In the previously mentioned studies on expert judgement, it is not clear whether or not the texts were also taken into consideration in item categorisation. Judging the items in isolation would challenge the validity of such an analysis. Besides, instead of asking the experts to categorise the items into subskills, they were given a list of descriptive operations and they were asked to mark what ever operation they used in answering the questions. They were also asked to identify the text span they read for each item. Thus, it was possible to see to what degree they read expeditiously and carefully. However, unlike Alderson (1990a, 1990b) and Hudson (1996), no implicational scale was expected to appear among the subskills since these were operations geared at different purposes of reading rather than cognitive skills that are easier and more difficult than one another.

The protocol analysis of the test takers was also supportive of the above mentioned assumption. It is evident in the data that test takers' reading processes were essentially shaped by the type of the item and the nature of the text. The test takers in this study adjusted their reading speed in line with the amount of information they needed to process, e.g., when they needed to understand a little such as in scanning, their reading was fast and comprehension was little, and when they needed to understand the text parts in detail, they read slowly and the comprehension process was deeper, as Carver (1997) suggests. It is true that answering test questions is highly complex and varies from reader to reader, therefore readers might employ different skills in tests as Alderson (2000) points out, but the analysis in this study has made it clear that by carefully controlling the item and text properties, it is possible to control the type of reading that should be deployed for the successful

completion of test items. This analysis exemplified how this could be achieved with several items. The defective items were also illustrative of the instances where such control was not successfully implemented. Close observations on the level of comprehension exhibited by the test takers also enabled the researcher to pin down the amount of unwanted variation in the data, which was considerably low.[76] Yet, it should be once more underlined here that these are reading operations, i.e. different reading styles that test takers use in their attempt to answer test questions and they do not presuppose, or rather guarantee comprehension. It would be wrong to assume that when a reader processes a text carefully, he or she is warranted full understanding of the text. Full comprehension is naturally very much dependent on the level of language proficiency of the test takers, on condition that such textual features as topic and cultural specificity, etc. are controlled. Therefore, assuming that the use of assigned skills by the test takers should lead to success as Li (1992, in Alderson 2000) does, would be faulty unless comprehension itself is also taken into consideration. The present study has been able to take care of this weakness in previous studies by considering the amount of understanding the test takers achieved. Therefore, what makes it difficult to prove the existence of reading skills is not that they are difficult to define as Hudson (1996) claims, but rather the fact that the reader might shift from one to another in the course of test taking. However, the operation or operations an item might trigger are observable, identifiable and controllable assuring that the properties of the text are carefully controlled and the items are carefully designed in accordance with skill specifications.

[76] These were the cases in which unfavourable processing was followed by correct response.

Additional support to the statistical existence of the reading subskills came from the fact that they appeared similarly in the factor analyses of the different versions of the test. Besides, in the correlation analysis between the BUEPT and the IELTS, it was seen that extraction of the scanning part from the analysis improved the correlation between the two test where the latter was analysed as comprising basically careful reading items.

All in all, it can be said that the present study has given support to the successful implementation of the reading framework by Urquhart and Weir (1998) in the BUEPT test, and therefore, to the construct validity of the test. Following the same line of reasoning, it can be added that the findings in this study are supportive of distinguishable, if not totally separable, reading subskills.

CHAPTER 5

CONCLUSION

5.1 Introduction

The purpose behind all assessment procedures is to infer - based on the test scores - the extent to which a test taker would perform an ability in real-life situations. In designing EAP reading tests, we are interested in inferring the reading ability of a test taker in situations in which he or she would read academic texts in English. The first step in doing this is to base the test on a construct definition, a construct framework, so that we can assess the extent to which our results can generalise beyond the testing situation. As Alderson (2000) points out, constructs are abstractions that we define for a specific assessment purpose and we may pick an aspect of the ability in line with our testing purpose. It is well supported in the field that we should operationalise the construct through test specifications, which in turn determine text and task design. Otherwise, if we are unable to define what makes up the construct, then it would be impossible to base tests on construct theories and we are left with 'reliability and psychometric validity' (Grabe, 2000). Then, it would not be possible to argue on what the test measures, and in the long run, it may not be possible to foresee the impact of the test on test taking populations.[1]

In the revision of the BUEPT reading test then, the first step was to determine a theoretically sound and practically applicable reading framework that would provide

[1] For example, it is the researcher's belief that this is why TOEFL is being revised. See Enright et al. (2000) for the new TOEFL reading framework.

skill definitions and ensure content relevance and representativeness. Test specifications were developed based on the framework and they determined the text selection and item generation procedures. Secondly, the main ideas in the texts on which the items would be formulated were determined by the text mapping procedure, which enabled the five test writers to reach a consensus on the items to be formulated, thus reducing the possible subjectivity the test writer might reflect in the test. The third step was to pilot the test versions and subject the data to statistical analysis focusing on score distributions and item characteristics through classical test theory. The results from the pilot administrations helped determine the test administration and scoring procedures, the range of item difficulty and usefulness. The weak items were identified and either through item exclusion or repair, the tests were reduced to their purged versions. These procedures improved the technical quality of the tests and minimised the construct irrelevant test variance. The experts' content analysis of the September 2000 version of the test helped the analysis of whether or not each item reflects the content defined by each dimension of the reading construct as defined in the framework. Thus, the content related evidence to construct validity has been provided. Secondly, verbal protocols of the test takers were analysed to investigate whether or not the test takers utilised the operations specified in the test specifications. It was confirmed that in the majority of the cases, the specified operations were used to arrive at the correct answers. The next step was to analyse the data from regular administrations of the tests. Descriptive statistics and item analysis helped the evaluation of the items and seriously defective items were excluded from the score calculations of the test takers. The data form the actual test administrations were subjected to PCA analysis to investigate the dimensions of the reading construct measured by the tests. Factor analysis provided evidence for the

structural aspect of construct validity of the tests. The factor structures from different versions were compared to find the constructs were generalisable across the test versions to a great extent. Finally, the external aspect of the BUEPT reading test was analysed through content analysis and correlation with the IELTS test, which suggested that the dimensions except for the scanning part overlap considerably. The findings from these investigations provided substantial support for the validity of the score interpretations based on the BUEPT reading test. As such, the study also gives support to the soundness and applicability of the Urquhart and Weir's (1998) framework.

5.2 Research Implications

Primarily, it can be claimed that the present study with its findings from various types of investigations suggests that in the tests of academic reading in EFL, a subsection tapping on expeditious reading operations should be included. Focusing exclusively on careful reading skill in tests will risk construct invalidity. It should also be pointed out that tests measuring only careful reading operations may have negative impact on teaching. If the aim is to have efficient readers, then teaching tasks should also include practice on expeditious reading operations. Tests that might have powerful effect on teaching, therefore, should tap on such skills, too. However, the value of scanning at the word recognition level should be considered in relation with the expected level of proficiency of the test takers.

Secondly, the study suggests that expert opinion should be taken systematically in relation to both the texts and the items considering several features of both. The use of content analysis scheme based on Bachman et al.'s (1995) is an alternative. Thirdly, it has been proven that verbal protocol data is very illuminating in terms of the operations used by the test takers. Therefore, it should be a systematic part of the test development procedure. However, the test taking process should be evaluated from several aspects such as the level of comprehension and test taking strategies. Mere categorisation of the operations by item type might not be very suggestive.

Next, when tests are correlated for external evidence for construct validity, the construct congruence should be evaluated carefully. The content analysis scheme used in this study has also been effective in that respect.

Lastly, it has been shown that construct validation is a comprehensive process and evidence from various sources is needed. Unless supportive evidence for the facets of construct validity is provided, such an endeavour would be incomplete.

5.3 Research Limitations and Suggestions for Future Research

The major weakness of this test revision and validation study is that due to time and resource limitations, expert judgement and verbal protocol data were collected after the September 2000 test was developed and administered. It was also not possible to collect such data for each version discussed in the study. Therefore, the validation claims do not generalise to all the versions. Had it been possible to use such information at the test development phase, much better tests would have been

developed. As Weir et al. (2000) suggest, expert judgement and test takers' introspection or retrospection should be a part of a priori validation and should be repeated for each version of tests.

Secondly, it was not possible to measure the statistical equivalence of the tests through the use of an anchor test. Therefore, a thorough statistical comparison of the test versions could not be done. However, the utmost care was given to improve the test versions at the piloting stage and the difficulty of the reading tests were balanced through item revision and exclusion.

Thirdly, in the present study, the data analysed included only the test takers' scores who were all undergraduate students from L1 Turkish background. However, a small group of post-graduate students from several backgrounds take the BUEPT tests, too. There is also increasing number of non-Turkish students taking the test every year. Therefore, the future investigations of the validity of the BUEPT test should include test takers from various backgrounds.

Last but not least, this investigation does not involve any arguments on the consequential aspect of construct validity. The major reason for this is that the test was discontinued after the fifth version administered in September 2001 and it was not possible to collect any data as to the intended and unintended consequences of the interpretations based on the scores from the BUEPT reading test. As mentioned before, the reasons for that were institutional rather than scientific. The testing office members of the time had initiated this revision process based on their perceptions of the need to adjust the test in line with the suggestions of the recent research in EAP

reading, language testing and validation. There were also comments from the departments on the inadequacy of the language proficiency of the students registering to first year courses. The decision on the revision was taken by the testing office members and with the support of the British Council, they were trained and prepared the new versions of the test. The curriculum committee, who were responsible for adjusting the teaching materials to the new test, were also invited to the training sessions given by Cyril Weir. Before the first version of the new test was administered, the training sessions were offered to the teachers of the school. It should also be pointed out that the reading test was not the only test revised but the listening and writing tests had also undergone certain changes. However, the revisions to the BUEPT tests were not welcomed by the majority of the experienced teachers, who had been teaching at the school for a considerable time. They showed reactions by not attending the test administrations and marking sessions. Some deliberately refused to use the new curriculum materials and would not be convinced by the lengthy reports explaining the statistical results of the tests. When the resistance accumulated, it became apparent that it would be for the benefit of the institution to turn back to the old system. This was a very illuminating experience in the sense that when institutions have long established traditions, homemade tests become a part of their teaching culture and it is not sensible and in fact practicable to impose sudden changes on them. What would be more fruitful is possibly to involve the teachers in both the test revision process and resulting curricular changes with a view to consulting and convincing them of the usefulness of the new methods. This is also presented as a suggestion to future test revision and validation studies.

REFERENCES

Afflerbach, P. (1990). The influence of prior knowledge on expert readers' main idea construction strategies. Reading Research Quarterly.25. (1) 31-46.

Alderson, J. C. (1990a). Testing reading comprehension skills (Part one). Reading in A Foreign Language. 6 (2). 425-38.

Alderson, J. C. (1990b). Testing reading comprehension skills (Part Two). Reading in A Foreign Language. 7 (1). 465-503.

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson and B. North (Eds.) Language testing in the 1990's: The communicative legacy (pp. 71-86). London: Modern English Publications and The British Council.

Alderson, J. C. (1993). Judgements in language testing. In D. Douglas and C. Chapelle, (Eds.), A new decade of language testing research: Selected papers from the 1990 language testing research colloquium (pp.46-57). Alexandria, Virginia: Teachers of English to Speakers of Other Languages (TESOL), Inc.

Alderson, J. C. (1999). Reading constructs and reading assessment. In M. Chalhoub-Deville (Ed.), Issues in computer-adaptive testing of reading proficiency. (pp.49-70). Cambridge: Cambridge University Press.

Alderson, J. C. (2000). Assessing reading. Cambridge: Cambridge University Press.

Alderson, J. C. and Buck, G. (1993) Standards in testing: A study of the practice of UK examination boards in EFL/ESL Testing. Language Testing. 10 (1). 1-26.

Alderson, J. C. and Clapham, C. (1992). Applied linguistics and language testing: A case study of the ELTS test. Applied Linguistics. 13 (2). 149-163.

Alderson, J. C., Clapham, C. and Wall, D. (1995). Language test construction and evaluation. Cambridge: Cambridge University Press.

Alderson, J. C. and Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. Language Testing. 13 (2). 280-297.

Alderson, J. C. and Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. Reading In A Foreign Language. 5. 253-70.

Alderson, J. C. and Urquart, A. H. (1985): The effect of students' academic discipline on their performance on ESP reading tests. Language Testing. 2. 192-204.

Alderson, J. C. and Wall, D. (1993). Does washback exist?. Applied Linguistics. 14 (2). 115-129.

Allan, A. (1992). Development and validation of a scale to measure test-wiseness in

EFL/ESL reading test takers. Language Testing. 9. 101-122.

Alptekin, C. (1991). Neuropsychological aspects of foreign language tests. Language Learning Journal. 3. 71-72.

Alptekin, C. (1999). The cognitive processes underlying reading comprehension in foreign language learning and their implications for assessment. Paper presented at Türk Silahlı Kuvvetleri 1. Yabancı Dil Sempozyumu. Kara Kuvvetleri Lisan Okulu Komutanlığı Istanbul. Turkey.

Alptekin, C. (2000). Assessing L2 reading in proficiency tests: Validity and variance concerns. Paper presented at the First EFL Conference on Testing: Approaches, Applications and Evaluations. Çankaya University. Ankara, Turkey.

Alptekin, C. (2004). Cultural familiarity in inferential and literal comprehension in L2 reading. Paper presented at the 6th Annual International Baccalaureate Day, Yüzyıl Işıl Schools.

ALTE Code of Practice and Quality (2002). http://www.dpb.dpu.dk/infodok/sprogforum/Espr23/saville.html

Anastasi, A., and Urbina, S. (1997). Psychological testing. Upper Saddle River, New Jersey: Prentice Hall.

Anderson, N. J . (1991). Individual differences in strategy use in second language reading and testing. The Modern Language Journal. 75 (5). 460-472.

Anderson, N. J., Bachman, L., Perkins, K., Cohen, A. (1991). An exploratory study into the validity of a reading comprehension test: Triangulation of data sources. Language Testing. 8 (1). pp: 41-66.

Angoff, W. H. (1988). Validity: An evolving concept. In Wainer and Braun (Eds.), Test validity (pp.19-32). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Bachman, L. F. (1989). Language testing-SLA research interfaces. In R. B. Kaplan (Ed.) Annual Review of Applied Linguistics. 9 (pp. 193-209). New York: Cambridge University Press.

Bachman, L. F. (1990a): Fundamental considerations in testing. Oxford: Oxford University Press.

Bachman, L. F. (1990b). Constructing measures and measuring constructs. In B. Harley, P. Allen, J. Cummins and M. Swain (Eds.). The development of second language proficiency (pp. 26-38). New York: Cambridge University Press.

Bachman, L. F. (1997). Generalizability theory. In C. Clapham and D. Corson (Eds.) Encyclopedia of language and education Volume 7: Language testing

and assessment (pp. 255-262). Dordrecht: Kluwer Academic Publishers.

Bachman, L. F. (1998). Appendix: Language testing – SLA research interfaces. In L. F. Bachman and A. D. Cohen (Eds.) Interfaces between second language acquisition and language testing research (pp. 177-195). Cambridge: Cambridge University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. Language Testing. 17 (1). 1-42.

Bachman, L. F. (2001). Designing and developing useful language tests. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara and K. O'Louglin, K. (Eds.) Experimenting with uncertainity: Essays in honour of Alan Davies.Cambridge: Cambridge University Press.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. Language Testing. 19 (4). 453-476.

Bachman, L. F. and Cohen, A. D. (Eds.) (1998). Interfaces between second language acquisition and language testing research. Cambridge: Cambridge University Press.

Bachman, L., Davidson, F. and Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. Language Testing. 13 (2). 125-150.

Bachman; L. F., Davidson, F., Ryan, K. and Choi, I. (1995). An investigation into the comparability of two tests of English as a foreign language: The Cambridge-TOEFL comparability study. Cambridge: Cambridge University Press.

Bachman, L. F., Kunnan, A., Vaniarjan, S., and Lynch, B. (1988). Task and ability analysis as a basis for examining content and construct comparability in two EFL batteries. Language Testing. 5 (2). 128-159.

Bachman, L. F. and Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. TESOL Quarterly. 16 (4). 449-465.

Bachman, L. F. and Palmer, A. S. (1996): Language testing in practice. Oxford: Oxford University Press.

Bae, J. and Bachman, L. F. (1998). A latent variable approach to listening and reading: Testing factorial invariance across two groups of children in the Korean/English two-way immersion program. Language Testing. 15 (3). 380-414.

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. Language Testing. 13 (2). 257-279.

Bailey, K. M and Brown, J. D. (1995). Language testing courses: What are they? In

A. Cumming and R. Berwick (Eds.) Validation in language testing (pp.236-256). Clevedon, Avon: Multilingual Matters Ltd.

Baker, E. L., O'Neil, H. F. and Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. American Psychologist. 48 (12). 1210-1218.

Barnett M. A. (1989). More than meets the eye: Foreign language reading: Theory and practice. New Jersey: Prentice Hall Regents.

Barr, R., Kamil, M.L., Mosenthal, P. and Pearson, P. D. (Eds.) (1991). Handbook of reading research, Vol II. New York: Longman.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. Educational Measurement: Issues and Practices. 17 (1). 10-17, 22.

Bensoussan, H. and Kreindler, I. (1990). Improving advanced reading comprehension in a foreign language: Summaries versus short-answer questions. Journal of Research in Reading. 13. 55-68.

Bernhart, E. B. (1991). Reading development in a second language: Theoretical, empirical and classroom perspectives. Norwood, New Jersey: Ablex Publishing Cooperation.

Brindley, G. (1998a). Describing language development? Rating scales and SLA. In L. F. Bachman and A. D. Cohen (Eds.) Interfaces between second language acquisition and language testing research (pp.112-140). Cambridge: Cambridge University Press.

Brindley, G. (1998b). Assessment in the AMEP: Current trends and future directions. Prospect. 13 (3). 59-72.

Britton, B. K. and Grasser, A. C. (1996). Models of understanding text. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Brown, J. D. (1991). A comprehensive criterion-referenced language testing project. In J. C. Alderson and B. North (Eds.) Language testing in the 1990's: The communicative legacy (pp. 163-184). London: Modern English Publications and The British Council.

Brown, J. D. (1996). Testing in language programs. Upper Saddle River, New Jersey: Prentice Hall Regents.

Brown, J. D. (1999). The relative importance of persons, items, subtests and languages to TOEFL test variance. Language Testing. 16 (2). 217-238.

Brown, J. D. and Hudson, T. (2002). Criterion-referenced language testing. Cambridge: Cambridge University Press.

Brown, J. D. and Ross, J. A. (1998). Decision dependability of subtests, tests and the overall TOEFL test battery. In M. Milanovic and N. Saville (Eds.), Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem (pp.231-265). Cambridge: Cambridge University Press.

Buck, G., Tatsuko, K. and Kostin, I. (1997). The sub-skills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. Language Learning 47 (3). 423-466.

Cambridge IELTS 2 (2000). Examination papers from the University of Cambridge Local Examinations Syndicate. Cambridge: Cambridge University Press.

Canale, M. and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguiatics. 1 (1). 1-47.

Carlson, P. A. and Larralde, V. (1995). Combining concept mapping and adaptive advice to teach reading. Journal of Universal Computer Science. 1 (3). 156-161.

Carr, T. H. and Levy, B. A. (Eds.) (1990). Reading and its development: Component skills approach. San Diego, California: Academic Press, Inc.

Carrell, P. L. (1988). Some causes of text-boundedness and schema interference in

ESL reading. In P. L. Carrell, J. Devine and D. E. Eskey (Eds.), Interactive approaches to second language reading (pp. 101-113). Cambridge: Cambridge University Press.

Carrell, P. (1990). Training formal schemata – replication results. In L. A. Arena (Ed.), Language proficiency: Defining, teaching and testing (pp. 85-152). Hillsdale: Plenum Press.

Carrell, P. (1992). Awareness of text structure: Effects on recall. Language Learning. 42 (1). 1-20.

Carrell, P. and Eisterhold, J. C. (1988). Schema theory and ESL reading pedagogy. In P. L. Carrell, J. Devine and D. E. Eskey (Eds.), Interactive approaches to second language reading (pp. 101-113). Cambridge: Cambridge University Press.

Carver, R. P. (1992). What do standardised tests of reading comprehension measure in terms of efficiency, accuracy and rate?'. Reading Research Quarterly. 27. 347-59.

Carver, R. P. (1997). Reading for one second, one minute, or one year from the perspective of raunding theory. Scientific Studies of Reading. 1 (1). 3-43.

Carver, R. P. (1998). Predicting reading level in grades 1 to 6 from listening level

and decoding level: Testing theory relevant to the simple view of reading. Reading and Writing: An Interdisciplinary Journal. 10. 121-154.

Chalhoub-Deville, M. (1997). Theoretical models, assessment framework and test construction. Language Testing. 14 (1). 3-22.

Chalhoub-Deville, M. (Ed.) (1999): Issues in computer-adaptive testing of reading proficiency. Cambridge: Cambridge University Press.

Chalhoub-Deville, M. and Deville, C. (1999). Computer adaptive testing in second language contexts. Annual Review of Applied Linguistics. 19. 273-299.

Chapelle, C. (1988). Field independence: A source of language variation? Language Testing. 7. 121-146.

Chapelle, C. (1998). Construct definition and validity inquiry: Implications for language testing. In L. F. Bachman and A. D. Cohen (Eds.) Interfaces between second language acquisition and language testing research (pp. 32-70). Cambridge: Cambridge University Press.

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

Chapelle, C. and Douglas, D. (1993). Foundations and directions for a new decade

of language testing. In D. Douglas and C. Chapelle (Eds.), A new decade of language testing research: Selected papers from the 1990 language testing research colloquium (pp. 1-22). Alexandria, Virginia: Teachers of English to Speakers of Other Languages (TESOL), Inc.

Chapelle, C., Grabe, W. and Berns, M. (1997). Communicative language proficiency: Definition and implications for TOEFL 2000. TOEFL Monograph Series MS-10. Princeton, New Jersey: Educational Testing Service.

Choi, I. C. and Bachman, L. (1992). An investigation into the adequacy of three IRT models for data from two EFL reading tests. Language Testing. 9 (1). 51-78.

Clapham, C. (1993). Is ESP justified? In D. Douglas and C. Chapelle (Eds.), A new decade of language testing research: Selected papers from the 1990 language testing research colloquium (pp. 257-271). Alexandria, Virginia: Teachers of English to Speakers of Other Languages (TESOL), Inc.

Clapham, C. (1996). The development of IELTS: A study of the effects of background knowledge on reading comprehension. Cambridge: Cambridge University Press.

Clapham, C. (1997). The academic modules: Reading. In C. Clapham and J. C. Alderson (Eds.) Constructing and trialing the IELTS test. IELTS Research

Report 3. London, Cambridge, Canberra: The British Council, University of Cambridge Local Examinations Syndicate, International Development Program of Australian Universities and Colleges.

Clapham, C. (2000). Assessment for academic purposes: Where next? System. 28 (4). 511-521.

Clapham, C. and Alderson, J. C. (1997). International English Language Testing System: Research report 3. London, Cambridge, Canberra: The British Council, University of Cambridge Local Examinations Syndicate, International Development Program of Australian Universities and Colleges.

Clapham, C. and Corson, D. (1997) (Ed.). Encyclopedia of language and education Volume 7: Language testing and assessment. Dordrecht: Kluwer Academic Publishers.

Coady, J. (1979). A psycholinguistic model of the ESL reader. In R. Mackay, B. Barkman and R. R. Jordan (Eds.), Reading in a second language: Hypotheses, organisation and practice (pp.5-12). Rowley, Massachusett: Newbury House.

Cohen, A. (1993). The role of instructions in testing summarising ability. In D. Douglas and C. Chapelle (Eds.) A new decade of language testing research: Selected papers from the 1990 language testing research colloquium (pp.132-160). Alexandria, Virginia: Teachers of English to Speakers of Other Languages (TESOL), Inc.

Cohen, A. (1994). English for academic purposes in Brazil: The use of summary tasks. In C. Hill and K. Parry (Eds.) From testing to assessment: English as an international language (pp.174-204). New York: Longman Group Limited.

Cohen, A. (1998a). Strategies and processes in test taking and SLA. In L. F. Bachman and A. D. Cohen (Eds.) Interfaces between second language acquisition and language testing research (pp.90-111). Cambridge: Cambridge University Press.

Cohen, A. D. (1998b). Strategies in learning and using a second language. New York: Adisson Wesley Longman Limited.

Council of Europe (2000). A common European framework of reference for languages: Learning, teaching and assessment. http://www.culture2.coe.int/portfolio/documents intro/common framework.html

Criper, C. and Davies, A. (1988). ELTS validation project report. ELTS Research report 1(i). The British Council and University of Cambridge Local Examinations Syndicate.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer. and H. I. Braun (Eds.), Test validity (pp. 3-17) Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Cumming, A. (1996). Introduction: The concept of validation in language testing. In

A. Cummin and R. Berwick (Eds.) Validation in language testing (pp. 1-14). Clevedon: Multilingual Matters Ltd.

Cumming, A. and Berwick, R. (Eds.) (1995). Validation in language testing. Clevedon, Avon: Multilingual Matters Ltd.

Cumming, A. and Mellow, D. (1995). An investigation into the validity of written indicators of second language proficiency. In A. Cumming and R. Berwick (Eds.) (1995). Validation in language testing (pp. 72-93). Clevedon, Avon: Multilingual Matters Ltd.

Cushing Weigle, S. and Lynch, B. (1995). Hypothesis testing in construct validation. In A. Cumming and R. Berwick (Eds.) (1995). Validation in language testing (pp. 58-71). Clevedon, Avon: Multilingual Matters Ltd.

Davies, F. B. (1968). Research in comprehension in reading. Reading Research Quarterly. 3. 499-545.

Davies, A. (1997). Demands of being professional in language testing. Language Testing. 14 (3). 328-339.

Deville, C. and Chalhoub-Deville, M. (1993). Modified scoring, traditional item analysis and SATO's caution index used to investigate the reading recall protocol. Language Testing. 10 (2). 117-132.

Douglas, D. (1998). Testing methods in context-based second language research. In L. F. Bachman and A. D. Cohen (Eds.) Interfaces between second language acquisition and language testing research (pp. 141-155). Cambridge: Cambridge University Press.

Douglas, D. (2000). Assessing languages for specific purposes. Cambridge: Cambridge University Press.

Elder, C. (1996). The effect of language background on foreign language test performance. Language Learning. 46 (2). 233-282.

Ellis, R. (1994). The study of second language acquisition. Oxford: Oxford University Press.

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., Schedl, M. (2000): TOEFL 2000 reading framework: A working paper. TOEFL Monograph Series MS-17. Princeton, NJ: Educational Testing Service.

Fortus, R., Coriat, R. and Fund, S. (1998). Prediction of item difficulty in the English subtest of Israel's inter-university psychometric entrance test. In A. J. Kunnan (Ed.). Validation in language assessment: Selected papers from the 17th Language Research Colloquium, Long Beach (pp.61-87). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Fouly, K., Bachman, L. and Cziko, G. A. (1990). The divisibility of language

competence: A confirmatory approach. Language Learning. 40 (1). 1-21.

Freedle, R. and Kostin, I. (1993a). The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: Main idea, inference, and supporting idea items. TOEFL Research Reports No. RR-93-44. Princeton, NJ: Educational Testing Service.

Freedle, R. and Kostin, I. (1993b). The prediction of TOEFL reading item difficulty: Implications for construct validity. Language Testing 10 (2). 133-170.

Fulcher, G. (1996a). Invalidating validity claims for the ACTFL oral rating scale. System. 24 (2). 163-172.

Fulcher, G. (1996b). Does thick description lead to smart tests? A data based approach to rating scale construction. Language Testing. 13 (1). 208-238.

Fulcher, G. (1997). An English language placement test: Issues in reliability and validity. Language Testing. 14 (2). 113-138.

Fulcher, G. (1998): Widdowson's model of communicative competence and the testing of reading: An exploratory study. System. 26. 281-302.

Fulcher, G. (1999a). Assessment in English for academic purposes: Putting content validity in its place. Applied Linguistics. 20 (2). 221-236.

Fulcher, G. (1999b). Ethics in language testing. http://taesig.8m.com/news1.html

Gordon, C. M. and Hanauer, D. (1995). The interaction between task and meaning construction in EFL reading comprehension tests. TESOL Quarterly. 29 (2). 299-324.

Grabe, W. (1991). Current developments in second language reading research. TESOL Quarterly. 25 (3). 375-406.

Grabe, W. (1999). Developments in reading research and their implications for computer-adaptive reading assessment. In M. Chalhoub-Deville (Ed.), Issues in computer-adaptive testing of reading proficiency (pp.11-47). Cambridge: Cambridge University Press.

Grabe, W. (2000). Reading research and its implications for reading assessment. In A. Kunnan (Ed.), Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida. Cambridge: Cambridge University Press.

Grabe, W. and Stoller, F. L. (2002). Teaching and researching reading. London: Pearson Education Limited.

Grant, R. (1993). Strategic training for using text headings to improve students' processing of content. Journal of Reading. 36 (6). 482-488.

Green, A. (1998). Verbal protocol analysis in language testing research: A handbook. Cambridge: Cambridge University Press.

Green, R. and Weir, C. (1998): SPSS guide for language testers. Unpublished manuscript. Centre for Applied Linguistics, University of Reading.

Guerrero, M. D. (2000). The unified validity of the four skills exam: Applying Messick's framework. Language Testing. 17 (4). 397-421.

Guthrie, J. T. and Kirsch, I. S. (1987). Distinctions between reading comprehension and locating information in text. Journal of Educational Psychology. 79. 220-297.

Hammadou, J. (1991). Interrelationships among prior knowledge, inference, and language proficiency in foreign language reading. The Modern Language Journal. 75 (1). 27-38.

Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. Language Testing. 14 (3). 295-303.

Hamp-Lyons, L. & Lumley, T. (2001). Assessing language for specific purposes. Language Testing. 18. (1). 127-132.

Hamp-Lyons, L. and Lynch, B. K. (1998). Perspectives on validity: A historical

analysis of language testing conference abstracts. In A. J. Kunnan, (Ed.). Validation in language assessment: Selected papers from the 17th Language Research Colloquium, Long Beach. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Harley, B., Allen, P., Cummins, J. and Swain, M. (Eds.) (1990). The nature of language proficiency. In B. Harley, P. Allen, J. Cummins and M. Swain (Eds.). The development of second language proficiency. (pp. 7-25) New York: Cambridge University Press.

Hatch, E. and Lazarton, A. (1991). The research manual: Design and statistics for applied linguistics. New York: Newbury House.

Hatcher, L. (1994): A step-by-step approach to using the SAS system for factor analysis and structural equation modeling. Cary, North Carolina: SAS.

Hegarty, M., Carpenter, P. A., Just, M. A. (1991). Diagrams in the comprehension of scientific texts. In R. Barr, M.L. Kamil, P. Mosenthal, and P. D. Pearson (Eds.) Handbook of reading research, Vol II. New York: Longman.

Henning, G. (1987): A Guide to language testing. Cambridge, Massachusetts: New Jersey: Newbury House Publishers.

Henning, G. (1992). Dimensionality and construct validity of language tests.

Language Testing. 9 (1). 1-11.

Henning, G., Anbar, M., Helm, C. E. and D'Arcy, S. J. (1993). Computer-assisted testing of reading comprehension: Comparisons among multiple-choice and open-ended scoring methods. In D. Douglas and C. Chapelle (Eds.) A new decade of language testing research: Selected papers from the 1990 language testing research colloquium (pp.123-131). Alexandria, Virginia: Teachers of English to Speakers of Other Languages (TESOL), Inc.

Henning, G. T., Hudson, T. and Turner, J. (1985). Item response theory and the assumption of unidimensionality for language tests. Language Testing. 2 (2). 141-154.

Hill, C. andParry K. (1992). The test at the gate: Models of literacy in reading assessment. TESOL Quarterly. 26 (3). 433-461.

Hoover, W. A. and Tunmer W. E. (1993). The components of reading. In G. B.Thompson W. E. Tunmer and T. Nicholson (Eds.), Reading acquisition process (pp.1-19). Clevedon: Multilingual Matters Ltd.

Hudson, T. (1991). A content comprehension approach to reading English for science and technology. TESOL Quarterly. 25 (1). 77-104.

Hudson, T. (1993). Testing the specificity of ESP reading skills. In D. Douglas and

C. Chapelle (Eds.) A new decade of language testing research: Selected papers from the 1990 language testing research colloquium (pp. 58-82). Alexandria, Virginia: Teachers of English to Speakers of Other Languages (TESOL), Inc.

Hudson, T. (1996). Assessing second language academic reading from a communicative competence perspective: Relevance for TEOFL 2000. TOEFL Monograph Series. MS-4. Princeton, New Jersey: Educational Testing Service.

Hughes, A. (1988). Introducing a needs based test of English language proficiency into an English medium university in Turkey. In A. Hughes (Ed.), Testing English for university study (pp.134-153). Hong Kong: Modern English Publications and the British Council.

Hughes, A. (1989). Testing for language teachers; Cambridge: Cambridge University Press.

Huhta, A. and Randell, E. (1995). Multiple-choice summary: A measure of text comprehension. In Cumming, A. and Berwick, R. (Eds.) (1996). Validation in language testing (pp. 94-110). Clevedon, Avon: Multilingual Matters Ltd.

ILTA Code of Ethics (2000). http://www.dundee.ac.uk/languagestudies/ltest/ilta/code.pdf. Also in E. Shohamy (2001). The power of tests: A critical perspective on the uses of language tests (pp. 163-171). Singapore: Pearson Education Asia Pte Ltd.

International Test Commission (2000). International Guidelines for Test Use. http://www.intestcom.org

Jafarpur, A. (2003). Is the test constructor a facet?. Language Testing. 20 (2). 57-87.

Johnson, J. L. and Plake, B. S. (1998). A historical comparison of validity standards and validity practices. Educational and Psychological Measurement. 58 (5). 736-753.

Just, M. A. and Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. Psychological Review. 87 (4). 329-354.

Just, M. A. and Carpenter, P. A. (1987). The psychology of reading and language comprehension. Newton, Massachusetts: Allyn and Bacon, Inc.

Just, M. A. and Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. Psychological Review. 99 (1). 122-149.

Khalifa, H. (1997): A study in the construct validation of the reading module of an EAP proficiency test battery: Validation from a variety of perspectives. Unpublished PhD Thesis; Reading University.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. Psychological Review. 95 (2). 163-182.

Kintsch; W. (1994). Text comprehension, memory, and learning. American Psychologist. 49 (4). 294-303.

Kintsch; W. (1998). Comprehension: A paradigm for cognition. New York: Cambridge University Press.

Kintsch, W. and Van Dijk T. A. (1978): Toward a model of text comprehension and production. Psychologigal Review. 85 (5). 363-394.

Kintsch, W. and Yarbrough, J. C. (1982). Role of rhetorical structure in text comprehension. Journal of Educational Psychology. 74 (6). 828-834.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: text organisation and response format. Language Testing. 19 (2). 193-220.

Kunnan, A. J. (1992). An investigation of a criterion-referenced test using G-theory, and factor and cluster analyses. Language Testing. 9 (1). 31-49.

Kunnan, A. J. (1993). The development of a criterion-referenced language test for non-native English speaking graduate students. Journal of English as a Foreign Language. 10 and 11. 118-135.

Kunnan, A. J. (1994). Modelling relationships among some test-taker characteristics and performance on EFL tests: An approach to construct validation. Language Testing. 11 (1). 225-250.

Kunnan, A. J. (1995). Test taker characteristics and test performance: A structural modeling approach. Cambridge: Cambridge University Press.

Kunnan, A. J. (Ed.) (1998a). Validation in language assessment: Selected papers from the 17th Language Research Colloquium, Long Beach. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Kunnan, A. J. (1998b). An introduction to structural equation modelling for language assessment research. Language Testing. 15 (3). 295-332.

Kunnan, A. J. (2000) (Ed.). Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida. Cambridge. Cambridge University Press.

Lee, J. F. and Musumeci, D. (1988). On hierarchies of reading skills and text types. The Modern Language Journal. 72. 173-187.

Lee, T. (1996). Taking a multifaceted view of the unidimensional measurement from Rasch analysis in language tests. In M. Milanovic and N. Saville (Eds.). Performance testing, cognition and assessment: Selected papers from The 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem (pp. 266-275). Cambridge: Cambridge University Press.

Lewkowicz, J. A. (1997). The integrated testing of a second language. In C. Clapham and D. Corson (Eds.), Encyclopedia of language and education, Volume 7: Language testing and assessment (pp.121-130). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Lin, Z. (2002). Discovering EFL learners' perception of prior knowledge and its roles in reading comprehension. Journal of Research in Reading. 25 (2). 172-190.

Linn, R. L., Baker, E. L. and Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher 20 (8). 5-21.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. Educational Researcher. 23 (9). 4-14.

Lumley, T. (1993a). The notion of subskills in reading comprehension tests: An EAP example. Language Testing. 10 (3). 211-234.

Lumley, T. (1993b). Reading comprehension subskills: Teachers' perceptions of

content in an EAP Test. Melbourne Papers In Language Testing. 2 (1). 25-55.

Lunzer, E., Waite, M. and Dolan, T. (1979): Comprehension and comprehension tests. In E. Lunzer and K. Gardner (Eds.), The effective use of reading (pp.37-71). London: Heinemann Educational.

Lynch, B. K. (1997). In search of the ethical test. Language Testing. 14 (3). 315-327.

Lynch, B. K. and McNamara (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessment of the ESL speaking skills of immigrants. Language Testing. 15 (2). 158-180.

Matsumoto, K. (1993). Verbal report data and introspective methods in second language research: State of the art. RELC Journal. 24 (1). 32-60.

Matthews, M. (1990). Skills, taxonomies and problems for the testing of reading. Reading in a Foreign Language. 7 (1). 511-517.

McNamara, T. F. (1990). Item Response Theory and the validation of an ESP test for health professionals. Language Testing. 7 (1). 52-75.

McNamara, T. F. (1991). Test dimensionality: IRT analysis of an ESP listening test. Language Testing. 8 (2). 139-159.

McNamara, T. F. (1995): Modelling performance: Opening Pandora's box. Applied Linguistics. 16 (2). 157-179.

McNamara, T. (1996). Measuring second language performance. New York: Addison Wesley Longman Limited.

Messick, S. A. (1988): The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer and H. I. Braun (Eds), Test validity (pp. 33-45). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Messick, S. A. (1989a): Validity. In R. L. Linn. Educational measurement (pp. 13-103). New York. American Council on Education. Mac Millian Publishing Company.

Messick, S. A. (1989b). Meaning and values in test validation: The science and ethics of assessment. Educational Researcher 18 (2). 5-11.

Messick, S. A. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher. 23 (2). 13-23.

Messick, S. A. (1995a). Validity of psychological assessment: Validation inferences from persons' responses and performance as scientific inquiry into score meaning. American Psychologist. 50 (9). 741-749.

Messick, S. (1995b). Standards of validity and the validity of standards in

performance assessment. Educational Measurement: Issues and Practices. 14 (4). 5-8.

Messick, S. A. (1996). Validity and washback in language testing. Language Testing 13 (2).

Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. Review of Educational Research. 62. (3) 229-258.

Moss, P. A. (1994). Can there be validity without reliability? Educational Researcher. 20 (2). 5-12.

Moss, P.A. (1995). Themes and variations in validity theory. Educational Measurement: Issues and Practices. 14 (2). 5-13.

Moss, P.A. (1996). Enlarging the dialogue in educational measurement: Voices from interpretative research traditions. Educational Researcher. 25 (1).20-28.

Moss, P.A. (1998). The role of consequences in validity theory. Educational Measurement: Issues and Practices. 17 (2). 6-12.

Nevo, N. (1989). Test taking strategies on a multiple-choice test of reading comprehension. Language Testing. 6 (2). 199-215.

North, B. and Schneider, G. (1998). Scaling descriptors for language proficiency scales. Language Testing 15 (2). 217-263.

Norton, B. and Stein, P. (1998). Why the 'Monkey Passage' bombed: Tests, genres and teaching. In A. J. Kunnan (Ed.). Validation in language assessment: Selected papers from the 17th Language Research Colloquium, Long Beach. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Norris, J. M., Brown, J. D., Hudson, T. & Yoshioka, J. (1998). Designing second language performance assessments. Honolulu: University of Hawai'i Press.

Oller, J. W. Jr. (1976). Evidence of a general language proficiency factor: An expectancy grammar. Die Neuren Sprachen. 76. 165-174.

Paris, S. G., Wasik, B. A. and Turner J. C. (1996). The development of strategic readers. In R. Barr, M.L. Kamil, P. Mosenthal, and P. D. Pearson (Eds.), Handbook of reading research, Vol II. (pp. 609-640). New York: Longman.

Perfetti, C. A. (1991). Representations and awareness in the acquisition of reading competence. In L. Rieben and C. A. Perfetti (Eds.) Learning to read: Basic research and its implications (pp. 33-44). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Perfetti, C. A. (1999). Cognitive research and the misconceptions of reading

education. In J. Oakhill and R. Beard (Eds.), Reading development and the teaching of reading (pp. 42-58). Oxford: Blackwell Publishers.

Perkins, K. (1992). The effect of passage topical structure types on ESL reading comprehension difficulty. Language Testing. 9. (2), 163-172.

Perkins, K. and Brutten, S. (1988). An item discriminability study of textually explicit, textually implicit and scriptally implicit questions. RELC Journal. 19. 1-11.

Perkins, K. and Brutten, S. (1992). The effect of processing depth on ESL reading comprehension. Journal of Research in Reading. 15 (2). 67-81.

Peterson, C. L., Caverly, D. C., Nicholson, S., O'Neal, S., Cusenbart, S. (2000). Building reading proficiency at the secondary level: A guide to resources. Texas: Southwest Educational Development Laboratory.

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. Language Testing. 20. (1). 26-56.

Pierce, B. N. (1992). Demystifying the TOEFL reading test. TESOL Quarterly. 26 (4). 665-689.

Pierce, B. N. (1994). The Test of English as a Foreign language: Developing items

for reading comprehension. In C. Hill and K. Parry (Eds.) From testing to assessment: English as an International Language (pp. 39-60). New York: Longman Group Limited.

Pritchard, R. (1990). The effects of multicultural schemata on reading processing strategies. Reading Research Quarterly. 25. 232-49.

Purpura, J. E. (1997). An analysis of the relationship between test takers' cognitive and metacognitive strategy use and second language test performance. Language Learning. 47 (2). 289-325.

Purpura, J. E. (1999): Learner strategy use and performance on language tests: A structural equation modeling approach. Cambridge: Cambridge University Press.

Rayner, K. and Pollatsek, A. (1989). The psychology of reading. Englewood Cliffs, New Jersey: Prentice-Hall.

Ridgway, T. (1997). Thresholds of the background knowledge effect in foreign language reading. Reading in a Foreign Language. 11. 151-68.

Riley, G. L. and Lee, J. F. (1996). A comparison of recall and summary protocols as measures of second language reading comprehension. Language Testing. 13 (2). pp: 173-190.

Rost, D. H. (1993). Assessing the different components of reading comprehension: Fact or fiction. Language Testing. 10. (1) 79-92.

Rupp, A. A., Garcia, P. and Jamieson, J. (2001). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. International Journal of Testing. 1 (3&4). 185-216.

Ryan, K. E. and Bachman, L. F. (1992). Differential item functioning on two tests of EFL proficiency. Language Testing. 9. 12-29.

Samuel, S. J and Kamil, M. L. (1988). Models of reading process. In P. L. Carell, J. Devine and D. E. Eskey (Eds.), Interactive approaches to second language reading. (pp. 22-36). Cambridge: Cambridge University Press.

Sarig, G. (1989). Testing meaning construction: Can we do it fairly?. Language Testing. 6. 77-94.

Saville, N. (2000). Investigating the impact of international language examinations. University of Cambridge Local Examinations Syndicate Research Notes. 2. http://www.cambridge-efl.org/rs_notes/0002/rs_notes2_3.cfm

Sawaki, Y. (2001). Comparability of conventional and computerised tests of reading in a second language. Language Learning & Technology. 5 (2). 38-59.

Schedl, M., Gordon, A., Carey, P. A. and Tang, K. L. (1996). An analysis of the

dimensionality of TOEFL reading comprehension items .TOEFL Research Reports No. RR-95-27. Princeton, NJ: Educational Testing Service.

Shih, M. (1992). Beyond comprehension exercises in the ESL academic reading class. TESOL Quarterly. 26 (2). 289-315.

Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair?. Language Testing. 14 (3). 340-349.

Shohamy, E. (2001). The power of tests: A critical perspective on the uses of language tests. Singapore: Pearson Education Asia Pte Ltd.

Shohamy, E., Donitsa-Schimidt, S. and Ferman, I. (1996). Test impact revisited: Washback effect over time. Language Testing. 13 (2). 298-317.

Shohamy, E. and Inbar, O. (1991). Validation of listening comprehension tests: The effect of text and question type. Language Testing. 8 (1). 23-40.

Skehan, P. (1991). Progress in language testing: the 1990s. In J. C Alderson and B. North (Eds.) Language testing in the 1990's: The communicative legacy (pp. 3-21). London: Modern English Publications and The British Council.

Skehan, P. (1998). A cognitive approach to language learning. Hong Kong: Oxford University Press.

Skehan, P. and Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.) Cognition and second language instruction (pp.183-205). Cambridge: Cambridge University Press.

Smith, F. (1994). Understanding reading: A psycholinguistic analysis of reading and learning to read. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Sparks, R. L., Artzer, M., Ganschow, L., Siebenhar, D., Plageman, M. and Patton, J. (1998). Differences in native-language skills, foreign-language aptitude, and foreign-language grades among high-, average-, and low-proficiency foreign-language learners: Two studies. Language Testing. 15 (2). 181-216.

Spearitt, D. (1972). Identification of sub-skills of reading comprehension by maximum likelihood factor analysis. Reading Research Quarterly. 8. 92-111.

Spolsky, B. (1997). The ethics of gatekeeping tests: What have we learned in a hundred years?. Language Testing. 14 (3). 242-247.

SPSS Base 10.0 User's Guide (1999): Chicago: SPSS Inc.

Stanovich, K. E. (1991). Changing models of reading and reading acquisition. In L. Rieben and C. A. Perfetti (Eds.) Learning to read: Basic research and its implications (pp. 19-31). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Stanovich, K. E. (1996). Word recognition: Changing perspectives. In R. Barr, M.L. Kamil, P. Mosenthal, and P. D. Pearson (Eds.), Handbook of reading research, Vol II (pp. 418-452). New York: Longman.

Stanovich, K. E. (2000). Progress in understanding reading: Scientific foundations and new frontiers. New York: Guilford Press.

Stanovich, K. E. and Stanovich, P. J. (1999). How research might inform the debate about early reading acquisiton. In J. Oakhill and R. Beard (Eds.), Reading development and the teaching of reading (pp. 12-41). Oxford: Blackwell Publishers.

Stemmer, B. (1992). An alternative approach to C-test validation. In R. Grotjahn (Ed.). Der C-test. Theoretische Grundlagen und Praktische Anwendungen. Bd. 1 (pp. 97-144). Bochum: Brockmeyer.

Storey, P. (1997). Examining the test-taking process: A cognitive perspective on the discourse cloze test. Language Testing. 14 (2). 214-231.

Tarone, E. (1998). Research on interlanguage variation. In L. F. Bachman and A. D. Cohen (Eds.) Interfaces between second language acquisition and language testing research (pp. 71-89). Cambridge: Cambridge University Press.

Tharu, J. (1993). Tests of English proficiency: The problem of standards. Journal of English as a Foreign Language 3/ 4. 59-78.

Upshur, J. A. and Turner, C. E. (1995). Constructing rating scales for second language tests. ELT Journal. 49 (1).

Urquhart, S. and Weir, C. (1998). Reading in a second language. New York: Longman.

Van Dijk, T. A. (1977). Semantic macro-structures and knowledge frames in discourse comprehension. In M. A. Just and P. A. Carpenter (Eds.), Cognitive processes in comprehension (pp.3-32). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Verhoeven, L. (1992). Assessment of bilingual proficiency. In L. Verhoeven and J. H.A.L. de Jong (Eds.) The construct of language proficiency (pp. 125-146) Amsterdam: John Benjamins Publishing Company.

Wall, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. Language Testing. 13 (2). 334-354.

Wall, D., Clapham, C. and Alderson, J.C. (1991). Validating tests in difficult circumstances. In J. C. Alderson and B. North (Eds.) Language testing in the 1990's: The communicative legacy (pp. 209-225). London: Modern English Publications and The British Council.

Wall, D., Clapham, C. and Alderson, J.C. (1994). Evaluating a placement test.

Language Testing. 11. 321-344.

Weir, C. (1988). The specification, realisation and validation of an English language proficiency test. In A. Hughes (Ed.), Testing English for university study (pp. 45-88). Hong Kong: Modern English Publications and the British Council.

Weir, C. (1993). Understanding and developing language tests. Hemel Hempstead: Prentice Hall International Ltd.

Weir, C. J. (1999, April 13): Lecture at Boğaziçi University, The School of Foreign Languages.

Weir, C. (2001). The formative and summative uses of language test data: Present concerns and future directions. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, and K. O'Louglin, (Eds.), Experimenting with uncertainity: Essays in honour of Alan Davies (pp. 117-125). Cambridge: Cambridge University Press.

Weir, C. J. , Hughes, A. and Porter, D. (1990). Reading skills: Hierarchies, implicational relationships and identifiability. Reading in a Foreign Language. 7. 505-10.

Weir, C., Huizhong, Y. and Yan, J. (Eds.) (2000). An empirical investigation of the componentiality of L2 reading in English for academic purposes. Cambridge: Cambridge University Press.

Weir, C. J. and Porter, D. (1994). The multidivisible or unitary nature of reading: The language tester between Scylla and Charybdis. Reading in a Foreign Language. 10. 1-19.

Widdowson, H. G. (2001). Communicative language testing: The art of the possible. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara and K. O'Louglin (Eds.) Experimenting with uncertainity: Essays in honour of Alan Davies. Cambridge: Cambridge University Press.

Wijgh, I. F. (1995). A communicative test in analysis: Strategies in reading authentic texts. In A. Cumming and R. Berwick (Eds.) Validation in language testing (pp.154-170). Clevedon, Avon: Multilingual Matters Ltd.

Wolf, D. F. (1993). A comparison of assessment tasks used to measure FL reading comprehension. The Modern Language Journal. 77 (4). 473-489.

Wu, W. M. and Stansfield, C. W. (2001). Towards authenticity of task in test development. Language Testing. 18 (29). 187-206.

Yong-Won, L. (2000). Identifying suspect item bundles for the detection of differential bundle functioning in an EFL reading comprehension test: A preliminary study. In A. J. Kunnan (Ed.), Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida. (pp. 105-127). Cambridge. Cambridge University Press.

Yong-Won, L. (2004). Examining passage-related local independence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. Language Testing. 21 (1). 74-100.

Young, R., Shermis, M. D., Brutten, S. R. and Perkins, K. (1996). From conventional to computer-adaptive testing of ESL reading comprehension. System. 24 (1). 23-40.

APPENDIX 2.1

Framework for Conducting a Strong Program of Construct Validation (Benson, 1998)

Substantive Stage
Theory-Based (including previous research and observation)
Generate theoretical and empirical definitions
Gather content-related evidence
Consider construct underrepresentation and construct irrelevancy

Structural Stage
Item/Subscale intercorrelations
Exploratory factor analysis
Confirmatory factor analysis
Generalisability theory
Multitrait-Multimethod matrix
Item response theory (including differential item functioning)

External stage
Multitrait-Multimethod matrix
Group differentiation
Experimental manipulation
  Existing or known groups
  Experimental manipulation
Correlations of tests with other tests (including criterion-related evidence)
Structural equation modelling

APPENDIX 2.2

Language Assessment Research Themes in Messick's Framework (Kunnan, 1998)

	Test Interpretation	Test Use
Evidential Basis	1. Proficiency components 2. Test dimensionality 3. Test-validation process 4. Test development: New test methods, rating scales, conditions, etc.	1. Test-taking processes 2. Test-taking strategies 3. Test-taker characteristics: Academic background, native language and culture, field in/dependence, differential item functioning (DIF) studies: native language and culture, gender, ethnicity, age, etc.
Consequential Basis	1. Value system differences: Test-taker and specialists' feedback	1. Social consequences and washback 2. Ethics, standards and equity 3. Alternatives

APPENDIX 2.3

Skill Taxonomies[1]

1. Davies (1968)

- Recalling word meanings
- Drawing inferences about the meaning of a word in context
- Finding answers to questions answered explicitly or in paraphrase
- Weaving together ideas in the content
- Drawing inferences from the content
- Recognising a writer's purpose, attitude, tone and mood
- Identifying writer's technique
- Following the structure of a passage

2. Lunzer et al. (1979)

- Word meaning
- Words in context
- Literal comprehension
- Drawing inferences from single strings
- Drawing inferences from multiple strings
- Interpretation of metaphor
- Finding salient or main ideas
- Forming judgements

3. Munby (1978)

- Recognising the script of a language
- Deducing the meaning and the use of unfamiliar lexical items
- Understanding explicitly stated information
- Understanding when not explicitly stated
- Understanding conceptual meaning
- Understanding the communicative value of sentences
- Understanding the relations within the sentence
- Understanding relations between parts of the text through lexical cohesion devices
- Interpreting text by going outside it
- Recognising indicators in discourse
- Identifying the main point of information in discourse
- Distinguishing the main idea from supporting detail
- Extracting salient points to summarise
- Selective extraction of relevant points from a text
- Basic reference skills
- Skimming
- Transcoding information to diagrammatic display

---

[1] Hughes (1989), Weir (1993) and Urquhart and Weir (1998) are exclusively EAP reading taxonomies

4. Hughes (1989)

Macro-skills

- Scanning text to locate specific information
- Skimming text to obtain the gist
- Identifying the stages of an argument
- Identifying examples presented in support of an argument

Micro-skills (underlying skills)

- Identifying referents of pronouns
- Using context to guess meaning of unfamiliar words
- Understanding relations between parts of text by recognising indicators in discourse, especially for the introduction, development, transition, and conclusion of ideas

5. Grabe (1991) (cited above)

1. Automatic recognition skills
2. Vocabulary and structural knowledge
3. Formal discourse structure knowledge
4. Content/world background knowledge
5. Synthesis and evaluation skills
6. Metacognitive knowledge and skills monitoring

6. Weir (1993)

1. **Scanning** for specific information in a text in order to

1.1 locate the specific parts of a text you are going to read, e. g. by using table of contents, key words, etc.

1.2 locate specific data encoded in a non-verbal form, e.g. in diagrams, graphs, etc.

1.3 locate specific data in reference works, e.g. words in dictionaries.

1.4 retrieve data already encountered during reading, e.g. to check a spelling, quotation, etc.

2. **Skimming** a text or parts of a text to quickly establish a general idea of the content in order to

2.1 help you to anticipate what it might contain and therefore better understand it when you read it more carefully.

2.2 decide whether it should be read before other texts because it is more appropriate, clearly written, concise, etc.

2.3 determine how much of it is relevant for your purpose(s) and should be read carefully.

2.4 review what you have already read in order to recall or clarify the main purpose.

3. **Reading** a text or part(s) of a text carefully to extract all the relevant information for the following purposes:

3.1 to carry out a written assignment e.g. dissertation, coursework.

3.2 to present a paper orally in a seminar.

3.3 to answer examination questions.

3.4 to correct your own written work.

4. **Reading** a text or part(s) of a text for background knowledge
4.1 of topic(s) covered by the course, e.g. as pre-course reading or preparation for lectures/seminars, etc.
4.2 of topic(s) related to but not covered by the course, e.g. keeping up-to-date with new developments in your field.

7. Urquhart and Weir (1998)

	Global	**Local**
Expeditious	**A.** Skimming quickly to establish discourse topic and main ideas. Search reading to locate quickly and understand information relevant to predetermined needs.	**B.** Scanning to locate specific information; symbol or group of symbols; names, dates, figures or words.
Careful	**C.** Reading carefully to establish accurate comprehension of the explicitly stated main ideas the author wishes to convey; propositional inferencing.	**D.** Understanding syntactic structure of sentence and clause. Understanding lexical and/or grammatical cohesion. Understanding lexis/deducing meaning of lexical items from morphology and context.

APPENDIX 3.1
THE TESTS

THE BUEPT READING TEST: The September Test – Released Version

Both the trial and September versions of the reading module of the BUEPT include the reading sections listed below. However, only the sections printed here could be released. For the copies of the rest of the test, the Testing Office of Boğaziçi University, School of Foreign Languages should be contacted.

Scanning:	The Pinch of Salt Solution
Search Reading:	New Perspectives on Child Education
	The Role of Immigrants in France
Careful Reading:	Can Animals Learn to Share, Cooperate, Punish and Show Empathy?
	Readiness for Learning

TEXTS

SCANNING: THE PINCH OF SALT SOLUTION

A United Nations Children Fund (UNICEF) survey carried out in Laos in 1993 revealed some shocking results. Ninety per cent of residents living in one community in the remote northern Luang Namtha province were found to be suffering from goitre – a swelling of the neck caused by the enlargement of the thyroid gland.

In most cases, goitre is the result of a lack of iodine in the diet, which is necessary for the production of thyroid hormones and essential for normal growth and development. When iodine is lacking, the thyroid gland enlarges in an attempt to increase hormone production. In milder cases of iodine deficiency, individuals feel sluggish and may have a reduced IQ. In the worst cases, a lack of iodine can result in cretinism – mental retardation, physical stuntedness, deafness muteness, paralysis – and can be fatal. Iodine-deficient women are more prone to miscarriages and stillbirths, and even if they give birth to an apparently healthy baby, there is a risk that it will die in infancy and that other disorders, such as poor eye-hand coordination, speech impediments or other neuromuscular disorders, may develop in early childhood.

Since food is the major source of iodine (water contains minute amounts), iodine deficiency tends to occur in areas where the soil lacks iodine – especially in hilly or flood-prone regions where iodine tends to be washed out of the soil – and in areas where the population does not have a varied diet or access to seafood, which is the richest source of iodine. In the West, where iodine is added to cattlefeed, dairy products provide most of the body's iodine requirements.

The landlocked position of the Lao People's Democratic Republic, as Laos is officially called, and its distance from the oceans where iodine is in such abundance, makes its population particularly vulnerable to iodine deficiency disorder (IDD). The heavy monsoon rainfalls from May to September leach iodine from the soils of

its largely mountainous terrain, and the poor and isolated rural population's staple diet of glutinous rice and little imported or processed food provides little of the micronutrients that the body needs. In addition, the Lao government, which is currently in the process of decentralising its economy, has limited resources to spend on the health care of its 4.3 million people, the vast majority of whom depend on subsistence farming.

The UNICEF surveys carried out two years ago indicated that Laos had the worst IDD rates in the world. As well as finding that 92 per cent of women, men and children in one Luang Namtha community had goitre, 32 per cent of adult factory workers in the capital Vientiane were found to be suffering from the disorder. Many adults in the city bear several scars around their necks, the legacy of successive goitre operations necessitated by iodine deficiency.

A team of technicians also tested the iodine content of urine from over 2,000 children in randomly selected schools across the country. This revealed that approximately 65 per cent of children were suffering from the most severe form of IDD and that only five per cent had a normal iodine intake.

The World Health Organisation (WHO) classifies goitre rates of 30 per cent as a 'severe' problem, 20 to 30 per cent as 'moderate', and five to 20 per cent as 'mild'. The global scale and severity of IDD has only recently been acknowledged. Today, WHO and UNICEF estimate that nearly 1.6 billion people – about 30 per cent of the world's population – are at risk of IDD, and that some 655 million people worldwide suffer from goitre, the most obvious sign of IDD. Every year, iodine deficiency in pregnant women is responsible for at least 60,000 miscarriages or stillbirths, and over 120,000 cases of cretinism. Even when born normally, young children whose diet is low in iodine may suffer from reduced intelligence. In this way, iodine deficiency locks entire communities into a cycle of poverty and underdevelopment, hampering economic progress in many developing nations.

But the solution – iodizing all salt supplies – is relatively simple and cheap. By giving people microscopic amounts of iodine, the equivalent of a teaspoon over a lifetime, IDD can be avoided and average intelligence boosted. One of the cheapest methods is to add iodine to the one commodity that is consumed by everyone – common salt. According to UNICEF estimates, the cost of salt iodization is just five US cents (about three pence) per person per year. In 1920, this method was successfully adopted by most industrialised countries, led by Switzerland and the USA, to eradicate the problem. Unfortunately, the developing world missed out.

In 1990, at the World Summit for Children in New York, WHO and UNICEF confronted political leaders about the serious problem of iodine deficiency in developing nations. Governments at the conference agreed to a programme that would 'iodize all salt for human and animal consumption in all countries where iodine deficiency disorders are a public health problem'. The year 2000 was set as the target date for the global elimination of IDD, while countries subsequently agreed to iodize at least 95 per cent of all salt supplies in each country by the end of 1995.

Of the 94 countries with IDD problems, 60 per cent are on track to have iodized 95 per cent of their salt supplies by the end of this year, and another 32 countries could

achieve this target shortly with 'an accelerated effort', says UNICEF. While Bangladesh, China, India and Tanzania, with nearly half of the world's people at risk, may not achieve this year's goal, they are very close to eradicating the problem. WHO and UNICEF are reasonably confident that IDD can be eradicated worldwide within the next two or three years.

Iodizing salt is comparatively straightforward when it can be done at a single location. In Syria, for instance, the procedure is simple because the Ministry of Industry is the sole producer of salt. Similarly, landlocked Bhutan, which imports all its salt from India, has seen child goitre rates fall from 60 per cent to 25 per cent since an iodizing plant was installed at the border a decade ago.

But when salt is produced by many small-scale entrepreneurs, or imported from a variety of sources, salt iodization becomes more complex. Bolivia successfully tackled the problem by setting up a company to publicly promote the benefits of iodized salt. As demand grew, 35 salt-producing companies iodized their stocks with subsidies provided by the government to keep prices down. Today, over 80 per cent of Bolivians have access to iodized salt, and the goitre rate has dropped to 20 per cent, a third of its original level.

In Bangladesh, where IDD stems not from inaccessibility to the sea but from the monsoons and floods that leach iodine from the soils, the government is supplying free iodizing machines and packaging equipment to all 265 of the country's salt-crushing factories. And in India, which has more than 10,000 small-scale salt producers, an advertising campaign has been launched to teach people the benefits of paying slightly more for iodized salt.

A relatively new iodizing technique is being used in the Central African Republic. Situated over 1,000 kilometres away from the nearest coast, goitre affects more than 60 per cent of its population. 'Diffusers' containing candle-like sticks of solidified iodine, which take a year to gradually dissolve, are being placed deep into the water source below village pumps.

In Laos, where only six factories produce the majority of the country's salt, iodizing all its salt supplies is a feasible goal. However, it will require proper legislation, enforcement, and public education. Capsules of iodized oil will be given to women of childbearing age living in high-risk areas to protect their children from cretinism and other defects, until iodized salt becomes widely available in rural villages.

The main salt factory east of Vientiane, which uses log fires to evaporate salt from brine that is pumped up from a nearby spring, has already installed a salt iodization spray-mix plant, a facility that is currently underutilised. By using new and cheaper packaging that still retains the salt's iodine content, the manufacturers hope to boost demand by producing more affordable iodized salt.

Laos' six major salt works produce more than half of the annual 24,000 tonnes of salt required for human and animal consumption. But these installations have the capacity to increase production to meet 98 per cent of the country's salt requirements. But until the iodization of all domestically produced salt becomes

mandatory, an information campaign in schools and through the media is educating people about the importance of iodine supplementation.

The cost of the salt iodization programme is small. UNICEF calculates that IDD could be eradicated in Laos for approximately US$500,000 spent over three years (approximately US$50,000 is currently spent every year on goitre operations in Laos) and the cost of sustaining the programme thereafter would be minimal – priceless in terms of human development.

SEARCH READING: NEW PERSPECTIVES ON CHILD EDUCATION

There are various notions of the place of media in the education of young children. At one extreme, some McLuhan enthusiasts might propose that all the world's people, including infants and young children, are really transistors being bombarded constantly by electrons, in which case all we need to do to turn on the young learner is develop the right circuitry. Others would argue that the young learner is such a fragile bit of humanity that the main prerequisite for learning is a personal and intimate relationship with other people, particularly his parents, his family, and his peer group; since media have nothing to contribute to positive personal feelings, they have no place in learning systems. We suggest that the best answer lies somewhere between these two extremes.

There are several bases for the proposals made later on as to what media can contribute to the education of younger children. First, there is a difference between mass media and media developed for learning purposes. This is not to argue that mass media messages fail to affect behavior. Obviously, advertising especially is effective in doing this. But commercial programs are seldom designed to produce the type of behavior educators are interested in. In this article the word media will be shorthand to signify stimuli designed to produce a certain type of educationally approved behavior.

Second, this article is built on the premise that all young children, including those who are disadvantaged, need experience which requires some restructuring on the part of the individual child. Play and creative activities which assist the child to develop self-discipline are necessary, but such activities alone are insufficient. It is necessary to push beyond natural growth stages in order to prepare for genuinely demanding intellectual tasks. This is not an endorsement of the so-called "pressure-cooker" approach; but something more than a neutral position is required concerning the stimulation and direction of a child's intellectual development.

Third, there is the distinct possibility that all cultures place far too much emphasis on verbal symbols. In fact, some of the developing nations have become aware that literacy programs can lead to discrimination between those who possess the symbols and those who do not. The writers accept the stress on language which many early childhood education specialists emphasize, but do so with some hesitancy. Man's future developments may place much greater reliance on capacity to manipulate a wide range of symbols, especially mathematical symbols, rather than, as now is the case, mainly verbal ones.

The place of media in developing more fully the capacity for handling of all types of symbols is obvious. We turn to some of the research and theory which point to the need for the in-depth study of media as they relate to the education of the child from birth through age five.

Media and the Very Young

Birth to 15 months: A number of research studies have been performed on animals featuring an almost complete absence of perceptual stimulation. Such deprivation at early stages has a permanent deleterious impact on the development of the animal. As attention has increasingly focused on the very young child in our society, many types of study have been made. For example, orphanage babies have been placed for some time in cribs covered with white sheets. When the cribs were uncovered, the infants were able to see only a white and figureless ceiling. Children reared in such an environment were found to be very much slower in developing the capacity to differentiate among various kinds of visual stimuli so necessary for beginning to read.

Studies have been carried on with infants where bumping or striking an apparatus caused movement of colorful items in the crib or threw various kinds of flashing lights on the ceiling. Infants repeatedly triggered the mechanism which produced the visual stimuli. Subsequent studies indicated the superiority of the subjects in tasks requiring visual discrimination.

Various types of experience are necessary, therefore, even in the earliest days after birth. Infants should probably be propped before television sets even if only commercial programs are being telecast in order to give appropriate visual stimulation. Better yet, special telecasts could be prepared which would deal not only with colors, shapes, and forms but textures as well.

Furthermore, attention should be given to affixing apparatus to the crib which can be activated by the swing of an arm or the kick of a foot. These mechanisms might produce static or moving visual displays immediately above the crib or on the ceiling.

In the earliest months of life, then, it would appear that much more attention should be given to providing visual stimuli than is now the case. Although much of early human life is given to the refinement of physical skills, there is strong evidence that a base is already being developed immediately following birth for future intellectual activity.

The Pre-linguistic Stage: A child begins to respond to language long before he can speak, of course. In order for the child to speak his native language with facility at a later stage, he must have exposure to spoken language at all stages, but particularly from one to three years of age. The earlier a child is exposed to a great deal of talk and conversation, the more he intuitively masters linguistic patterns.

In has been known for some time that reading to an infant helps him to learn to read early. It is pretty well accepted by researchers that the disadvantaged child is often slow to read because he fails to separate the "talk" which surrounds him from the

other noises. Some research indicates that as much talk goes on in the disadvantaged as in the advantaged home, but the infant's attention is not focused on the human communication.

Other research indicates that mechanical devices are valuable in exposing an infant to linguistic patterns. Again, television can be used at this stage. Radio, particularly FM with its greater clarity, could serve. Even more to the point is the use of records and transcriptions, and most of all of tapes. With the advent of low-cost cassette tape playback machines and their ease of recording, it is even possible for a family to record its conversations for replay to provide additional contact with familiar voices and consistent linguistic patterns.

Aural discrimination training begun at birth should be continued into year one, and the period from one to three should make generous provision for methodical linguistic input.

Two and a half to Five: From birth to three years was a period in the child's life when major activities were predominantly of an internal nature. Therefore, media which require a limited restructuring of stimuli by the child could be used. Beginning around three, the child requires an environment which is interactive in nature, feeding back in such a way that he must begin to manipulate abstractions in order to develop his own cognitive structure.

Because developments in the cognitive area require interaction, either human or through media, it can be argued that some type of more definitely structured schooling should be initiated around age three to complement what the home should do. In many instances the school may find it necessary to do the entire job. The type of material and equipment needed will be similar, whether education is carried on in the home or in the school.

The media chosen for cognitive development presumably will need to be of a programmed type, so that a response made by the child to visual or auditory stimuli will present alternatives to which the child will make further responses. Most of the equipment currently available for young children tends to be one-way presentation devices. Needed additions to such equipment will display colored visuals on a small screen and auditory stimuli by means of a record. Further, they will require the child to make certain responses; when a response is correct, further new material will be presented. At this stage of human learning it is evident that equipment, along with the required programs, becomes more complex and thus more expensive. It also becomes evident that breakthroughs in the improved use of equipment are necessary. For example, such pieces of equipment as the computer are excellent for model cognitive development. The potential of the computer for interaction and interplay may, for all practical purposes, be virtually unlimited.

One of the manufacturers of children's toys has marketed a device which requires that the child assemble blocks in a certain order to produce a correct linguistic pattern in audio form. Greater attention needs to be given to the development of such toys, games, and instructional materials. They require the active participation of the child in developing various patterns related to cognitive structure.

One of the newer developments in school construction is the creation of an environment which is related to what is being learned. For example if a teacher wishes to create the environment in which an Eskimo lives, she merely pushes a button. The appropriate setting is called up from the computer bank and visually "sprayed" on the walls of the classroom. It is also possible to provide the auditory stimuli to accompany the visual setting and, if required, the odors as well! There is no reason why children should not be able to call up from the data bank any setting which they desire. Thus the schoolroom of the future is likely to provide a total environment for young children enabling them and their teachers to create a psychedelic setting in order to expedite learning.

In addition to initiating a cognitive structure, the child at this stage also begins to recognize himself as an individual and begins to sense his status and role in regard to others with whom he associates. For the development of the affective area, interaction plays an important role at this stage as well. The child begins to understand that others expect much or little of him, that they accept or reject him, that they believe in him or doubt his capabilities. There is a need, therefore, to begin to give attention to the affective area - the child's feelings. Otherwise, the child's self-image and ego will weaken through lack of positive attention.

The child of this age should see himself and others in the best possible way. First of all, he needs to know how he appears, as well as the features of his friends. The use of mirrors should probably be more widespread in schools for such small children. In addition, the teacher should make more frequent use of the 35mm and 8mm cameras and of tape recorders to help the young children to know themselves and to gain confidence in their own abilities. Further, the camera and the tape recorder might become the child's instruments for capturing the world, confirming and reinforcing his own ideas and enabling him to communicate his ideas to others. Based on such information teachers can begin to make preliminary identification of those individuals who may need attention of various kinds.

Conclusion

Enough is known about the significance of the first several years of life for later intellectual, physical, and emotional development to warrant major research and the study of development needs during these years. Nothing completely takes the place of the warmth, concern, and activity of a good home. However, even in a good home, and later in a good nursery school, provision for the appropriate use of media must be made if the fullest cognitive and affective development of each child is to be achieved. This article attempts to suggest a few of many approaches which might be utilized. Much more special equipment will need to be developed for use with children from birth to five, because most of what we now have was developed for older children. Special equipment and material designs to evoke particular behaviors will certainly make possible greater leaps forward during this stage in the child's life, which only now is beginning to be recognized as his most critical and formative years.

CAREFUL READING: CAN ANIMALS LEARN TO SHARE, COOPERATE, PUNISH, AND SHOW EMPATHY?

In a zoo in Chicago, a 3-year-old boy fell into a gorilla enclosure and was knocked unconscious. Within moments, Binti Jua, a female gorilla picked up the boy, and gently put him down in front of the caretaker's door.

Most reports suggested that Binti rescued the boy because she felt empathy for him. Although there is no ambiguity about what the gorilla did, there are a lot of questions about why. Was she concerned about his well-being? Would she have acted in the same way toward a conscious boy, a cat, or a bag of potato chips?

Studies have shown that children don't fully grasp the distinction between a dead being and a live one until they are almost 10 years old. And to date, no study of ape intelligence comes close to showing that gorillas have the mental sophistication of a 10 year-old human. We can only guess why Binti did what she did. And one incident is not enough to warrant conclusions. But Binti's actions do raise the public and scientific interest as to what degree other animals possess the mental traits humans have. Can other creatures share, cooperate, punish cheaters, show empathy, and act altruistically?

In a 1988 study, the ethologist Stammbach set up an experiment with macaque monkeys in order to test their ability to rein in aggressive behavior and act cooperatively. First each monkey was trained to press a lever on a machine so as to receive a popcorn treat and then subgroups were created. A low-ranking member in each subgroup was trained to press a set of levers in a specific sequence that caused the machine to deliver enough popcorn for three individuals. During the training, the machine began releasing popcorn only to the low-ranking specialist.

At first, high-ranking individuals threatened low-ranking individuals to keep them away from the dispenser altogether. Then the high-ranking individuals learned that the low-ranking individuals had a unique skill, so they followed them to the machine and waited to grab all the popcorn. Before long the low-ranking specialists stopped operating the machine. But their strike didn't last long. Some higher-ranking individuals changed their behavior. Rather than chasing specialists away or eating all their popcorn, they began to inhibit their aggression. They approached peacefully and allowed the lower-ranking specialists to eat a portion of the popcorn. However, this change in behavior had no impact on their dominance rank within the group. Specialists kept their low rank but were allowed a moment at the high table when their skills were of use to the royalty.

Other experiments have found that monkeys even have a rudimentary sense of ownership and respect for property. The space that a territory owner defends is like its property, and an intruder's respect reveals its acknowledgement of ownership and property rights. In a 1991 study, ethologists Kummer and Cords tested macaques that had something other macaques wanted: a see-through tube filled with raisins. The tube was either fixed to a wall or freestanding. If it was freestanding, it was attached to a long or a short piece of rope, or no rope at all. A subordinate animal was allowed

first crack at the tube in all the various placements. Then researchers observed how the more dominant individuals reacted. Consistently, dominants took ownership of fixed tubes more often than free tubes, and took over free tubes when the subordinates failed to carry them. Staying close to the tube and looking at it were not sufficient cues of ownership from the dominant's perspective. A dominant macaque would appear to inhibit its impulse to grab the tube if a subordinate held it close to its body. Here, then, is an intriguing example of how inhibition plays a crucial role in maintaining social conventions among monkeys.

But in any social situation with conventions, individuals often find that it pays to break the rules. Would such rule-breakers be punished? To explore this possibility, an experiment was done on some rhesus monkeys. Unlike macaques, which don't share food, rhesus monkeys tend to call out when they find food. In the study, lone individuals were presented with a small stash of food. Their first response was to look around, presumably to decide if there were enemies near. A few individuals waited and then moved cautiously toward the food. Only half the discoverers called out. When they were detected by other group members, some were aggressively attacked. The initial suspicion was that those who were being attacked were lower-ranking than those who were not. Surprisingly, both high and low-ranking individuals were attacked. Whether or not they were attacked seemed to depend on their vocal behavior. Silent discoverers who were caught with food were attacked more often and more severely than those who cried out. It was as if individuals were being punished for being inappropriately silent, for deceptively withholding information about a rich food source.

Thus research indicates that animals can inhibit their impulses and punish those who violate community rules. But what about empathy? What about Binti? Unless we can establish that animals understand the thoughts and feelings of others, we cannot assume that their behavior is moral. Codes of moral behavior are founded on beliefs of right and wrong. How we form those beliefs is based on an idea of justice, a consideration of how particular actions affect others. And to understand how our behavior affects others requires empathy.

One experiment was designed by Miller to see if a monkey could interpret another monkey's facial expression, a presumed indicator of emotion. First, a researcher trained rhesus monkeys to pull a lever to avoid getting shocked after hearing a specific sound. Then one of the monkeys - the "actor" - was put in a room with a lever and a live television image of a second animal - the "receiver" - that was both out of sight and earshot. The receiver was exposed to the sound that indicated a shock was coming but lacked a lever to avoid it.

The assumption underlying this experiment was that the receiver would hear the sound, anticipate the shock, and show fear on its face. If the actor understood the receiver's facial expressions, then it would use this information to pull its lever. If the actor failed, both animals received a shock. Because shock trials were presented randomly, and neither animal could hear the other, there was no way to predict the timing of a response except by using the receiver's image in the monitor. As it turned out, the actor pulled the lever significantly more when the receiver heard the sound. The actor was able to read the receiver's facial expressions. Moreover, the animals

seemed to behave cooperatively: to avoid the shock, the receiver gave a signal and the actor read the receiver's signal.

Did the receivers intend to provide information to the actors? Was this a cooperative effort? The receivers, must have felt helpless and afraid. But to establish that they were signaling the actors, one would have to demonstrate that they were aware of the actors' presence. And, given the design of the experiment, they certainly were not. Rather, each receiver's response was elicited by the sound. It seems likely that the actors picked up on a change in the activity of the receivers, one that was consistent enough to predict the shock. But using an expression to predict a response is not the same as seeing the expression as an indication of another's emotions at the time.

This experiment left many loose ends. Although it is clear that rhesus monkeys can learn to avoid shock by attending to a facial expression, we don't know if this response is motivated by empathy, which is necessary for altruism. One has to feel what it would be like to be someone else. We don't know whether the actors were even aware of the receivers' feelings. From the actors' perspective, all that mattered was that the image displayed on the video monitor functioned as a reliable predictor of shock. A better experiment would have allowed the actors to see what was happening to the receiver but restrict the shock to the receiver alone.

In a 1964 study, Masserman ran a different experiment, again with rhesus monkeys. An actor was trained to pull one of two chains to receive its food in response to a brief flash of blue or red light. Next, a receiver was housed nearby, where the actor could see it. The experimenter then changed the consequences of responding to the color of the flash. Pulling in response to one delivered food; pulling in response to the other delivered both food to the actor and a severe shock to the receiver. Most actors pulled the chain delivering the shock far less often than the chain delivering food only. Two of the 15 actors even stopped pulling both chains for between 5 to 12 days. When the actors were paired with new receivers, most continued to refrain from pulling the chain that delivered the shock. And pairs that knew each other well tended to show more altruistic behavior than pairs that were unfamiliar.

What is remarkable about this experiment is that some monkeys refrained from eating to avoid injuring another. Perhaps the actors empathized, imagining what it would be like to receive the shock. Alternatively, perhaps seeing another monkey grimace in pain is unpleasant or threatening, and rhesus monkeys will do whatever they can to avoid unpleasant conditions. Or perhaps the actor worried that one day it might be the recipient of a shock. Although refraining from eating appears to be a response of empathy or sympathy, it may actually be a selfish response.

As the experiments show, animals are by no means robots driven solely by instinctual responses. They are sensitive to their social and ecological environments, and under certain conditions they can inhibit one response and favor another. Moreover, they can punish others and sometimes alleviate another's pain. But no experiment to date has provided evidence that animals are aware of others' beliefs or intentions. And without such awareness, there can be no ethical judgment.

QUESTIONS, ANSWERS AND INSTRUCTIONS: PILOT VERSION

The answers are given in boldface.

Scanning: The Pinch of Salt Solution

* This part of the Reading Test is aimed at testing your ability to read a text quickly in order to locate specific information. You are not required to read the whole text .
* Read the questions first and then find the answers by reading the text quickly. Write your answers in the spaces provided. Give short and precise answers.
* The questions are not in the same order the information appears in the text.
* You have 10 minutes to complete this part. At the end of 10 minutes the answer sheet of the Scanning part will be collected.

1. What percentage of residents living in Luang Namtha were found to be suffering from goitre? **Ninety (per cent)**
2. Which country has the worst IDD rates? **Laos (Lao People's Democratic Republic)**
3. To what percentage has the goitre rate dropped in Bolivia after iodization of salt? **20 (percent)**
4. What is the cause of 60,000 miscarriages in the world? **iodine deficiency (in pregnant women)**
5. According to WHO, what percentage of goitre rate creates a severe problem? **30 (percent)**
6. When was iodizing salt successfully adopted as a method by countries such as Switzerland and the USA? **(in) 1920**
7. What was the target date set for global elimination of IDD? **(the year) 2000**
8. What is the approximate amount of money spent on goitre operations in Laos each year? **(US$) 50,000**
9. How much salt is consumed by humans and animals in Laos annually? **24,000 tonnes**
10. What kind of products provide most of the body's iodine requirements in the West? **dairy products**
11. In which country is the Ministry of Industry the sole producer of salt? **Syria**

Skimming: New Perspectives on Child Education

* This part of the Reading Test aims at testing your ability to go through a text quickly in order to get the general idea. You are not required to read the text in detail in order to answer the question.
* You have 5 minutes to answer this question. At the end of 5 minutes this answer sheet will be collected.

1. Read the text quickly and circle the option that best expresses the general idea of the whole text.

a) New types of media programs should be designed for better education.

b) Interaction is necessary to achieve cognitive and affective development of children under five.

•**c)** Use of media designed for educational purposes is important for the development of a child under five.

d) Use of media designed for educational purposes is important in the development of a child in later years.

e) Play and creative activities provide adequate stimulation for the intellectual development of children.

Search Reading 1: New Perspectives on Child Education

* This part of the Reading Test is aimed at testing your ability to read quickly and selectively to find important information and ideas.
* First, locate the part of the text which provides the necessary information. Then, read carefully to answer each question.
* The questions are in the order the information appears in the text.
* Write your answers in the spaces provided. Give precise answers. You have 15 minutes for Search Reading 1 and you have 15 minutes for Search Reading 2. At the end of 30 minutes this booklet will be collected.

2. Mass media are effective on behavior to a certain extent; however, not all programs are adequately developed to generate the kind of **educationally approved behavior / behavior educators are interested in.**

3. What should be done to promote a child's mental development so that he can handle challenging problems?
(it is necessary) to push beyond natural growth stages .

4. The authors claim that children should be skilled in coping with not only written language but also **a wide range of symbols / all types of symbols.**

5. According to the text, a child brought up in a colorless environment will probably have difficulty in learning how to read because he has been deprived of **visual stimuli / various kinds of visual stimuli / perceptual stimulation** .

6. Listening to recorded material or to people in the immediate environment, the child becomes familiar with **linguistic patterns** of his mother tongue before he is able to speak.

7. The use of two-way presentation equipment requires the active involvement of a child. What would such equipment help in children?
(to manipulate abstractions in order) to develop his own cognitive structure/ cognitive development (structure) / active participation of the child in

developing various patterns related to cognitive structure / cognitive development

8. There is stage in which the child starts seeing himself as a separate individual. Thus, it is very important at this stage to show **positive attention / attention to the affective are/ attention to his (the child's) feelings.**

Careful Reading 1: Can Animals Learn to Share, Cooperate, Punish, and Show Empathy?

* This part of the Reading Test is aimed at testing your ability to read a text carefully.
* The questions are in the order the information appears in the text.
* Read the text and answer the following questions in the spaces provided Give precise answers.
* You have 25 minutes for Careful Reading 1 and 25 minutes for Careful Reading 2. At the end of 50 minutes this booklet will be collected.

1. The story of Binti is significant in that it makes laymen and scientists think about whether we can observe in animals the **mental traits** that are thought to be human-specific.

2. According to Stammbach's experiment, the peaceful approach of the high-ranking macaque monkeys toward the lower ones, which is a change in their behavior, can be attributed to their capability to **rein in aggressive behavior / act cooperatively / inhibit aggression**.

3. Kummer and Cord observed that it is through **inhibition / respect for property** that the dominant macaque monkeys control their instinct of grabbing the tubes held tightly by the subordinates, thus acting accordingly with the social rules of the community.

4. Rhesus monkeys who find food and remain silent are seen as **rule breakers / deceptively holding info / those who violate community rules** and punished accordingly.

5. In the experiment carried out by Miller, receivers did not know that actors in the other room could all see the signals that receivers were giving them. For the above reason, the author claims that the receiver and the actor could not have been behaving **cooperatively.**

6. In the experiment carried out by Miller, the existence of empathy might have been proven if only the **receiver** got the shock and if the **actor** actually saw the suffering of the other.

7. In Masserman's study it looks as if monkeys **emphathized** with other monkeys by not eating the food but there are other alternative explanations to their behavior. Therefore, until a study proves that animals are conscious of the **beliefs / intentions** of others, such studies will remain inconclusive.

QUESTIONS: SEPTEMBER VERSION

Scanning: The Pinch of Salt Solution

1. Which country has the worst IDD rates?
2. To what percentage has the goitre rate dropped in Bolivia after iodization of salt?
3. What is the cause of 60,000 miscarriages in the world?
4. According to WHO, what percentage of goitre rate creates a severe problem?
5. When was iodizing salt successfully adopted as a method by countries such as Switzerland and the USA?
6. What was the target date set for global elimination of IDD?
7. What is the approximate amount of money spent on goitre operations in Laos each year?
8. How much salt is consumed by humans and animals in Laos annually?
9. What kind of products provide most of the body's iodine requirements in the West?
10. In which country is the Ministry of Industry the sole producer of salt?

Skimming: New Perspectives on Child Education

1. Read the text quickly and circle the option that best expresses the general idea of the whole text.

a) New types of equipment should be designed to help the early cognitive and affective development of children.
b) All types of interaction with others are necessary to achieve cognitive and affective development of children under five.
c) Use of media designed for educational purposes plays a significant role in the cognitive and affective development of a child under five.
d) Use of media designed for educational purposes in the early years of childhood plays a significant role in the intellectual development of a child in later years.
e) Play and creative activities provide adequate stimulation for the intellectual development of children under five.

Search Reading 1: New Perspectives on Child Education

2. Mass media are effective on behavior to a certain extent; however, not all programs are adequately developed to generate the kind of _____ _____ .

3. What should be done to promote a child's mental development so that he can handle challenging problems?

_____

4. The authors claim that children should be skilled in coping with _____ in order to keep up with the advances humankind is likely to make.

5. Listening to recorded material or to people in the immediate environment, the child becomes familiar with _____ of his mother

tongue before he is able to speak.

6. The use of two-way presentation equipment requires the active involvement of a child. What would such equipment help in children?

Careful Reading 1: Can Animals Learn to Share, Cooperate, Punish, and Show Empathy?

1. The story of Binti is significant in that it makes laymen and scientists think about whether we can observe in animals the _____ that are thought to be human-specific.

2. According to Stammbach's experiment, the peaceful approach of the high-ranking macaque monkeys toward the lower ones, which is a change in their behavior, can be attributed to their capability to _____.

3. Kummer and Cord observed that it is through _____ that the dominant macaque monkeys control their instinct of grabbing the tubes held tightly by the subordinates, thus acting accordingly with the social rules of the community.

4. In the experiment carried out by Miller, the existence of empathy might have been proven if only the _____ got the shock and if the _____ actually saw the suffering of the other.

5. In Masserman's study it looks as if monkeys _____ with other monkeys by not eating the food but there are other alternative explanations to their behavior. Therefore, until a study proves that animals are conscious of the _____ of others, such studies will remain inconclusive.

THE IELTS TEST

READING

READING PASSAGE 1

You should spend about 20 minutes on Questions 1–13 which are based on Reading Passage 1 below.

Green Wave Washes Over Mainstream Shopping

Research in Britain has shown that 'green consumers' continue to flourish as a significant group amongst shoppers. This suggests that politicians who claim environmentalism is yesterday's issue may be seriously misjudging the public mood.

A report from Mintel, the market research organisation, says that despite recession and financial pressures, more people than ever want to buy environmentally friendly products and a 'green wave' has swept through consumerism, taking in people previously untouched by environmental concerns. The recently published report also predicts that the process will repeat itself with 'ethical' concerns, involving issues such as fair trade with the Third World and the social record of businesses. Companies will have to be more honest and open in response to this mood.

Mintel's survey, based on nearly 1,000 consumers, found that the proportion who look for green products and are prepared to pay more for them has climbed from 53 per cent in 1990 to around 60 per cent in 1994. On average, they will pay 13 per cent more for such products, although this percentage is higher among women, managerial and 'armchair greens'; they said they care about environmental issues but their concern does not affect their spending habits. Only 10 per cent say they do not care about green issues.

Four in ten people are 'ethical spenders', buying goods which do not, for example, involve dealings with oppressive regimes. This figure is the same as in 1990, although the number of 'armchair ethicals' has risen from 28 to 35 per cent and only 22 per cent say they are unconcerned now, against 30 per cent in 1990. Hughes claims that in the twenty-first century, consumers will be encouraged to think more about the entire history of the products and professional groups and those aged 35 to 44.

Between 1990 and 1994 the proportion of consumers claiming to be unaware of or unconcerned about green issues fell from 18 to 10 per cent but the number of green spenders among older people and manual workers has risen substantially. Regions such as Scotland have also caught up with the south of England in their environmental concerns. According to Mintel, the image of green consumerism as associated in the past with the more eccentric members of society has virtually disappeared. The consumer research manager for Mintel, Angela Hughes, said it had become firmly established as a mainstream market. She explained that as far as the average person is concerned environmentalism has not 'gone off the boil'. In fact, it has spread across a much wider range of consumer groups, ages and occupations.

Mintel's 1994 survey found that 13 per cent of consumers are 'very dark green', nearly always buying environmentally friendly products, 28 per cent are 'dark green', trying 'as far as possible' to buy such products, and 21 per cent are 'pale green' – tending to buy green products if they see them. Another 26 per cent are services they buy, including the policies of the companies that provide them and that this will require a greater degree of honesty with consumers.

Among green consumers, animal testing is the top issue – 48 per cent said they would be deterred from buying a product it if had been tested on animals – followed by concerns regarding irresponsible selling, the ozone layer, river and sea pollution, forest destruction, recycling and factory farming. However, concern for specific issues is lower than in 1990, suggesting that many consumers feel that Government and business have taken on the environmental agenda.

Questions 1–6

Do the following statements agree with the claims of the writer of Reading Passage 1?
In boxes 1–6 on your answer sheet write

YES *if the statement agrees with the claims of the writer*
NO *if the statement contradicts the claims of the writer*
NOT GIVEN *if it is impossible to say what the writer thinks about this*

1 The research findings report commercial rather than political trends.

2 Being financially better off has made shoppers more sensitive to buying 'green'.

3 The majority of shoppers are prepared to pay more for the benefit of the environment according to the research findings.

4 Consumers' green shopping habits are influenced by Mintel's findings.

5 Mintel have limited their investigation to professional and managerial groups.

6 Mintel undertakes market surveys on an annual basis.

Questions 7–9

Choose the appropriate letters A–D and write them in boxes 7–9 on your answer sheet.

7 Politicians may have 'misjudged the public mood' because ...

A they are pre-occupied with the recession and financial problems.
B there is more widespread interest in the environment agenda than they anticipated.
C consumer spending has increased significantly as a result of 'green' pressure.
D shoppers are displeased with government policies on a range of issues.

8 What is Mintel?

A an environmentalist group
B a business survey organisation
C an academic research team
D a political organisation

9 A consumer expressing concern for environmental issues without actively supporting such principles is ...

A an 'ethical spender'.
B a 'very dark green' spender.
C an 'armchair green'.
D a 'pale green' spender.

Questions 10–13

Complete the summary using words from the box below.
Write your answers in boxes 10–13 on your answer sheet.

NB There are more answers than spaces, so you will not use them all.

The Mintel report suggests that in future companies will be forced to practise greater ... **(10)** ... in their dealings because of the increased awareness amongst ... **(11)** ... of ethical issues. This prediction is supported by the growth in the number of ... (12) ... identified in the most recent survey published. As a consequence, it is felt that companies will have to think more carefully about their ... (13) ... .

environmental research	armchair ethicals
honesty and openness	environmentalists
ethical spenders	consumers
politicians	political beliefs
social awareness	financial constraints
social record	

READING PASSAGE 2

You should spend about 20 minutes on Questions 14–26 which are based on Reading Passage 2 below.

A There is a great concern in Europe and North America about declining standards of literacy in schools. In Britain, the fact that 30 per cent of 16 year olds have a reading age of 14 or less has helped to prompt massive educational changes. The development of literacy has far-reaching effects on general intellectual development and thus anything which impedes the development of literacy is a serious matter for us all. So the hunt is on for the cause of the decline in literacy. The search so far has focused on socio-economic factors, or the effectiveness of 'traditional' versus 'modern' teaching techniques.

B The fruitless search for the cause of the increase in illiteracy is a tragic example of the saying 'They can't see the wood for the trees'. When teachers use picture books, they are simply continuing a long-established tradition that is accepted without question. And for the past two decades, illustrations in reading primers have become increasingly detailed and obtrusive, while language has become impoverished – sometimes to the point of extinction.

C Amazingly, there is virtually no empirical evidence to support the use of illustrations in teaching reading. On the contrary, a great deal of empirical evidence shows that pictures interfere in a damaging way with all aspects of learning to read. Despite this, from North America to the Antipodes, the first books that many school children receive are totally without text.

D A teacher's main concern is to help young beginner readers to develop not only the ability to recognise words, but the skills necessary to understand what these words mean. Even if a child is able to read aloud fluently, he or she may not be able to understand much of it: this is called 'barking at text'. The teacher's task of improving comprehension is made harder by influences outside the classroom. But the adverse effects of such things as television, video games, or limited language experiences at home, can be offset by experiencing 'rich' language at school.

E Instead, it is not unusual for a book of 30 or more pages to have only one sentence full of repetitive phrases. The artwork is often marvellous, but the pictures make the language redundant, and the children have no need to imagine anything when they read such books. Looking at a picture actively prevents children younger than nine from creating a mental image, and can make it difficult for older children. In order to learn how to comprehend, they need to practise making their own meaning in response to text. They need to have their innate powers of imagination trained.

F As they grow older, many children turn aside from books without pictures, and it is a situation made more serious as our culture becomes more visual. It is hard to wean children off picture books when pictures have played a major part throughout their formative reading experiences, and when there is competition for their attention from so many other sources of entertainment. The least intelligent are most vulnerable, but tests show that even intelligent children are being affected. The response of educators has been to extend the use of pictures in books and to simplify the language, even at senior levels. The Universities of Oxford and Cambridge recently held joint conferences to discuss the noticeably rapid decline in literacy among their undergraduates.

G Pictures are also used to help motivate children to read because they are beautiful and eye-catching. But motivation to read should be provided by listening to stories well read, where children imagine in response to the story. Then, as they start to read, they have this experience to help them understand the language. If we present pictures to save children the trouble of developing these creative skills, then I think we are making a great mistake.

H Academic journals ranging from educational research, psychology, language learning, psycholinguistics, and so on cite experiments which demonstrate how detrimental pictures are for beginner readers. Here is a brief selection:

I The research results of the Canadian educationalist Dale Willows were clear and consistent: pictures affected speed and accuracy and the closer the pictures were to the words, the slower and more inaccurate the child's reading became. She claims that when children come to a word they already know, then the pictures are unnecessary and distracting. If they do not know a word and look to the picture for a clue to its meaning, they may well be misled by aspects of the pictures which are not closely related to the meaning of the word they are trying to understand.

J Jay Samuels, an American psychologist, found that poor readers given no pictures learnt significantly more words than those learning to read with books with pictures. He examined the work of other researchers who had reported problems with the use of pictures and who found that a word without a picture was superior to a word plus a picture. When children were given words and pictures, those who seemed to ignore the pictures and pointed at the words learnt more words than the children who pointed at the pictures, but they still learnt fewer words than the children who had no illustrated stimuli at all.

Questions 14–17

*Choose the appropriate letters **A–D** and write them in boxes 14–17 on your answer sheet.*

14 Readers are said to 'bark' at a text when ...

A they read too loudly.
B there are too many repetitive words.
C they are discouraged from using their imagination.
D they have difficulty assessing its meaning.

15 The text suggests that ...

A pictures in books should be less detailed.
B pictures can slow down reading progress.
C picture books are best used with younger readers.
D pictures make modern books too expensive.

16 University academics are concerned because ...

A young people are showing less interest in higher education.
B students cannot understand modern academic texts.
C academic books are too childish for their undergraduates.
D there has been a significant change in student literacy.

17 The youngest readers will quickly develop good reading skills if they ...

A learn to associate the words in a text with pictures.
B are exposed to modern teaching techniques.
C are encouraged to ignore pictures in the text.
D learn the art of telling stories.

Questions 18–21

Do the following statements agree with the information given in Reading Passage 2?
In boxes 18–21 on your answer sheet write

YES	*if the statement agrees with the information*
NO	*if the statement contradicts the information*
NOT GIVEN	*if there is no information about this in the passage*

18 It is traditionally accepted that children's books should contain few pictures.

19 Teachers aim to teach both word recognition and word meaning.

20 Older readers are having difficulty in adjusting to texts without pictures.

21 Literacy has improved as a result of recent academic conferences.

Questions 22–25

Reading Passage 2 has ten paragraphs, A–J. Which paragraphs state the following information? Write the appropriate letters A–J in boxes 22–25 on your answer sheet.

NB There are more paragraphs than summaries, so you will not use them all.

22 The decline of literacy is seen in groups of differing ages and abilities.

23 Reading methods currently in use go against research findings.

24 Readers able to ignore pictures are claimed to make greater progress.

25 Illustrations in books can give misleading information about word meaning.

Question 26

From the list below choose the most suitable title for the whole of Reading Passage 2. Write the appropriate letter A–E in box 26 on your answer sheet.

A The global decline in reading levels
B Concern about recent educational developments
C The harm that picture books can cause
D Research carried out on children's literature
E An examination of modern reading styles

READING PASSAGE 3

You should spend about 20 minutes on Questions 27–40 which are based on Reading Passage 3 below.

IN SEARCH OF THE HOLY GRAIL

It has been called the Holy Grail of modern biology. Costing more than £2 billion, it is the most ambitious scientific project since the Apollo programme that landed a man on the moon. And it will take longer to accomplish than the lunar missions, for it will not be complete until early next century. Even before it is finished, according to those involved, this project should open up new understanding of, and new treatments for, many of the ailments that afflict humanity. As a result of the Human Genome Project, there will be new hope of liberation from the shadows of cancer, heart disease, auto-immune diseases such as rheumatoid arthritis, and some psychiatric illnesses.

The objective of the Human Genome Project is simple to state, but audacious in scope: to map and analyse every single gene within the double helix of humanity's DNA[1]. The project will reveal a new human anatomy – not the bones, muscles and sinews, but the complete genetic blueprint for a human being. Those working on the Human Genome Project claim that the new genetical anatomy will transform medicine and reduce human suffering in the twenty-first century. But others see the future through a darker glass, and fear that the project may open the door to a world peopled by Frankenstein's monsters and disfigured by a new eugenics[2].

The genetic inheritance a baby receives from its parents at the moment of conception fixes much of its later development, determining characteristics as varied as whether it will have blue eyes or suffer from a life-threatening illness such as cystic fibrosis. The human genome is the compendium of all these inherited genetic instructions. Written out along the double helix of DNA are the chemical letters of the genetic text. It is an extremely long text, for the human genome contains more than 3 billion letters. On the printed page it would fill about 7,000 volumes. Yet, within little more than a decade, the position of every letter and its relation to its neighbours will have been tracked down, analysed and recorded.

Considering how many letters there are in the human genome, nature is an excellent proof-reader. But sometimes there are mistakes. An error in a single 'word' – a gene – can give rise to the crippling condition of cystic fibrosis, the commonest genetic disorder among Caucasians. Errors in the genetic recipe for haemoglobin, the protein that gives blood its characteristic red colour and which carries oxygen from the lungs to the rest of the body, give rise to the most common single-gene disorder in the world: thalassaemia. More than 4,000 such single-gene defects are known to afflict humanity. The majority of them are fatal; the majority of the victims are children.

None of the single-gene disorders is a disease in the conventional sense, for which it would be possible to

administer a curative drug: the defect is pre-programmed into every cell of the sufferer's body. But there is hope of progress. In 1986, American researchers identified the genetic defect underlying one type of muscular dystrophy. In 1989, a team of American and Canadian biologists announced that they had found the site of the gene which, when defective, gives rise to cystic fibrosis. Indeed, not only had they located the gene, they had analysed the sequence of letters within it and had identified the mistake responsible for the condition. At the least, these scientific advances may offer a way of screening parents who might be at risk of transmitting a single-gene defect to any children that they conceive. Foetuses can be tested while in the womb, and if found free of the genetic defect, the parents will be relieved of worry and stress, knowing that they will be delivered of a baby free from the disorder.

In the mid-1980s, the idea gained currency within the scientific world that the techniques which were successfully deciphering disorder-related genes could be applied to a larger project: if science can learn the genetic spelling of cystic fibrosis, why not attempt to find out how to spell 'human'? Momentum quickly built up behind the Human Genome Project and its objective of 'sequencing' the entire genome – writing out all the letters in their correct order.

But the consequences of the Human Genome Project go far beyond a narrow focus on disease. Some of its supporters have made claims of great extravagance – that the Project will bring us to understand, at the most fundamental level, what it is to be human. Yet many people are concerned that such an emphasis on humanity's genetic constitution may distort our sense of values, and lead us to forget that human life is more than just the expression of a genetic program written in the chemistry of DNA.

If properly applied, the new knowledge generated by the Human Genome Project may free humanity from the terrible scourge of diverse diseases. But if the new knowledge is not used wisely, it also holds the threat of creating new forms of discrimination and new methods of oppression. Many characteristics, such as height and intelligence, result not from the action of genes alone, but from subtle interactions between genes and the environment. What would be the implications if humanity were to understand, with precision, the genetic constitution which, given the same environment, will predispose one person towards a higher intelligence than another individual whose genes were differently shuffled?

Once before in this century, the relentless curiosity of scientific researchers brought to light forces of nature in the power of the atom, the mastery of which has shaped the destiny of nations and overshadowed all our lives. The Human Genome Project holds the promise that, ultimately, we may be able to alter our genetic inheritance if we so choose. But there is the central moral problem: how can we ensure that when we choose, we choose correctly? That such a potential is a promise and not a threat? We need only look at the past to understand the danger.

Glossary

[1] *DNA* — *Deoxyribonucleic acid, molecules responsible for the transference of genetic characteristics.*

[2] *eugenics* — *The science of improving the qualities of the human race, especially the careful selection of parents.*

Questions 27–32

Complete the sentences below (Questions 27–32) with words taken from Reading Passage 3.
Use NO MORE THAN THREE WORDS OR A NUMBER for each answer.
Write your answers in boxes 27–32 on your answer sheet.

Example — *Answer*

The passage compares the genetic instructions in DNA to ... chemical letters.

27 The passage compares the Project in scale to the ..... .

28 The possible completion date of the Project is ..... .

29 To write out the human genome on paper would require ..... books.

30 A genetic problem cannot be treated with drugs because strictly speaking it is not a ..... .

31 Research into genetic defects had its first success in the discovery of the cause of one form of ..... .

32 The second success of research into genetic defects was to find the cause of ..... .

Questions 33–40

Classify the following statements as representing

A *the writer's fears about the Human Genome Project*
B *other people's fears about the Project reported by the writer*
C *the writer's reporting of facts about the Project*
D *the writer's reporting of the long-term hopes for the Project*

Write the appropriate letters A–D in boxes 33–40 on your answer sheet.

33 The Project will provide a new understanding of major diseases.

34 All the components which make up DNA are to be recorded and studied.

35 Genetic monsters may be created.

36 The correct order and inter-relation of all genetic data in all DNA will be mapped.

37 Parents will no longer worry about giving birth to defective offspring.

38 Being 'human' may be defined solely in terms of describable physical data.

39 People may be discriminated against in new ways.

40 From past experience humans may not use this new knowledge wisely.

THE IELTS TEST – Answer Key

ACADEMIC READING

Each question correctly answered scores 1 mark.

Reading Passage 1, Questions 1–13

1 YES
2 NO
3 YES
4 NOT GIVEN
5 NO
6 NOT GIVEN
7 B
8 B
9 C
10 honesty and openness
11 consumers
12 armchair ethicals
13 social record

Reading Passage 2, Questions 14–26

14 D
15 B
16 D
17 C
18 NO
19 YES
20 YES
21 NOT GIVEN
22 F
23 C
24 J
25 I
26 C

Reading Passage 3, Questions 27–40

27 Apollo (space) programme
28 (early) next century
29 7,000
30 disease
31 muscular dystrophy
32 cystic fibrosis
33 D
34 C
35 B
36 C
37 D
38 B
39 A
40 A

APPENDIX 3.2

SPECIFICATION OF OPERATIONS AND PERFORMANCE CONDITIONS FOR THE BUEPT READING TEST

SCANNING

Purpose: Looking quickly through a text, not necessarily following the linearity of the text, to locate a specific symbol or group of symbols: e.g. a particular word, phrase, name, figure, date, etc.
Operationalisations: Looking for (matching) specific words or phrases, figures, dates and names.
Focus: Local
Text coverage: Most of the text is ignored.
Rate of reading: Rapid inspection of the text with only occasional closer inspection.
Direction of processing: Sequencing in the text is not observed.
Relationship with the underlying process: Surface level processing of a text. Mainly bottom-up processing. Finding a match between what is sought and what is given in a text, very little information is processed for long-term retention or even for immediate understanding.
Nature of the text: Texts written for non-specialist audience. Texts involving factual details such as several names, figures, numbers, dates, references, key words, etc.
Sources of texts: Chapters from textbooks, academic journal articles, semi-academic journal articles.
Rhetorical organisation: Informative, descriptive texts with explicit text structure and subsections divided by subtitles.
Propositional features
Lexical range: Normally no technical jargon (glossed when appears). Academic, semi-technical words.
Topic area: Does not require background knowledge. Familiar, generally accessible texts. All topic areas.
Illocutionary features: To inform, to explain, to describe.
Channel of presentation: Normally textual. Some texts might contain graphics.
Text length: 1500-2000 words.
Speed of processing: 1minute per question.
Number of texts: 1
Number of questions: 10.
Order of questions: Not sequential.
Instructions: Clearly written in English.
Question format: Short-answer WH questions.
Weighting: 0.5 point for each item. All items equally weighted.

SKIMMING

Purpose: Processing a text selectively to get the main idea(s) and the discourse topic as efficiently as possible-which might involve both expeditious and careful reading

to establish a general sense of the text
to quickly establish a macropropositional structure without decoding all the text
to decide the relevance of texts to established needs.

Operationalisations: Reading title and subtitles quickly, reading abstract carefully (when applies), reading introductory and concluding paragraphs carefully, reading first and last sentences of each paragraph carefully, skimming for frequently occurring words or phrases.
Focus: Both global and local.
Text coverage: Selective reading to establish important propositions (macrostructure) of a text.
Rate of reading: Rapid with some careful reading.
Direction of processing: Sequencing observed.
Relationship with the underlying process: Interactive process involving both top-down and bottom-up processing.
Nature of the text: Academic texts written for non-specialist audience.
Sources of texts: Chapters from textbooks, academic journal articles, semi-academic journal articles.
Rhetorical organisation: Expository texts that may involve information, comparison, causation, argumentation, problem-solving. Texts with reasonable number of main ideas and explicit structure. The texts must have at least some of the following features that facilitate selective reading.

Subtitles, headings
Initial summary or abstract
Introductory and concluding paragraphs
Helpful first and last sentences in paragraphs
Discourse markers (conjunctions, connectors, etc.)
Markers of importance ('this is crucial', 'the main aspect…', etc.)
Repeated key content words
Summarising non-verbal information (charts, graphs, etc.)
Clear text structure, arguments clearly stated

Propositional features
Lexical range: Normally no technical jargon (glossed when appears). Academic, semi-technical words.
Topic area: Does not require background knowledge. Mostly humanities and social sciences, economy, management.
Illocutionary features: To inform, to explain, to argue, to persuade, to advise.
Channel of presentation: Normally textual. Some texts might contain graphs.
Text length: Approximately 2000 words.
Speed of processing: 5 minutes.
Number of texts: 1 (Only with the first text of search reading).
Number of questions: 1
Instructions: Clearly written in English.
Question format: Multiple-choice.
Weighting: 1 point.

SEARCH READING

Purpose: Locating information on predetermined topic(s) (e.g., questions set on main ideas in a text). This normally goes beyond mere matching of words (as in scanning). The process is rapid and selective but is likely involve careful reading once relevant information has been located.
Operationalisations: Keeping alert for words in the same or related semantic field with the topic of search. Using formal/textual knowledge for locating information. Using titles and subtitles. Reading abstract (when applies), reading first and last sentences of each paragraph carefully.
Focus: Both global and local.
Text coverage: Selecting information relevant to predetermined topic(s).
Rate of reading: Rapid with careful reading when information is located.
Direction of processing: Sequencing not always observed.
Relationship with the underlying process: Interactive process involving both top-down and bottom-up processing. There is more observance of linearity and sequencing as compared with scanning. Involves top-down processing when the formal/textual knowledge is used. The periods of close attention to the text tend to be more frequent and longer than scanning. Bottom-up processing is involved when close attention is paid to the selected part(s) of the text.
Nature of the text: Academic texts written for non-specialist audience.
Sources of texts: Chapters from textbooks, academic journal articles, semi-academic journal articles.
Rhetorical organisation: Expository texts that may involve information, comparison, causation, argumentation, problem-solving . Texts with reasonable number of main ideas and explicit structure. The texts must have at least some of the following features which facilitate selective reading.

Subtitles, headings
Initial summary or abstract
Introductory and concluding paragraphs
Helpful first and last sentences in paragraphs
Discourse markers (conjunctions, connectors, etc.)
Markers of importance ('this is crucial', 'the main aspect…', etc.)
Repeated key content words
Summarising non-verbal information (charts, graphs, etc.)
Clear text structure, arguments clearly stated

Propositional features
Lexical range: Normally no technical jargon (glossed when appears). Academic semi-technical words.
Topic area: Does not require background knowledge. Mostly humanities and social sciences, economy, management.
Illocutionary features: To inform, to explain, to argue, to persuade, to advise.
Channel of presentation: Normally textual. Some texts might contain graphs.
Text length: Approximately 2000 words.
Speed of processing: 3 minutes per question.
Number of texts: 2
Number of questions: 5 questions on each text. 10 questions at total.
Order of questions: Not necessarily sequential.
Instructions: Clearly written in English

Question format: Short-answer WH questions and/or sentence completion questions, which require candidates to write down answers in spaces provided on the question paper.
Weighting: 1 point for each item. All items equally weighted.

CAREFUL READING AT THE GLOBAL LEVEL

Purpose: Processing a text carefully and thoroughly in order to comprehend main idea(s) and supporting information. Decoding the whole text in order to comprehend it all or to establish a macrostructure for the text.
Operationalisations: Separating explicitly stated idea(s) from supporting detail(s). Distinguishing generalisations from examples, facts from opinions. Understanding the development of an argument and/or logical organisation (sequence, causation, purpose, thesis-antithesis, evidence, justification, condition, concession, problem-solution, evaluation, etc.). Generating a representation of a text as a whole. Understanding of implicitly stated arguments/Making propositional inferences (without recourse to knowledge from outside the text).
Focus: Both global and local.
Text coverage: Reading from beginning to end.
Rate of reading: Reading the whole text carefully.
Direction of processing: Linear and sequential, with regressions if needed.
Relationship with the underlying process: Mainly text-based bottom-up sequential process with limited top-down process.
Nature of the text: Academic texts written for non-specialist audience.
Sources of texts: Chapters from textbooks, academic journal articles.
Rhetorical organisation: Expository texts in analytic and critical nature that may involve information, causation, argumentation, problem-solution, evaluation, etc. Texts with reasonable number of main ideas. Texts should be propositionally more demanding involving more abstract argumentation. The text structure may not necessarily be clear or overt.
Propositional features
Lexical range: Wider lexical range. Academic. Technical jargon glossed when appears.
Topic area: Necessary information should be contained within the text when the topic is not generally accessible. Mostly humanities and social sciences, economy, management.
Illocutionary features: To inform, to explain, to argue, to persuade, to advise.
Channel of presentation: Normally textual. Some texts might contain graphs.
Text length: Approximately 1000-1500 words.
Speed of processing: 5 minutes per question.
Number of texts: 2
Number of questions: 5 questions on each text. 10 questions at total.
Instructions: Clearly written in English
Question format: Short-answer WH questions and/or sentence completion questions, which require candidates to write down answers in spaces provided on the question paper.
Weighting: 1 point for each item. All items equally weighted.

Changes Made To Test Specifications: Effective from June 2001 test, the following changes were made in the test specifications.

Skimming
Eliminated from the test.

Search reading
Text length: 2800-3000 words.
Number of texts: 1
Number of questions: 7 questions

Careful reading
Text length: 1500-2000 words.
Number of texts: 1
Number of questions: 7 questions

APPENDIX 3.3
CONTENT ANALYSIS SCHEME

YOUR NAME: .....

Dear Participant,
You are going to do a content analysis of the reading parts of two proficiency tests, namely IELTS (International English Language Testing Systems) and BUEPT (Bogaziçi University English Proficiency Test) using the scheme given below. First, you will concentrate on the test and text features, then you will evaluate the questions. It is of utmost importance that you take the tests as you would normally do under exam conditions obeying the time limits. Otherwise, it may not be possible to analyse the type of reading operations that are required by particular texts and questions. Therefore, please follow the instructions carefully. If you cannot, please let the researcher know that. I appreciate the value of your help in this quite time-consuming and tiresome undertaking. Thank you very much.

Aylin Ünaldı

INSTRUCTIONS:

1. First, read the explanations and definitions in this scheme carefully.
2. Take the IELTS test as you would do under routine exam conditions observing the time limits strictly. However, mark the part(s) (underline/write the question number) of the text that you have read when answering the questions as this may be of help to you later on.
3. Check your answers using the keys.
4. Respond to Part I: Test Rubric Characteristics
5. Go back to the first text, read it quickly once more and respond to the questions in Part II:
   Text Characteristics. (Make sure that you are using the scale correctly in these parts.Refer to explanations frequently because the scaling may be counter-intuitive at times)
6. Move onto Part III: Item Characteristics. Read the test questions once more. Try to remember how you behaved when you answered them and using the tables given evaluate them in terms of:
   a. Operations
   b. Text Span
   c. Overall Difficulty
7. After you finish, move on to the next text and repeat the same procedure. Do the same for all texts and questions.
8. Take the BUEPT test the next day as you will do under routine exam conditions observing the time limits carefully.
9. Repeat the procedures given above for all the texts and questions of BUEPT.

Adapted from Bachman et al. (1995), Khalifa (1997), Urquhart and Weir (1998).

EXPLANATIONS AND DEFINITIONS

PART I - TEST RUBRIC CHARACTERISTICS: These consist of the characteristics that specify how the test takers are expected to proceed in taking the test.

1. **Test Organisation (TO):** Are the description and the relative importance (weigths) of the parts of the test clearly given? Is there a clear sequencing among the parts? Rate the explicitness of the text organisation using the scale given below (1 for clear, 5 for not clear test organisation). Indicate any problems you observe.

2. **Time Allocation (TA):** Is the amount of time per part, passage and item sufficient? Rate the time allocation using the scale given below (1 for sufficient, 5 for insufficient time allocation). Name the parts, passages, items that you think have not been allocated sufficient time explicitly on the comment lines.

3. **Instructions (I):** Are specification of procedures (how the responses are to be recorded on the test booklet or answer sheet, the order in which the parts are to be taken), specification of tasks (instructions that indicate the test taker how s/he is to arrive at the answer), criteria for scoring (detailed descriptions for correctness in the keys) clear? Rate the clarity of the instructions using the scale given below (1 for clear, 5 for not clear instructions). Indicate any problems you observe.

PART II - TEXT CHARACTERISTICS: These refer to the information contained in a given text and it is largely these characteristics that determine the comprehensibility of the input.

1. **Nature of the text (NT)**: Indicate the nature of the text using the given list: instructions, sign, message, leaflet, brouchure, advertisement, newspaper article, magazine article, research/journal article, text book article, other.

2. **Grammar/Syntax (GR):** Does the text include many passive verbs, compound and complex sentences, embeddings (relative and noun clauses, gerundive, infinitival complement structures, etc.)? Rate the the overall complexity of the text using the scale given below (1 for simple, basic syntax, 5 for the text in which the structures listed above frequently occur).

3. **Vocabulary (VOC):** Does the text include many infrequent, specialised, ambiguous words? Rate the complexity of the vocabulary used in the text using the scale given below (1 for basic, frequent vocabulary, 5 for complex, sophisticated vocabulary).

4. **Cohesion (COH):** Throughout the text, are the relations between ideas (sentences, clauses and parts of the text in which different ideas presented) explicitly marked through reference, conjunctions, lexical/phrasal connectors, etc.? Is the ordering of old and new information explicitly signalled? Rate the explicitness of the cohesion in the text using the scale given below (1 for explicit, clear textual ties, 5 for not explicit relations).

5. **Rhetorical Organisation (RO):** Is there an explicit rhetorical organisation in the text? Mark as it applies: (1) narration (2) description (3) information (4) comparison and contrast (5) classification (6) process analysis (7) discursive (8) argumentation (9) other:
   Does the text have a clear line of argument running through it? Is the text clearly organised into sections with an introduction, topic sentence, support sentences and conclusion? Rate the explicitness of the rhetorical organisation of the text using the scale given below (1 for explicit, clear text structure, 5 for not explicit organisation).

6. **Degree of Contextualisation(DC):** Are the arguments in the text supported by a wide range of meaningful linguistic, paralinguistic and situational cues in the context? Is the new information expressed in the text rich with familiar or known information that is relevant (contextualised/context embedded)? Rate the contextualisation of the language use in the text using the scale given below (1 for high ratio, 5 for low ratio of contextual information to new information).

7. **Distribution of the New Information (DNI):** Is the new information in the text distributed over a relatively short space (compact) or long space (diffuse)? Rate the distribution of information in the text using the scale given below (1 for relatively little amount of new information, 5 for dense new information presented in the text).

8. **Type of Information(TI):** Is the information in the text concrete and factual or is much of the information abstract, symbolic and counterfactual or? Rate the type of information in the text using the scale given below (1 for relatively more concrete, 5 for relatively more abstract information presented in the text).

9. **Topic Specificity(TS):** Is the topic of the text of general interest or is it subject specific and require background knowledge on the part of the reader? Rate topic specificity using the scale given below (1 for non-specific topic, 5 for topic that may require background knowledge).

10. **Cultural Specificity (CS):** Is the topic of the text culture-free or is it loaded with specific cultural content? Rate cultural specificity using the scale given below (1 for non-specific content, 5 for culture specific content).

11. **Overall Difficulty (OD):** How do you rate the overall difficulty of the passage? Rate difficulty of the passage using the scale given below (1 for easy, 5 for difficult text).

PART III - ITEM CHARACTERISTICS:

OPERATIONS: These refer to the operation(s)/skill(s) that are expected to contribute to arriving at the correct answer. Therefore, the test taker will normally use one or more of these operations while answering the questions.

O1. rapidly looking for figures, dates, names, etc in the text.
O2. rapidly inspecting the text (and go back and forth in it) to locate the answer.
O3. matching the exact key words/phrases in the question and in the text.

O4. matching the key words/phrases in the question with their synonyms / paraphrases in the text.
O5. using my own knowledge of how the text is structured (knowledge of formal text structure) in order to locate the answers.
O6. using the title, subtitles, section headings and the first and last sentences of the paragraphs.
O7. reading the abstract/introduction and conclusion carefully.
O8. reading carefully to confirm the answer after deciding the location of information.
O9. forming a summary of the main ideas/text topic in mind.
O10. reading slowly and carefully for detailed understanding of explicitly stated ideas in the text (when there are the same key word/words in both the question and the text).
O11. reading slowly and carefully for detailed understanding of an idea in the text (when there are no key words occuring in both text and question).
O12. reading a part of a text more than once in order to understand it.
O13. focusing on pronouns, discourse markers, grammar, etc.
O14. deducing the meaning of a word from the context.
O15. dealing with relatively uncommon vocabulary.

TEXT SPAN: This is related with the relationship of the item to the passage in terms of the amount of text that should be processed for successful comprehension.

TS1. no relationship to the passage; item can be answered without reference to the passage, *or* relationship of item to passage is not clear
TS2. relates to a specific part of the passage, and requires only localised understanding of that part
TS3. relates to several specific parts of the passage, or requires test taker to relate one part of the passage to several others
TS4. item relates to the entire passage, and requires an understanding of the entire passage
TS5. requires test taker to relate information in passage to the real world, outside the text

TEST: IELTS

Read the instructions and respond to the questions by circling a number in the scales given. Mention/explain any problems, important details or your personal observations on the comment lines.

TEST RUBRIC CHARACTERISTICS:

1. **Test Organisation (TO):**
   (clear) 1 2 3 4 5 (not clear)
   Comments:_____
   _____

2. **Time Allocation (TA):**
   (sufficient) 1 2 3 4 5 (insufficient)
   Comments:_____
   _____

3. **Instructions (I):**
   (clear) 1 2 3 4 5 (not clear)
   Comments:_____
   _____

TEXT CHARACTERISTICS: IELTS READING PASSAGE 1: GREEN WAVE WASHES OVER ... pp: 83-84 (582 words)

1. **Nature of the text (NT):** Mark as it applies (you can mark more than one): (1) instructions (2) sign (3) message (4) leaflet (5) brochure (6) advertisement (7) newspaper article (8) magazine article (9) research/journal article (10) text book article (11) other: _____

2. **Grammar/Syntax (GR):**
   (basic) 1 2 3 4 5 (complex)
   Comments:_____

3. **Vocabulary (VOC):**
   (frequent) 1 2 3 4 5 (less frequent)
   Comments:_____

4. **Cohesion (COH):**
   (explicit) 1 2 3 4 5 (not explicit)
   Comments:_____

5. **Rhetorical Organisation (RO):** Mark as it applies:(1) narration (2) description (3) information (4)comparison and contrast (5) classification (6) process analysis (7) discursive (8) argumentation
   (9) other: _____
   Degree of rhetorical organisation explicitness:
   (explicit) 1 2 3 4 5 (not explicit)
   Comments:_____

6. **Degree of Contextualisation(DC):**
   (context embedded) 1 2 3 4 5 (context reduced)
   Comments:_____

7. **Distribution of the New Information (DNI):**
   (diffused) 1 2 3 4 5 (compact)
   Comments:_____

8. **Type of Information(TI):**
(concrete) 1 2 3 4 5 (abstract)
Comments:_____

9. **Topic Specificity(TS):**
(not specific) 1 2 3 4 5 (highly specific)
Comments:_____

10. **Cultural Specificity (CS):**
(not specific) 1 2 3 4 5 (highly specific)
Comments:_____

11. **Overall Difficulty (OD):**
(easy) 1 2 3 4 5 (difficult)

Comments:_____

ITEM CHARACTERISTICS: IELTS READING PASSAGE 1: GREEN WAVE WASHES OVER … pp: 83-84 (582 words)
q: question O: operation TS: text span

Mark (X) the operation(s)/skill(s) that you think have contributed to arriving at the correct answer. (You can choose more than one)

GWWMS	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	q11	q12	q13
O1													
O2													
O3													
O4													
O5													
O6													
O7													
O8													
O9													
O10													
O11													
O12													
O13													
O14													
O15													

Mark (X) the text span that you think should be processed in answering the question. (Choose one)

GWWMS	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	q11	q12	q13
TS1													
TS2													
TS3													
TS4													
TS5													

TEXT CHARACTERISTICS: IELTS READING PASSAGE 2: pp: 87-88 (863 words)

1. **Nature of the text (NT):** Mark as it applies (you can mark more than one): (1) instructions (2) sign (3) message (4) leaflet (5) brochure (6) advertisement (7) newspaper article (8) magazine article (9) research/journal article (10) text book article (11) other:

   _____

2. **Grammar/Syntax (GR):**
   (basic) 1 2 3 4 5 (complex)
   Comments:_____

3. **Vocabulary (VOC):**
   (frequent) 1 2 3 4 5 (less frequent)
   Comments:_____

4. **Cohesion (COH):**
   (explicit) 1 2 3 4 5 (not explicit)
   Comments:_____

5. **Rhetorical Organisation (RO):** Mark as it applies:(1) narration (2) description (3) information (4)comparison and contrast (5) classification (6) process analysis (7) discursive (8) argumentation
   (9) other: _____
   Degree of rhetorical organisation explicitness:
   (explicit) 1 2 3 4 5 (not explicit)
   Comments:_____

6. **Degree of Contextualisation(DC):**
   (context embedded) 1 2 3 4 5 (context reduced)
   Comments:_____

7. **Distribution of the New Information (DNI):**
   (diffused) 1 2 3 4 5 (compact)
   Comments:_____

8. **Type of Information(TI):**
   (concrete) 1 2 3 4 5 (abstract)
   Comments:_____

9. **Topic Specificity(TS):**
   (not specific) 1 2 3 4 5 (highly specific)
   Comments:_____

10. **Cultural Specificity (CS):**
    (not specific) 1 2 3 4 5 (highly specific)
    Comments:_____

11. **Overall Difficulty (OD):**
    (easy) 1 2 3 4 5 (difficult)

Comments:_____

ITEM CHARACTERISTICS: IELTS READING PASSAGE 2: pp: 87-88 (863 words)
q: question O: operation TS: text span

Mark (X) the operation(s)/skill(s) that you think have contributed to arriving at the correct answer. (You can choose more than one)

2nd text	q14	q15	q16	q17	q18	q19	q20	q21	q22	q23	q24	q25	q26
O1													
O2													
O3													
O4													
O5													
O6													
O7													
O8													
O9													
O10													
O11													
O12													
O13													
O14													
O15													

Mark (X) the text span that you think should be processed in answering the question. (Choose one)

2nd text	q14	q15	q16	q17	q18	q19	q20	q21	q22	q23	q24	q25	q26
TS1													
TS2													
TS3													
TS4													
TS5													

TEXT CHARACTERISTICS: IELTS READING PASSAGE 3: IN SEARCH OF THE HOLY GRAIL pp: 91-92 (1004 words)

1. **Nature of the text (NT):** Mark as it applies (you can mark more than one): (1) instructions (2) sign (3) message (4) leaflet (5) brochure (6) advertisement (7) newspaper article (8) magazine article (9) research/journal article (10) text book article (11) other: _____

2. **Grammar/Syntax (GR):**
   (basic) 1 2 3 4 5 (complex)
   Comments: _____

3. **Vocabulary (VOC):**
   (frequent) 1 2 3 4 5 (less frequent)
   Comments: _____

4. **Cohesion (COH):**
   (explicit) 1 2 3 4 5 (not explicit)
   Comments: _____

5. **Rhetorical Organisation (RO):** Mark as it applies:(1) narration (2) description (3) information (4)comparison and contrast (5) classification (6) process analysis (7) discursive (8) argumentation
   (9) other: _____
   Degree of rhetorical organisation explicitness:
   (explicit) 1 2 3 4 5 (not explicit)
   Comments: _____

6. **Degree of Contextualisation(DC):**
   (context embedded) 1 2 3 4 5 (context reduced)
   Comments: _____

7. **Distribution of the New Information (DNI):**
   (diffused) 1 2 3 4 5 (compact)
   Comments: _____

8. **Type of Information(TI):**
   (concrete) 1 2 3 4 5 (abstract)
   Comments: _____

9. **Topic Specificity(TS):**
   (not specific) 1 2 3 4 5 (highly specific)
   Comments: _____

10. **Cultural Specificity (CS):**
    (not specific) 1 2 3 4 5 (highly specific)
    Comments: _____

11. **Overall Difficulty (OD):**
    (easy) 1 2 3 4 5 (difficult)

Comments: _____

ITEM CHARACTERISTICS: IELTS READING PASSAGE 3: IN SEARCH OF THE HOLY GRAIL pp: 91-92 (1004 words)
q: question O: operation TS: text span

Mark (X) the operation(s)/skill(s) that you think have contributed to arriving at the correct answer. (You can choose more than one)

ISHG	q27	q28	q29	q30	q31	q32	q33	q34	q35	q36	q37	q38	q39	q40
O1														
O2														
O3														
O4														
O5														
O6														
O7														
O8														
O9														
O10														
O11														
O12														
O13														
O14														
O15														

Mark (X) the text span that you think should be processed in answering the question. (Choose one)

ISHG	q27	q28	q29	q30	q31	q32	q33	q34	q35	q36	q37	q38	q39	q40
TS1														
TS2														
TS3														
TS4														
TS5														

How do you rate the overall difficulty of the sections below? Circle a number.

IELTS READING PASSAGE 1: GREEN WAVE WASHES OVER MAINSTREAM SHOPPING pp 83-84

Questions 1-6:
(very easy) 1 2 3 4 5 (very difficult)

Questions 7-9
(very easy) 1 2 3 4 5 (very difficult)

Questions 10-13
(very easy) 1 2 3 4 5 (very difficult)

IELTS READING PASSAGE 2: pp. 87-88

Questions 14-17
(very easy) 1 2 3 4 5 (very difficult)

Questions 18-21
(very easy) 1 2 3 4 5 (very difficult)

Questions 22-25
(very easy) 1 2 3 4 5 (very difficult)

Question 26
(very easy) 1 2 3 4 5 (very difficult)

IELTS READING PASSAGE 3: IN SEARCH OF HOLY GRAIL pp. 91-92

Questions 27-32
(very easy) 1 2 3 4 5 (very difficult)

Questions 33-40
(very easy) 1 2 3 4 5 (very difficult)

TEST: BUEPT

Read the instructions and respond to the questions by circling a number in the scales given. Mention/explain any problems, important details or your personal observations on the comment lines.

TEST RUBRIC CHARACTERISTICS

1. **Test Organisation (TO):**
   (clear) 1 2 3 4 5 (not clear)
   Comments:_____
   _____

2. **Time Allocation (TA):**
   (sufficient) 1 2 3 4 5 (insufficient)
   Comments:_____
   _____

3. **Instructions (I):**
   (clear) 1 2 3 4 5 (not clear)
   Comments:_____
   _____

TEXT CHARACTERISTICS: BUEPT READING PASSAGE 1: THE PINCH OF SALT SOLUTION pp: 1-3 (1763 words)

1. **Nature of the text (NT):** Mark as it applies (you can mark more than one): (1) instructions (2) sign (3) message (4) leaflet (5) brochure (6) advertisement (7) newspaper article (8) magazine article (9) research/journal article (10) text book article (11) other:
   _____

2. **Grammar/Syntax (GR):**
   (basic) 1 2 3 4 5 (complex)
   Comments:_____

3. **Vocabulary (VOC):**
   (frequent) 1 2 3 4 5 (less frequent)
   Comments:_____

4. **Cohesion (COH):**
   (explicit) 1 2 3 4 5 (not explicit)
   Comments:_____

5. **Rhetorical Organisation (RO):** Mark as it applies:(1) narration (2) description (3) information (4)comparison and contrast (5) classification (6) process analysis (7) discursive (8) argumentation
   (9) other: _____
   Degree of rhetorical organisation explicitness:
   (explicit) 1 2 3 4 5 (not explicit)
   Comments:_____

6. **Degree of Contextualisation(DC):**
   (context embedded) 1 2 3 4 5 (context reduced)
   Comments:_____

7. **Distribution of the New Information (DNI):**
   (diffused) 1 2 3 4 5 (compact)
   Comments:_____

8. **Type of Information(TI):**
(concrete) 1 2 3 4 5 (abstract)
Comments:_____

9. **Topic Specificity(TS):**
(not specific) 1 2 3 4 5 (highly specific)
Comments:_____

10. **Cultural Specificity (CS):**
(not specific) 1 2 3 4 5 (highly specific)
Comments:_____

11. **Overall Difficulty (OD):**
(easy) 1 2 3 4 5 (difficult)

Comments:_____

ITEM CHARACTERISTICS: BUEPT READING PASSAGE 1: THE PINCH OF SALT SOLUTION pp: 1-3 (1763 words)
q: question O: operation TS: text span

Mark (X) the operation(s)/skill(s) that you think have contributed to arriving at the correct answer. (You can choose more than one)

VTE	q1	q2	q3	q4	q5	Q6	q7	q8	q9	q10
O1										
O2										
O3										
O4										
O5										
O6										
O7										
O8										
O9										
O10										
O11										
O12										
O13										
O14										
O15										

Mark (X) the text span that you think should be processed in answering the question. (Choose one)

VTE	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10
TS1										
TS2										
TS3										
TS4										
TS5										

TEXT CHARACTERISTICS: BUEPT READING PASSAGE 2: NEW PERSPECTIVES ON CHILD EDUCATION pp: 4-7 (2474 words)

1. **Nature of the text (NT):** Mark as it applies (you can mark more than one): (1) instructions (2) sign (3) message (4) leaflet (5) brochure (6) advertisement (7) newspaper article (8) magazine article (9) research/journal article (10) text book article (11) other: _____

2. **Grammar/Syntax (GR):**
   (basic) 1 2 3 4 5 (complex)
   Comments: _____

3. **Vocabulary (VOC):**
   (frequent) 1 2 3 4 5 (less frequent)
   Comments: _____

4. **Cohesion (COH):**
   (explicit) 1 2 3 4 5 (not explicit)
   Comments: _____

5. **Rhetorical Organisation (RO):** Mark as it applies:(1) narration (2) description (3) information (4)comparison and contrast (5) classification (6) process analysis (7) discursive (8) argumentation
   (9) other: _____
   Degree of rhetorical organisation explicitness:
   (explicit) 1 2 3 4 5 (not explicit)
   Comments: _____

6. **Degree of Contextualisation(DC):**
   (context embedded) 1 2 3 4 5 (context reduced)
   Comments: _____

7. **Distribution of the New Information (DNI):**
   (diffused) 1 2 3 4 5 (compact)
   Comments: _____

8. **Type of Information(TI):**
   (concrete) 1 2 3 4 5 (abstract)
   Comments: _____

9. **Topic Specificity(TS):**
   (not specific) 1 2 3 4 5 (highly specific)
   Comments: _____

10. **Cultural Specificity (CS):**
    (not specific) 1 2 3 4 5 (highly specific)
    Comments: _____

11. **Overall Difficulty (OD):**
    (easy) 1 2 3 4 5 (difficult)

Comments: _____

ITEM CHARACTERISTICS: <u>BUEPT READING PASSAGE 2:</u> NEW PERSPECTIVES ON CHILD EDUCATION pp: 4-7 (2474 words)
q: question O: operation TS: text span

Mark (X) the operation(s)/skill(s) that you think have contributed to arriving at the correct answer. (You can choose <u>more than one</u>)

NPCE	q1	q2	q3	q4	q5	q6
O1						
O2						
O3						
O4						
O5						
O6						
O7						
O8						
O9						
O10						
O11						
O12						
O13						
O14						
O15						

Mark (X) the text span that you think should be processed in answering the question. (Choose <u>one</u>)

NPCE	q1	q2	q3	q4	q5	q6
TS1						
TS2						
TS3						
TS4						
TS5						

TEXT CHARACTERISTICS: BUEPT READING PASSAGE 3: CAN ANIMALS LEARN TO SHARE, COOPERATE ... pp: 8-10 (1990 words)

1. **Nature of the text (NT):** Mark as it applies (you can mark more than one): (1) instructions (2) sign (3) message (4) leaflet (5) brochure (6) advertisement (7) newspaper article (8) magazine article (9) research/journal article (10) text book article (11) other: _____

2. **Grammar/Syntax (GR):**
   (basic) 1 2 3 4 5 (complex)
   Comments:_____

3. **Vocabulary (VOC):**
   (frequent) 1 2 3 4 5 (less frequent)
   Comments:_____

4. **Cohesion (COH):**
   (explicit) 1 2 3 4 5 (not explicit)
   Comments:_____

5. **Rhetorical Organisation (RO):** Mark as it applies:(1) narration (2) description (3) information (4)comparison and contrast (5) classification (6) process analysis (7) discursive (8) argumentation
   (9) other: _____
   Degree of rhetorical organisation explicitness:
   (explicit) 1 2 3 4 5 (not explicit)
   Comments:_____

6. **Degree of Contextualisation(DC):**
   (context embedded) 1 2 3 4 5 (context reduced)
   Comments:_____

7. **Distribution of the New Information (DNI):**
   (diffused) 1 2 3 4 5 (compact)
   Comments:_____

8. **Type of Information(TI):**
   (concrete) 1 2 3 4 5 (abstract)
   Comments:_____

9. **Topic Specificity(TS):**
   (not specific) 1 2 3 4 5 (highly specific)
   Comments:_____

10. **Cultural Specificity (CS):**
    (not specific) 1 2 3 4 5 (highly specific)
    Comments:_____

11. **Overall Difficulty (OD):**
    (easy) 1 2 3 4 5 (difficult)

Comments:_____

ITEM CHARACTERISTICS: <u>BUEPT READING PASSAGE 3:</u> CAN ANIMALS LEARN TO SHARE, COOPERATE … pp: 8-10 (1990 words)
q: question O: operation TS: text span

Mark (X) the operation(s)/skill(s) that you think have contributed to arriving at the correct answer. (You can choose <u>more than one</u>)

Caltsc..	q1	q2	q3	q4	q5
O1					
O2					
O3					
O4					
O5					
O6					
O7					
O8					
O9					
O10					
O11					
O12					
O13					
O14					
O15					

Mark (X) the text span that you think should be processed in answering the question. (Choose <u>one</u>)

CALTSC..	q1	q2	q3	q4	q5
TS1					
TS2					
TS3					
TS4					
TS5					

How do you rate the overall difficulty of the sections below? Circle a number.

<u>BUEPT READING PASSAGE 1:</u> THE PINCH OF SALT SOLUTION pp: 1-3

Questions 1-10:
(very easy) 1 2 3 4 5 (very difficult)

<u>BUEPT READING PASSAGE 2:</u> NEW PERSPECTIVES ON CHILD EDUCATION pp: 4-7

Question 1:
(very easy) 1 2 3 4 5 (very difficult)

Questions 2-6:
(very easy) 1 2 3 4 5 (very difficult)

<u>BUEPT READING PASSAGE 3:</u> CAN ANIMALS LEARN TO SHARE, COOPERATE ….. pp: 8-10

Questions: 1-5:
(very easy) 1 2 3 4 5 (very difficult)

APPENDIX 3.4
VERBAL PROTOCOL ANALYSIS SCHEME

OPERATIONS: These refer to the operation(s)/skill(s) that are expected to contribute to arriving at the correct answer. Therefore, the test taker will normally use one or more of these operations while answering the questions.

O1. rapidly looking for figures, dates, names, etc in the text.
O2. rapidly inspecting the text (and go back and forth in it) to locate the answer.
O3. matching the exact key words/phrases in the question and in the text.
O4. matching the key words/phrases in the question with their synonyms / paraphrases in the text.
O5. using my own knowledge of how the text is structured (knowledge of formal text structure) in order to locate the answers.
O6. using the title, subtitles and section headings, the first and last sentences of the paragraphs.
O7. reading the abstract/introduction and conclusion carefully.
O8. reading carefully to confirm the answer after deciding the location of information.
O9. forming a summary of the main ideas/text topic in mind.
O10. reading slowly and carefully for detailed understanding of explicitly stated ideas in the text (when there are the same key word/words in both the question and the text).
O11. reading slowly and carefully for detailed understanding of an idea in the text (when there are no key words occuring in both text and question).
O12. reading a part of a text more than once in order to understand it.
O13. focusing on pronouns, discourse markers, grammar, etc.
O14. deducing the meaning of a word from the context.
O15. dealing with relatively uncommon vocabulary.

TEXT SPAN: This is originally related with the relationship of the item to the passage in terms of the amount of text that should be processed for successful comprehension. Here, text span refers to the part of the text that the test taker processed in order to arrive at the answer.

TS1. no relationship to the passage; item can be answered without reference to the passage, *or* relationship of item to passage is not clear
TS2*.relates to a specific part of the passage (one sentence or less than a sentence), and requires only localised understanding of that part
TS2. relates to a specific part of the passage (more than a sentence but a paragraph at most) and requires only localised understanding of that part
TS3. relates to several specific parts of the passage, or requires test taker to relate one part of the passage to several others
TS4. item relates to the entire passage, and requires an understanding of the entire passage
TS5. requires test taker to relate information in passage to the real world, outside the text

TEST TAKING STRATEGIES

tts1. using the order of the question(s) in the text as a clue to locate the text span where the answer is possibly located / skipping the parts where former questions are located.

e.g.: 'The answer to second question should be somewhere here between the first and second questions so I will check this part.'
'The first question must be somewhere on the first page.'
'The third question is about Kummer and Cord's experiment. So the answer to the second question must be before that.'

tts2. matching the text and the question at word/phrase level to extract the answer after locating the text span where the answer is possibly located without substantial understanding (differs from scanning in that TTS2 is not a fast process of looking for specific information but a slow and careful matching of the linguistic items in the question and the text).

e.g.: 'The question asks for some 'ability to do something' and the text says 'capability to grasp'. 'Ability' is synomymous with 'capability'. I think the answer should be 'grasp'.
'There are two blanks in the question and at the beginning of this paragraph, two concepts are given: 'experience' and 'action'. These should be the answers.'

tts3. using grammatical clues to extract the answer from the text

e.g.: 'I thought I should find a verb to fill in this blank. I picked up the verb of this sentence because I thought the answer is here.'

tts4. using text knowledge (organisation of the information in the paragraph) as a clue to locate the answer without substantial comprehension.

e.g.: 'The answer should be located in this paragraph because it is talking about preschool children and the question is about preschool children, too. There is only one thing discussed here in this paragraph and it is 'reversibility'. And the question is asking for something like a concept. I thought the answer should be 'reversibility'.
'The purpose of an experiment is given at the beginning of a paragraph and the result towards the end. Therefore, I went to the last lines for the result because I think the question asks the finding.

OBSERVATIONS

OBS1. understood the question/ Yes(Y), No(N), Partially(P)
OBS2. located the part of the text that contained the answer correctly/Yes(Y), No(N)
OBS3. understood the part of the text that contained (or assumed to contain) the answer correctly/ Yes(Y), No(N), Partially(P)
OBS4. answered the question correctly/ Yes(Y), No(N)
Unanswered: The test taker did not provide or did not have time to provide an answer to the question/ (-)

VERBAL PROTOCOL CHECKLIST

Subject:

Operations and Text Span

Sub...	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10	O11	O12	O13	O14	O15	TS1	TS2	TS3	TS4	TS5
SC1																				
SC2																				
SC3																				
SC4																				
SC5																				
SC6																				
SC7																				
SC8																				
SC9																				
SC10																				
SR1																				
SR2																				
SR3																				
SR4																				
SR5																				
SR6																				
SR7																				
SR8																				
SR9																				
SR10																				
SR11																				
CR1																				
CR2																				
CR3																				
CR4																				
CR5																				
CR6																				
CR7																				
CR8																				
CR9																				
CR10																				

Test Taking Strategies and Observations

Sub...	tts1	tts2	tts3	tts4	tts5	tts6	OBS1	OBS2	OBS3	OBS4
SC1										
SC2										
SC3										
SC4										
SC5										
SC6										
SC7										
SC8										
SC9										
SC10										
SR1										
SR2										
SR3										
SR4										
SR5										
SR6										
SR7										
SR8										
SR9										
SR10										
SR11										
CR1										
CR2										
CR3										
CR4										
CR5										
CR6										
CR7										
CR8										
CR9										
CR10										

APPENDIX 4.1

Correct Responses By The Test Takers

	SC/10	SR/11		CR/10	
		SRI/6	SRII/5	CRI/5	CRII/5
S1	10	1	1	3	3
S2	9	2	5	4	3
S3	10	3	4	5	2
S4	8	4	3	5	5
S5	7	1	4	1	4
S6	8	4	4	4	4
S7	10	2	2	2	3
S8	9	6	5	5	5
S9	9	1	4	1	3
S10	8	2	5	4	5
S11	7	4	4	5	4
S12	9	2	4	2	4
S13	10	3	3	2	3
S14	9	4	2	3	4
S15	10	4	5	5	4

APPENDIX 4.2
September 2000 – Pilot Version
Normality Tests and Graphs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
total readi	81	100.0%	0	.0%	81	100.0%

Descriptives

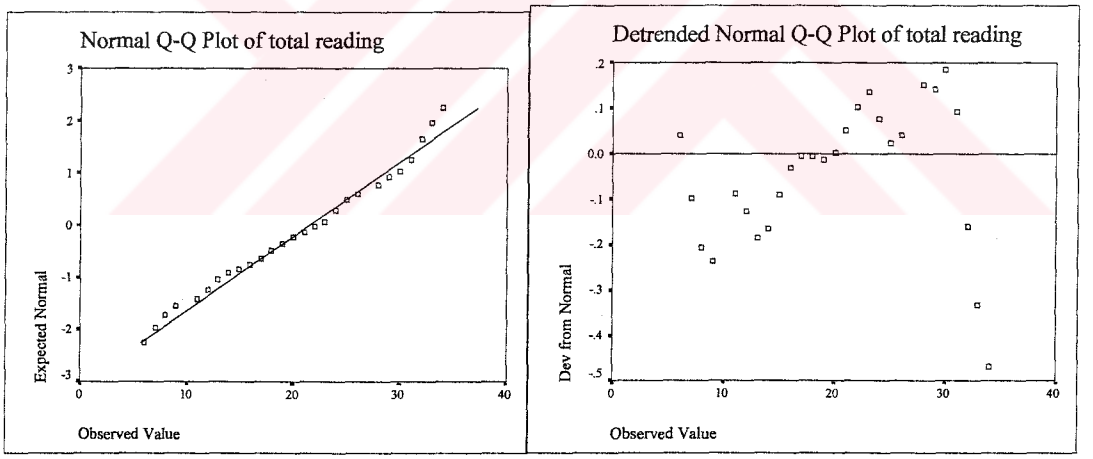
			Statistic	Std. Error
total reading	Mean		21.5062	.7792
	95% Confidence Interval for Mean	Lower Bound	19.9555	
		Upper Bound	23.0568	
	5% Trimmed Mean		21.6728	
	Median		22.0000	
	Variance		49.178	
	Std. Deviation		7.0127	
	Minimum		6.00	
	Maximum		34.00	
	Range		28.00	
	Interquartile Range		10.0000	
	Skewness		-.253	.267
	Kurtosis		-.726	.529

Tests of Normality

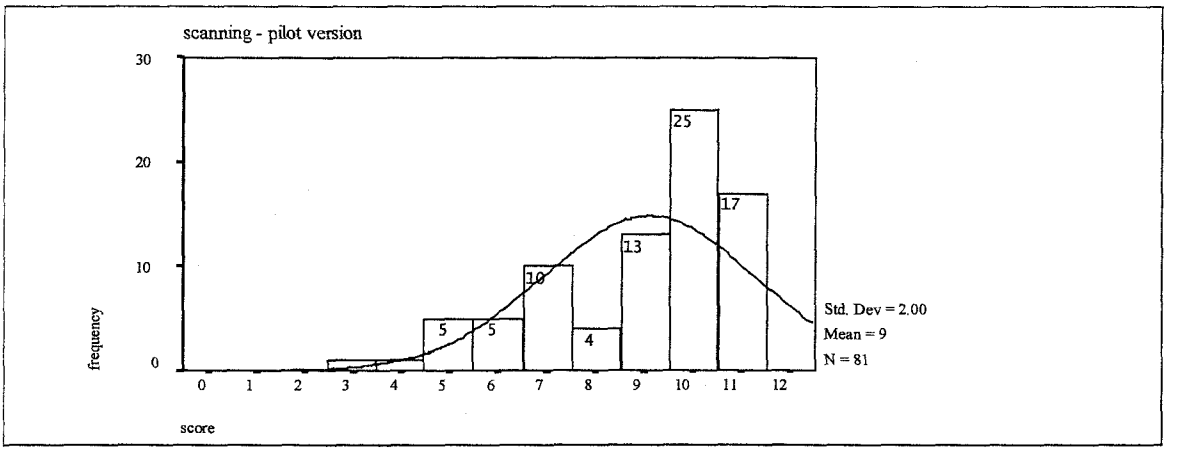
	Kolmogorov-Smirnov [a]		
	Statistic	df	Sig.
total reading	.083	81	.200*

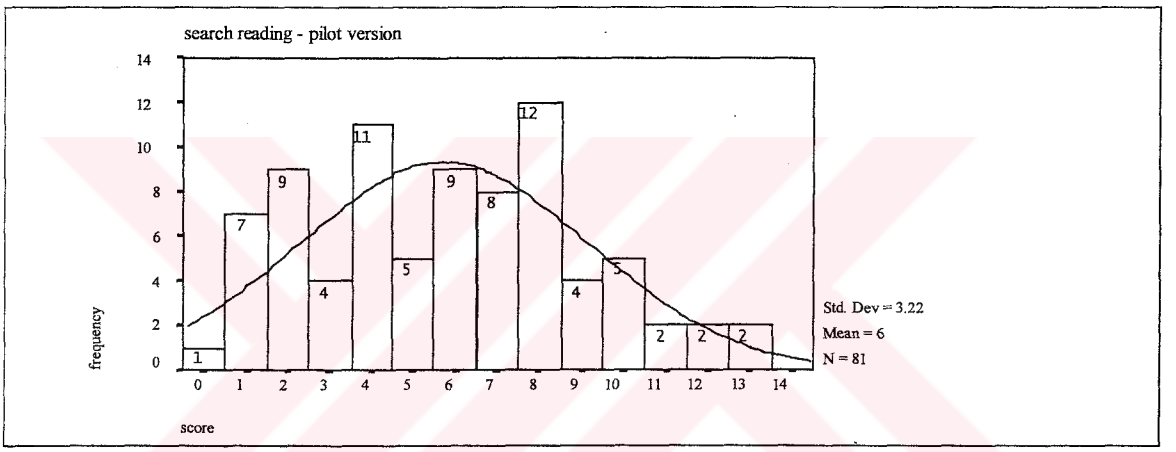
*. This is a lower bound of the true significance.

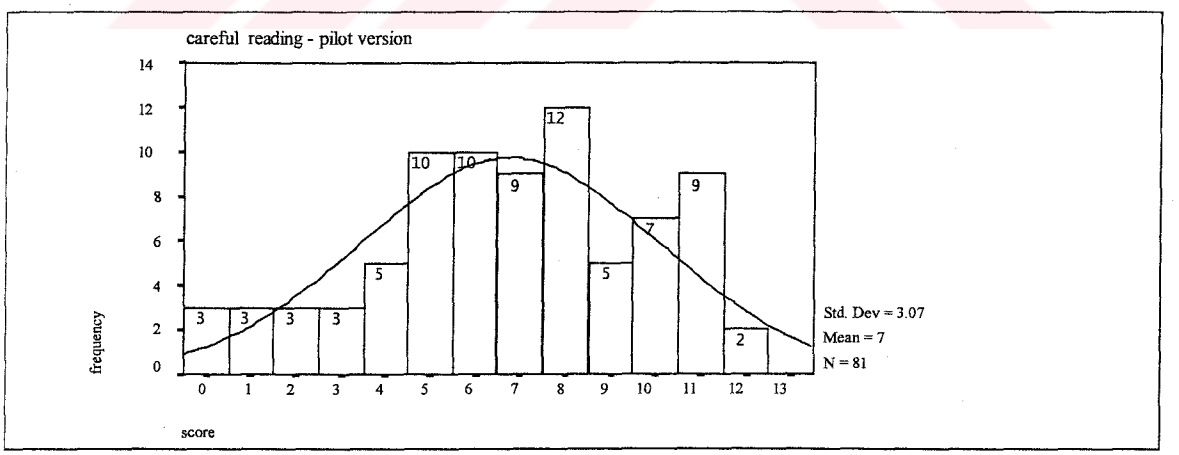
a. Lilliefors Significance Correction



APPENDIX 4.3
September 2000 – Pilot Version
Score Distribution Graphs by Subtest







APPENDIX 4.4
September 2000 – Pilot Version
Normality Tests and Graphs by Subtests

Scanning

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
total scanning	81	100.0%	0	.0%	81	100.0%

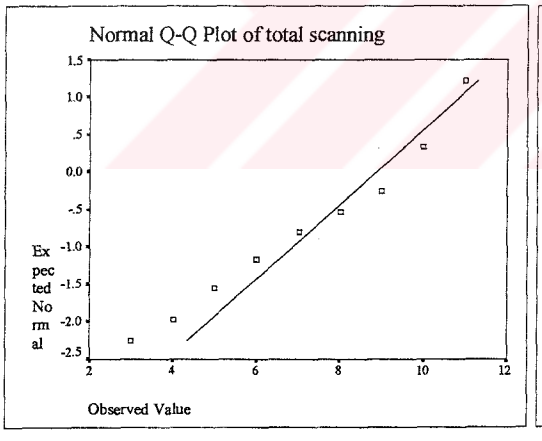
Descriptives

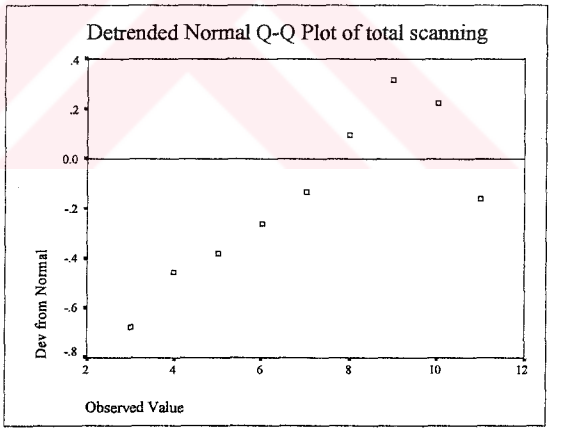
			Statistic	Std. Error
total scanning	Mean		8.8642	.2227
	95% Confidence Interval for Mean	Lower Bound	8.4209	
		Upper Bound	9.3075	
	5% Trimmed Mean		9.0014	
	Median		10.0000	
	Variance		4.019	
	Std. Deviation		2.0047	
	Minimum		3.00	
	Maximum		11.00	
	Range		8.00	
	Interquartile Range		3.0000	
	Skewness		-.935	.267
	Kurtosis		-.019	.529

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
total scanning	.233	81	.000

a. Lilliefors Significance Correction





Search Reading

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
total search rea	81	100.0%	0	.0%	81	100.0%

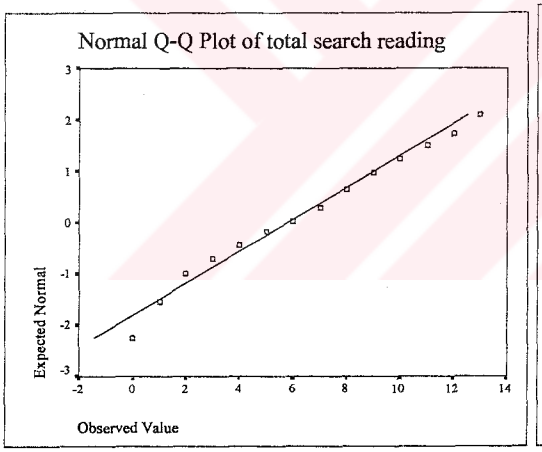
Descriptives

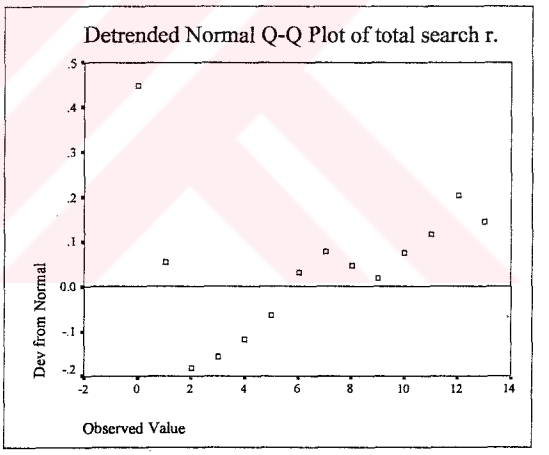
			Statistic	Std. Error
total search reading	Mean		5.8025	.3576
	95% Confidence Interval for Mean	Lower Bound	5.0907	
		Upper Bound	6.5142	
	5% Trimmed Mean		5.7119	
	Median		6.0000	
	Variance		10.360	
	Std. Deviation		3.2188	
	Minimum		.00	
	Maximum		13.00	
	Range		13.00	
	Interquartile Range		5.0000	
	Skewness		.201	.267
	Kurtosis		-.712	.529

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
total search reading	.107	81	.022

a. Lilliefors Significance Correction





Careful Reading

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
total careful reading	81	100.0%	0	.0%	81	100.0%

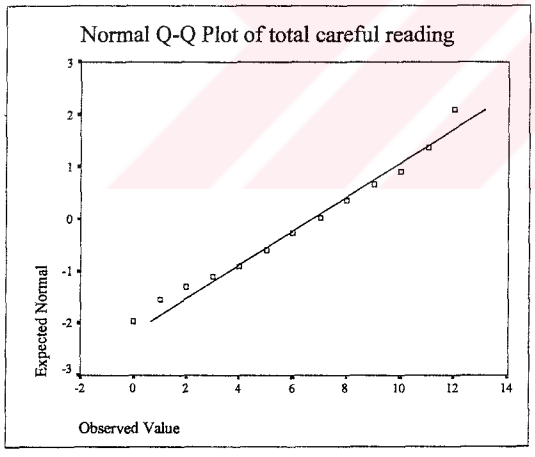
Descriptives

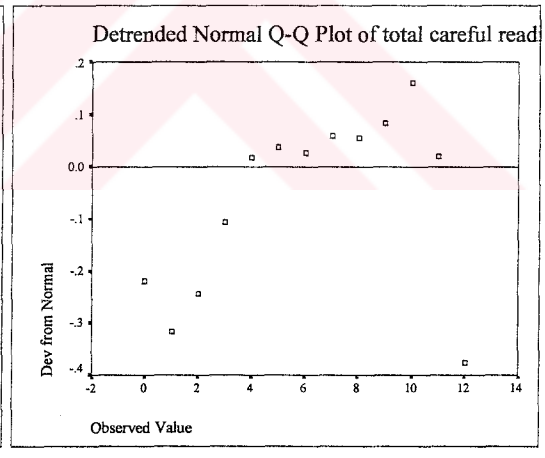
			Statistic	Std. Error
total careful reading	Mean		6.7284	.3416
	95% Confidence Interval for Mean	Lower Bound	6.0486	
		Upper Bound	7.4081	
	5% Trimmed Mean		6.8230	
	Median		7.0000	
	Variance		9.450	
	Std. Deviation		3.0741	
	Minimum		.00	
	Maximum		12.00	
	Range		12.00	
	Interquartile Range		4.0000	
	Skewness		-.346	.267
	Kurtosis		-.501	.529

Tests of Normality

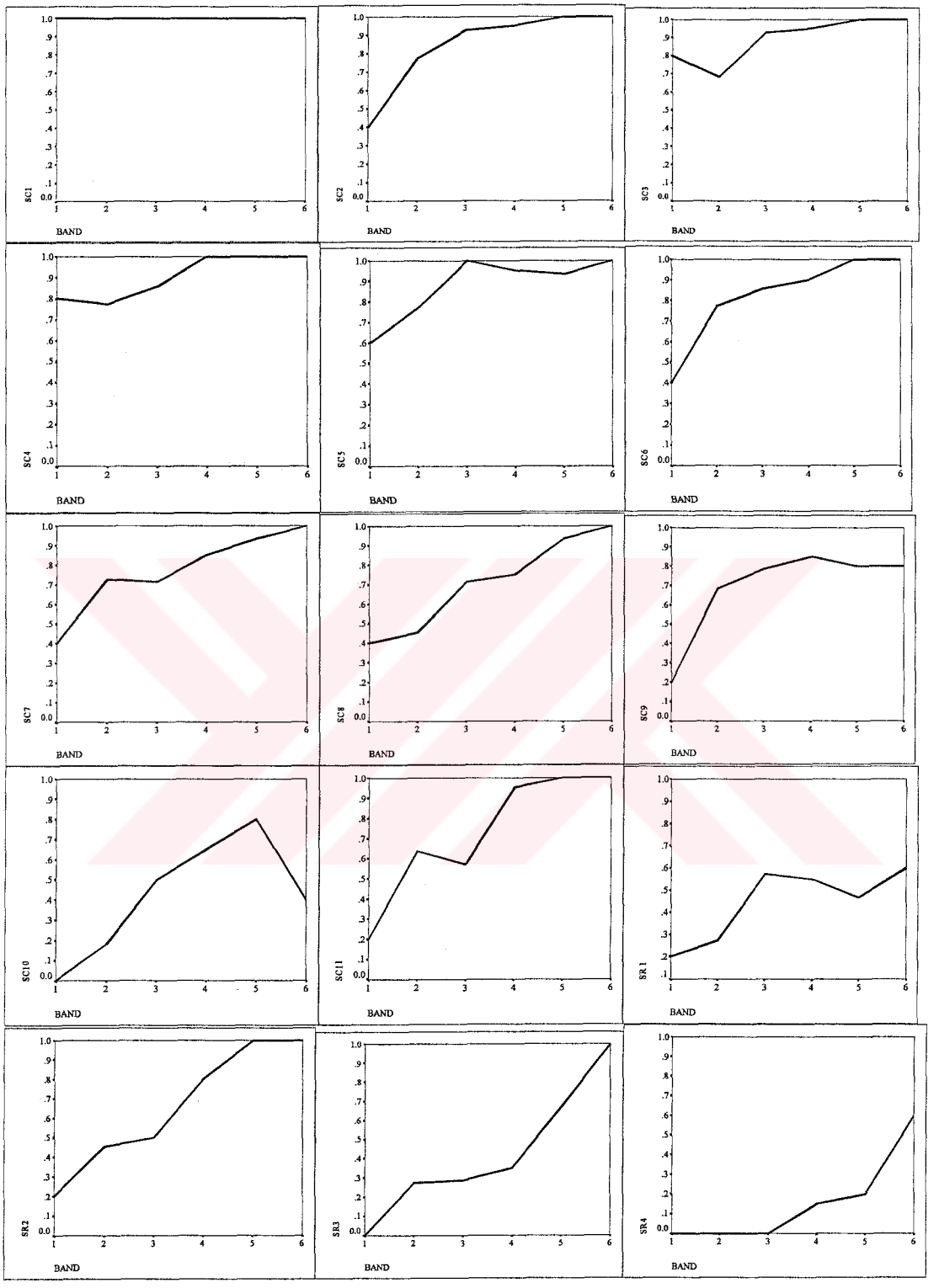
	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
total careful reading	.093	81	.083

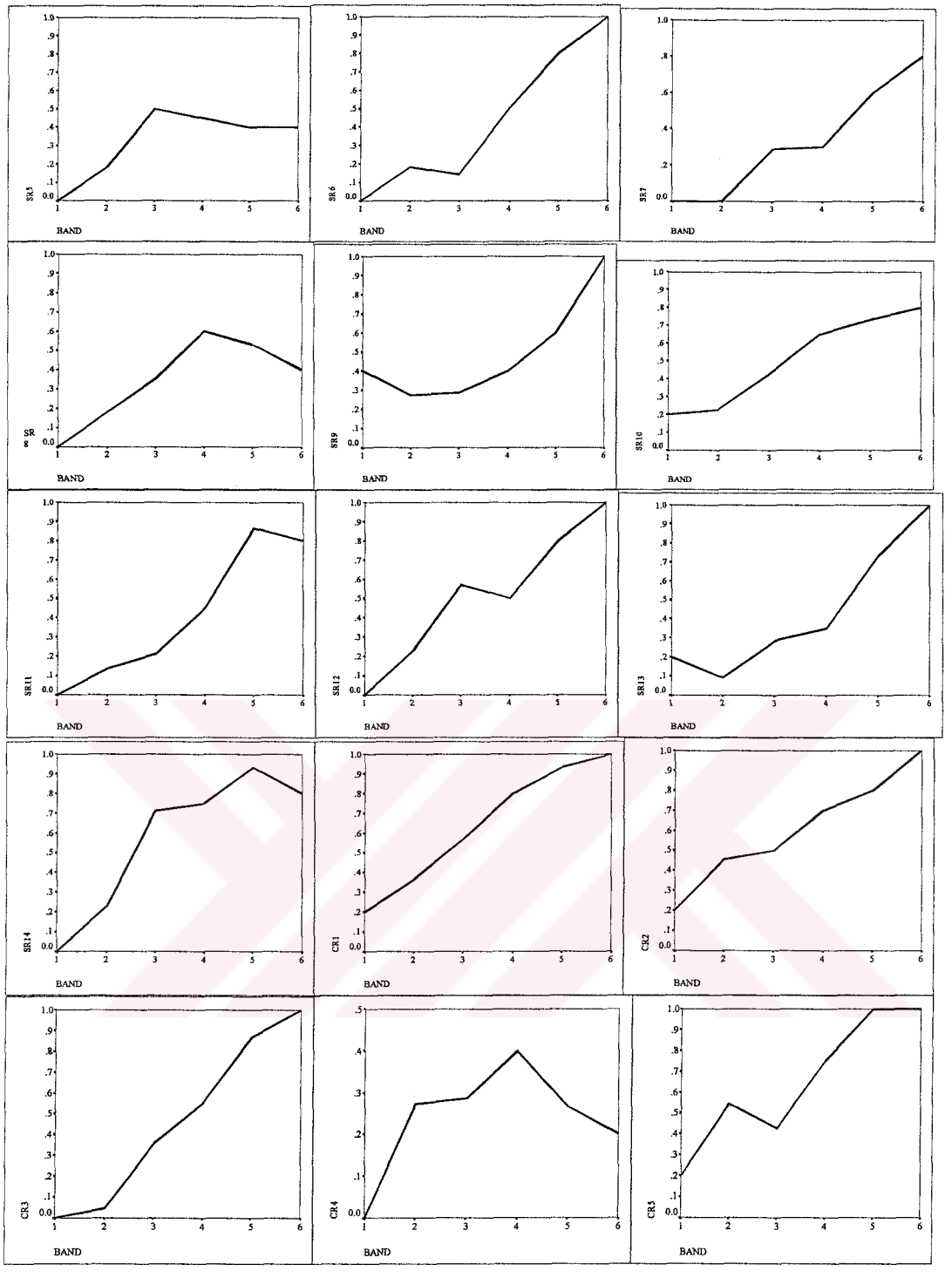
a. Lilliefors Significance Correction

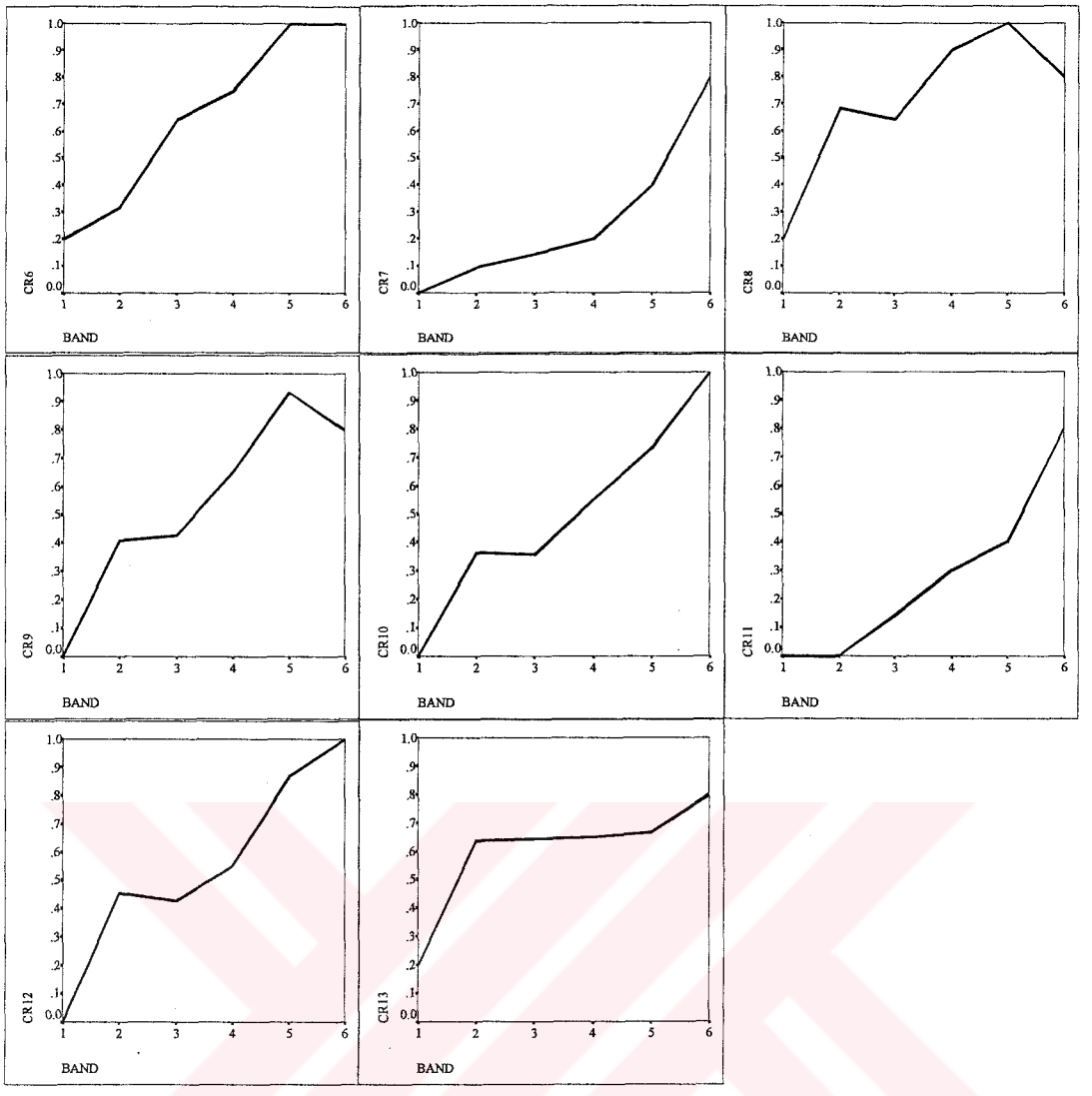




APPENDIX 4.5
September 2000 – Pilot Version
Band Score Graphs




SR5
SR6
SR7
SR8
SR9
SR10
SR11
SR12
SR13
SR14
CR1
CR2
CR3
CR4
CR5
BAND


CR6
CR7
CR8
CR9
CR10
CR11
CR12
CR13
BAND
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
0.0
1
2
3
4
5
6

APPENDIX 4.6
September 2000 Test
Normality Tests and Graphs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Total Reading	341	100.0%	0	.0%	341	100.0%

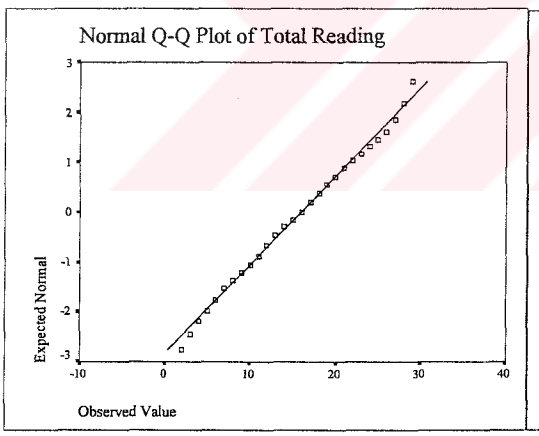
Descriptives

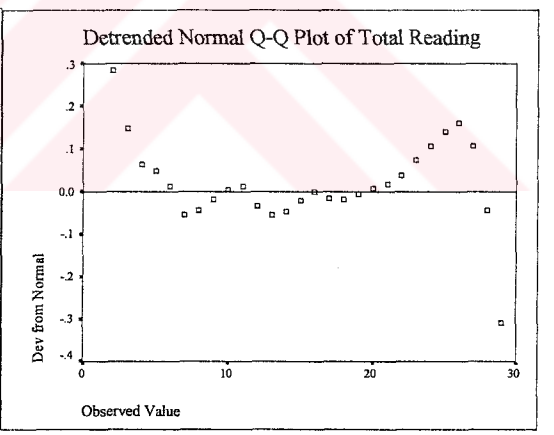
			Statistic	Std. Error
TOTALR	Mean		15.9501	.3057
	95% Confidence Interval for Mean	Lower Bound	15.3489	
		Upper Bound	16.5514	
	5% Trimmed Mean		15.9282	
	Median		16.0000	
	Variance		31.865	
	Std. Deviation		5.6449	
	Minimum		2.00	
	Maximum		29.00	
	Range		27.00	
	Interquartile Range		8.0000	
	Skewness		.076	.132
	Kurtosis		-.410	.263

Tests of Normality

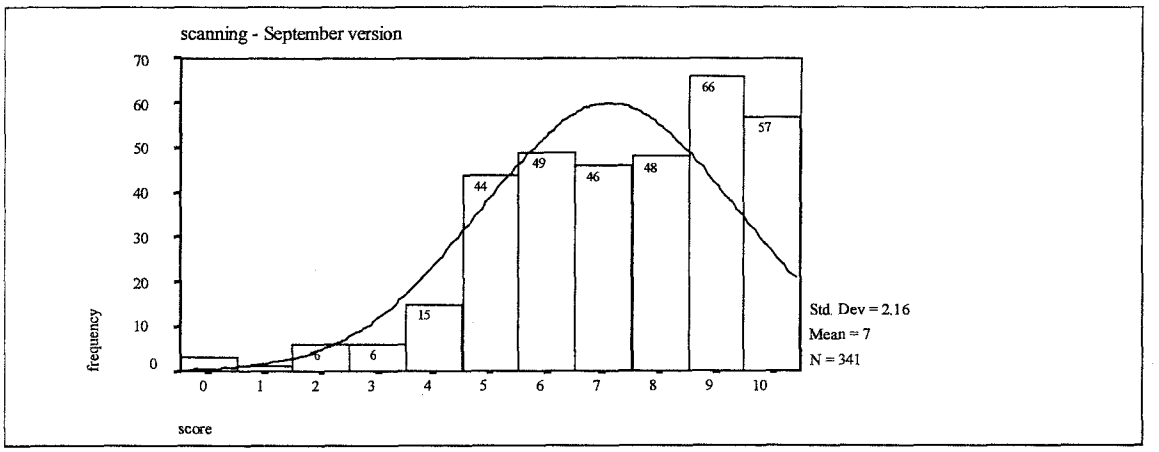
	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
TOTALR	.052	341	.029

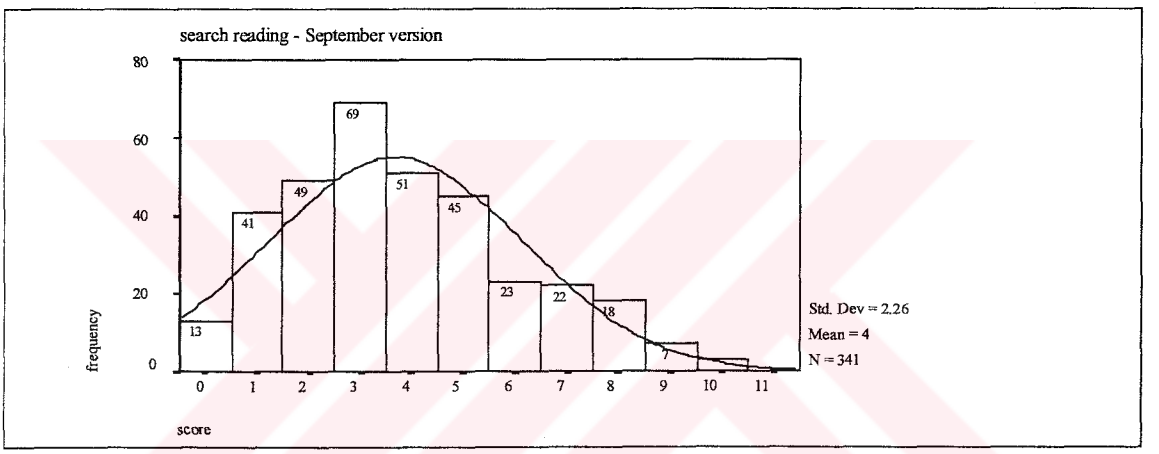
a. Lilliefors Significance Correction

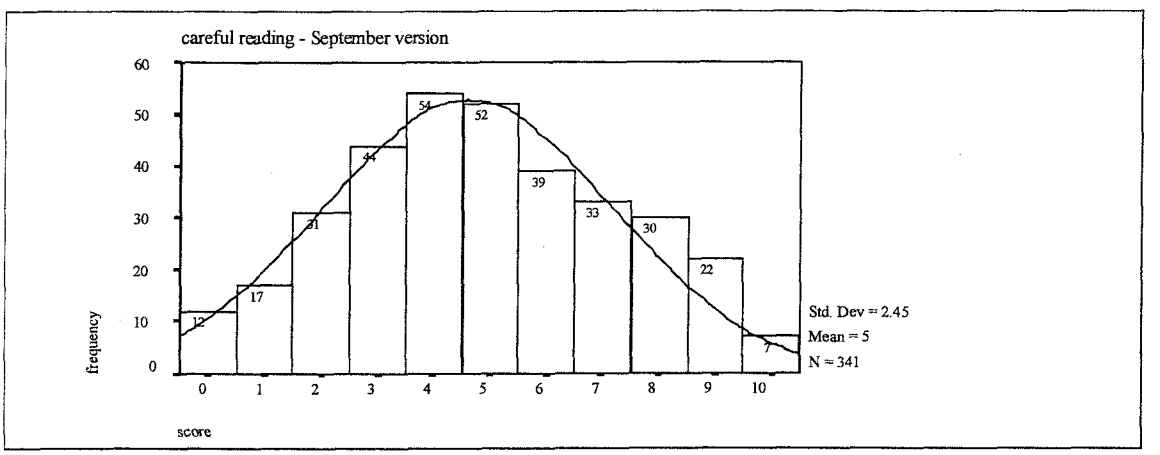




APPENDIX 4.7
September 2000 Test
Score Distribution Graphs by Subtest







APPENDIX 4.8
September 2000 Test
Normality Tests and Graphs by Subtests

Scanning

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
scanning to	341	100.0%	0	.0%	341	100.0%

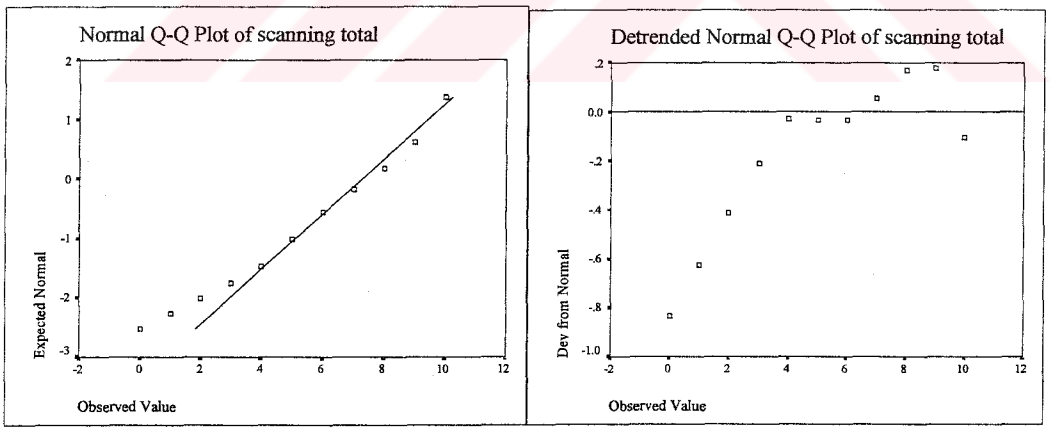
Descriptives

			Statistic	Std. Error
scanning total	Mean		7.2581	.1172
	95% Confidence Interval for Mean	Lower Bound	7.0276	
		Upper Bound	7.4885	
	5% Trimmed Mean		7.3943	
	Median		8.0000	
	Variance		4.680	
	Std. Deviation		2.1634	
	Minimum		.00	
	Maximum		10.00	
	Range		10.00	
	Interquartile Range		3.0000	
	Skewness		-.674	.132
	Kurtosis		.146	.263

Tests of Normality

	Kolmogorov-Smirnov [a]		
	Statistic	df	Sig.
scanning total	.150	341	.000

a. Lilliefors Significance Correction



Search Reading

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
search r. total	341	100.0%	0	.0%	341	100.0%

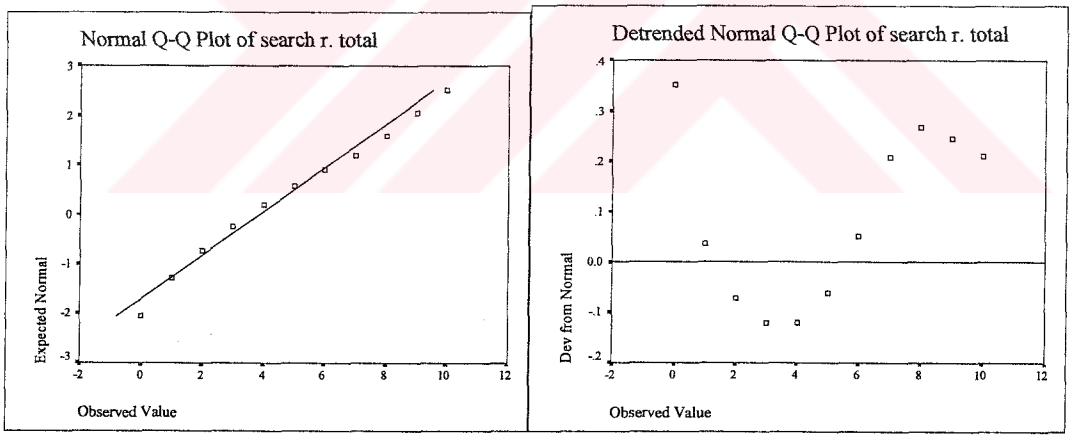
Descriptives

			Statistic	Std. Error
search r. total	Mean		3.8240	.1224
	95% Confidence Interval for Mean	Lower Bound	3.5833	
		Upper Bound	4.0648	
	5% Trimmed Mean		3.7489	
	Median		3.0000	
	Variance		5.110	
	Std. Deviation		2.2606	
	Minimum		.00	
	Maximum		10.00	
	Range		10.00	
	Interquartile Range		3.0000	
	Skewness		.517	.132
	Kurtosis		-.314	.263

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
search r. total	.147	341	.000

a. Lilliefors Significance Correction



Careful Reading

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
careful r. total	341	100.0%	0	.0%	341	100.0%

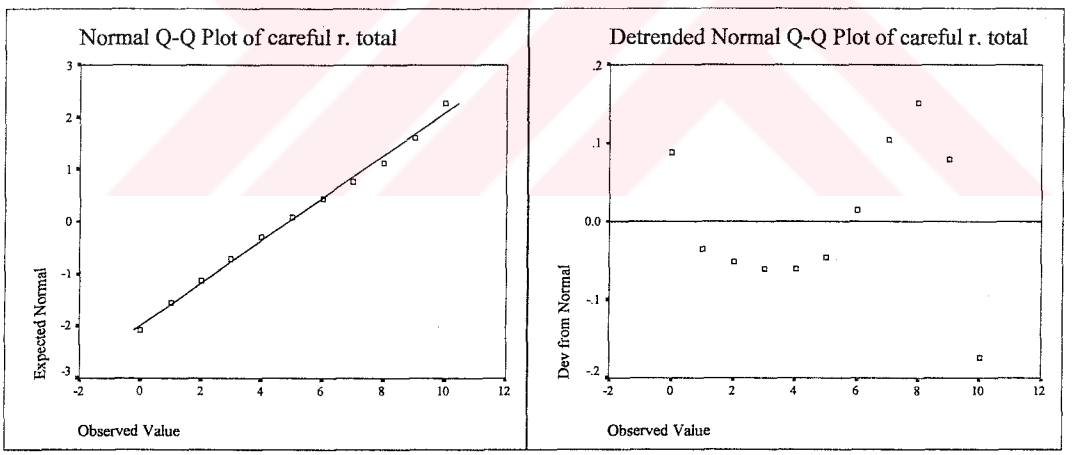
Descriptives

			Statistic	Std. Error
careful r. total	Mean		4.8680	.1327
	95% Confidence Interval for Mean	Lower Bound	4.6069	
		Upper Bound	5.1291	
	5% Trimmed Mean		4.8697	
	Median		5.0000	
	Variance		6.009	
	Std. Deviation		2.4513	
	Minimum		.00	
	Maximum		10.00	
	Range		10.00	
	Interquartile Range		4.0000	
	Skewness		.075	.132
	Kurtosis		-.696	.263

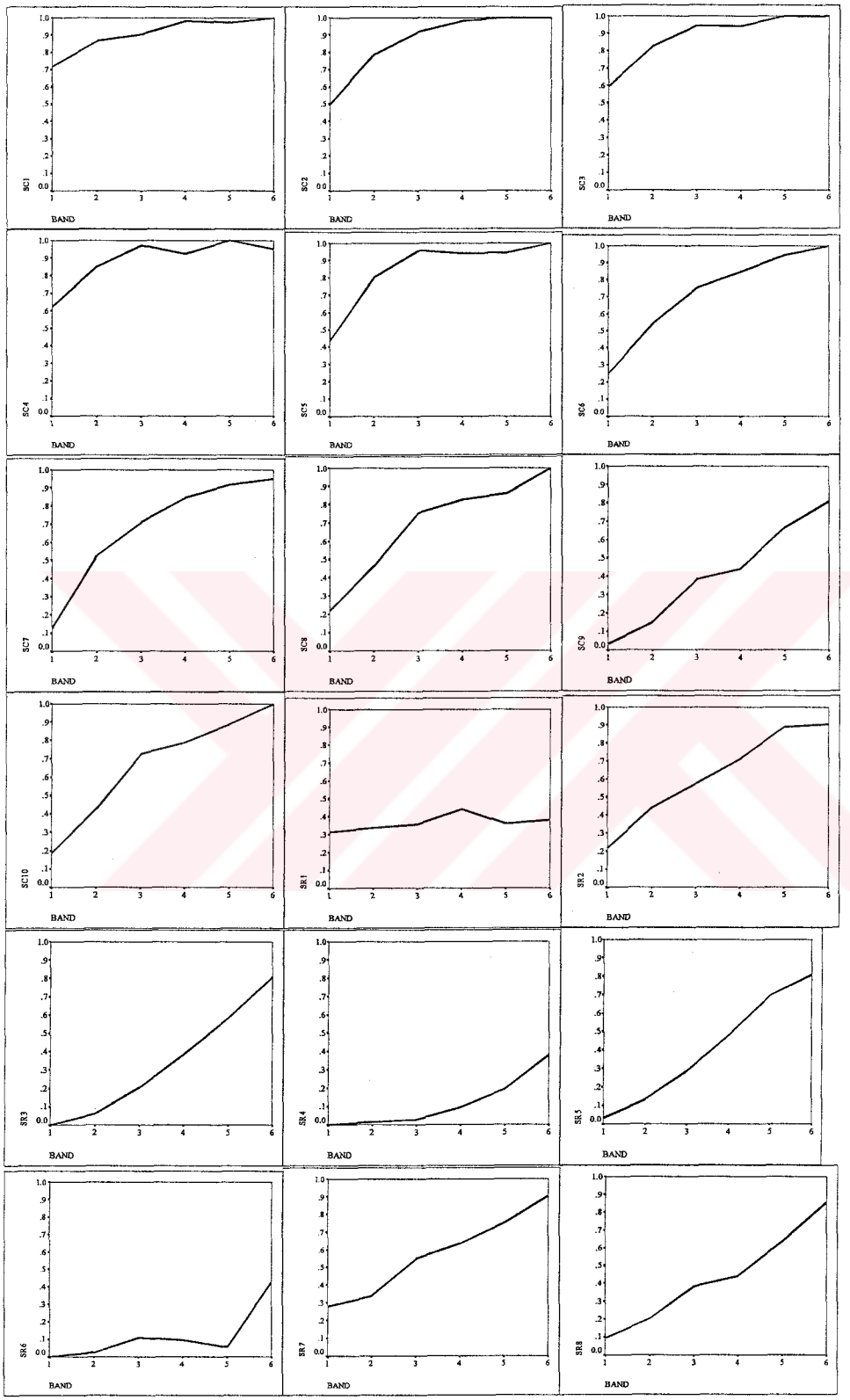
Tests of Normality

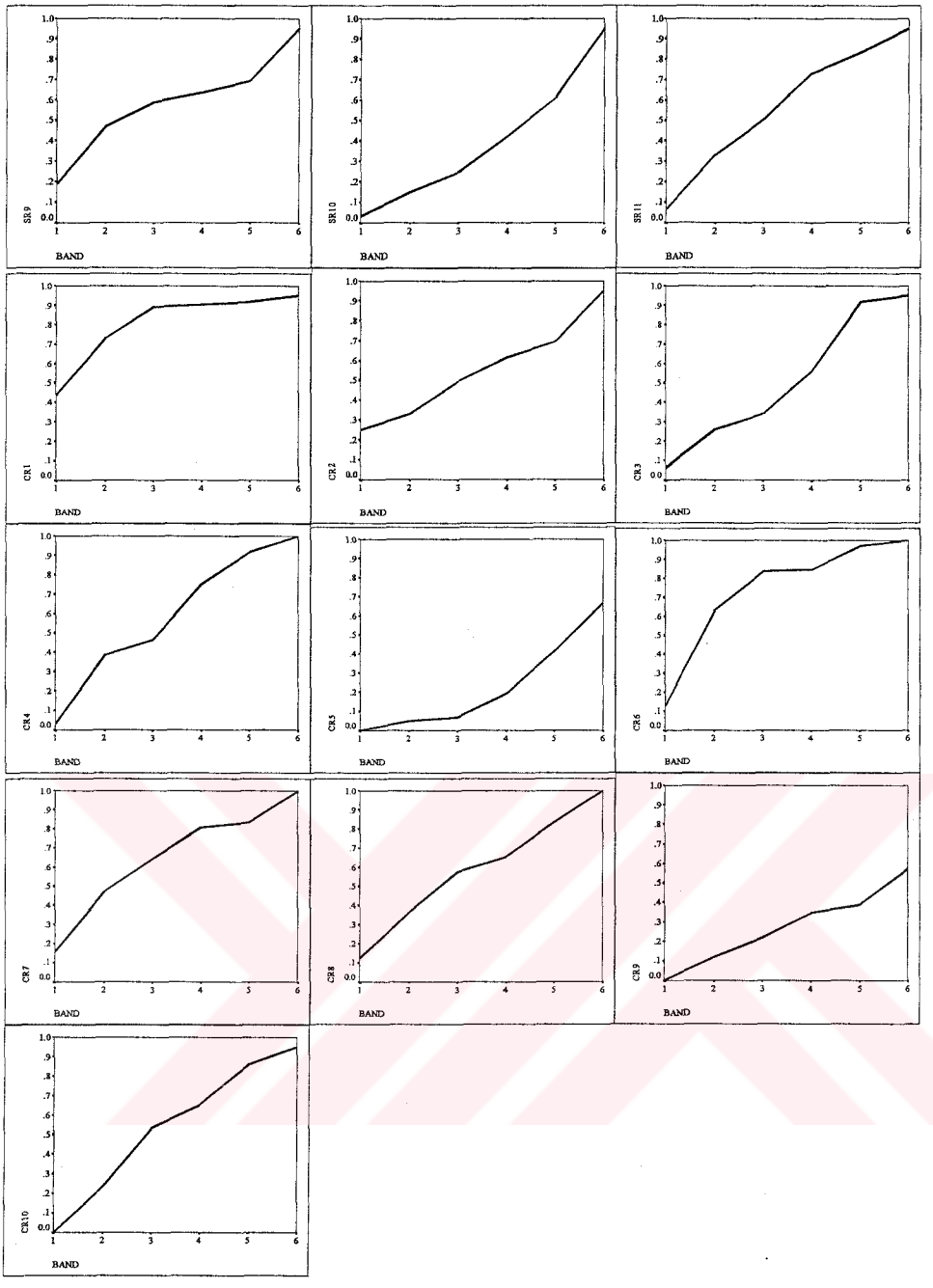
	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
careful r. total	.102	341	.000

a. Lilliefors Significance Correction



APPENDIX 4.9
September 2000 Test
Band Score Graphs



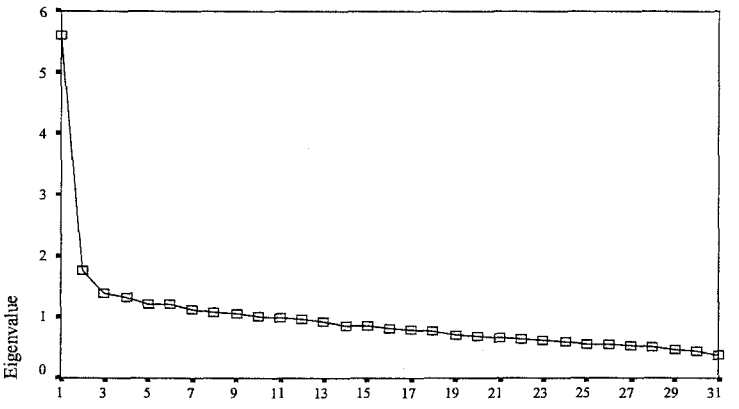

SR9
SR10
SR11
CR1
CR2
CR3
CR4
CR5
CR6
CR7
CR8
CR9
CR10
BAND

APPENDIX 4.10
PCA: September 2000 Test – Whole Set

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,843
Bartlett's Test of Sphericity	Approx. Chi-Square	1752,698
	df	465
	Sig.	,000

Scree Plot



Communalities

	Initial	Extraction
SC1	1,000	,508
SC2	1,000	,343
SC3	1,000	,651
SC4	1,000	,560
SC5	1,000	,559
SC6	1,000	,496
SC7	1,000	,592
SC8	1,000	,608
SC9	1,000	,585
SC10	1,000	,540
SR1	1,000	,542
SR2	1,000	,543
SR3	1,000	,521
SR4	1,000	,596
SR5	1,000	,610
SR6	1,000	,484
SR7	1,000	,571
SR8	1,000	,502
SR9	1,000	,485
SR10	1,000	,503
SR11	1,000	,497
CR1	1,000	,619
CR2	1,000	,721
CR3	1,000	,580
CR4	1,000	,504
CR5	1,000	,515
CR6	1,000	,547
CR7	1,000	,452
CR8	1,000	,517
CR9	1,000	,487
CR10	1,000	,514

Extraction Method: Principal Component Analysis.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5,597	18,054	18,054	5,597	18,054	18,054
2	1,761	5,682	23,736	1,761	5,682	23,736
3	1,379	4,448	28,183	1,379	4,448	28,183
4	1,326	4,279	32,462	1,326	4,279	32,462
5	1,208	3,895	36,358	1,208	3,895	36,358
6	1,202	3,879	40,237	1,202	3,879	40,237
7	1,119	3,609	43,846	1,119	3,609	43,846
8	1,087	3,505	47,351	1,087	3,505	47,351
9	1,054	3,400	50,751	1,054	3,400	50,751
10	1,020	3,289	54,040	1,020	3,289	54,040
11	,983	3,170	57,210			
12	,961	3,100	60,310			
13	,921	2,969	63,279			
14	,860	2,774	66,054			
15	,850	2,743	68,796			
16	,809	2,608	71,405			
17	,789	2,546	73,951			
18	,769	2,481	76,431			
19	,707	2,282	78,713			
20	,679	2,191	80,904			
21	,655	2,112	83,016			
22	,648	2,091	85,107			
23	,615	1,985	87,092			
24	,598	1,928	89,020			
25	,554	1,787	90,807			
26	,545	1,758	92,565			
27	,536	1,729	94,294			
28	,504	1,627	95,921			
29	,453	1,460	97,381			
30	,436	1,405	98,786			
31	,376	1,214	100,000			

Extraction Method: Principal Component Analysis.

Component Matrix [a]

Vari	1	2	3	4	5	6	7	8	9	10
SC1	,229	-,015	,252	-,054	,283	,497	-,040	,138	,184	,091
SC2	,397	,131	,213	,005	,136	,064	-,301	,088	-,018	-,031
SC3	,363	,225	,409	,191	-,031	-,076	,074	,202	-,406	,218
SC4	,281	,265	,337	,038	,041	-,370	-,046	,164	,266	,239
SC5	,360	,453	-,376	-,002	,190	,121	-,063	,128	,068	,083
SC6	,451	,299	-,082	-,064	,026	,137	-,344	-,129	-,130	,146
SC7	,471	,471	-,015	,010	,046	-,277	,073	-,025	,139	-,211
SC8	,446	,565	-,017	-,088	-,045	-,184	,178	-,068	-,068	-,072
SC9	,456	,267	-,123	,070	-,232	,328	,131	-,125	,240	-,184
SC10	,501	,403	-,130	,102	-,083	,259	,096	-,071	-,071	,078
SR1	-,019	-,007	,425	,042	,082	,020	,270	-,075	,522	,038
SR2	,401	-,162	,228	,087	-,033	,179	,191	-,291	-,152	,345
SR3	,527	-,089	,121	-,153	-,197	,316	,181	-,096	,007	,131
SR4	,333	-,274	-,086	-,547	,072	-,051	,298	,047	-,067	-,004
SR5	,492	-,089	-,012	-,199	-,025	-,125	,428	,070	,011	-,340
SR6	,284	-,162	-,140	,013	,332	,320	-,211	,275	,048	-,148
SR7	,333	-,082	-,092	,320	-,185	,199	,214	,455	-,049	-,118
SR8	,415	-,257	-,263	,274	-,332	-,023	,051	,075	,031	,011
SR9	,310	-,154	-,311	,270	-,233	-,260	-,220	-,044	,135	,069
SR10	,515	-,231	,215	,237	-,193	,023	-,167	,019	-,059	-,111
SR11	,525	-,176	,188	,318	,106	-,111	-,027	,095	,094	-,107
CR1	,335	-,233	-,176	,283	,340	-,167	,288	,228	,026	,251
CR2	,352	-,034	-,060	-,364	-,324	-,168	-,123	,342	,092	,431
CR3	,534	-,211	,225	-,355	-,123	-,011	-,230	-,025	,046	-,041
CR4	,554	-,159	,166	,070	,058	-,131	-,196	-,129	-,014	-,251
CR5	,474	-,029	,135	-,247	-,125	-,051	-,157	,189	-,112	-,344
CR6	,495	-,168	,024	,181	,219	-,116	-,064	-,416	,019	-,026
CR7	,461	-,139	-,073	-,047	,072	-,015	,083	-,118	-,429	,044
CR8	,455	-,214	-,270	-,152	-,024	,013	-,138	-,053	,349	,155
CR9	,348	-,057	-,164	-,123	,527	-,155	,001	,108	-,084	,011
CR10	,576	-,077	-,205	-,023	,093	-,052	,068	-,321	,051	,111

Extraction Method: Principal Component Analysis.
a. 10 components extracted.

Component Transformation Matrix

Component	1	2	3	4	5	6	7	8	9	10
1	,571	,482	,381	,282	,273	,231	,229	,159	,103	-,031
2	-,221	,830	-,193	-,226	-,213	-,175	,003	-,032	-,306	,062
3	,385	-,229	,174	-,016	-,227	-,250	,039	,112	-,599	,531
4	,161	,008	,056	-,657	,534	,223	-,398	-,142	-,110	,117
5	,010	,010	-,099	-,002	-,442	,701	-,333	,433	-,046	,052
6	-,253	,013	,326	-,067	,184	-,399	-,236	,750	,060	-,102
7	-,484	,080	,269	,516	,281	,189	-,279	-,226	-,192	,374
8	-,118	-,092	-,526	,103	,469	,187	,418	,343	-,378	-,019
9	-,026	,077	-,238	-,094	,043	-,047	,145	,149	,581	,737
10	-,372	-,075	,512	-,384	-,114	,284	,590	-,007	-,041	,054

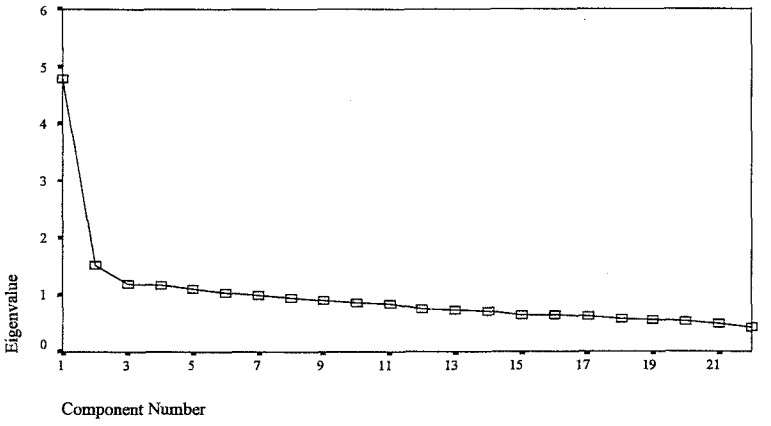
Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

APPENDIX 4.11
PCA: September 2000 Test – Purged Set

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.855
Bartlett's Test of Sphericity	Approx. Chi-Square	1242.981
	df	231
	Sig.	.000

Scree Plot



Communalities

	Initial	Extraction
SC6	1.000	.475
SC7	1.000	.666
SC8	1.000	.647
SC9	1.000	.486
SC10	1.000	.605
SR2	1.000	.438
SR3	1.000	.523
SR5	1.000	.406
SR7	1.000	.545
SR8	1.000	.530
SR9	1.000	.669
SR10	1.000	.574
SR11	1.000	.477
CR1	1.000	.573
CR2	1.000	.688
CR3	1.000	.631
CR4	1.000	.504
CR6	1.000	.544
CR7	1.000	.425
CR8	1.000	.387
CR9	1.000	.567
CR10	1.000	.453

Extraction Method: Principal Component Analysis.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.794	21.790	21.790	4.794	21.790	21.790
2	1.533	6.967	28.757	1.533	6.967	28.757
3	1.182	5.373	34.130	1.182	5.373	34.130
4	1.165	5.296	39.426	1.165	5.296	39.426
5	1.105	5.025	44.451	1.105	5.025	44.451
6	1.030	4.683	49.134	1.030	4.683	49.134
7	1.002	4.554	53.688	1.002	4.554	53.688
8	.946	4.301	57.989			
9	.904	4.110	62.099			
10	.862	3.920	66.019			
11	.835	3.795	69.814			
12	.749	3.403	73.217			
13	.726	3.298	76.516			
14	.702	3.193	79.709			
15	.649	2.952	82.661			
16	.645	2.931	85.592			
17	.625	2.842	88.434			
18	.565	2.570	91.004			
19	.547	2.486	93.490			
20	.523	2.375	95.865			
21	.491	2.230	98.095			
22	.419	1.905	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix [a]

	Component						
	1	2	3	4	5	6	7
SC6	.436	-.323	-.054	-.141	.057	-.166	-.356
SC7	.461	-.489	.274	.053	.077	-.257	.254
SC8	.433	-.626	.158	.029	.033	-.097	.179
SC9	.473	-.361	-.182	.264	-.067	.083	-.134
SC10	.494	-.429	-.069	.182	-.069	.267	-.250
SR2	.421	.119	-.136	-.119	-.393	.242	-.023
SR3	.532	-.039	-.397	-.076	-.192	.194	.008
SR5	.485	-.019	.091	.084	-.038	.064	.387
SR7	.345	.120	-.078	.558	-.039	.175	.249
SR8	.448	.287	-.142	.419	.127	.023	-.184
SR9	.345	.196	.030	.250	.410	-.452	-.277
SR10	.532	.247	-.160	.078	-.268	-.305	.184
SR11	.534	.239	.144	.099	-.234	-.145	.164
CR1	.354	.310	.450	.205	.199	.257	.035
CR2	.339	.004	-.403	-.147	.597	.089	.157
CR3	.538	.109	-.283	-.432	.050	-.108	.222
CR4	.558	.145	.099	-.161	-.173	-.325	-.016
CR6	.519	.185	.266	-.171	-.176	-.125	-.305
CR7	.469	.120	.091	-.160	.009	.286	-.273
CR8	.470	.183	-.160	-.132	.277	.055	.100
CR9	.342	.041	.456	-.285	.198	.309	.157
CR10	.604	.001	.124	-.129	.076	.113	-.195

Extraction Method: Principal Component Analysis.
a. 7 components extracted.

Component Transformation Matrix

Component	1	2	3	4	5	6	7
1	.572	.492	.305	.362	.351	.238	.166
2	.380	-.435	-.710	.110	.249	.237	.179
3	.011	-.293	.335	-.554	.691	-.094	.097
4	-.199	.114	.049	-.369	-.210	.799	.357
5	-.546	-.190	.119	.561	.243	-.029	.526
6	-.429	.335	-.301	.084	.477	.274	-.550
7	.088	-.567	.428	.304	-.085	.406	-.473

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

APPENDIX 4.12
PCA: September 2000 Test – Individual Subtests

Scanning

Communalities

	Initial	Extraction
SC1	1.000	.720
SC2	1.000	.410
SC3	1.000	.535
SC4	1.000	.458
SC5	1.000	.407
SC6	1.000	.352
SC7	1.000	.519
SC8	1.000	.586
SC9	1.000	.484
SC10	1.000	.489

Extraction Method: PCA

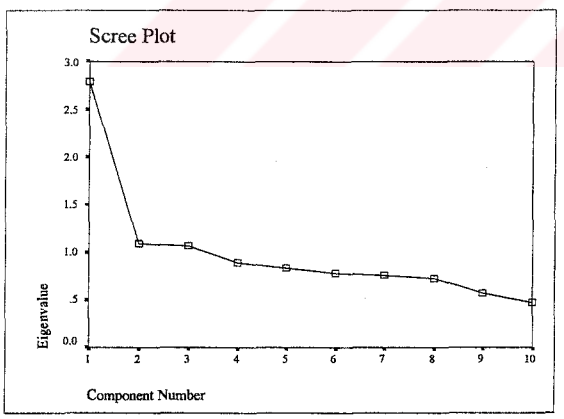
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.780
Bartlett's Test of Sphericity	Approx. Chi-Square	421.274
	df	45
	Sig.	.000

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.791	27.908	27.908	2.791	27.908	27.908
2	1.096	10.956	38.864	1.096	10.956	38.864
3	1.074	10.742	49.606	1.074	10.742	49.606
4	.894	8.941	58.547			
5	.837	8.365	66.912			
6	.781	7.808	74.720			
7	.759	7.585	82.305			
8	.724	7.244	89.550			
9	.571	5.714	95.264			
10	.474	4.736	100.000			

Extraction Method: Principal Component Analysis.



Component Matrix[a]

	Component		
	1	2	3
SC1	.238	.514	.631
SC2	.436	-.045	.468
SC3	.434	-.392	.440
SC4	.383	-.522	.198
SC5	.563	.218	-.205
SC6	.562	.187	.033
SC7	.648	-.235	-.212
SC8	.671	-.239	-.281
SC9	.533	.407	-.185
SC10	.650	.208	-.153

Extraction Method: PCA.
a. 3 components extracted.

Component Transformation Matrix

Component	1	2	3
1	.845	.513	.153
2	.343	-.738	.582
3	-.411	.439	.799

Extraction Method: PCA Rotation Method: Varimax with Kaiser Normalization.

Search reading

Communalities

	Initial	Extraction
SR1	1.000	.638
SR2	1.000	.299
SR3	1.000	.470
SR4	1.000	.701
SR5	1.000	.486
SR6	1.000	.123
SR7	1.000	.250
SR8	1.000	.446
SR9	1.000	.426
SR10	1.000	.490
SR11	1.000	.489

Extraction Method: PCA

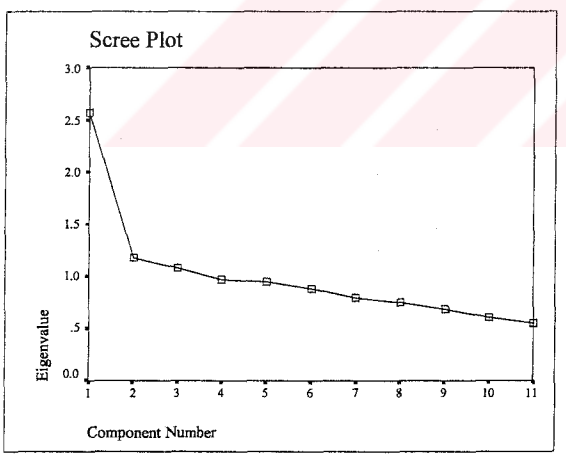
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.752
Bartlett's Test of Sphericity	Approx. Chi-Square	345.594
	df	55
	Sig.	.000

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.564	23.314	23.314	2.564	23.314	23.314
2	1.178	10.707	34.021	1.178	10.707	34.021
3	1.077	9.789	43.810	1.077	9.789	43.810
4	.969	8.812	52.622			
5	.945	8.588	61.210			
6	.886	8.052	69.261			
7	.794	7.219	76.481			
8	.746	6.781	83.262			
9	.685	6.228	89.490			
10	.604	5.488	94.978			
11	.552	5.022	100.000			

Extraction Method: PCA



Component Matrix[a]

	Component		
	1	2	3
SR1	-.018	.234	.764
SR2	.470	.168	.223
SR3	.592	.306	.164
SR4	.405	.609	-.407
SR5	.543	.358	-.251
SR6	.344	.061	-.022
SR7	.446	-.227	.016
SR8	.545	-.341	-.181
SR9	.362	-.502	-.207
SR10	.641	-.217	.177
SR11	.614	-.178	.283

Extraction Method: PCA

a. 3 components extracted.

Component Transformation Matrix

Component	1	2	3
1	.765	.615	.192
2	-.643	.710	.288
3	.040	-.344	.938

Extraction Method: PCA Rotation Method: Varimax with Kaiser Normalization.

Careful reading

Communalities

	Initial	Extraction
CR1	1.000	.412
CR2	1.000	.434
CR3	1.000	.569
CR4	1.000	.382
CR5	1.000	.443
CR6	1.000	.474
CR7	1.000	.298
CR8	1.000	.289
CR9	1.000	.269
CR10	1.000	.434

Extraction Method: PCA

KMO and Bartlett's Test

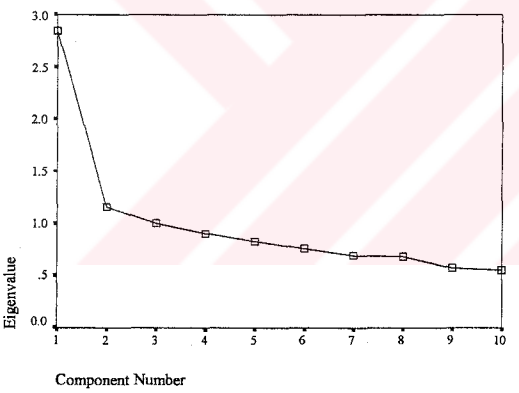
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.801
Bartlett's Test of Sphericity	Approx. Chi-Square	424.597
	df	45
	Sig.	.000

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.843	28.434	28.434	2.843	28.434	28.434
2	1.160	11.602	40.036	1.160	11.602	40.036
3	1.000	9.998	50.034			
4	.902	9.022	59.057			
5	.826	8.264	67.320			
6	.763	7.634	74.954			
7	.699	6.993	81.947			
8	.680	6.801	88.749			
9	.575	5.754	94.503			
10	.550	5.497	100.000			

Extraction Method: PCA

Scree Plot



Component Matrix[a]

	Component	
	1	2
CR1	.394	.507
CR2	.404	-.520
CR3	.629	-.417
CR4	.618	.016
CR5	.524	-.411
CR6	.573	.381
CR7	.534	.115
CR8	.526	-.110
CR9	.446	.264
CR10	.620	.223

Extraction Method: PCA

a. 2 components extracted.

Component Transformation Matrix

Component	1	2
1	.745	.667
2	.667	-.745

Extraction Method: PCA Rotation Method: Varimax with Kaiser Normalization.

APPENDIX 4.13
PCA: September 2000 Test – Half-Set I

Communalities

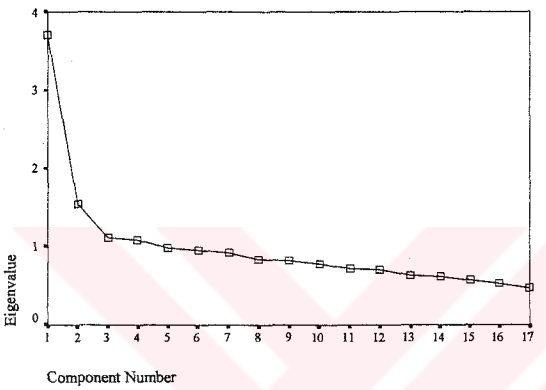
	Initial	Extraction
SC5	1,000	,449
SC6	1,000	,343
SC7	1,000	,475
SC8	1,000	,552
SC9	1,000	,413
SC10	1,000	,465
SR1	1,000	,699
SR7	1,000	,421
SR8	1,000	,491
SR9	1,000	,284
SR10	1,000	,447
SR11	1,000	,492
CR6	1,000	,394
CR7	1,000	,307
CR8	1,000	,315
CR9	1,000	,427
CR10	1,000	,481

Extraction Method: PCA

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,824
Bartlett's Test of Sphericity	Approx. Chi-Square	800,129
	df	136
	Sig.	,000

Scree Plot



Component Matrix[a]

	Component			
	1,000	2,000	3,000	4,000
SC5	,462	-,424	-,019	-,236
SC6	,500	-,258	,025	-,160
SC7	,528	-,405	-,009	,180
SC8	,503	-,528	-,016	,138
SC9	,517	-,197	-,304	,117
SC10	,575	-,320	-,173	,046
SR1	-,060	,039	,111	,826
SR7	,343	,184	-,515	,060
SR8	,449	,417	-,306	-,148
SR9	,367	,299	-,129	-,208
SR10	,487	,392	-,179	,157
SR11	,512	,343	,020	,335
CR6	,488	,237	,316	,013
CR7	,452	,161	,274	-,036
CR8	,454	,218	,117	-,218
CR9	,358	,018	,544	-,051
CR10	,613	,119	,303	-,018

Extraction Method: Principal Component Analysis.
a. 4 components extracted.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,704	21,785	21,785	3,704	21,785	21,785
2	1,544	9,085	30,870	1,544	9,085	30,870
3	1,120	6,591	37,461	1,120	6,591	37,461
4	1,088	6,400	43,862	1,088	6,400	43,862
5	,985	5,794	49,656			
6	,957	5,631	55,287			
7	,922	5,426	60,713			
8	,831	4,890	65,603			
9	,824	4,845	70,449			
10	,784	4,613	75,062			
11	,717	4,218	79,280			
12	,699	4,111	83,391			
13	,635	3,737	87,128			
14	,623	3,666	90,793			
15	,568	3,339	94,132			
16	,534	3,141	97,273			
17	,464	2,727	100,000			

Extraction Method: Principal Component Analysis.

Component Transformation Matrix

Component	1	2	3	4
1	,644	,584	,495	-,012
2	-,738	,314	,592	,081
3	-,184	,744	-,636	,087
4	,084	-,084	,013	,993

Extraction Method: PCA Rotation Method: Varimax with Kaiser Normalization.

APPENDIX 4.14
PCA: September 2000 Test – Half-Set II

Communalities

	Initial	Extraction
SC5	1.000	.533
SC6	1.000	.417
SC7	1.000	.576
SC8	1.000	.606
SC9	1.000	.452
SC10	1.000	.539
SR1	1.000	.539
SR2	1.000	.383
SR3	1.000	.498
SR4	1.000	.510
SR5	1.000	.564
SR6	1.000	.579
CR1	1.000	.520
CR2	1.000	.366
CR3	1.000	.592
CR4	1.000	.337
CR5	1.000	.405

Extraction Method: PCA

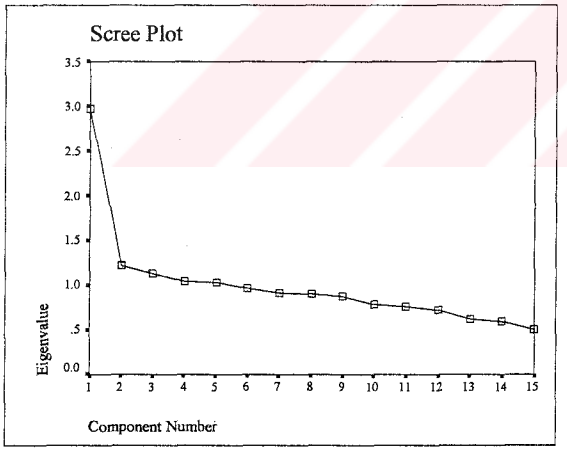
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampli Adequacy.		,771
Bartlett's Test of Sphericity	Approx. Chi-Square	485,121
	df	105
	Sig.	,000

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,969	19,790	19,790	2,969	19,790	19,790
2	1,225	8,170	27,960	1,225	8,170	27,960
3	1,126	7,507	35,467	1,126	7,507	35,467
4	1,039	6,927	42,394	1,039	6,927	42,394
5	1,028	6,854	49,248	1,028	6,854	49,248
6	,968	6,453	55,702			
7	,909	6,059	61,761			
8	,899	5,993	67,753			
9	,874	5,825	73,578			
10	,784	5,228	78,806			
11	,754	5,025	83,831			
12	,715	4,770	88,601			
13	,617	4,116	92,717			
14	,588	3,923	96,639			
15	,504	3,361	100,000			

Extraction Method: Principal Component Analysis.



Component Matrix[a]

	Component				
	1	2	3	4	5
SC1	.315	.105	.520	-.019	.242
SC2	.402	.316	.031	-.328	.253
SC3	.412	.538	.028	-.085	-.319
SC4	.314	.519	-.276	.186	-.201
SR1	.029	.357	.163	.488	.598
SR2	.434	.107	.318	.172	-.206
SR3	.575	-.074	.110	.125	.162
SR4	.460	-.525	-.047	.310	.039
SR5	.536	-.278	-.006	.209	-.052
SR6	.313	-.205	.367	-.491	.010
CR1	.302	-.054	.386	.270	-.504
CR2	.438	-.030	-.483	.145	-.086
CR3	.634	-.098	-.218	-.016	.244
CR4	.576	.031	-.032	-.211	-.033
CR5	.553	-.054	-.265	-.291	.102

Extraction Method: Principal Component Analysis.
a. 5 components extracted.

Component Transformation Matrix

Component	1	2	3	4	5
1	.646	.595	.325	.301	.179
2	-.760	.492	.346	.224	.104
3	.002	.048	-.217	.595	-.772
4	-.065	-.082	-.569	.564	.589
5	.026	-.629	.636	.432	.116

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

APPENDIX 4.15
September 2000 Test
Subtest-Factor Correlations (subsections)

September 2000 test – whole set: Subtest – Factor correlations (subsections)

Factors	Eigen value	% of varience (cum: 54.04%)	skimming (SR1)	scanning I (SC 1- 5)	scanning II (SC 6-10)	search r. 1 (SR 2-6)	search r. 2 (SR 7-11)	careful r. I (CR 1-5)	careful r. II (CR 6-10)
F1	5.597	18.054	.007	**.412***	.186*	.238*	**.535***	**.568***	**.418***
F2	1.761	5.682	-.044	**.311***	**.932***	.151*	.097	.117*	.257*
F3	1.379	4.448	.077	.185*	.168*	**.537***	.163*	.158*	**.426***
F4	1.326	4.279	.029	-.096	.066	**.549***	-.017	**.340***	.238*
F5	1.208	3.895	-.052	.000	.110*	.211*	**.720***	.154*	-.002
F6	1.202	3.879	-.027	.158*	.004	.113*	.172*	.171*	**.463***
F7	1.119	3.609	-.045	**.312***	.078	.090	.125*	**.476***	.071
F8	1.087	3.505	.080	**.347***	.097	.211*	-.031	.107*	.060
F9	1.054	3.400	.017	-.418*	.044	.051	.138*	.073	**.355***
F10	1.020	3.289	**.722***	.237*	-.046	.045	-.028	-.036	-.155*

September 2000 test – purged set: Subtest – Factor correlations (subsections)

Factors	Eigen value	% of varience (cum: 53.7%)	scanning II (SC 5-10)	search r. (-SR1,4,6)	Careful r. (-CR5)	search r. I (SR 2,3,5)	search r. II (SR 7-11)	careful r. I (CR 1-4)	careful r. II (CR 6-10)
F1	4.794	21.790	.106*	**.628***	**.478***	**.498***	**.552***	**.434***	**.403***
F2	1.533	6.967	**.661***	**.315***	.276*	**.418***	.169*	.100	**.356***
F3	1.182	5.373	**.679***	.079	.113*	.082	.057	.145*	.060
F4	1.165	5.296	.090	.249*	**.518***	.290*	.159*	**.621***	**.314***
F5	1.105	5.025	.086	.133*	**.592***	.149*	.089	.316*	**.681***
F6	1.030	4.683	.048	**.560***	.021	.297*	**.586***	.082	-.034
F7	1.002	4.554	.124*	.185*	.158*	-.301*	**.448***	.158*	.121*

Pearson Correlation: * *correlation is significant at 0.01level (2-tailed) / * correlation is significant at 0.05 level (2-tailed).

APPENDIX 4.16
January 2000 – Pilot Version
Normality Tests and Graphs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
total reading score	152	100.0%	0	.0%	152	100.0%

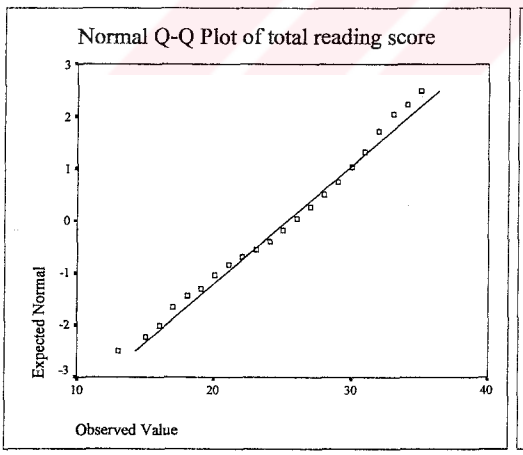
Descriptives

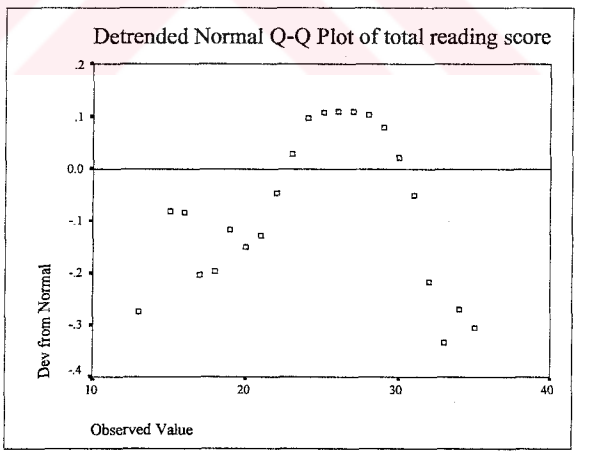
			Statistic	Std. Error
total reading scor	Mean		25.2895	.3617
	95% Confidence Interval for Mean	Lower Bound	24.5748	
		Upper Bound	26.0042	
	5% Trimmed Mean		25.3918	
	Median		26.0000	
	Variance		19.889	
	Std. Deviation		4.4597	
	Minimum		13.00	
	Maximum		35.00	
	Range		22.00	
	Interquartile Range		7.0000	
	Skewness		-.372	.197
	Kurtosis		-.449	.391

Tests of Normality

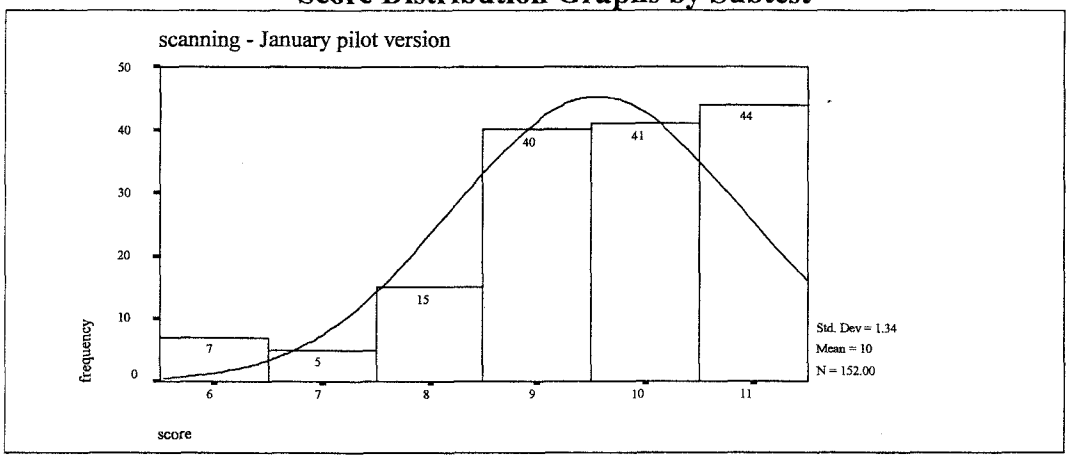
	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
total reading score	.093	152	.003

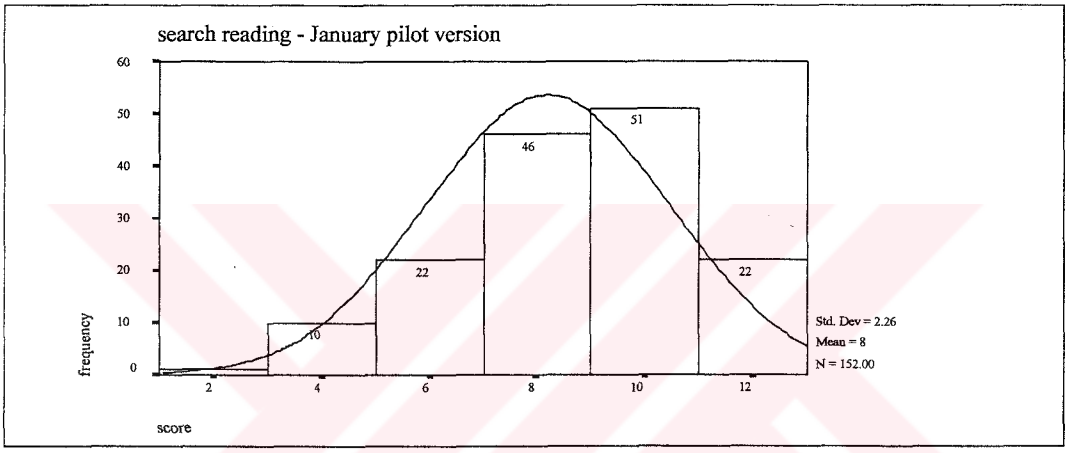
a. Lilliefors Significance Correction

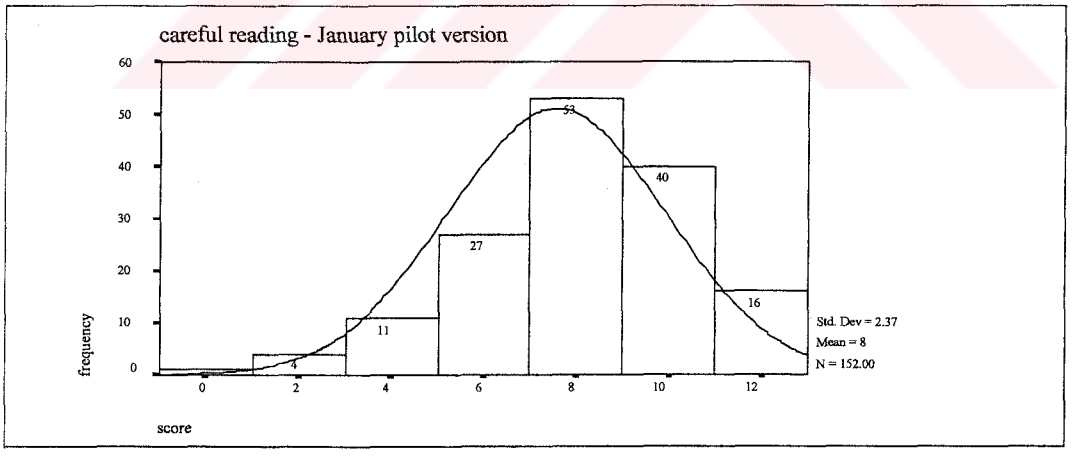




APPENDIX 4.17
January 2000 – Pilot Version
Score Distribution Graphs by Subtest







APPENDIX 4.18
January 2000 – Pilot Version
Normality Tests and Graphs by Subtests

Scanning

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
total scanning	152	100.0%	0	.0%	152	100.0%

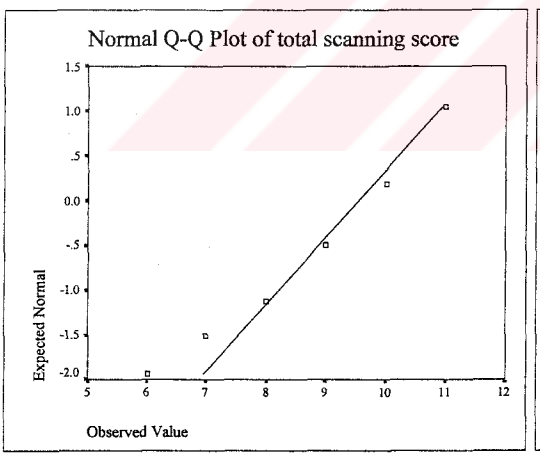
Descriptives

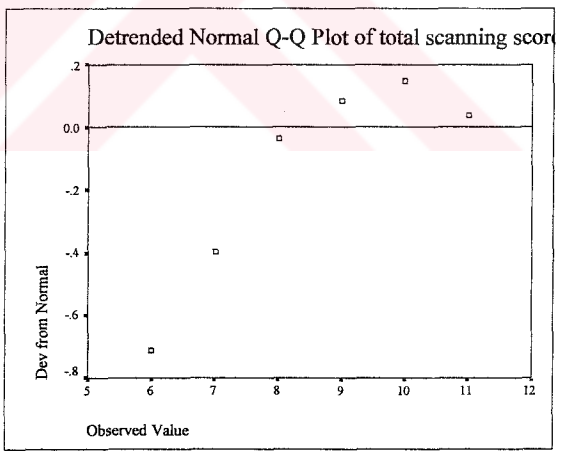
			Statistic	Std. Error
total scanning	Mean		9.5461	.1084
	95% Confidence Interval for Mean	Lower Bound	9.3319	
		Upper Bound	9.7602	
	5% Trimmed Mean		9.6579	
	Median		10.0000	
	Variance		1.786	
	Std. Deviation		1.3364	
	Minimum		6.00	
	Maximum		11.00	
	Range		5.00	
	Interquartile Range		2.0000	
	Skewness		-.893	.197
	Kurtosis		.406	.391

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
total scanning	.192	152	.000

a. Lilliefors Significance Correction





Search Reading

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
total search reading	152	100.0%	0	.0%	152	100.0%

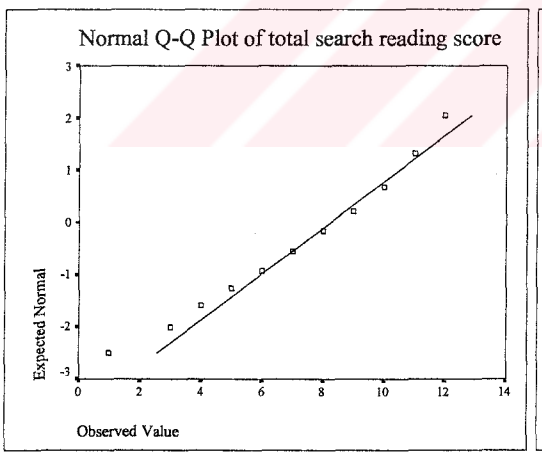
Descriptives

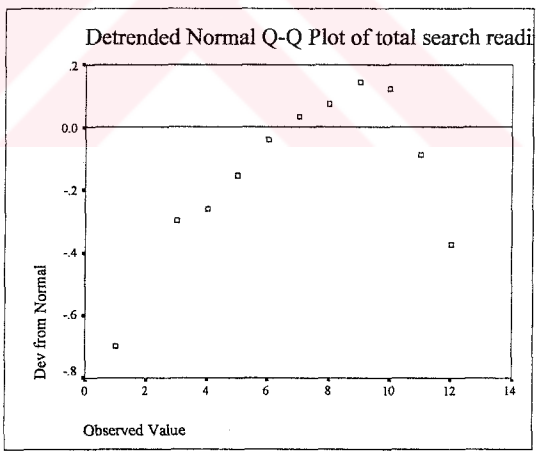
			Statistic	Std. Error
total search reading	Mean		8.1842	.1833
	95% Confidence Interval for Mean	Lower Bound	7.8221	
		Upper Bound	8.5463	
	5% Trimmed Mean		8.2749	
	Median		8.0000	
	Variance		5.105	
	Std. Deviation		2.2594	
	Minimum		1.00	
	Maximum		12.00	
	Range		11.00	
	Interquartile Range		3.0000	
	Skewness		-.567	.197
	Kurtosis		-.118	.391

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
total search reading	.138	152	.000

a. Lilliefors Significance Correction





Careful Reading

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
total careful reading	152	100.0%	0	.0%	152	100.0%

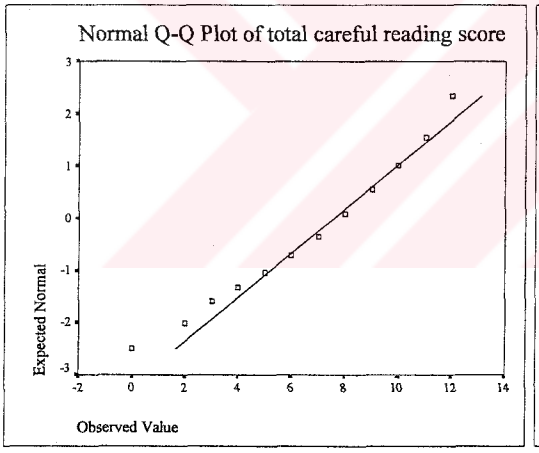
Descriptives

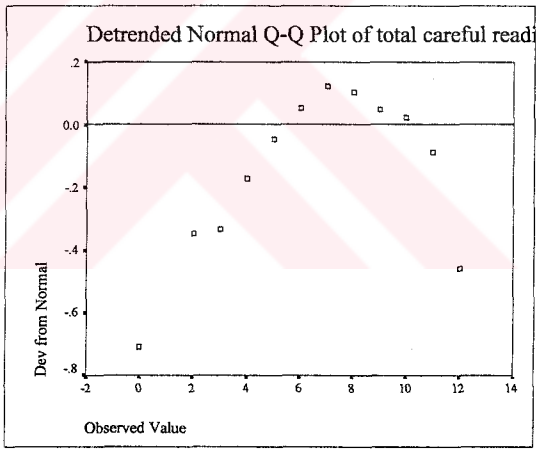
			Statistic	Std. Error
total careful reading	Mean		7.5592	.1922
	95% Confidence Interval for Mean	Lower Bound	7.1796	
		Upper Bound	7.9389	
	5% Trimmed Mean		7.6579	
	Median		8.0000	
	Variance		5.612	
	Std. Deviation		2.3690	
	Minimum		.00	
	Maximum		12.00	
	Range		12.00	
	Interquartile Range		3.0000	
	Skewness		-.579	.197
	Kurtosis		.113	.391

Tests of Normality

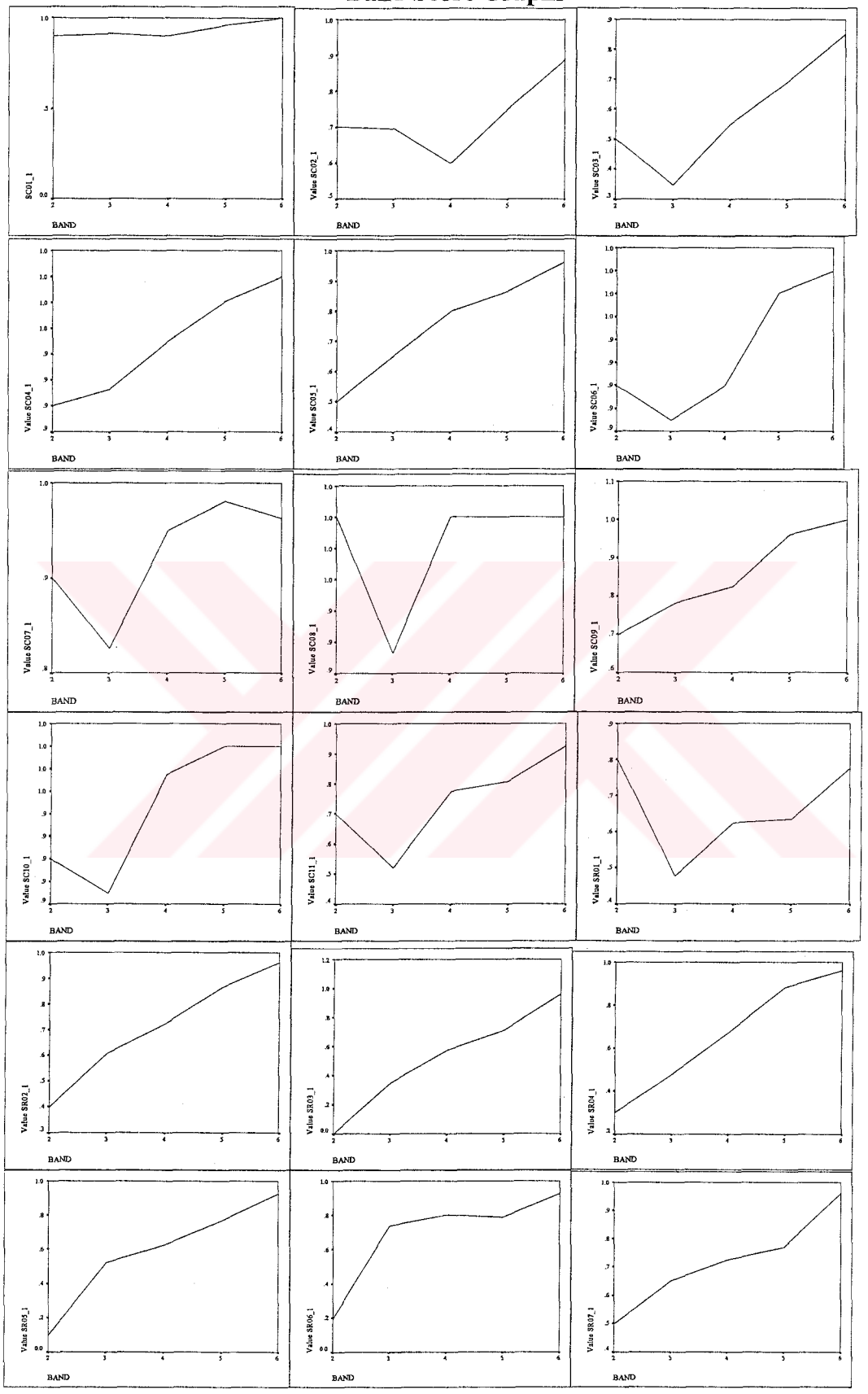
	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
total careful reading	.140	152	.000

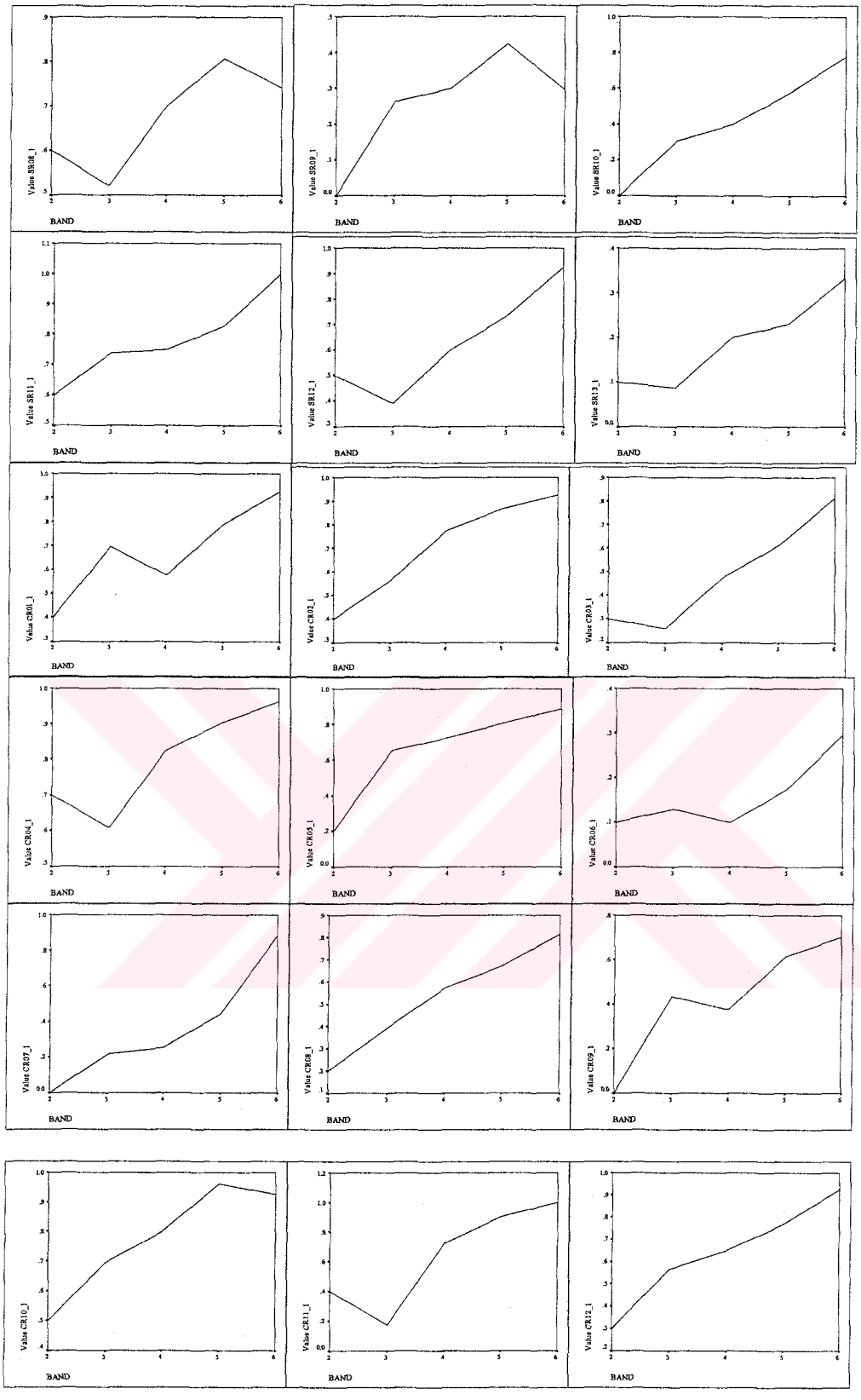
a. Lilliefors Significance Correction





APPENDIX 4.19
January 2001 – Pilot Version
Band Score Graphs




Value SR08_1
Value SR09_1
Value SR10_1
Value SR11_1
Value SR12_1
Value SR13_1
Value CR01_1
Value CR02_1
Value CR03_1
Value CR04_1
Value CR05_1
Value CR06_1
Value CR07_1
Value CR08_1
Value CR09_1
Value CR10_1
Value CR11_1
Value CR12_1
BAND

APPENDIX 4.20
January 2001 Test
Normality Tests and Graphs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
total reading sc	650	100.0%	0	.0%	650	100.0%

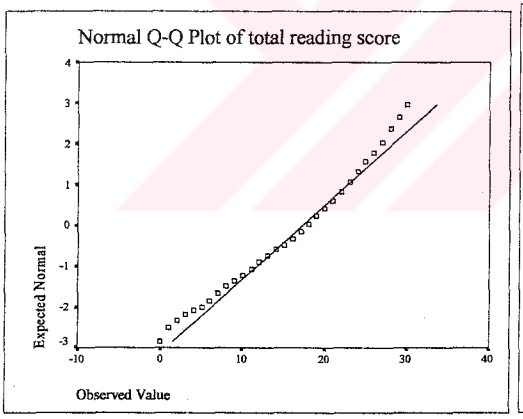
Descriptives

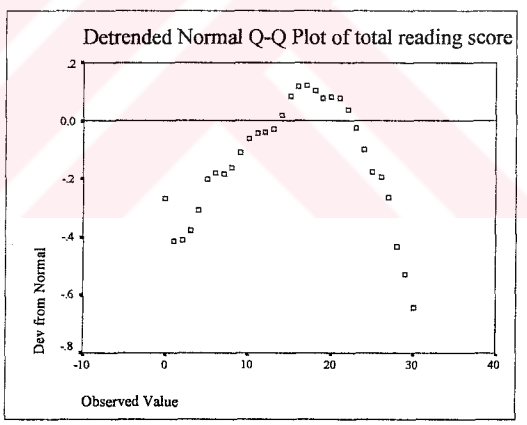
			Statistic	Std. Error
total reading scor	Mean		17.1662	.2172
	95% Confidence Interval for Mean	Lower Bound	16.7396	
		Upper Bound	17.5927	
	5% Trimmed Mean		17.3333	
	Median		18.0000	
	Variance		30.672	
	Std. Deviation		5.5382	
	Minimum		.00	
	Maximum		30.00	
	Range		30.00	
	Interquartile Range		8.0000	
	Skewness		-.474	.096
	Kurtosis		-.044	.191

Tests of Normality

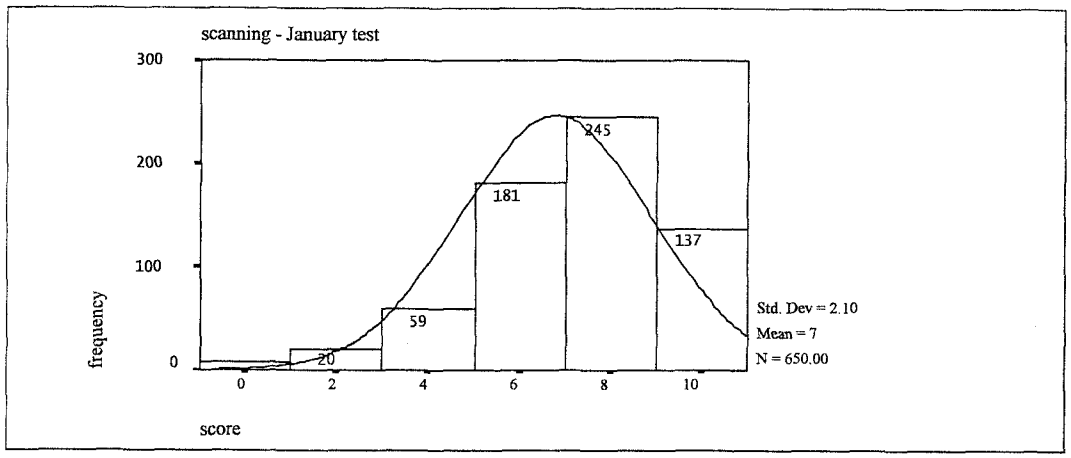
	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
total reading score	.086	650	.000

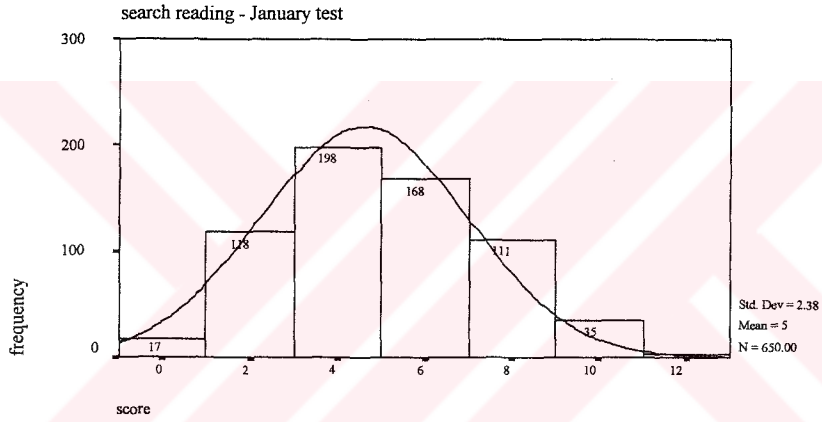
a. Lilliefors Significance Correction

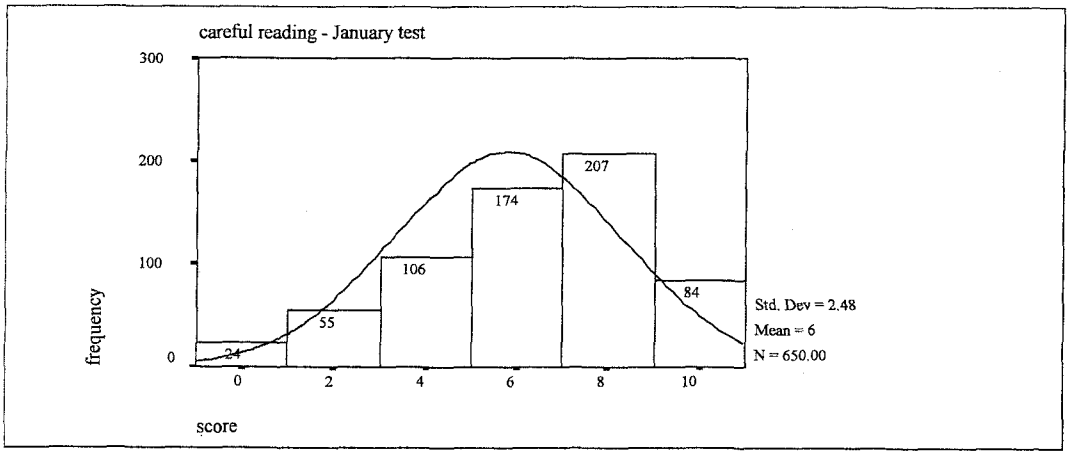




APPENDIX 4.21
January 2001 Test
Score Distribution Graphs by Subtest







APPENDIX 4.22
January 2001 Test
Normality Tests and Graphs by Subtest

Scanning

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
scanning	650	100.0%	0	.0%	650	100.0%

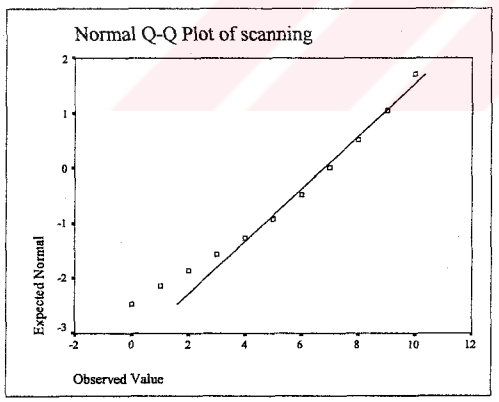
Descriptives

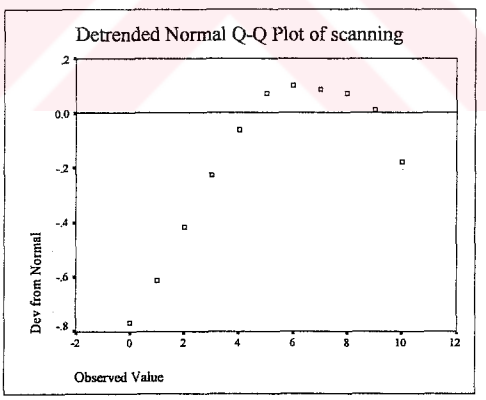
			Statistic	Std. Error
scanning	Mean		6.7846	8.242E-02
	95% Confidence Interval for Mean	Lower Bound	6.6228	
		Upper Bound	6.9465	
	5% Trimmed Mean		6.8983	
	Median		7.0000	
	Variance		4.416	
	Std. Deviation		2.1014	
	Minimum		.00	
	Maximum		10.00	
	Range		10.00	
	Interquartile Range		2.0000	
	Skewness		-.717	.096
	Kurtosis		.559	.191

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
scanning	.133	650	.000

a. Lilliefors Significance Correction





Search Reading

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
search reading	650	100.0%	0	.0%	650	100.0%

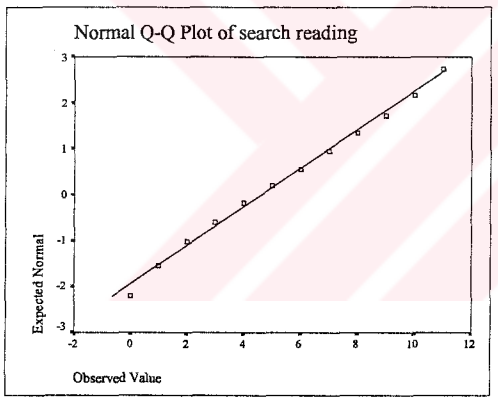
Descriptives

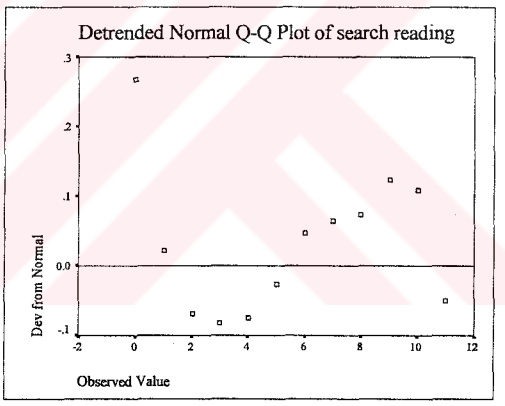
			Statistic	Std. Error
search reading	Mean		4.6015	9.330E-02
	95% Confidence Interval for Mean	Lower Bound	4.4183	
		Upper Bound	4.7847	
	5% Trimmed Mean		4.5538	
	Median		4.0000	
	Variance		5.658	
	Std. Deviation		2.3786	
	Minimum		.00	
	Maximum		11.00	
	Range		11.00	
	Interquartile Range		3.0000	
	Skewness		.223	.096
	Kurtosis		-.523	.191

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
search reading	.112	650	.000

a. Lilliefors Significance Correction





Careful Reading

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
careful readi	650	100.0%	0	.0%	650	100.0%

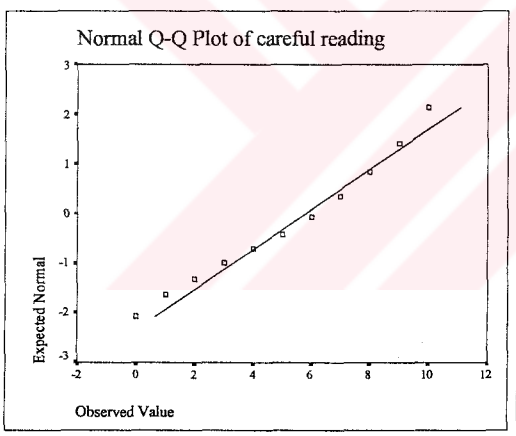
Descriptives

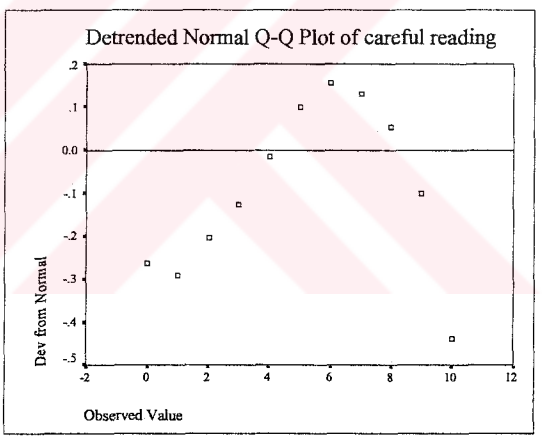
			Statistic	Std. Error
careful reading	Mean		5.7800	9.715E-02
	95% Confidence Interval for Mean	Lower Bound	5.5892	
		Upper Bound	5.9708	
	5% Trimmed Mean		5.8735	
	Median		6.0000	
	Variance		6.135	
	Std. Deviation		2.4769	
	Minimum		.00	
	Maximum		10.00	
	Range		10.00	
	Interquartile Range		4.0000	
	Skewness		-.519	.096
	Kurtosis		-.426	.191

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
careful reading	.142	650	.000

a. Lilliefors Significance Correction





APPENDIX 4.23
PCA: January 2001 Test – Whole Set

Communalities

	Initial	Extraction
SR1	1,000	,662
SR2	1,000	,332
SR3	1,000	,423
SR4	1,000	,468
SR5	1,000	,469
SR6	1,000	,424
SR7	1,000	,597
SR8	1,000	,562
SR9	1,000	,439
SR10	1,000	,641
SR11	1,000	,476
CR1	1,000	,416
CR2	1,000	,433
CR3	1,000	,406
CR4	1,000	,555
CR5	1,000	,430
CR6	1,000	,420
CR7	1,000	,429
CR8	1,000	,372
CR9	1,000	,435
CR10	1,000	,375
SC1	1,000	,564
SC2	1,000	,504
SC3	1,000	,501
SC4	1,000	,477
SC5	1,000	,495
SC6	1,000	,629
SC7	1,000	,681
SC8	1,000	,391
SC9	1,000	,651
SC10	1,000	,460

Extraction Method: PCA

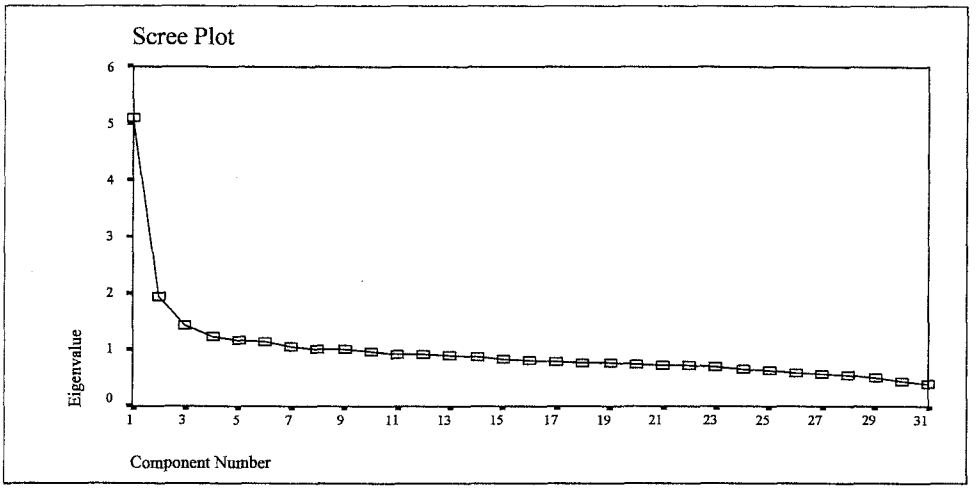
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of S Adequacy.		,861
Bartlett's Test Sphericity	Approx. Chi-Squ	72,226
	df	465
	Sig.	,000

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5,094	16,433	16,433	5,094	16,433	16,433
2	1,941	6,260	22,693	1,941	6,260	22,693
3	1,438	4,639	27,331	1,438	4,639	27,331
4	1,237	3,991	31,323	1,237	3,991	31,323
5	1,161	3,747	35,069	1,161	3,747	35,069
6	1,156	3,728	38,797	1,156	3,728	38,797
7	1,057	3,409	42,206	1,057	3,409	42,206
8	1,024	3,304	45,510	1,024	3,304	45,510
9	1,009	3,254	48,763	1,009	3,254	48,763
10	,966	3,115	51,878			
11	,929	2,998	54,876			
12	,926	2,988	57,864			
13	,894	2,885	60,749			
14	,874	2,819	63,568			
15	,846	2,729	66,296			
16	,827	2,668	68,964			
17	,794	2,561	71,525			
18	,780	2,517	74,043			
19	,773	2,494	76,537			
20	,748	2,413	78,950			
21	,737	2,377	81,327			
22	,717	2,314	83,641			
23	,696	2,246	85,886			
24	,665	2,147	88,033			
25	,632	2,040	90,073			
26	,592	1,909	91,982			
27	,584	1,884	93,866			
28	,557	1,797	95,663			
29	,509	1,643	97,307			
30	,441	1,424	98,730			
31	,394	1,270	100,000			

Extraction Method: Principal Component Analysis.



Component Matrix

	Component								
	1	2	3	4	5	6	7	8	9
SR1	.220	.103	.221	.193	.063	.108	-.510	.172	.460
SR2	.454	-.208	-.197	-.016	-.097	-.135	-.096	-.003	.081
SR3	.447	-.223	-.357	-.035	-.040	.047	.141	-.058	.132
SR4	.516	-.217	-.211	-.021	.076	-.079	-.116	-.275	.091
SR5	.532	-.212	-.118	-.157	.025	-.138	-.022	-.096	.272
SR6	.411	-.193	-.145	-.072	.099	-.398	-.102	-.029	.110
SR7	.300	-.061	-.202	.057	.383	.234	-.122	.484	-.095
SR8	.447	-.121	-.270	-.201	.240	.077	.058	.407	-.045
SR9	.399	-.026	-.004	-.200	.354	.203	.112	-.233	.073
SR10	.136	-.025	-.220	-.149	-.057	.712	-.006	-.188	-.078
SR11	.226	-.218	-.310	.030	-.364	.117	.052	.164	.322
CR1	.420	-.254	.190	.165	-.081	-.151	-.153	-.167	-.176
CR2	.549	-.233	.217	-.009	-.007	-.115	-.100	-.076	-.040
CR3	.429	-.239	.140	-.062	-.363	.050	-.016	.057	.056
CR4	.446	-.085	.183	.268	-.319	.095	.210	.233	-.186
CR5	.488	-.202	.135	-.004	-.292	.216	.006	-.021	-.013
CR6	.188	-.003	.382	-.313	.227	-.045	.199	.084	.203
CR7	.453	-.115	.347	-.207	.183	.090	-.014	.067	-.028
CR8	.425	-.084	.174	-.125	-.089	.077	-.094	.174	-.292
CR9	.315	-.056	.329	-.061	.150	.003	.436	-.081	.040
CR10	.412	-.173	.372	.060	.020	-.009	-.033	-.014	-.177
SC1	.297	.178	.078	.389	.068	.138	.287	-.257	.339
SC2	.286	.056	-.153	.336	-.010	-.325	.298	.295	-.015
SC3	.268	.036	.046	.418	.277	.056	-.382	-.127	-.095
SC4	.417	.019	-.263	.297	.066	.087	.102	-.208	-.282
SC5	.409	.354	.083	.377	.098	.138	.095	.082	.098
SC6	.432	.589	.020	-.126	-.258	.009	-.072	.051	.070
SC7	.477	.643	.049	-.127	-.123	-.002	-.079	.024	.006
SC8	.455	.234	-.187	.038	.194	-.147	.149	-.016	-.101
SC9	.463	.589	-.065	-.256	-.072	-.115	-.016	-.049	-.005
SC10	.442	.203	-.193	-.158	-.059	-.132	-.149	-.157	-.305

Extraction Method: Principal Component Analysis.
a.9 components extracted.

Component Transformation Matrix

Component	1	2	3	4	5	6	7	8	9
1	,563	,541	,430	,301	,272	,186	,046	,045	,047
2	-,376	-,333	,812	,231	-,054	-,036	-,048	-,152	,068
3	-,436	,497	-,064	-,016	-,331	,545	-,138	-,216	,294
4	-,124	,140	-,285	,726	-,017	-,474	-,282	-,128	,190
5	,119	-,364	-,214	,155	,462	,400	,021	-,635	,076
6	-,344	,094	-,092	,214	,241	-,005	,843	,180	,132
7	-,148	-,099	-,100	,424	-,022	,429	-,123	,335	-,683
8	-,344	,086	,007	-,189	,707	,014	-,411	,369	,184
9	,254	-,413	-,073	,207	-,204	,313	-,044	,477	,589

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

APPENDIX 4.24
PCA: January 2001 Test – Purged Set

Communalities

	Initial	Extraction
SC4	1.000	.473
SC5	1.000	.441
SC6	1.000	.609
SC7	1.000	.682
SC8	1.000	.352
SC9	1.000	.641
SC10	1.000	.338
SR2	1.000	.373
SR3	1.000	.350
SR4	1.000	.426
SR5	1.000	.468
SR6	1.000	.335
SR7	1.000	.489
SR8	1.000	.437
SR9	1.000	.357
CR1	1.000	.321
CR2	1.000	.412
CR3	1.000	.332
CR4	1.000	.549
CR5	1.000	.397
CR7	1.000	.444
CR8	1.000	.236
CR9	1.000	.321
CR10	1.000	.359

Extraction Method: PCA

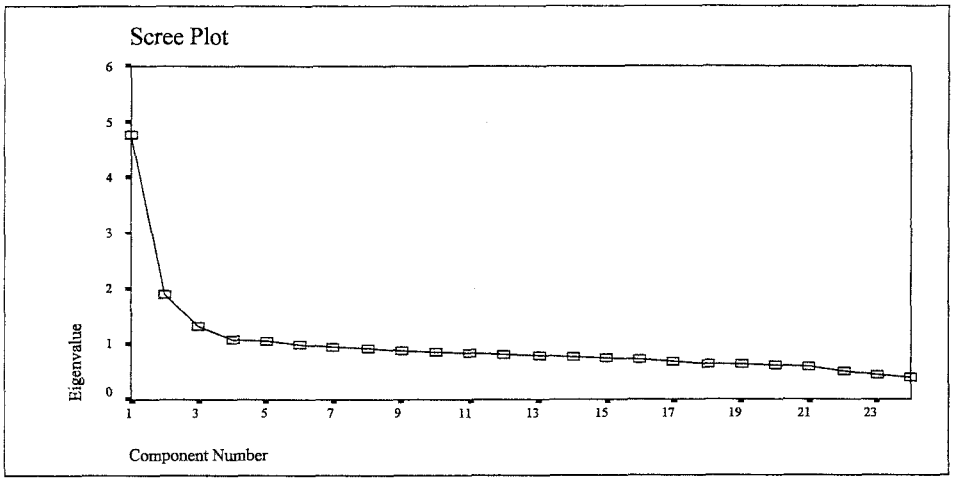
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.867
Bartlett's Test of Sphericity	Approx. Chi-Square	2496.119
	df	276
	Sig.	.000

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.773	19.889	19.889	4.773	19.889	19.889
2	1.895	7.895	27.784	1.895	7.895	27.784
3	1.325	5.520	33.304	1.325	5.520	33.304
4	1.090	4.542	37.846	1.090	4.542	37.846
5	1.059	4.414	42.260	1.059	4.414	42.260
6	.992	4.135	46.395			
7	.957	3.988	50.382			
8	.917	3.822	54.204			
9	.889	3.705	57.910			
10	.861	3.587	61.497			
11	.841	3.506	65.003			
12	.815	3.395	68.399			
13	.785	3.271	71.669			
14	.766	3.190	74.860			
15	.747	3.112	77.972			
16	.734	3.060	81.031			
17	.688	2.867	83.898			
18	.648	2.700	86.598			
19	.640	2.666	89.264			
20	.610	2.543	91.807			
21	.599	2.496	94.303			
22	.516	2.151	96.454			
23	.452	1.882	98.336			
24	.399	1.664	100.000			

Extraction Method: Principal Component Analysis.



Component Matrix [a]

	Component				
	1	2	3	4	5
SC4	.413	.021	-.236	-.106	.484
SC5	.392	.325	.133	.180	.364
SC6	.434	.606	.141	-.136	-.124
SC7	.478	.657	.114	-.050	-.080
SC8	.460	.247	-.210	.157	.104
SC9	.466	.615	-.040	-.071	-.197
SC10	.452	.233	-.183	-.215	-.027
SR2	.461	-.198	-.225	-.267	.010
SR3	.445	-.187	-.333	-.082	.002
SR4	.519	-.207	-.274	-.150	-.126
SR5	.539	-.200	-.173	-.036	-.326
SR6	.424	-.185	-.276	-.166	-.131
SR7	.299	-.058	-.274	.469	.317
SR8	.454	-.084	-.318	.347	.050
SR9	.401	-.025	-.095	.365	-.231
CR1	.420	-.265	.227	-.139	-.063
CR2	.557	-.240	.183	-.020	-.102
CR3	.440	-.226	.244	-.159	-.046
CR4	.441	-.100	.316	-.174	.462
CR5	.493	-.192	.207	-.205	.181
CR7	.460	-.131	.261	.314	-.220
CR8	.436	-.080	.194	.026	.046
CR9	.319	-.074	.297	.338	-.103
CR10	.418	-.199	.361	.123	-.006

Extraction Method: PCA

a. 5 components extracted.

Component Transformation Matrix

Component	1	2	3	4	5
1	.564	.447	.548	.312	.292
2	-.361	.869	-.332	.043	-.056
3	-.559	.058	.664	-.442	.218
4	-.419	-.125	.290	.768	-.367
5	-.254	-.164	-.254	.340	.854

Extraction Method: PCA Rotation Method: Varimax with Kaiser Normalization.

APPENDIX 4.25
PCA: January 2001 Test – Subtests

Scanning

Communalities

	Initial	Extraction
SC1	1,000	,491
SC2	1,000	,224
SC3	1,000	,268
SC4	1,000	,526
SC5	1,000	,490
SC6	1,000	,674
SC7	1,000	,705
SC8	1,000	,396
SC9	1,000	,649
SC10	1,000	,642

Extraction Method: PCA

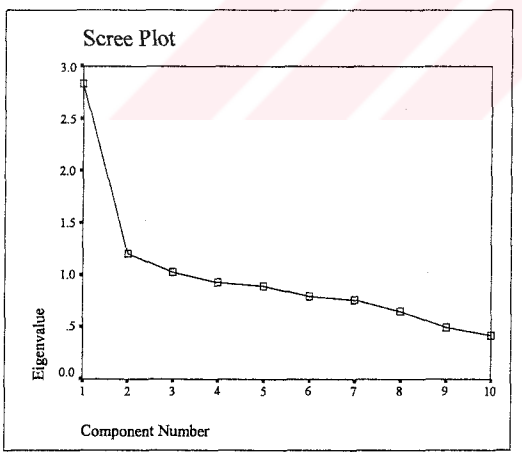
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.787
Bartlett's Test of Sphericity	Approx. Chi-Square	936.906
	df	45
	Sig.	.000

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Componen	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,834	28,342	28,342	2,834	28,342	28,342
2	1,207	12,069	40,411	1,207	12,069	40,411
3	1,023	10,234	50,646	1,023	10,234	50,646
4	,930	9,301	59,947			
5	,894	8,944	68,891			
6	,789	7,893	76,784			
7	,760	7,597	84,380			
8	,653	6,533	90,913			
9	,492	4,918	95,831			
10	,417	4,169	100,000			

Extraction Method: Principal Component Analysis.



Component Matrix

	Component		
	1	2	3
SC1	.363	.366	-.475
SC2	.304	.349	.098
SC3	.262	.418	-.154
SC4	.408	.501	.328
SC5	.548	.267	-.345
SC6	.679	-.407	-.217
SC7	.747	-.338	-.181
SC8	.543	.202	.246
SC9	.713	-.360	.104
SC10	.503	-.048	.622

Extraction Method: PCA

Component Transformation Matrix

Component	1	2	3
1	,771	,507	,386
2	-,619	,453	,641
3	-,150	,733	-,663

Extraction Method: PCA Rotation Method: Varimax with Kaiser Normalization.

Search Reading

Communalities

	Initial	Extraction
SR1	1,000	,397
SR2	1,000	,353
SR3	1,000	,490
SR4	1,000	,413
SR5	1,000	,427
SR6	1,000	,475
SR7	1,000	,455
SR8	1,000	,372
SR9	1,000	,298
SR10	1,000	,626
SR11	1,000	,384

Extraction Method: PCA

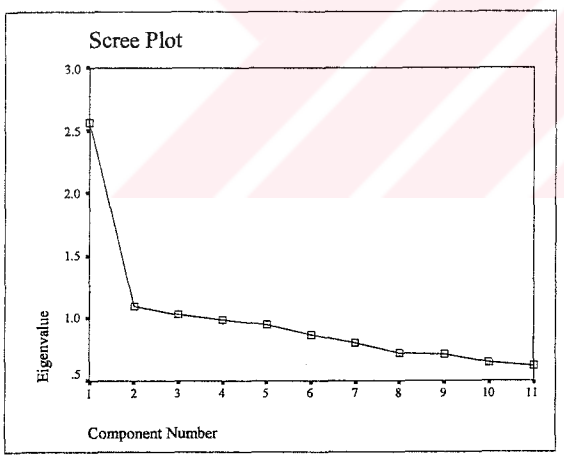
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.790
Bartlett's Test of Sphericity	Approx. Chi-Square	606.524
	df	55
	Sig.	.000

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,564	23,310	23,310	2,564	23,310	23,310
2	1,096	9,968	33,278	1,096	9,968	33,278
3	1,029	9,357	42,635	1,029	9,357	42,635
4	,987	8,970	51,605			
5	,954	8,671	60,276			
6	,866	7,876	68,152			
7	,802	7,294	75,446			
8	,721	6,558	82,004			
9	,713	6,479	88,483			
10	,651	5,919	94,402			
11	,616	5,598	100,000			

Extraction Method: Principal Component Analysis.



Component Matrix

	Component		
	1	2	3
SR1	.170	-.181	.579
SR2	.557	-.205	-.016
SR3	.585	.073	-.378
SR4	.626	-.127	-.071
SR5	.631	-.160	-.059
SR6	.520	-.450	.046
SR7	.382	.291	.474
SR8	.571	.175	.122
SR9	.438	.215	.244
SR10	.204	.764	-.001
SR11	.342	.162	-.491

Extraction Method: PCA

Component Transformation Matrix

Component	1	2	3
1	,848	,501	,174
2	-,524	,738	,426
3	-,085	,453	-,888

Extraction Method: PCA Rotation Method Varimax with Kaiser Normalization.

Careful Reading

Communalities

	Initial	Extraction
CR1	1,000	,311
CR2	1,000	,404
CR3	1,000	,394
CR4	1,000	,348
CR5	1,000	,375
CR6	1,000	,637
CR7	1,000	,399
CR8	1,000	,257
CR9	1,000	,329
CR10	1,000	,313

Extraction Method: PCA

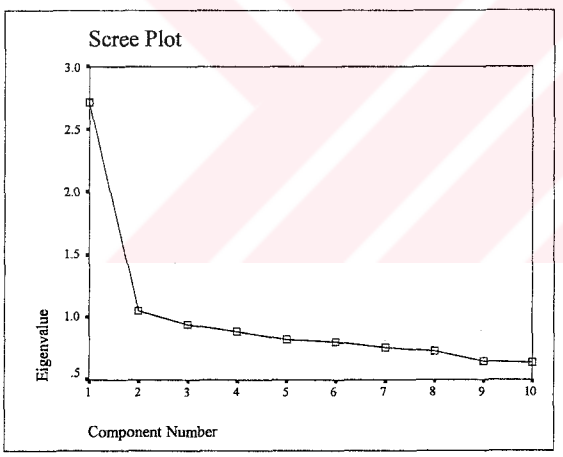
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.828
Bartlett's Test of Sphericity	Approx. Chi-Square	653.659
	df	45
	Sig.	.000

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,716	27,157	27,157	2,716	27,157	27,157
2	1,051	10,505	37,662	1,051	10,505	37,662
3	,942	9,416	47,078			
4	,886	8,859	55,937			
5	,825	8,248	64,184			
6	,802	8,016	72,200			
7	,761	7,607	79,807			
8	,732	7,322	87,129			
9	,649	6,491	93,619			
10	,638	6,381	100,000			

Extraction Method: Principal Component Analysis.



Component Matrix

	Component	
	1	2
CR1	.530	-.173
CR2	.636	.009
CR3	.539	-.322
CR4	.534	-.251
CR5	.576	-.206
CR6	.279	.748
CR7	.557	.298
CR8	.505	-.037
CR9	.408	.402
CR10	.559	-.013

Extraction Method: PCA

Component Transformation Matrix

Component	1	2
1	,880	,475
2	-,475	,880

Extraction Method: PCA Rotation Method
Varimax with Kaiser Normalization.

APPENDIX 4.26
PCA: January 2001 Test – Half-Set I

Communalities

	Initial	Extraction
SC6	1.000	.611
SC7	1.000	.684
SC8	1.000	.366
SC9	1.000	.651
SC10	1.000	.317
SR1	1.000	.154
SR2	1.000	.314
SR3	1.000	.366
SR4	1.000	.443
SR5	1.000	.437
SR6	1.000	.316
CR1	1.000	.312
CR2	1.000	.383
CR3	1.000	.361
CR4	1.000	.435
CR5	1.000	.432

Extraction Method: PCA

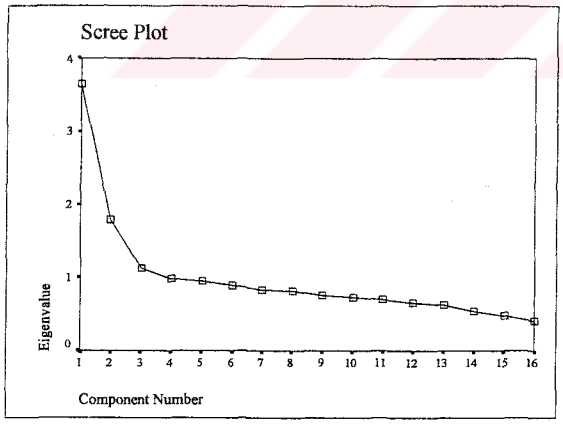
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.829
Bartlett's Test of Sphericity	Approx. Chi-Square	1650.825
	df	120
	Sig.	.000

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.648	22.800	22.800	3.648	22.800	22.800
2	1.795	11.220	34.020	1.795	11.220	34.020
3	1.139	7.116	41.136	1.139	7.116	41.136
4	.982	6.136	47.272			
5	.951	5.942	53.214			
6	.893	5.584	58.799			
7	.833	5.206	64.005			
8	.818	5.113	69.118			
9	.765	4.779	73.897			
10	.731	4.566	78.463			
11	.704	4.402	82.865			
12	.665	4.156	87.021			
13	.631	3.945	90.966			
14	.551	3.441	94.408			
15	.481	3.009	97.417			
16	.413	2.583	100.000			

Extraction Method: Principal Component Analysis.



Component Matrix

	Component		
	1.000	2.000	3.000
SC6	.488	-.588	.164
SC7	.512	-.638	.121
SC8	.458	-.211	-.334
SC9	.516	-.612	-.102
SC10	.489	-.221	-.169
SR1	.214	-.067	.322
SR2	.497	.240	-.101
SR3	.450	.225	-.337
SR4	.538	.251	-.301
SR5	.564	.242	-.244
SR6	.455	.235	-.232
CR1	.420	.297	.218
CR2	.547	.264	.123
CR3	.455	.254	.298
CR4	.434	.129	.480
CR5	.502	.218	.364

Extraction Method: PCA

Component Transformation Matrix

Component	1	2	3
1	.664	.525	.533
2	.423	-.851	.311
3	-.617	-.019	.787

Extraction Method: PCA Rotation Method: Varimax with Kaiser Normalization.

APPENDIX 4.27
PCA: January 2001 Test – Half-Set II

Communalities

	Initial	Extraction
SC6	1.000	.620
SC7	1.000	.673
SC8	1.000	.490
SC9	1.000	.641
SC10	1.000	.310
SR1	1.000	.748
SR7	1.000	.361
SR8	1.000	.470
SR9	1.000	.436
SR10	1.000	.738
SR11	1.000	.387
CR6	1.000	.408
CR7	1.000	.447
CR8	1.000	.264
CR9	1.000	.375
CR10	1.000	.399

Extraction Method: PCA

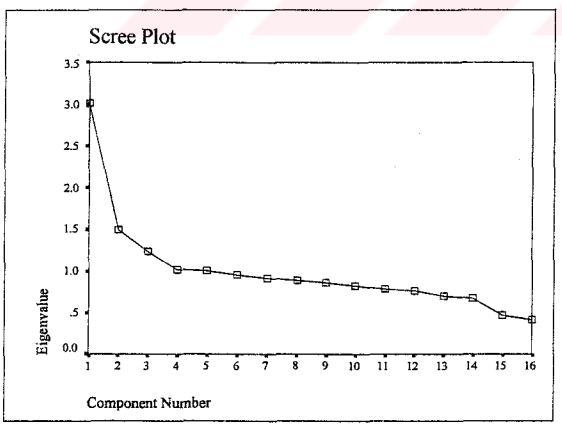
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.774
Bartlett's Test of Sphericity	Approx. Chi-Square	145.172
	df	120
	Sig.	.000

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.013	18.832	18.832	3.013	18.832	18.832
2	1.498	9.360	28.192	1.498	9.360	28.192
3	1.231	7.695	35.887	1.231	7.695	35.887
4	1.015	6.345	42.232	1.015	6.345	42.232
5	1.008	6.301	48.533	1.008	6.301	48.533
6	.957	5.979	54.513			
7	.917	5.728	60.241			
8	.886	5.536	65.777			
9	.862	5.390	71.167			
10	.825	5.159	76.325			
11	.782	4.888	81.214			
12	.757	4.733	85.947			
13	.697	4.354	90.301			
14	.669	4.182	94.483			
15	.469	2.934	97.417			
16	.413	2.583	100.000			

Extraction Method: Principal Component Analysis.



Component Matrix

	Component				
	1	2	3	4	5
SC6	.600	-.492	-.037	.029	.126
SC7	.674	-.460	-.068	-.008	.053
SC8	.511	-.037	.132	-.088	-.450
SC9	.668	-.435	-.036	-.055	-.017
SC10	.512	-.136	.150	-.003	-.082
SR1	.237	.047	-.157	.717	.388
SR7	.293	.295	.398	.170	-.003
SR8	.447	.348	.368	-.021	-.110
SR9	.420	.289	.068	-.376	.173
SR10	.141	.105	.440	-.315	.644
SR11	.141	.160	.458	.339	-.131
CR6	.234	.224	-.465	-.081	.283
CR7	.458	.399	-.266	-.004	.089
CR8	.417	.252	-.065	.149	.007
CR9	.314	.304	-.319	-.241	-.157
CR10	.367	.372	-.255	.133	-.207

Extraction Method: PCA

Component Transformation Matrix

Component	1	2	3	4	5
1	.767	.524	.347	.125	.039
2	-.641	.636	.403	.151	-.016
3	-.002	-.553	.726	.365	-.186
4	-.035	-.125	.347	-.469	.802
5	-.020	-.023	-.265	.780	.566

Extraction Method: PCA Rotation Method: Varimax with Kaiser Normalization.

APPENDIX 4.28
June 2001 – Pilot Version
Normality Tests and Graphs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
total reading	71	100.0%	0	.0%	71	100.0%

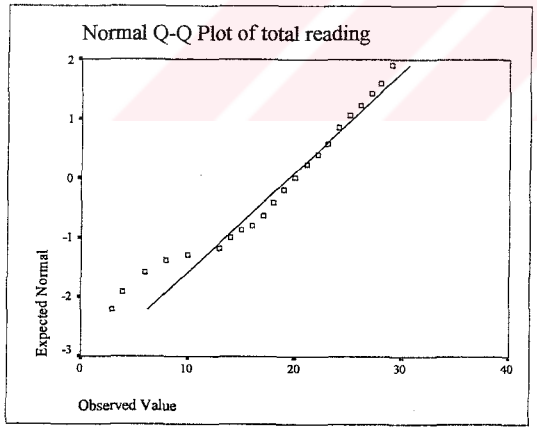
Descriptives

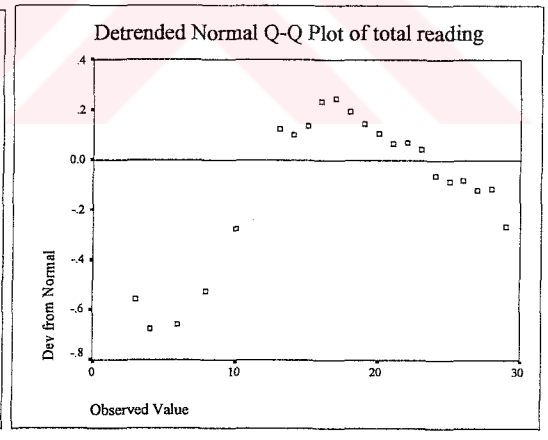
			Statistic	Std. Error
total reading	Mean		19.2817	.7004
	95% Confidence Interval for Mean	Lower Bound	17.8847	
		Upper Bound	20.6787	
	5% Trimmed Mean		19.5665	
	Median		20.0000	
	Variance		34.834	
	Std. Deviation		5.9020	
	Minimum		3.00	
	Maximum		29.00	
	Range		26.00	
	Interquartile Range		6.0000	
	Skewness		-.848	.285
	Kurtosis		.724	.563

Tests of Normality

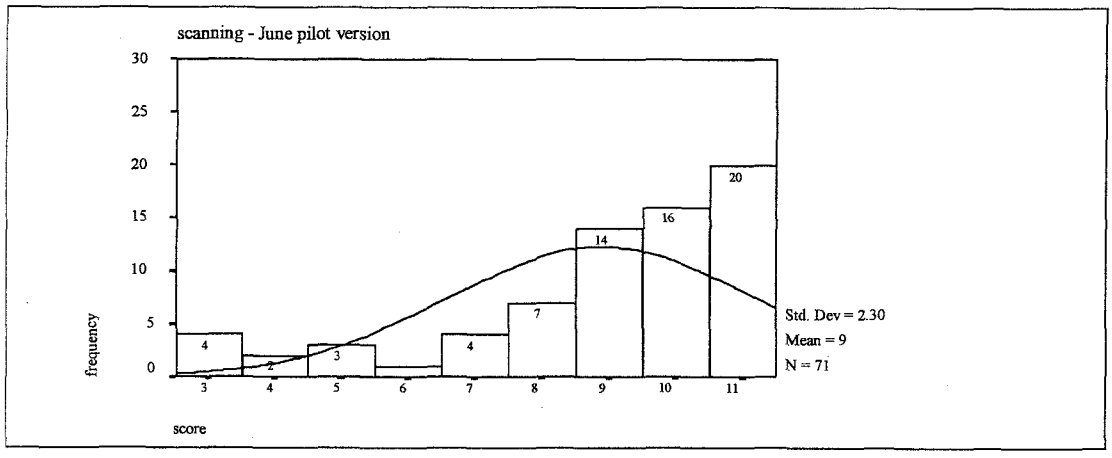
	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
total reading	.124	71	.009

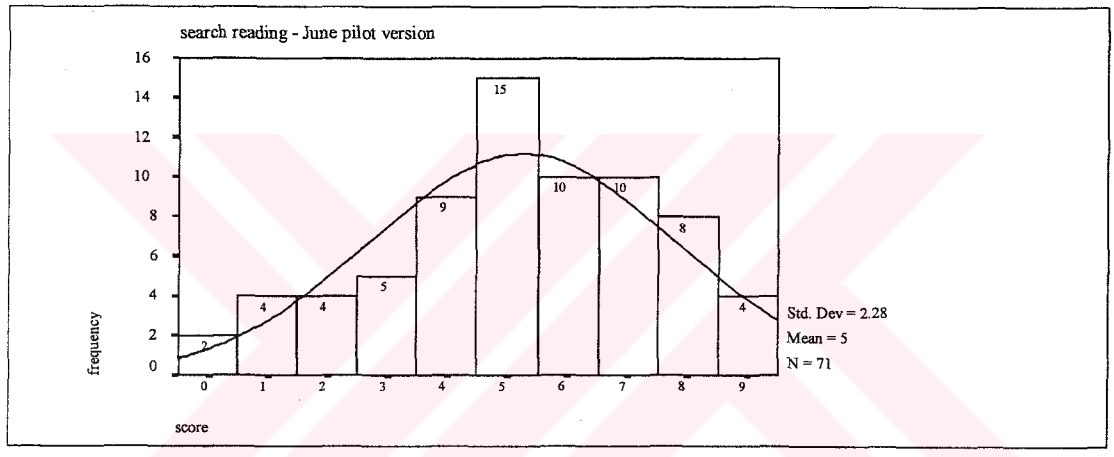
a. Lilliefors Significance Correction

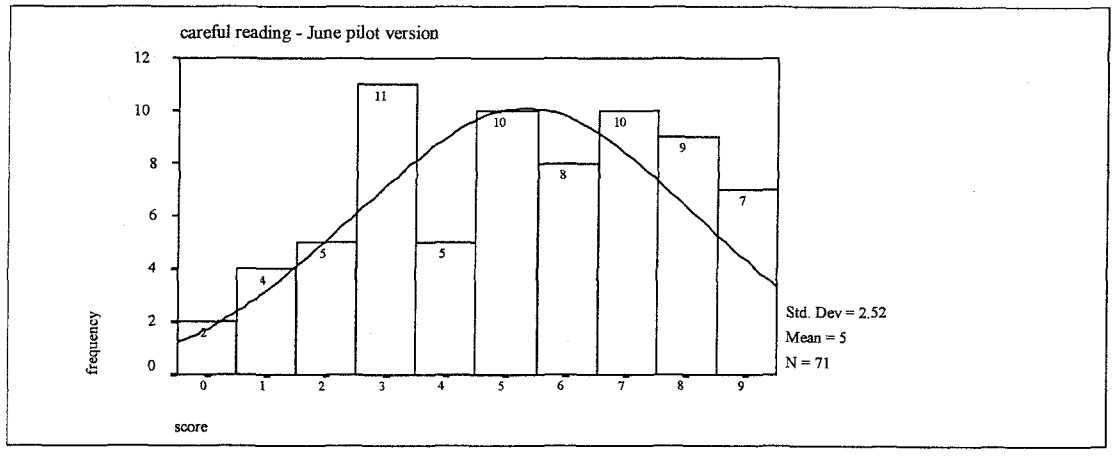




APPENDIX 4.29
June 2001 – Pilot Version
Score Distribution Graphs by Subtest







APPENDIX 4.30
June 2001 – Pilot Version
Normality Tests and Graphs by Subsets

Scanning

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
total scanning	71	100.0%	0	.0%	71	100.0%

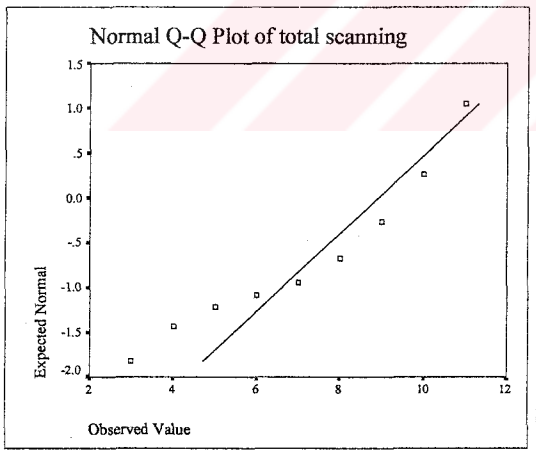
Descriptives

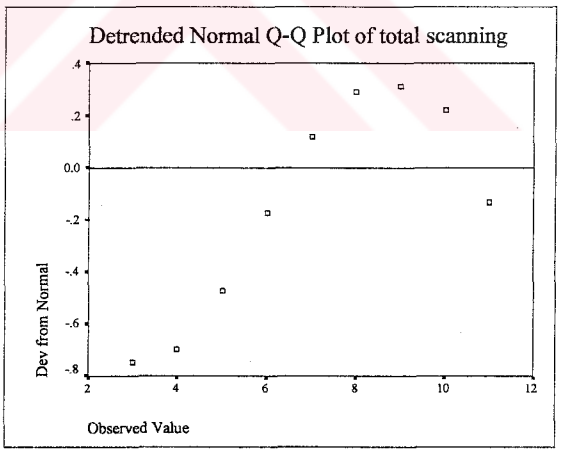
			Statistic	Std. Error
total scanning	Mean		8.8873	.2725
	95% Confidence Interval for Mean	Lower Bound	8.3438	
		Upper Bound	9.4308	
	5% Trimmed Mean		9.0970	
	Median		10.0000	
	Variance		5.273	
	Std. Deviation		2.2963	
	Minimum		3.00	
	Maximum		11.00	
	Range		8.00	
	Interquartile Range		3.0000	
	Skewness		-1.300	.285
	Kurtosis		.886	.563

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
total scanning	.224	71	.000

a. Lilliefors Significance Correction





Search Reading

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
total search reading	71	100.0%	0	.0%	71	100.0%

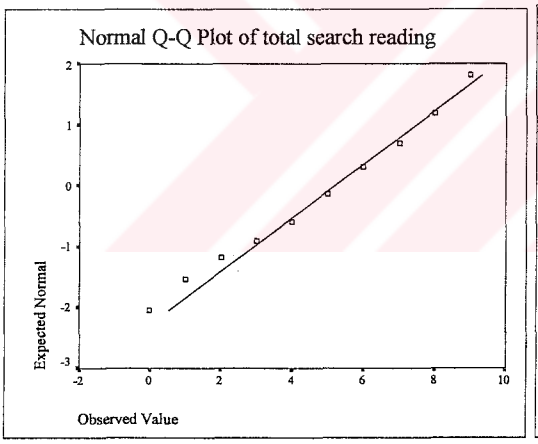
Descriptives

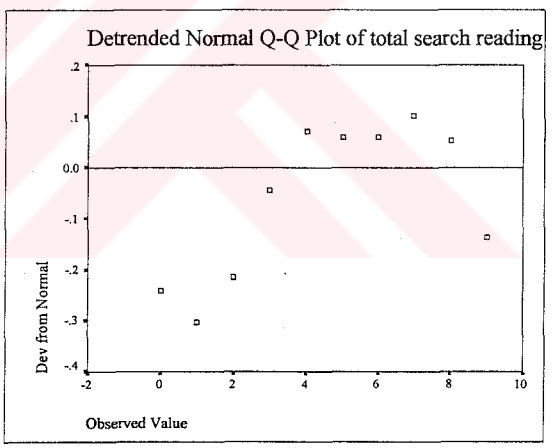
			Statistic	Std. Error
total search reading	Mean		5.1831	.2701
	95% Confidence Interval for Mean	Lower Bound	4.6444	
		Upper Bound	5.7218	
	5% Trimmed Mean		5.2347	
	Median		5.0000	
	Variance		5.180	
	Std. Deviation		2.2760	
	Minimum		.00	
	Maximum		9.00	
	Range		9.00	
	Interquartile Range		3.0000	
	Skewness		-.375	.285
	Kurtosis		-.421	.563

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
total search reading	.130	71	.005

a. Lilliefors Significance Correction





Careful Reading

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
total careful reading	71	100.0%	0	.0%	71	100.0%

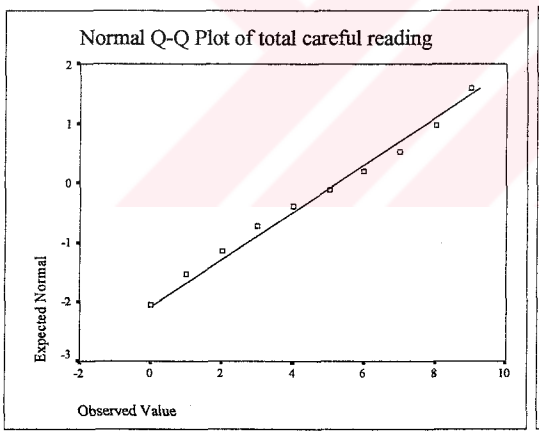
Descriptives

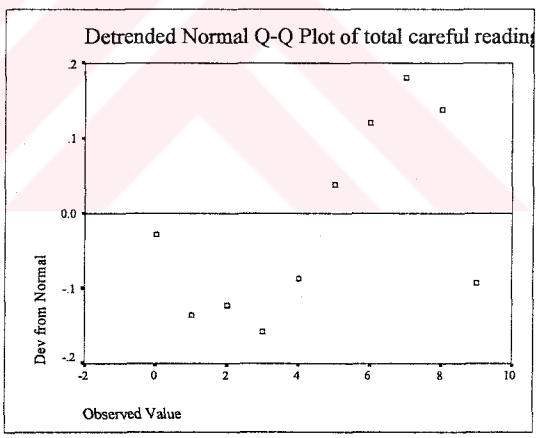
			Statistic	Std. Error
total careful reading	Mean		5.2113	.2995
	95% Confidence Interval for Mean	Lower Bound	4.6139	
		Upper Bound	5.8086	
	5% Trimmed Mean		5.2660	
	Median		5.0000	
	Variance		6.369	
	Std. Deviation		2.5237	
	Minimum		.00	
	Maximum		9.00	
	Range		9.00	
	Interquartile Range		4.0000	
	Skewness		-.218	.285
	Kurtosis		-.972	.563

Tests of Normality

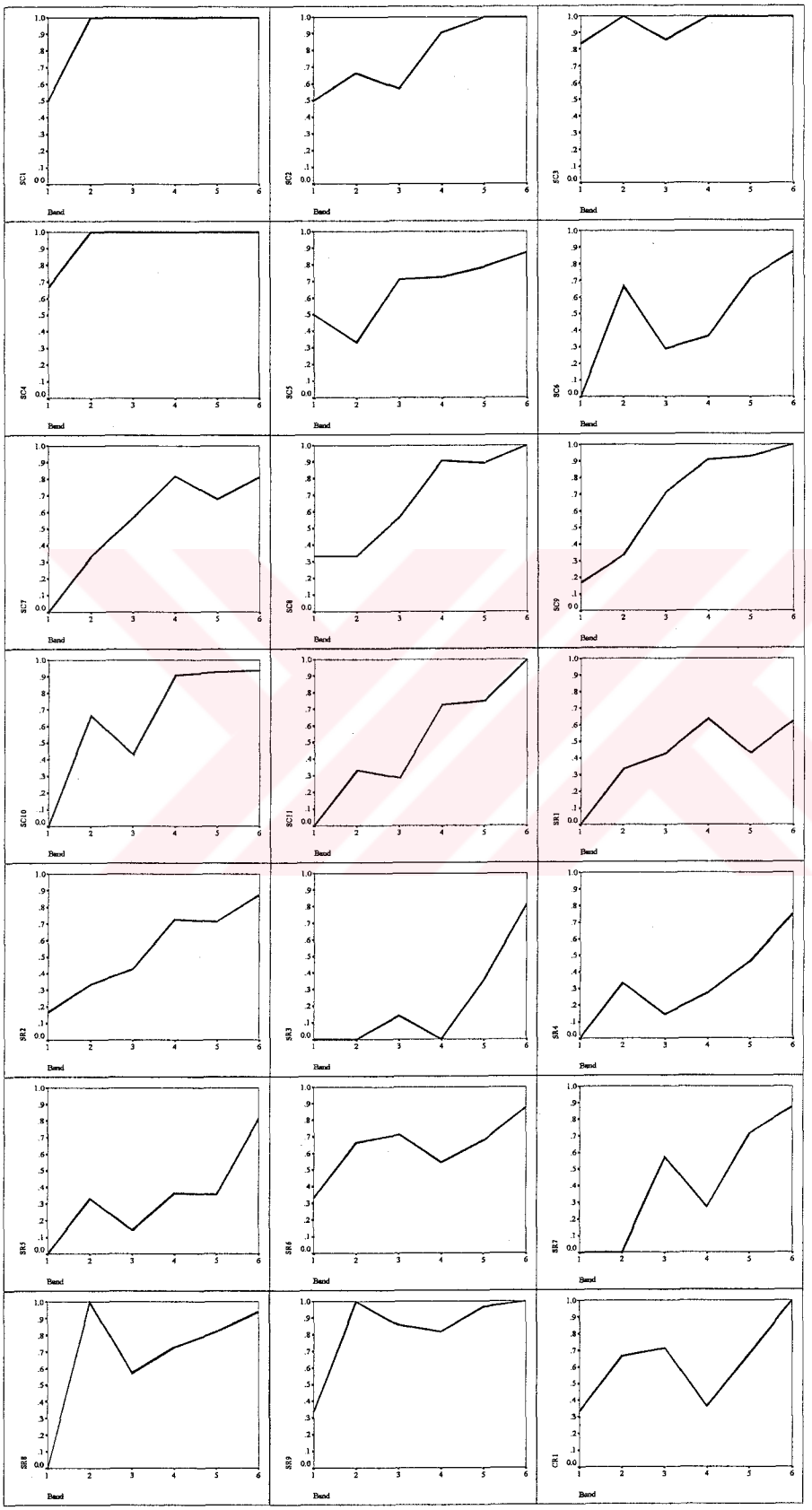
	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
total careful reading	.127	71	.006

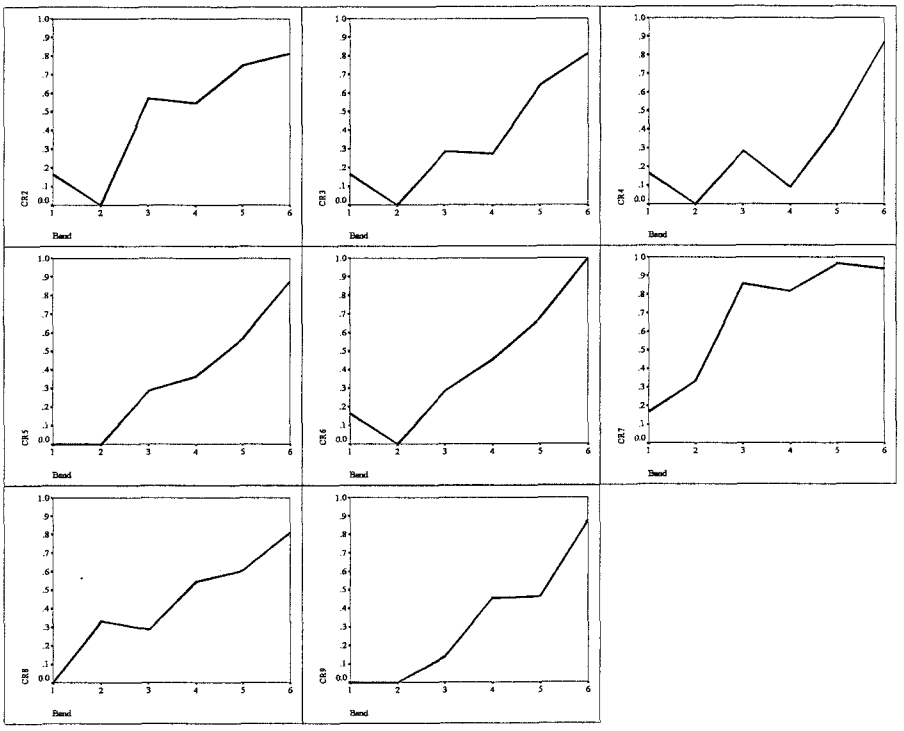
a. Lilliefors Significance Correction





APPENDIX 4.31
June 2001 – Pilot Version
Band Score Graphs




CR2
CR3
CR4
CR5
CR6
CR7
CR8
CR9
Band

APPENDIX 4.32
June 2001 Test
Normality Tests and Graphs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
totalreading	1100	100.0%	0	.0%	1100	100.0%

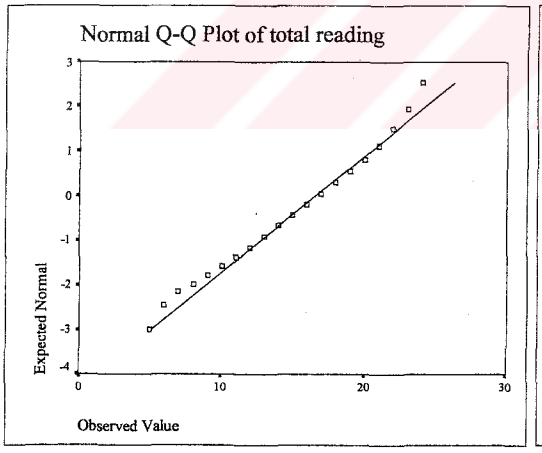
Descriptives

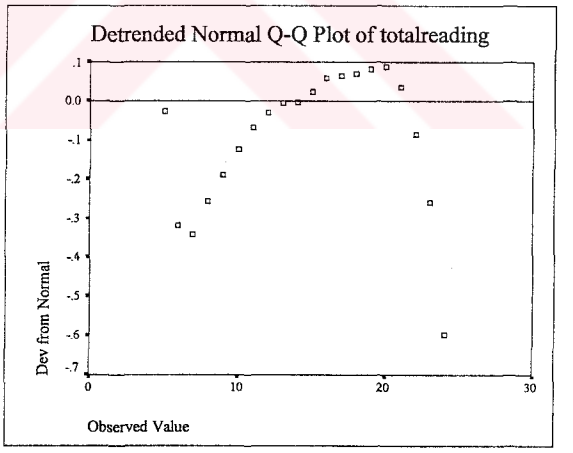
			Statistic	Std. Error
totalreading	Mean		16.5527	.1152
	95% Confidence Interval for Mean	Lower Bound	16.3267	
		Upper Bound	16.7787	
	5% Trimmed Mean		16.6737	
	Median		17.0000	
	Variance		14.595	
	Std. Deviation		3.8203	
	Minimum		5.00	
	Maximum		24.00	
	Range		19.00	
	Interquartile Range		5.0000	
	Skewness		-.399	.074
	Kurtosis		-.202	.147

Tests of Normality

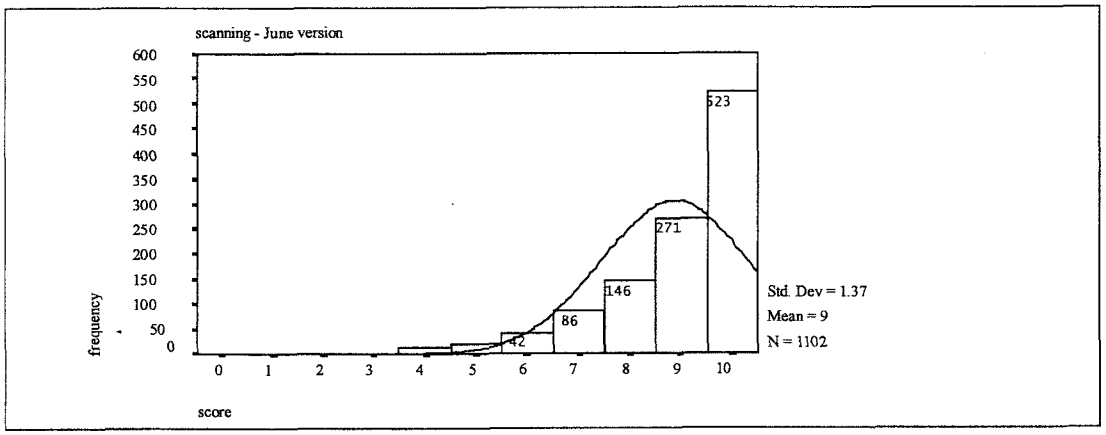
	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
totalreading	.079	1100	.000

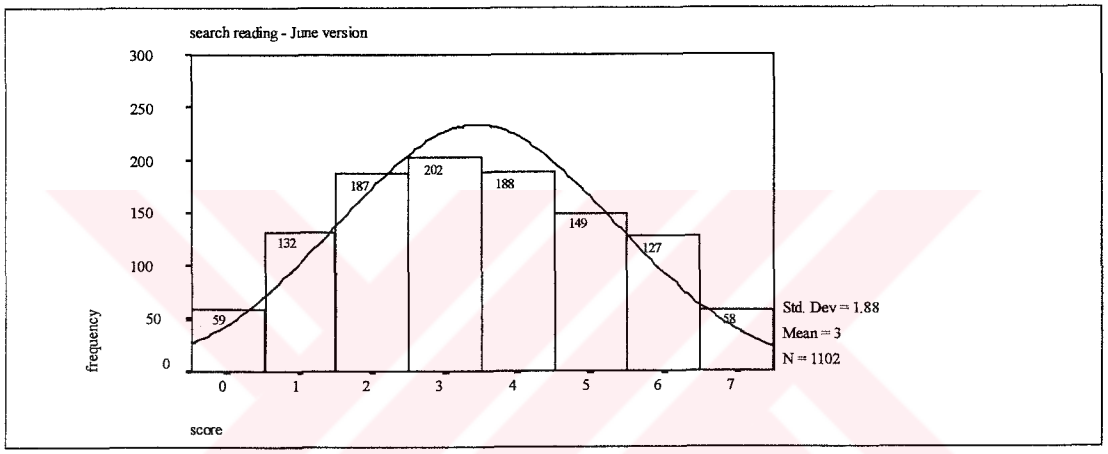
a. Lilliefors Significance Correction

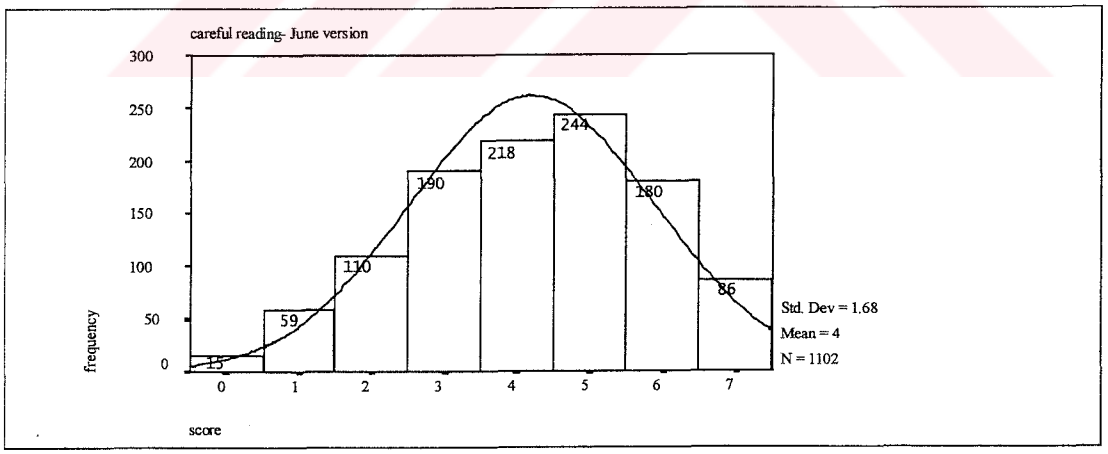




APPENDIX 4.33
June 2001 Test
Score Distribution Graphs by Subtest







APPENDIX 4.34
June 2001 Test
Normality Tests and Graphs by Subtest

Scanning

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
total scannin	1100	100.0%	0	.0%	1100	100.0%

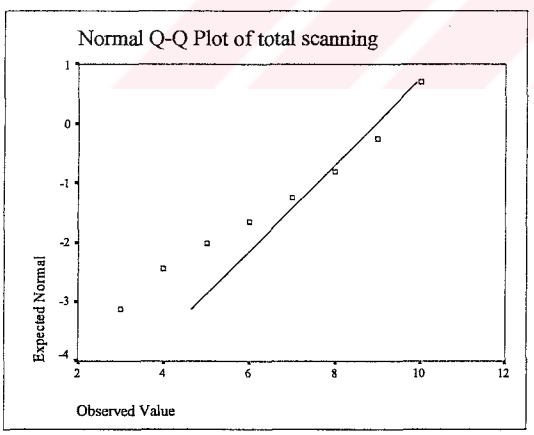
Descriptives

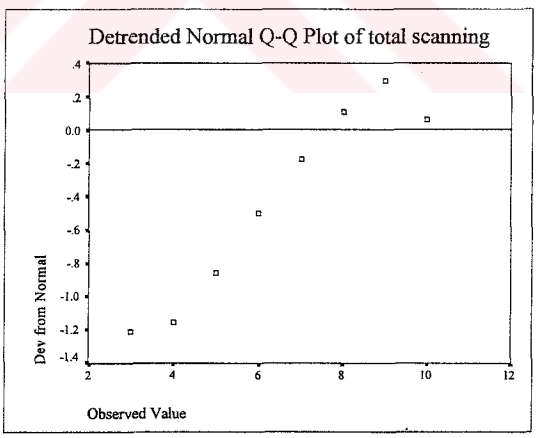
			Statistic	Std. Error
total scanning	Mean		8.9336	.041
	95% Confidence Interval for Mean	Lower Bound	8.8525	
		Upper Bound	9.0148	
	5% Trimmed Mean		9.0869	
	Median		9.0000	
	Variance		1.880	
	Std. Deviation		1.3711	
	Minimum		3.00	
	Maximum		10.00	
	Range		7.00	
	Interquartile Range		2.0000	
	Skewness		-1.469	.074
	Kurtosis		1.866	.147

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
total scanning	.256	1100	.000

a. Lilliefors Significance Correction





Search Reading

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
total search readi	1100	100.0%	0	.0%	1100	100.0%

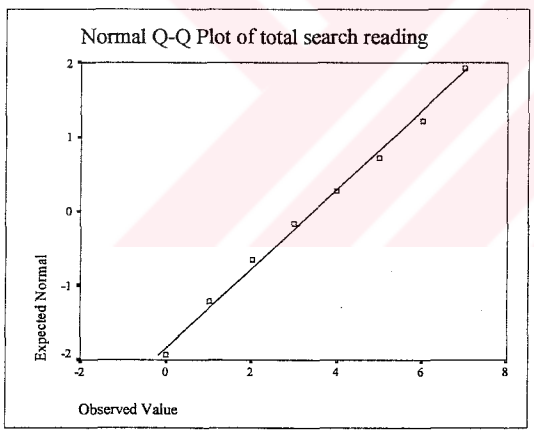
Descriptives

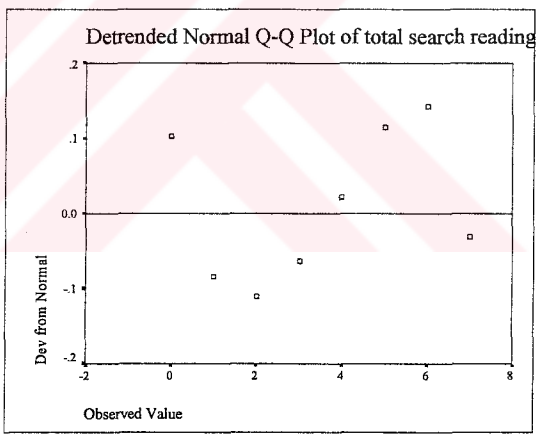
			Statistic	Std. Error
total search readin	Mean		3.4255	5.675E-02
	95% Confidence Interval for Mean	Lower Bound	3.3141	
		Upper Bound	3.5368	
	5% Trimmed Mean		3.4172	
	Median		3.0000	
	Variance		3.542	
	Std. Deviation		1.8821	
	Minimum		.00	
	Maximum		7.00	
	Range		7.00	
	Interquartile Range		3.0000	
	Skewness		.083	.074
	Kurtosis		-.864	.147

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
total search reading	.119	1100	.000

a. Lilliefors Significance Correction





Careful Reading

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
total careful reading	1100	100.0%	0	.0%	1100	100.0%

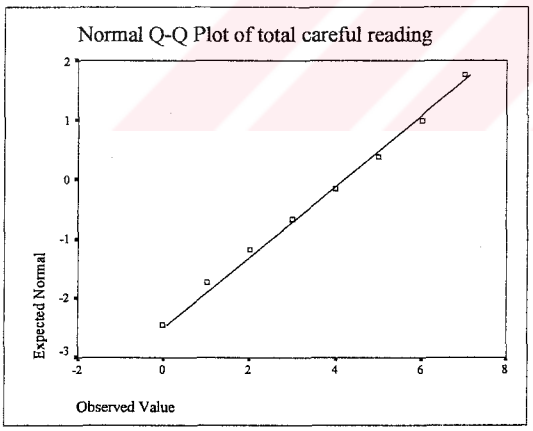
Descriptives

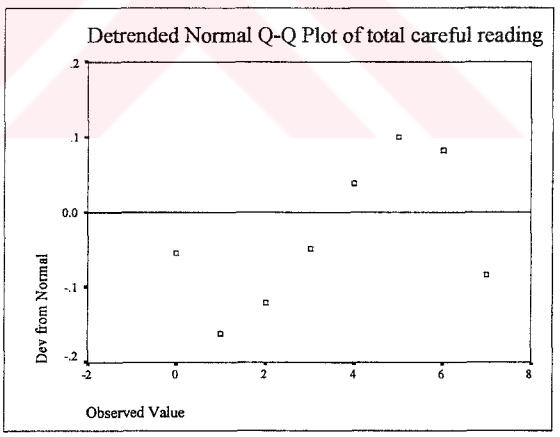
			Statistic	Std. Error
total careful reading	Mean		4.1936	5.060E-02
	95% Confidence Interval for Mean	Lower Bound	4.0943	
		Upper Bound	4.2929	
	5% Trimmed Mean		4.2303	
	Median		4.0000	
	Variance		2.817	
	Std. Deviation		1.6784	
	Minimum		.00	
	Maximum		7.00	
	Range		7.00	
	Interquartile Range		2.0000	
	Skewness		-.275	.074
	Kurtosis		-.580	.147

Tests of Normality

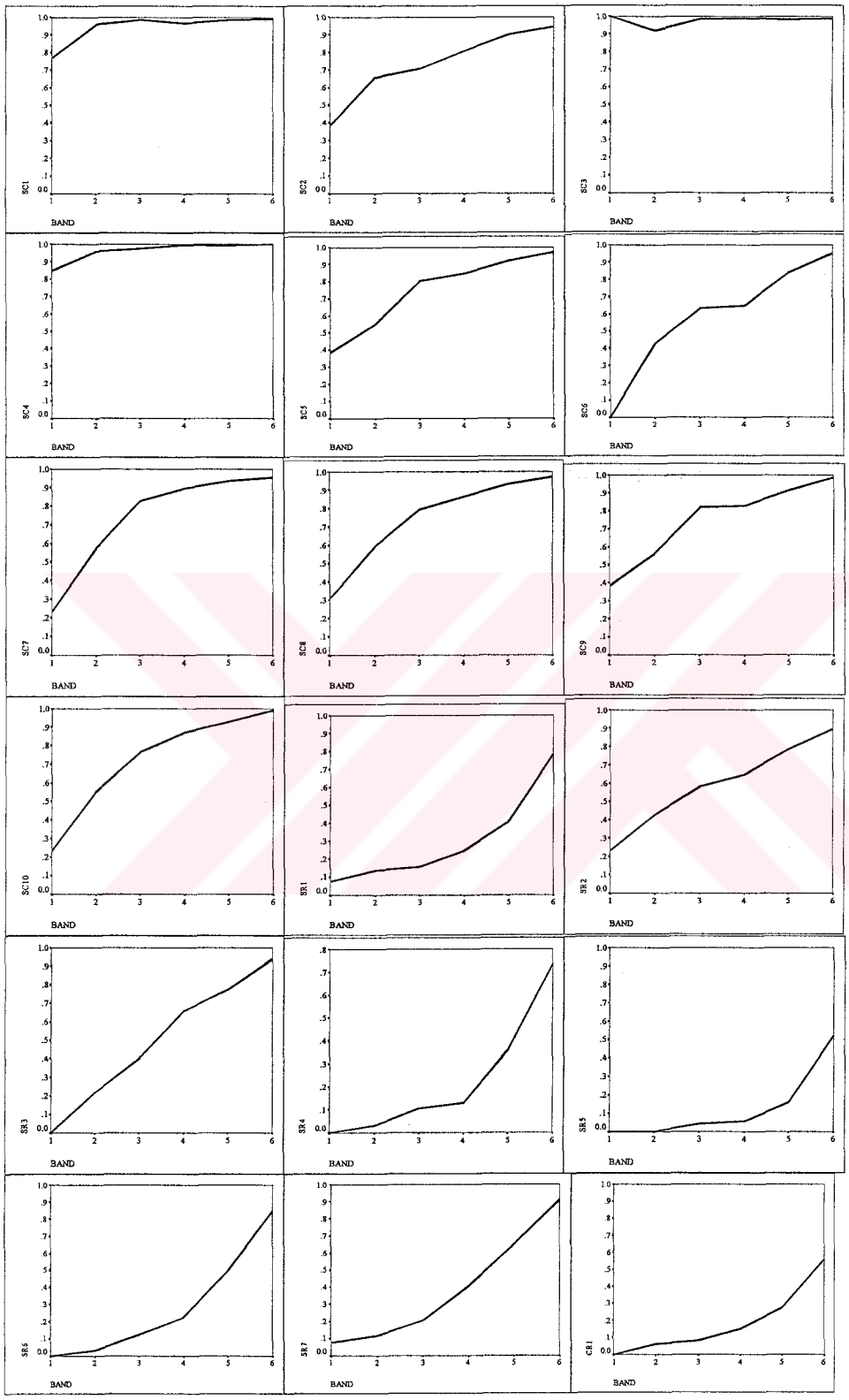
	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
total careful reading	.146	1100	.000

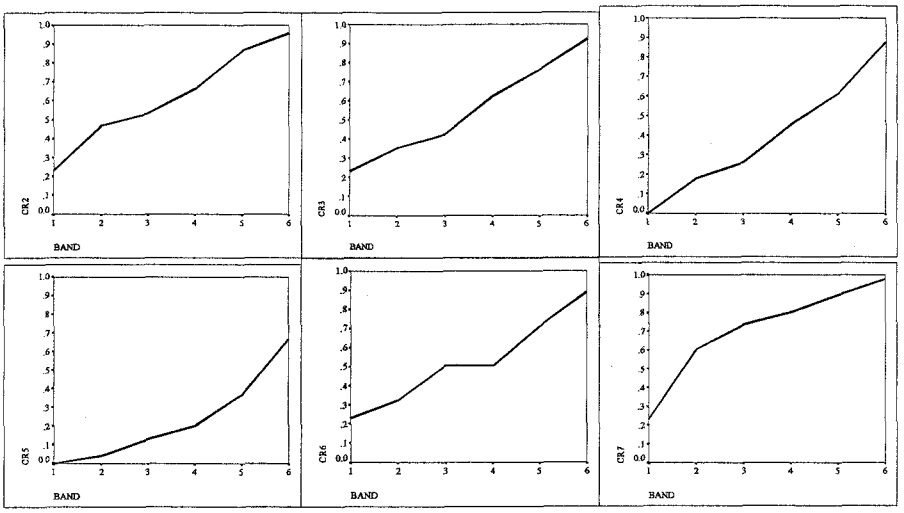
a. Lilliefors Significance Correction





APPENDIX 4.35
June 2001 Test
Band Score Graphs




CR2
CR3
CR4
CR5
CR6
CR7
BAND

APPENDIX 4.36
PCA: June 2001 Test – Whole Set

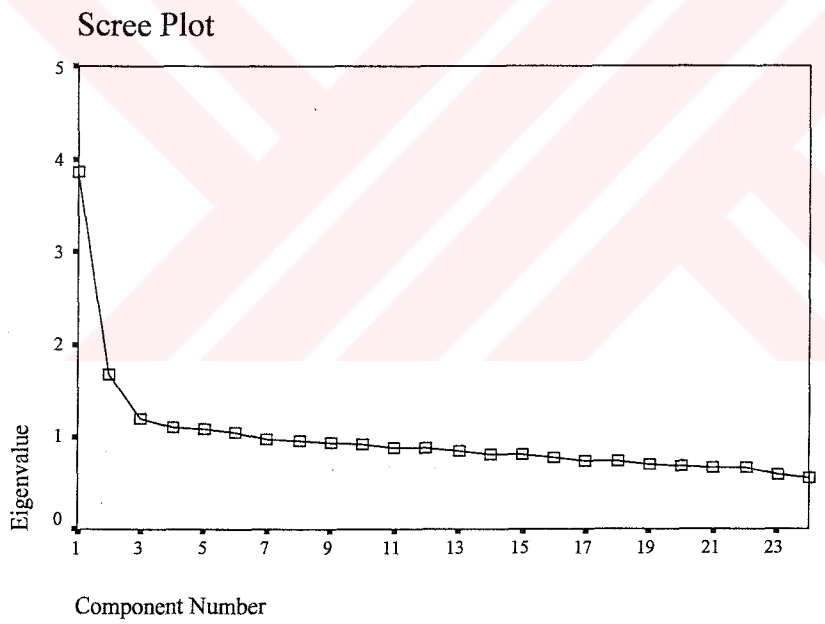
Communalities

	Initial	Extraction
SC1	1,000	,685
SC2	1,000	,541
SC3	1,000	,527
SC4	1,000	,420
SC5	1,000	,375
SC6	1,000	,306
SC7	1,000	,311
SC8	1,000	,459
SC9	1,000	,448
SC10	1,000	,521
SR1	1,000	,427
SR2	1,000	,549
SR3	1,000	,364
SR4	1,000	,377
SR5	1,000	,379
SR6	1,000	,444
SR7	1,000	,445
CR1	1,000	,387
CR2	1,000	,310
CR3	1,000	,277
CR4	1,000	,403
CR5	1,000	,262
CR6	1,000	,404
CR7	1,000	,312

Extraction Method: PCA

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,852
Bartlett's Test of Sphericity	Approx. Chi-Square	2642,258
	df	276
	Sig.	,000



Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,853	16,056	16,056	3,853	16,056	16,056
2	1,665	6,937	22,993	1,665	6,937	22,993
3	1,189	4,952	27,945	1,189	4,952	27,945
4	1,103	4,595	32,540	1,103	4,595	32,540
5	1,081	4,505	37,045	1,081	4,505	37,045
6	1,043	4,344	41,389	1,043	4,344	41,389
7	,970	4,041	45,430			
8	,950	3,960	49,390			
9	,931	3,877	53,267			
10	,912	3,800	57,067			
11	,887	3,694	60,761			
12	,880	3,668	64,429			
13	,844	3,516	67,945			
14	,808	3,365	71,310			
15	,804	3,349	74,659			
16	,775	3,230	77,889			
17	,741	3,089	80,978			
18	,727	3,030	84,008			
19	,702	2,926	86,935			
20	,680	2,833	89,768			
21	,666	2,773	92,541			
22	,662	2,758	95,299			
23	,583	2,430	97,729			
24	,545	2,271	100,000			

Extraction Method: Principal Component Analysis.

Component Matrix

	Component					
	1	2	3	4	5	6
SC1	,106	,132	,330	-,218	,650	-,279
SC2	,255	,192	,008	,156	,457	,453
SC3	,083	,221	,281	-,434	-,318	,320
SC4	,144	,313	,312	,407	,083	,178
SC5	,354	,388	,175	-,224	,072	-,114
SC6	,397	,224	-,032	,150	,168	,217
SC7	,336	,349	-,207	,096	-,068	-,140
SC8	,359	,510	-,208	-,105	-,019	-,126
SC9	,364	,382	-,211	-,155	-,303	-,095
SC10	,426	,522	-,173	,053	,016	-,185
SR1	,451	-,135	,269	,138	-,120	-,315
SR2	,309	,039	,140	,619	-,213	-,060
SR3	,499	-,034	,118	,117	-,133	,263
SR4	,528	-,252	,127	,032	-,123	-,056
SR5	,459	-,171	,311	-,190	-,035	-,068
SR6	,605	-,126	,182	-,147	-,031	,077
SR7	,572	-,119	,195	-,075	-,052	,238
CR1	,374	-,259	-,181	,190	,147	-,300
CR2	,418	-,271	-,075	-,094	,215	,025
CR3	,429	-,203	-,047	-,035	,095	-,199
CR4	,488	-,250	-,259	-,039	-,050	,177
CR5	,429	-,164	,035	-,072	-,117	-,177
CR6	,368	-,204	-,412	-,007	,106	,215
CR7	,351	-,080	-,378	-,154	,102	,077

Component Transformation Matrix

Component	1	2	3	4	5	6
1	,748	,437	,398	,297	,050	,023
2	-,382	,761	-,323	,308	-,246	,128
3	,451	-,312	-,682	,267	-,294	,275
4	-,066	-,077	-,255	,548	,650	-,450
5	-,217	-,171	,254	,344	,320	,799
6	-,199	-,313	,377	,572	-,571	-,259

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

APPENDIX 4.37
PCA: June 2001 Test – Purged Set I

Communalities

	Initial	Extraction
SC5	1.000	.381
SC6	1.000	.207
SC7	1.000	.304
SC8	1.000	.491
SC9	1.000	.351
SC10	1.000	.500
SR1	1.000	.318
SR3	1.000	.302
SR4	1.000	.384
SR6	1.000	.492
SR7	1.000	.439
CR2	1.000	.336
CR3	1.000	.284
CR4	1.000	.388
CR5	1.000	.231
CR6	1.000	.468

Extraction Method: PCA

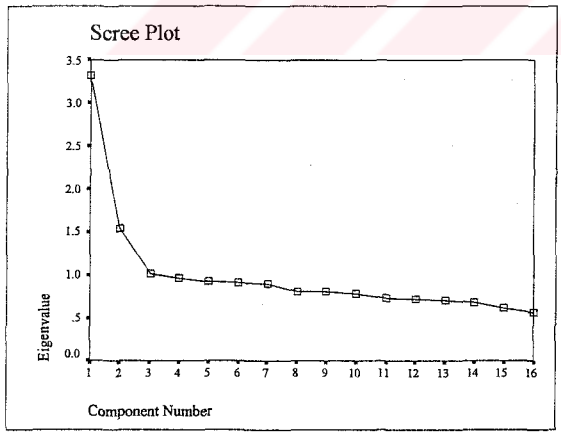
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Samplin Adequacy.		.841
Bartlett's Test of Sphericity	Approx. Chi-Square	924.965
	df	120
	Sig.	.000

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.323	20.767	20.767	3.323	20.767	20.767
2	1.539	9.619	30.386	1.539	9.619	30.386
3	1.015	6.343	36.729	1.015	6.343	36.729
4	.966	6.038	42.766			
5	.924	5.775	48.541			
6	.913	5.707	54.248			
7	.901	5.630	59.878			
8	.816	5.097	64.975			
9	.805	5.031	70.007			
10	.779	4.867	74.874			
11	.733	4.584	79.458			
12	.715	4.466	83.924			
13	.709	4.431	88.354			
14	.681	4.256	92.610			
15	.621	3.880	96.490			
16	.562	3.510	100.000			

Extraction Method: Principal Component Analysis.



Component Matrix

	Component		
	1	2	3
SC5	.379	.373	-.314
SC6	.397	.198	-.102
SC7	.351	.385	.181
SC8	.384	.540	.229
SC9	.404	.407	.150
SC10	.439	.554	.020
SR1	.441	-.151	-.317
SR3	.505	-.089	-.196
SR4	.538	-.283	-.122
SR6	.610	-.150	-.313
SR7	.584	-.184	-.255
CR2	.427	-.301	.251
CR3	.435	-.217	.219
CR4	.498	-.264	.264
CR5	.432	-.178	.112
CR6	.371	-.203	.538

Component Transformation Matrix

Component	1	2	3
1	.685	.499	.531
2	-.251	.846	-.471
3	-.684	.189	.705

Extraction Method: PCA Rotation Method: Varimax with Kaiser Normalization.

APPENDIX 4.38
PCA: June 2001 Test – Individual Subtests

Scanning

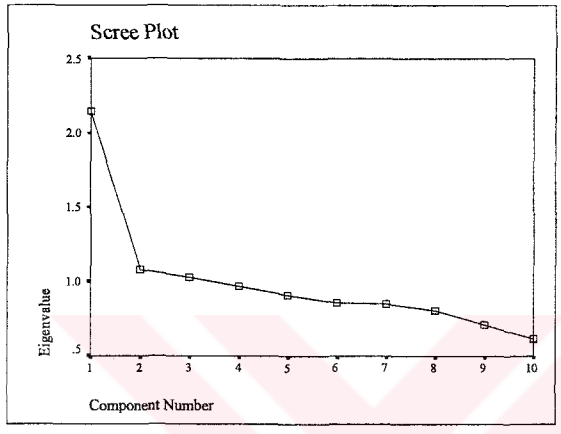
Communalities

	Initial	Extraction
SC1	1,000	,548
SC2	1,000	,464
SC3	1,000	,639
SC4	1,000	,219
SC5	1,000	,428
SC6	1,000	,273
SC7	1,000	,340
SC8	1,000	,425
SC9	1,000	,460
SC10	1,000	,467

Extraction Method: PCA

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.740
Bartlett's Test of Sphericity	Approx. Chi-Square	640.123
	df	45
	Sig.	.000



Component Matrix

	Component		
	1	2	3
SC1	.181	.650	.304
SC2	.334	.500	-.319
SC3	.187	-.074	.774
SC4	.291	.345	-.124
SC5	.525	.087	.380
SC6	.466	.167	-.166
SC7	.501	-.210	-.212
SC8	.625	-.168	-.078
SC9	.540	-.410	.011
SC10	.676	-.093	-.045

Extraction Method: PCA

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,148	21,477	21,477	2,148	21,477	21,477
2	1,082	10,822	32,299	1,082	10,822	32,299
3	1,033	10,330	42,629	1,033	10,330	42,629
4	,974	9,740	52,369			
5	,911	9,110	61,479			
6	,860	8,603	70,083			
7	,851	8,506	78,588			
8	,806	8,058	86,646			
9	,715	7,146	93,792			
10	,621	6,208	100,000			

Extraction Method: Principal Component Analysis.

Component Transformation Matrix

Component	1	2	3
1	,849	,432	,304
2	-,488	,861	,141
3	-,201	-,268	,942

Extraction Method: PCA Rotation Method: Varimax with Kaiser Normalization.

Search reading – one component extracted

Communalities

	Initial	Extraction
SR1	1.000	.301
SR2	1.000	.118
SR3	1.000	.324
SR4	1.000	.373
SR5	1.000	.331
SR6	1.000	.470
SR7	1.000	.422

Extraction Method: PCA

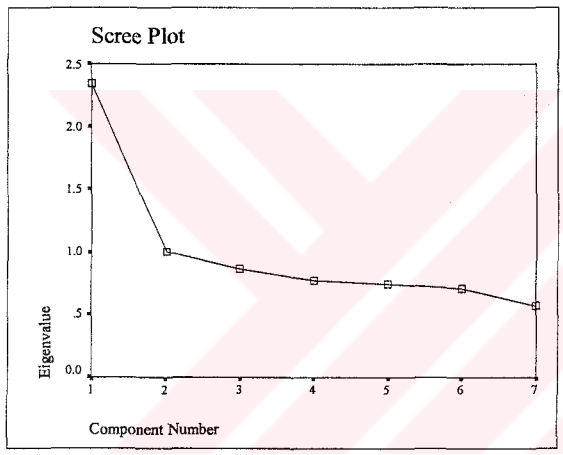
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Samplir Adequacy.		.776
Bartlett's Test of Sphericity	Approx. Chi-Square	826.455
	df	21
	Sig.	.000

Total Variance Explained

	Initial Eigenvalues		
Component	Total	% of Variance	Cumulative %
1	2,340	33,425	33,425
2	,998	14,255	47,680
3	,866	12,377	60,057
4	,772	11,029	71,087
5	,746	10,656	81,742
6	,706	10,091	91,833
7	,572	8,167	100,000

Extraction Method: PCA



Search reading – two components extracted

Communalities

	Initial	Extraction
SR1	1.000	.347
SR2	1.000	.787
SR3	1.000	.372
SR4	1.000	.374
SR5	1.000	.423
SR6	1.000	.565
SR7	1.000	.470

Extraction Method: PCA

Careful reading

Communalities

	Initial	Extraction
CR1	1.000	.264
CR2	1.000	.311
CR3	1.000	.295
CR4	1.000	.384
CR5	1.000	.245
CR6	1.000	.301
CR7	1.000	.200

Extraction Method: PCA

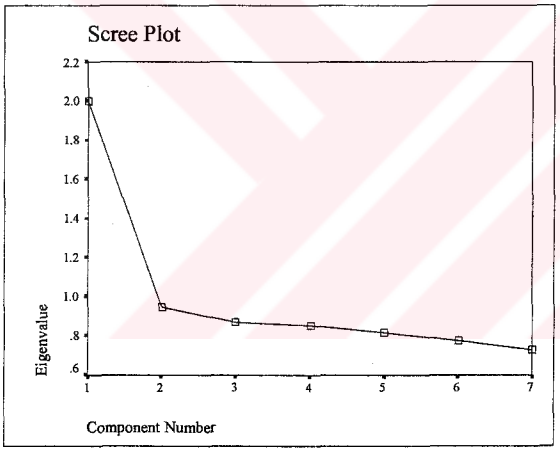
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.754
Bartlett's Test of Sphericity	Approx. Chi-Square	461.556
	df	21
	Sig.	.000

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.000	28.567	28.567	2.000	28.567	28.567
2	.949	13.561	42.128			
3	.868	12.401	54.529			
4	.854	12.205	66.734			
5	.819	11.702	78.437			
6	.778	11.110	89.547			
7	.732	10.453	100.000			

Extraction Method: Principal Component Analysis.



Careful reading - two components extracted

Communalities

	Initial	Extraction
CR1	1.000	.308
CR2	1.000	.370
CR3	1.000	.436
CR4	1.000	.400
CR5	1.000	.329
CR6	1.000	.446
CR7	1.000	.661

Extraction Method: PCA

APPENDIX 4.39
PCA: June 2001 Test – Purged Set II

Communalities

	Initial	Extraction
SC5	1.000	.336
SC6	1.000	.197
SC7	1.000	.286
SC8	1.000	.449
SC9	1.000	.348
SC10	1.000	.497
SR1	1.000	.323
SR3	1.000	.268
SR4	1.000	.363
SR5	1.000	.432
SR6	1.000	.469
SR7	1.000	.405
CR1	1.000	.301
CR2	1.000	.315
CR3	1.000	.261
CR4	1.000	.380
CR5	1.000	.215
CR6	1.000	.428

Extraction Method: PCA

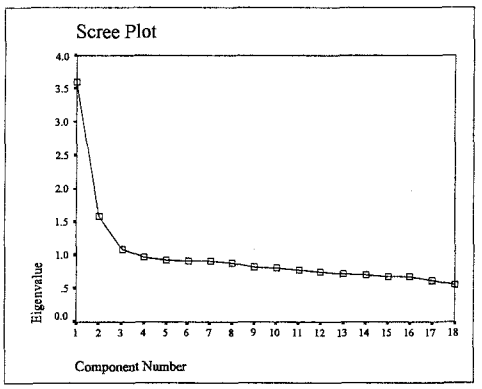
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampli Adequacy.		.857
Bartlett's Test of Sphericity	Approx. Chi-Square	243.355
	df	153
	Sig.	.000

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.606	20.034	20.034	3.606	20.034	20.034
2	1.584	8.802	28.836	1.584	8.802	28.836
3	1.082	6.013	34.850	1.082	6.013	34.850
4	.974	5.413	40.262			
5	.926	5.143	45.405			
6	.915	5.085	50.490			
7	.905	5.025	55.515			
8	.886	4.920	60.435			
9	.827	4.592	65.027			
10	.814	4.522	69.549			
11	.775	4.308	73.857			
12	.755	4.197	78.054			
13	.717	3.981	82.035			
14	.709	3.936	85.971			
15	.678	3.764	89.735			
16	.675	3.752	93.487			
17	.620	3.444	96.931			
18	.552	3.069	100.000			

Extraction Method: Principal Component Analysis.



Component Matrix

	Component		
	1	2	3
SC5	.355	.398	-.228
SC6	.382	.223	.045
SC7	.331	.400	.128
SC8	.356	.556	.115
SC9	.370	.444	.120
SC10	.409	.572	.050
SR1	.455	-.138	-.311
SR3	.497	-.041	-.139
SR4	.544	-.241	-.092
SR5	.471	-.181	-.422
SR6	.620	-.118	-.266
SR7	.581	-.129	-.224
CR1	.386	-.216	.325
CR2	.430	-.257	.253
CR3	.440	-.185	.184
CR4	.495	-.211	.301
CR5	.438	-.149	.035
CR6	.371	-.166	.513

Component Transformation Matrix

Component	1	2	3
1	.688	.479	.546
2	-.279	.868	-.410
3	-.670	.130	.731

Extraction Method: PCA Rotation Method: Varimax with Kaiser Normalization.

APPENDIX 4.40
September 2001 – Pilot Version
Normality Tests and Graphs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
total reading	75	100.0%	0	.0%	75	100.0%

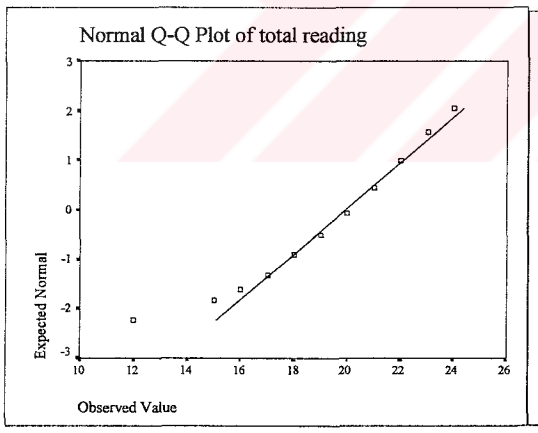
Descriptives

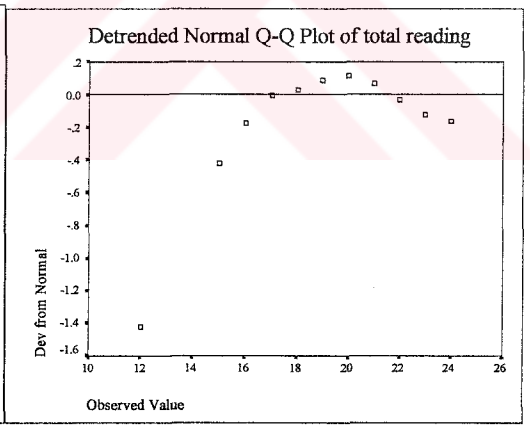
			Statistic	Std. Error
total reading	Mean		19.8933	.2501
	95% Confidence Interval for Mean	Lower Bound	19.3950	
		Upper Bound	20.3917	
	5% Trimmed Mean		19.9963	
	Median		20.0000	
	Variance		4.691	
	Std. Deviation		2.1659	
	Minimum		12.00	
	Maximum		24.00	
	Range		12.00	
	Interquartile Range		2.0000	
	Skewness		-.810	.277
	Kurtosis		1.477	.548

Tests of Normality

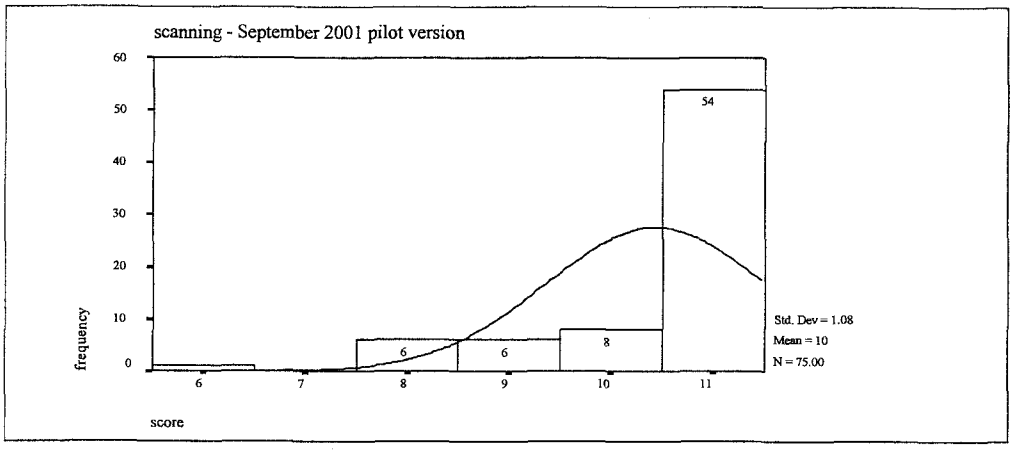
	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
total reading	.146	75	.000

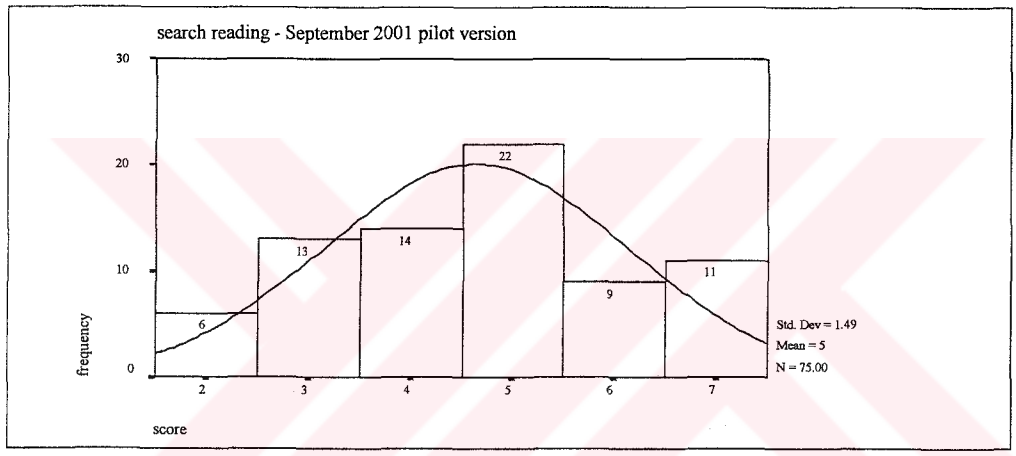
a. Lilliefors Significance Correction

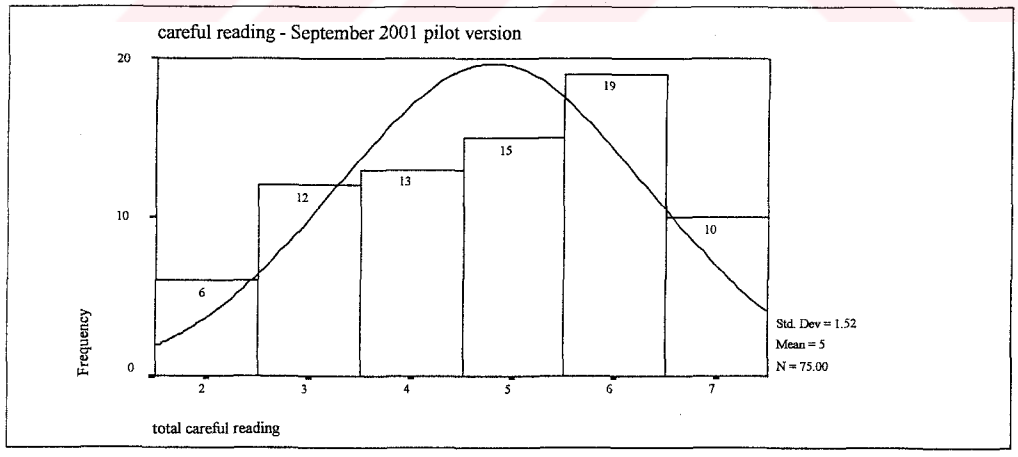




APPENDIX 4.41
September 2001 – Pilot Version
Score Distribution Graphs by Subtest







APPENDIX 4.42
September 2001 – Pilot Version
Normality Tests and Graphs by Subsets

Scanning

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
scanning	75	100.0%	0	.0%	75	100.0%

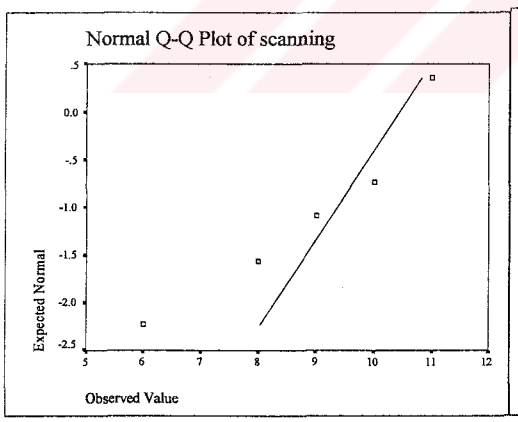
Descriptives

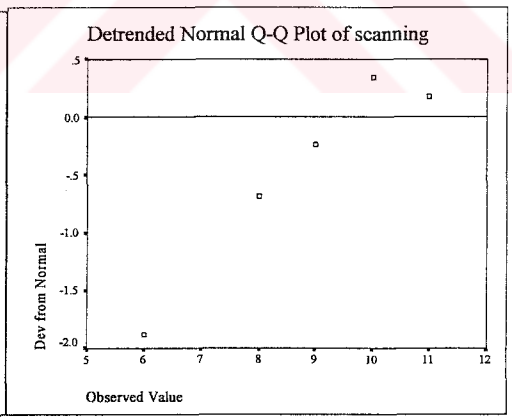
			Statistic	Std. Error
scanning	Mean		10.4267	.1247
	95% Confidence Interval for Mean	Lower Bound	10.1781	
		Upper Bound	10.6752	
	5% Trimmed Mean		10.5593	
	Median		11.0000	
	Variance		1.167	
	Std. Deviation		1.0802	
	Minimum		6.00	
	Maximum		11.00	
	Range		5.00	
	Interquartile Range		1.0000	
	Skewness		-1.987	.277
	Kurtosis		3.580	.548

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
scanning	.422	75	.000

a. Lilliefors Significance Correction





Search Reading

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
search reading	75	100.0%	0	.0%	75	100.0%

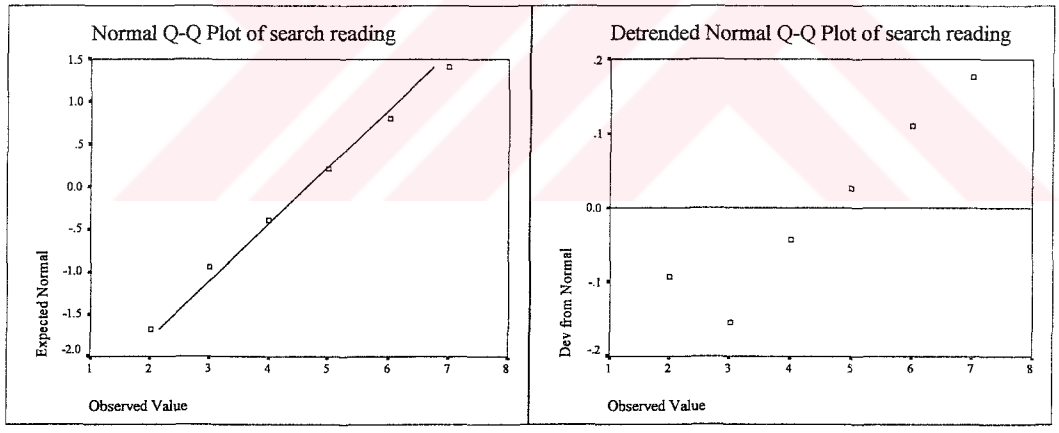
Descriptives

			Statistic	Std. Error
search reading	Mean		4.6400	.1715
	95% Confidence Interval for Mean	Lower Bound	4.2982	
		Upper Bound	4.9818	
	5% Trimmed Mean		4.6556	
	Median		5.0000	
	Variance		2.206	
	Std. Deviation		1.4854	
	Minimum		2.00	
	Maximum		7.00	
	Range		5.00	
	Interquartile Range		3.0000	
	Skewness		-.013	.277
	Kurtosis		-.837	.548

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
search reading	.156	75	.000

a. Lilliefors Significance Correction



Careful Reading

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
careful reading	75	100.0%	0	.0%	75	100.0%

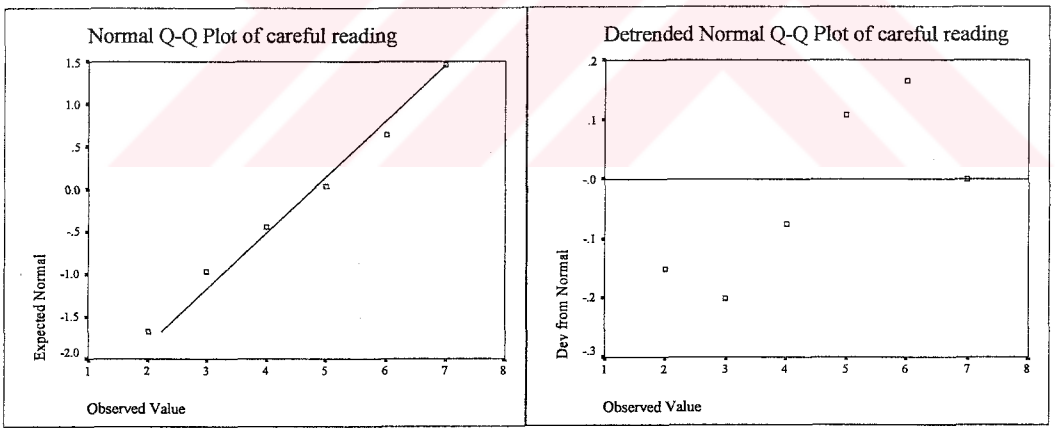
Descriptives

			Statistic	Std. Error
careful reading	Mean		4.7867	.1753
	95% Confidence Interval for Mean	Lower Bound	4.4373	
		Upper Bound	5.1360	
	5% Trimmed Mean		4.8185	
	Median		5.0000	
	Variance		2.305	
	Std. Deviation		1.5183	
	Minimum		2.00	
	Maximum		7.00	
	Range		5.00	
	Interquartile Range		2.0000	
	Skewness		-.246	.277
	Kurtosis		-1.001	.548

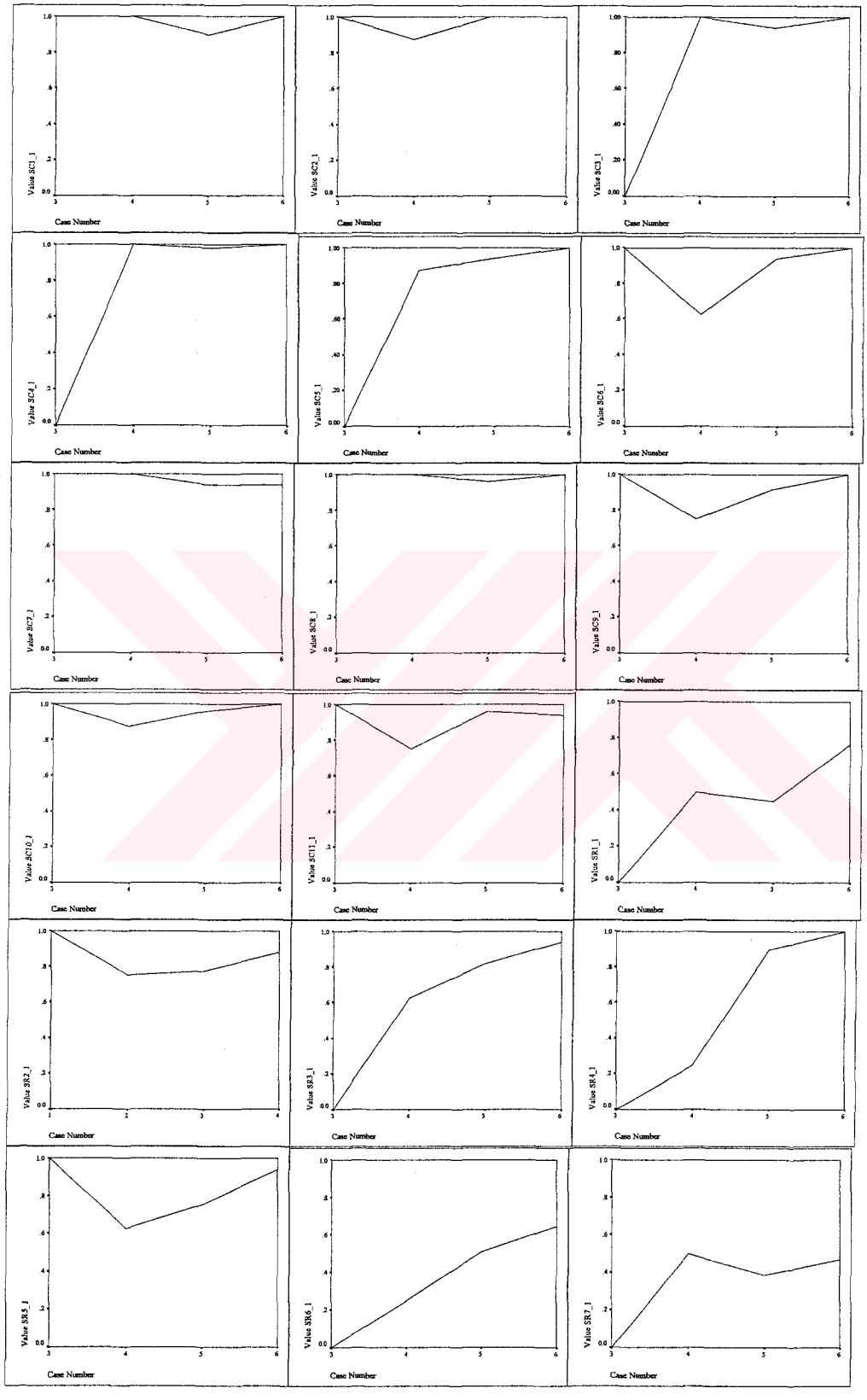
Tests of Normality

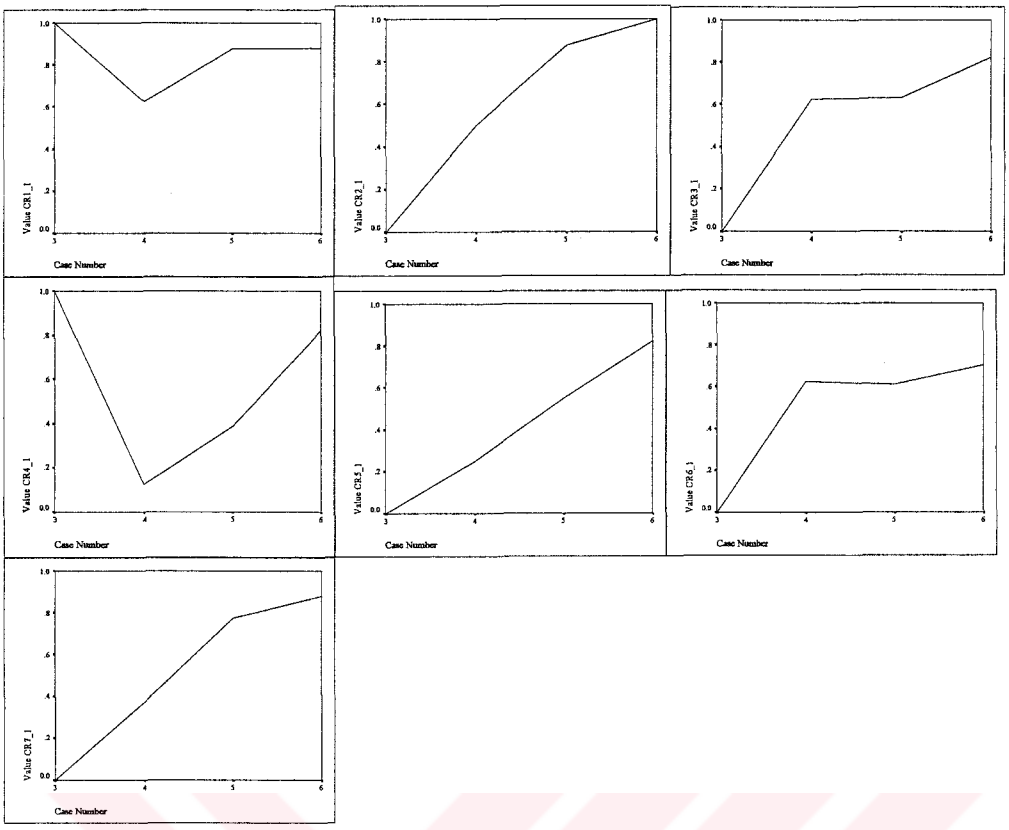
	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
careful reading	.175	75	.000

a. Lilliefors Significance Correction



APPENDIX 4.43
September 2001 – Pilot Version
Band Score Graphs




Value CR1_1
Case Number
Value CR2_1
Case Number
Value CR3_1
Case Number
Value CR4_1
Case Number
Value CR5_1
Case Number
Value CR6_1
Case Number
Value CR7_1
Case Number

APPENDIX 4.44
September 2001 Test
Normality Tests and Graphs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
total reading	719	100.0%	0	.0%	719	100.0%

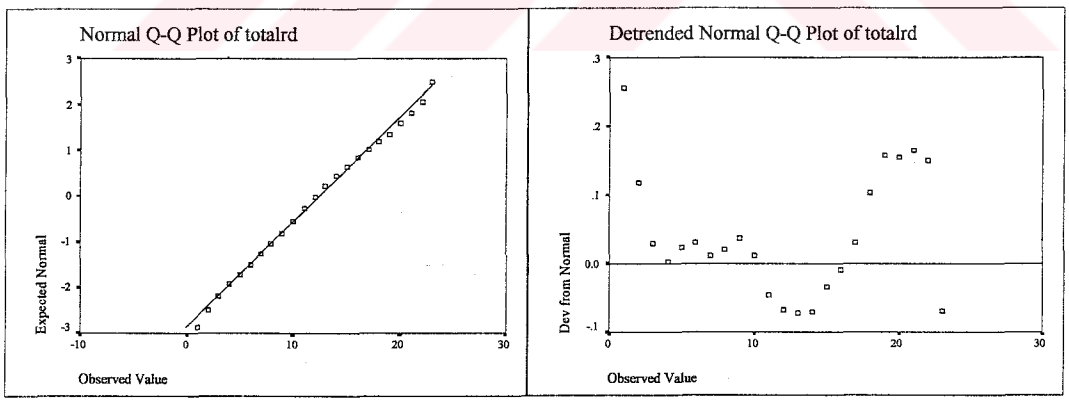
Descriptives

			Statistic	Std. Error
total reading	Mean		12.3936	.1628
	95% Confidence Interval for Mean	Lower Bound	12.0739	
		Upper Bound	12.7133	
	5% Trimmed Mean		12.3632	
	Median		12.0000	
	Variance		19.064	
	Std. Deviation		4.3662	
	Minimum		1.00	
	Maximum		23.00	
	Range		22.00	
	Interquartile Range		5.0000	
	Skewness		.166	.091
	Kurtosis		-.181	.182

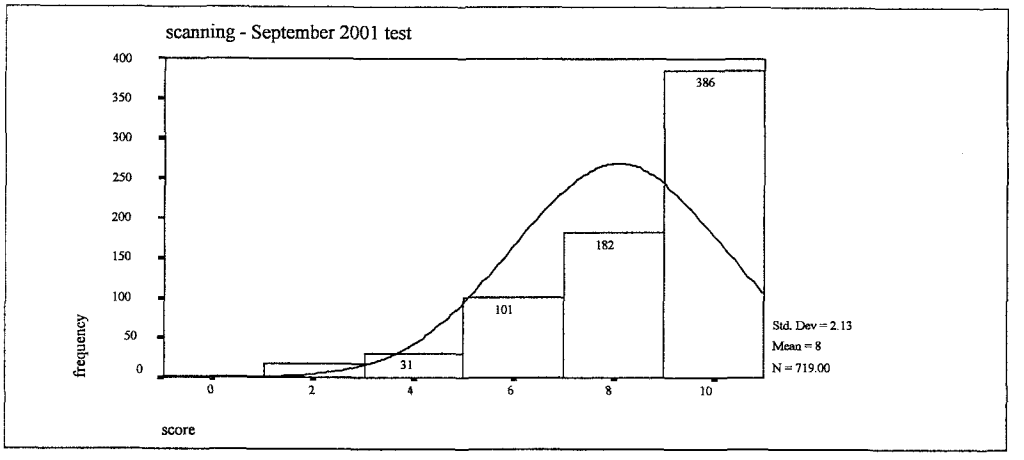
Tests of Normality

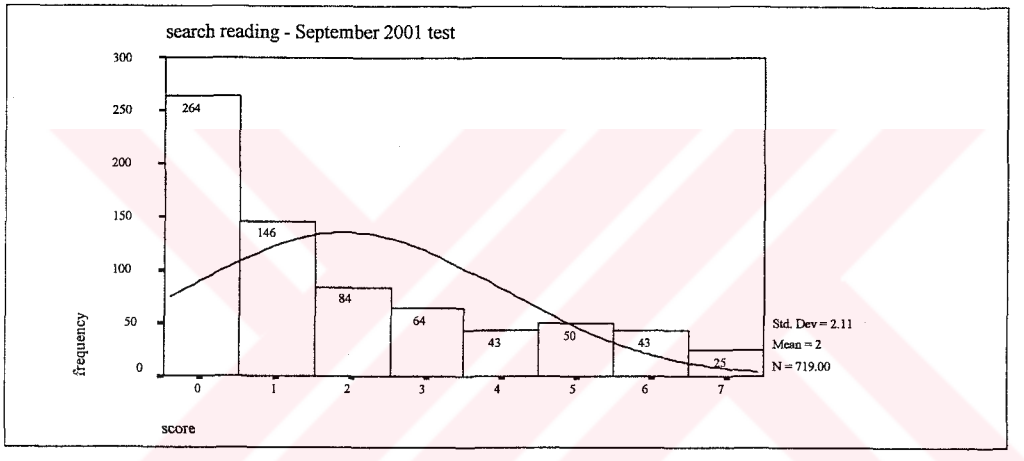
	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
totalrd	.080	719	.000

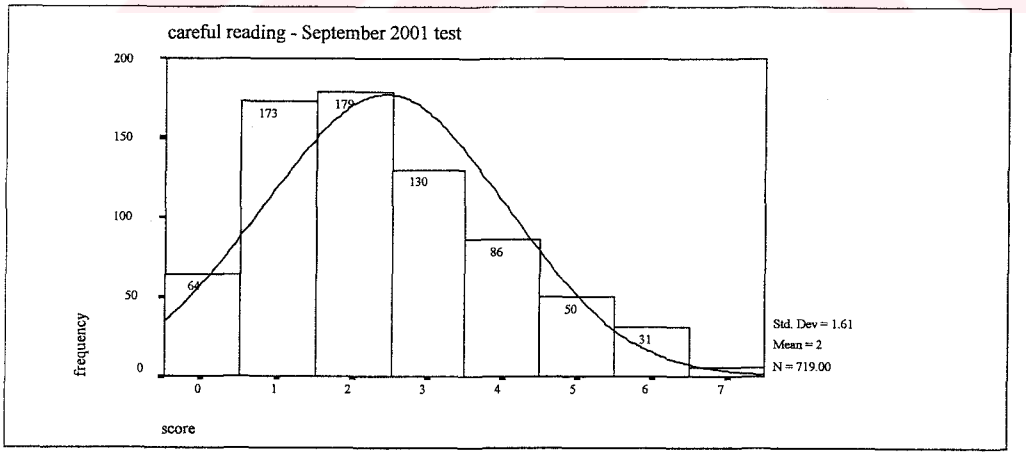
a. Lilliefors Significance Correction



APPENDIX 4.45
September 2001 Test
Score Distribution Graphs by Subtest







APPENDIX 4.46
September 2001 Test
Normality Tests and Graphs by Subtest

Scanning

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
scanning	719	100.0%	0	.0%	719	100.0%

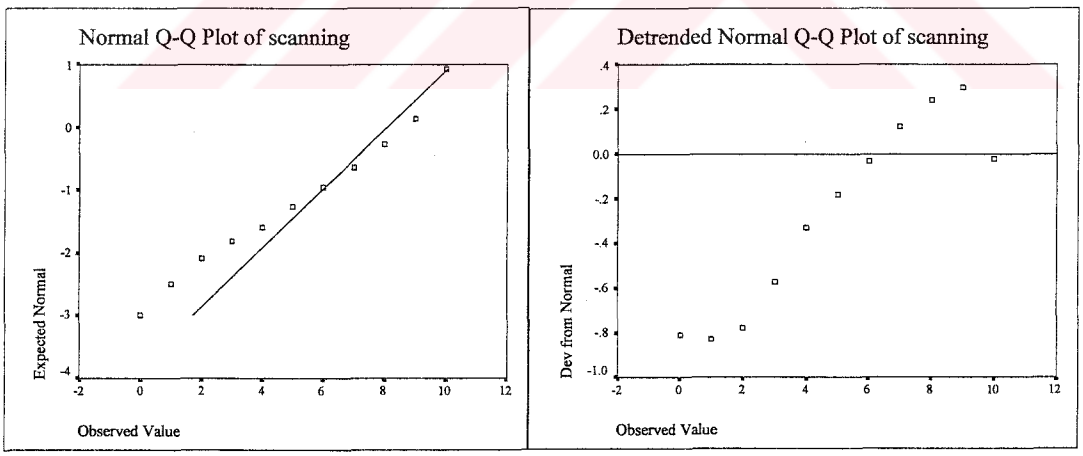
Descriptives

			Statistic	Std. Error
scanning	Mean		8.0765	.079
	95% Confidence Interval for Mean	Lower Bound	7.9209	
		Upper Bound	8.2321	
	5% Trimmed Mean		8.2842	
	Median		9.0000	
	Variance		4.516	
	Std. Deviation		2.1252	
	Minimum		.00	
	Maximum		10.00	
	Range		10.00	
	Interquartile Range		3.0000	
	Skewness		-1.191	.091
	Kurtosis		.937	.182

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
scanning	.205	719	.000

a. Lilliefors Significance Correction



Search Reading

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
search r.	719	100.0%	0	.0%	719	100.0%

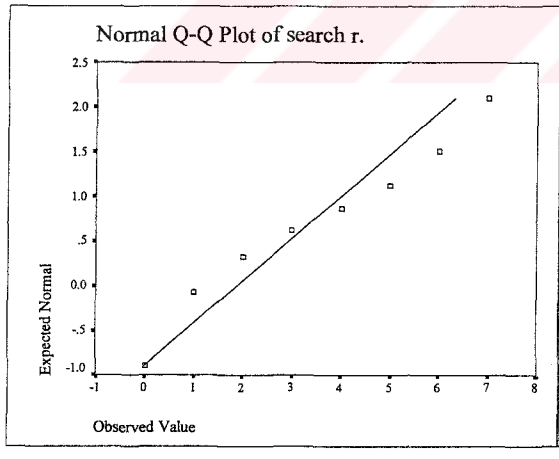
Descriptives

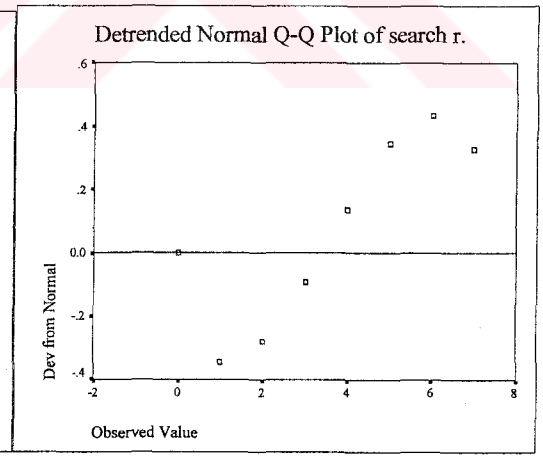
			Statistic	Std. Error
search r.	Mean		1.8929	.0786
	95% Confidence Interval for Mean	Lower Bound	1.7385	
		Upper Bound	2.0473	
	5% Trimmed Mean		1.7313	
	Median		1.0000	
	Variance		4.447	
	Std. Deviation		2.1087	
	Minimum		.00	
	Maximum		7.00	
	Range		7.00	
	Interquartile Range		3.0000	
	Skewness		.954	.091
	Kurtosis		-.290	.182

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
search r.	.234	719	.000

a. Lilliefors Significance Correction





Careful Reading

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
careful r.	719	100.0%	0	.0%	719	100.0%

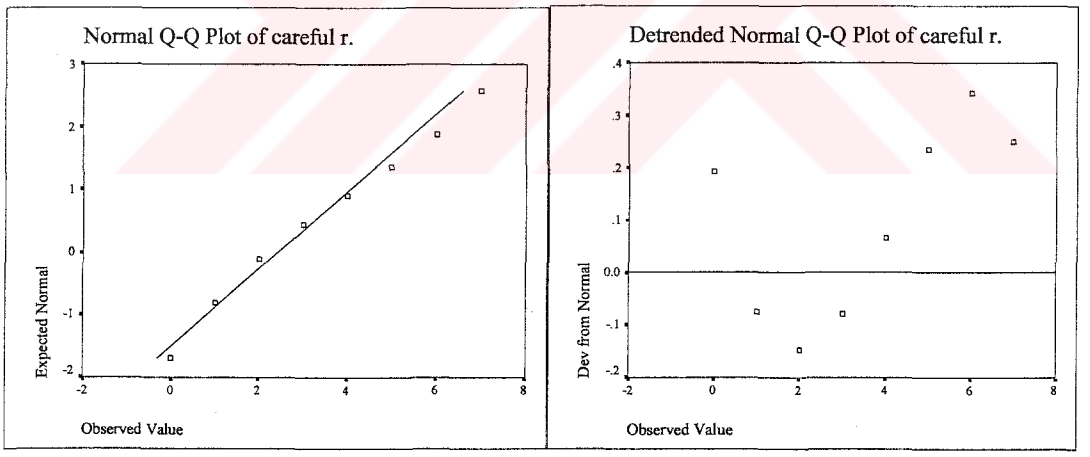
Descriptives

			Statistic	Std. Error
careful r.	Mean		2.4242	.0602
	95% Confidence Interval for Mean	Lower Bound	2.3061	
		Upper Bound	2.5423	
	5% Trimmed Mean		2.3510	
	Median		2.0000	
	Variance		2.604	
	Std. Deviation		1.6137	
	Minimum		.00	
	Maximum		7.00	
	Range		7.00	
	Interquartile Range		2.0000	
	Skewness		.593	.091
	Kurtosis		-.243	.182

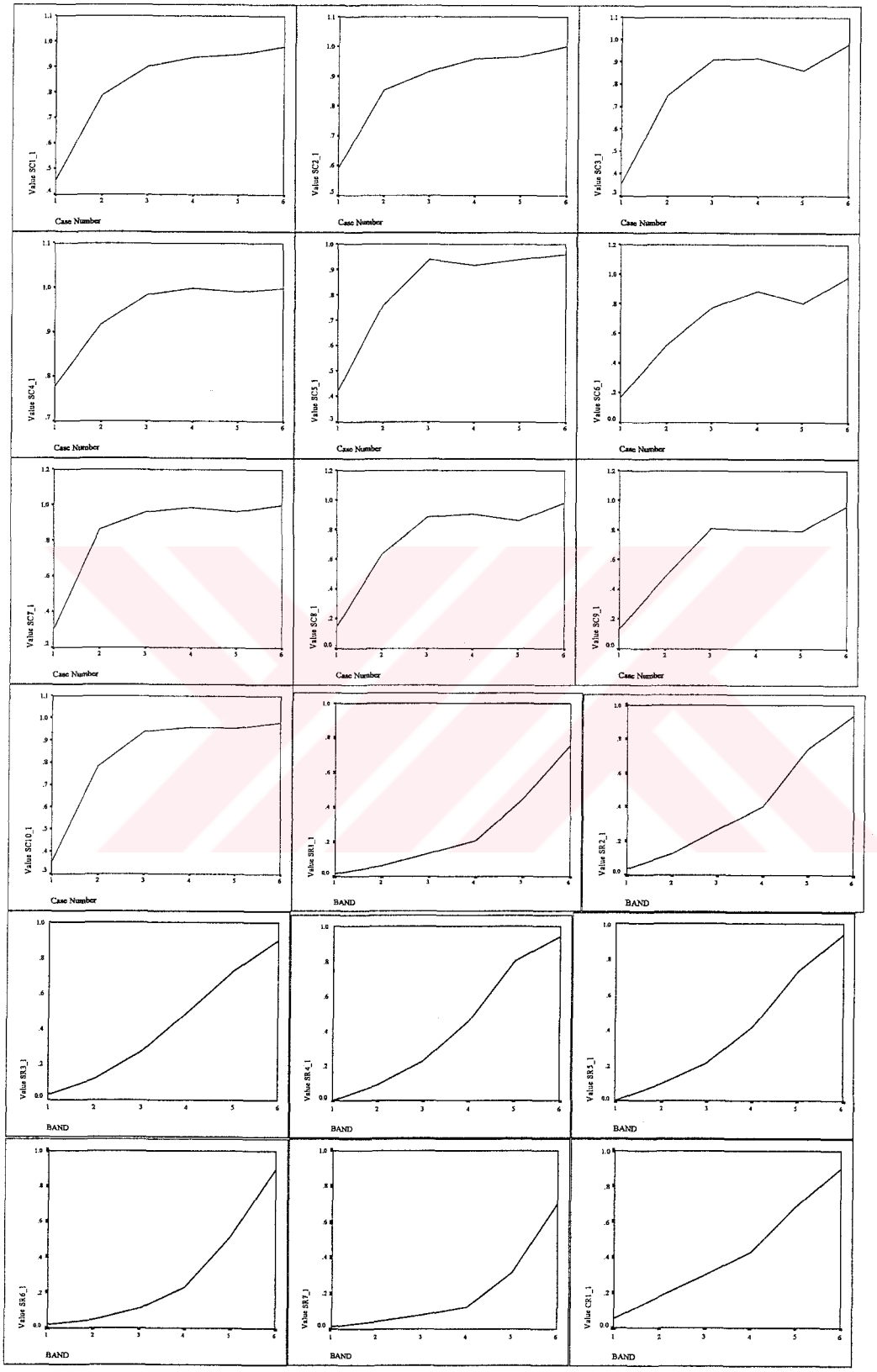
Tests of Normality

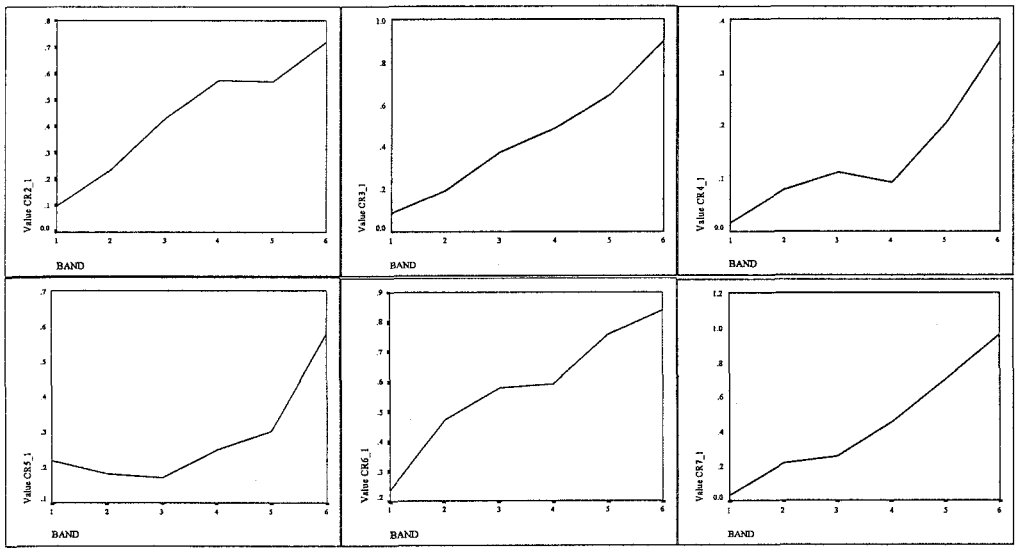
	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
careful r.	.182	719	.000

a. Lilliefors Significance Correction



APPENDIX 4.47
September 2001 Test
Band Score Graphs




Value CR2_1
BAND
Value CR3_1
BAND
Value CR4_1
BAND
Value CR5_1
BAND
Value CR6_1
BAND
Value CR7_1
BAND

APPENDIX 4.48
PCA: September 2001 Test – Whole Set

Communalities

	Initial	Extraction
SC1	1,000	,571
SC2	1,000	,370
SC3	1,000	,392
SC4	1,000	,372
SC5	1,000	,449
SC6	1,000	,409
SC7	1,000	,424
SC8	1,000	,678
SC9	1,000	,541
SC10	1,000	,679
SR1	1,000	,392
SR2	1,000	,497
SR3	1,000	,480
SR4	1,000	,573
SR5	1,000	,530
SR6	1,000	,512
SR7	1,000	,336
CR1	1,000	,391
CR2	1,000	,416
CR3	1,000	,295
CR4	1,000	,340
CR5	1,000	,404
CR6	1,000	,381
CR7	1,000	,358

Extraction Method: PCA

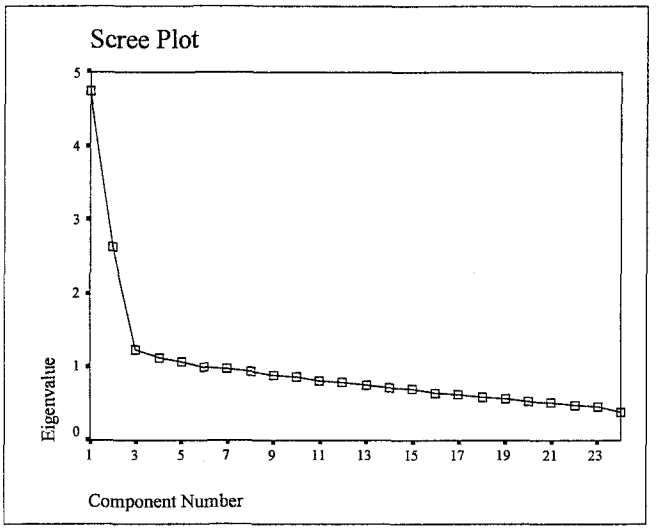
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,870
Bartlett's Test of Sphericity	Approx. Chi-Square	3305,368
	df	276
	Sig.	,000

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4,746	19,774	19,774	4,746	19,774	19,774
2	2,621	10,921	30,695	2,621	10,921	30,695
3	1,238	5,158	35,853	1,238	5,158	35,853
4	1,118	4,657	40,510	1,118	4,657	40,510
5	1,068	4,449	44,960	1,068	4,449	44,960
6	,996	4,150	49,109			
7	,976	4,066	53,175			
8	,931	3,880	57,055			
9	,884	3,683	60,738			
10	,868	3,615	64,353			
11	,812	3,382	67,735			
12	,785	3,271	71,005			
13	,752	3,134	74,140			
14	,719	2,998	77,138			
15	,699	2,914	80,051			
16	,648	2,698	82,749			
17	,617	2,571	85,320			
18	,595	2,480	87,799			
19	,565	2,352	90,152			
20	,536	2,233	92,384			
21	,512	2,132	94,516			
22	,475	1,980	96,496			
23	,450	1,877	98,373			
24	,391	1,627	100,000			

Extraction Method: Principal Component Analysis.



Component Matrix

	Component				
	1	2	3	4	5
SC1	,314	,397	,517	-,196	,095
SC2	,271	,188	,245	,323	,311
SC3	,300	,467	,157	-,178	,165
SC4	,268	,288	,205	,418	-,012
SC5	,362	,358	,425	-,083	-,050
SC6	,405	,370	-,031	,312	,096
SC7	,394	,481	,113	,098	,120
SC8	,422	,523	-,428	-,024	-,207
SC9	,425	,539	-,238	,021	-,111
SC10	,381	,529	-,470	-,039	-,176
SR1	,520	-,332	,043	,004	-,095
SR2	,619	-,302	,021	,059	-,137
SR3	,605	-,282	,061	-,070	-,160
SR4	,671	-,272	,061	-,127	-,174
SR5	,654	-,276	,079	-,029	-,138
SR6	,614	-,304	,109	,046	-,171
SR7	,480	-,207	,047	,209	-,130
CR1	,488	-,221	-,115	-,026	,301
CR2	,322	,023	-,192	-,284	,442
CR3	,462	-,050	-,169	-,209	,083
CR4	,223	-,229	-,178	,046	,452
CR5	,125	-,232	-,178	,437	,335
CR6	,258	-,009	,047	-,506	,238
CR7	,516	-,249	-,121	,081	,091

Component Transformation Matrix

Component	1	2	3	4	5
1	,796	,378	,292	,295	,226
2	-,504	,651	,518	-,132	,191
3	,094	-,604	,707	-,311	,172
4	,033	-,001	-,374	-,369	,850
5	-,319	-,263	,079	,814	,400

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

APPENDIX 4.49
PCA: September 2001 Test – Purged Set

Communalities

	Initial	Extraction
SC6	1.000	.327
SC7	1.000	.358
SC8	1.000	.611
SC9	1.000	.547
SC10	1.000	.588
SR1	1.000	.391
SR2	1.000	.496
SR3	1.000	.469
SR4	1.000	.542
SR5	1.000	.521
SR6	1.000	.502
SR7	1.000	.321
CR1	1.000	.336
CR2	1.000	.412
CR3	1.000	.274
CR4	1.000	.271
CR6	1.000	.373
CR7	1.000	.329

Extraction Method: PCA

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.881
Bartlett's Test of Sphericity	Approx. Chi-Square	2663.676
	df	153
	Sig.	.000

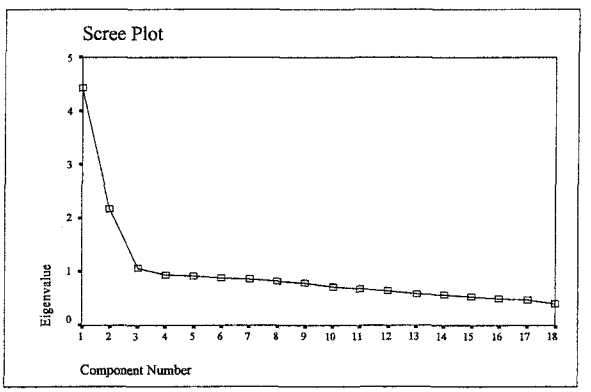
Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.422	24.569	24.569	4.422	24.569	24.569
2	2.174	12.076	36.645	2.174	12.076	36.645
3	1.072	5.958	42.602	1.072	5.958	42.602
4	.944	5.245	47.848			
5	.924	5.133	52.981			
6	.888	4.932	57.913			
7	.863	4.795	62.708			
8	.827	4.592	67.300			
9	.786	4.365	71.665			
10	.714	3.968	75.633			
11	.682	3.791	79.425			
12	.635	3.528	82.953			
13	.596	3.314	86.266			
14	.567	3.153	89.419			
15	.531	2.951	92.370			
16	.500	2.775	95.145			
17	.469	2.606	97.752			
18	.405	2.248	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix

	Component		
	1	2	3
SC6	.344	.439	-.127
SC7	.311	.496	-.125
SC8	.367	.690	.012
SC9	.351	.645	-.090
SC10	.324	.694	.044
SR1	.561	-.257	-.098
SR2	.660	-.199	-.142
SR3	.643	-.199	-.130
SR4	.707	-.187	-.087
SR5	.689	-.187	-.109
SR6	.649	-.228	-.169
SR7	.504	-.126	-.224
CR1	.518	-.121	.229
CR2	.319	.091	.550
CR3	.476	.050	.212
CR4	.257	-.181	.415
CR6	.255	.013	.555
CR7	.550	-.145	.079



Component Transformation Matrix

Component	1	2	3
1	.861	.380	.339
2	-.396	.918	-.023
3	-.320	-.115	.941

Extraction Method: PCA Rotation Method:
Varimax with Kaiser Normalization.

APPENDIX 4.50
PCA: September 2001 Test – Individual Subtests

Component structure for the scanning subtest is rotated with varimax rotation. For search and careful reading data, rotation was not used for the obvious reason that search reading was unidimensional and the component structure of careful reading was bi-componential with clear distribution.

Scanning

Communalities

	Initial	Extraction
SC1	1.000	.619
SC2	1.000	.501
SC3	1.000	.560
SC4	1.000	.460
SC5	1.000	.409
SC6	1.000	.418
SC7	1.000	.444
SC8	1.000	.683
SC9	1.000	.551
SC10	1.000	.691

Extraction Method: PCA

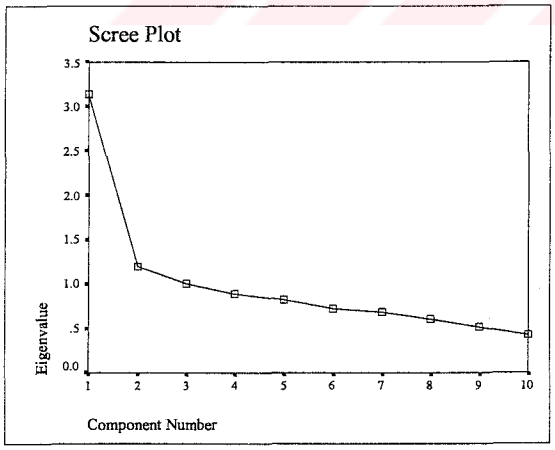
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.811
Bartlett's Test of Sphericity	Approx. Chi-Square	1219.676
	df	45
	Sig.	.000

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.134	31.338	31.338	3.134	31.338	31.338
2	1.203	12.029	43.367	1.203	12.029	43.367
3	1.001	10.010	53.377	1.001	10.010	53.377
4	.889	8.894	62.271			
5	.819	8.191	70.462			
6	.725	7.248	77.710			
7	.679	6.789	84.499			
8	.606	6.064	90.563			
9	.515	5.148	95.711			
10	.429	4.289	100.000			

Extraction Method: Principal Component Analysis.



Component Matrix

	Component		
	1	2	3
SC1	.503	.487	-.358
SC2	.337	.368	.503
SC3	.544	.180	-.482
SC4	.411	.210	.497
SC5	.507	.389	-.010
SC6	.558	.022	.326
SC7	.631	.136	-.165
SC8	.667	-.484	.065
SC9	.687	-.276	-.060
SC10	.646	-.522	-.011

Extraction Method: PCA

Component Transformation Matrix

Component	1	2	3
1	.694	.582	.424
2	-.720	.554	.419
3	.009	-.596	.803

Extraction Method: PCA Rotation Method Varimax with Kaiser Normalization.

Search Reading

Communalities

	Initial	Extraction
SR1	1.000	.398
SR2	1.000	.491
SR3	1.000	.493
SR4	1.000	.579
SR5	1.000	.542
SR6	1.000	.511
SR7	1.000	.299

Extraction Method: PCA

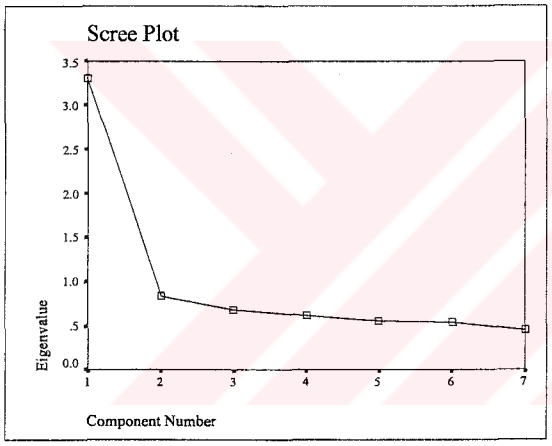
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampli Adequacy.		.876
Bartlett's Test of Sphericity	Approx. Chi-Square	313.865
	df	21
	Sig.	.000

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.312	47.312	47.312	3.312	47.312	47.312
2	.838	11.977	59.290			
3	.687	9.819	69.109			
4	.614	8.772	77.881			
5	.559	7.990	85.871			
6	.536	7.664	93.535			
7	.453	6.465	100.000			

Extraction Method: Principal Component Analysis.



Careful Reading

Communalities

	Initial	Extraction
CR1	1.000	.468
CR2	1.000	.281
CR3	1.000	.386
CR4	1.000	.233
CR5	1.000	.689
CR6	1.000	.386
CR7	1.000	.414

Extraction Method: PCA

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampli Adequacy.		.692
Bartlett's Test of Sphericity	Approx. Chi-Square	241.320
	df	21
	Sig.	.000

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.820	26.003	26.003	1.820	26.003	26.003
2	1.036	14.799	40.802	1.036	14.799	40.802
3	.924	13.200	54.002			
4	.898	12.828	66.829			
5	.870	12.431	79.260			
6	.779	11.127	90.387			
7	.673	9.613	100.000			

Extraction Method: Principal Component Analysis.