THE UNIVERSITY of EDINBURGH

Title	Design, development and validation of a specific purpose test of English for Taiwanese tour guides
Author	Kao, Chung-yao
Qualification	PhD
Year	2001

Digitisation Notes:

- Appendix 1 is omitted, with the Appendices beginning at Appendix 2

The design, development and validation of a specific purpose test of English for Taiwanese tour guides

Chung-yao Kao

Ph.D.

University of Edinburgh

2001

Declaration

I hereby declare that this thesis has been composed by myself and is entirely my own work.

Chung-yao Kao

Abstract

The aim of this thesis is to design, develop and validate a specific purpose test of English for Taiwanese tour guides (TG Test) on a trial basis. The existing TG test is a general proficiency test, and has a number of drawbacks with reference to the measurement of language use ability as a tour guide at work. The reasons for the use of a specific purpose test is that (1) performance varies according to the language use contexts, (2) a specific purpose test is clearer about the language use domain of interest, i.e., how Taiwanese use English as tour guides in their work, and (3) test performance is interpreted from the perspective of the test user.

Central to testing language for specific purposes (LSP) is the notion that test content and test methods are derived from an analysis of a specific language use domain. Task authenticity, directness in test method, and criterion-referencing in terms of interpreting test performance are related to LSP test design and development. Analysis of language use contexts and the workplace is a critical feature in LSP test design. Collaboration with field experts is important in LSP test development.

Test validation concerns the process of gathering evidence in support of the claim that the test measures the abilities it proposes for a given test purpose. Validity is viewed as a unitary concept (Messick, 1989). Validity inquiry includes collecting evidence of test interpretation, as well as examining the consequences of the test interpretation and test use. Messick's "facets of validity inquiry" will be the framework for the TG test validation. Preliminary results showed that the new TG test was empirically valid and broadly acceptable to the test users.

In this thesis, discussion of issues of the theory and practice of LSP testing is presented in Chapter 1. In Chapter 2, the research questions are presented, and procedures for the design, development and validation of the TG test are outlined. Chapter 3 presents the discussion of what is involved in the three language use dimensions of assessment, i.e., listening ability, speaking ability and grammatical competence. Test specifications and the tests are presented. Development of the rating scale, rater training and criteria to ensure the usefulness of the TG test are also contained in this chapter. In Chapter 4, test administrations and test results with the use of classical test analysis and Rasch analysis are presented and discussed. Chapter 5 is focused on the discussion of evidence collected to support the claims of the TG test. The thesis concludes with a discussion of areas needing to be improved. Future directions for the TG test as well as LSP testing in the Taiwanese foreign language use contexts are also considered.

Acknowledgements

My gratitude goes to my supervisors Mr B. Parkinson and Dr. G. Ferguson, without whose help and advice the completion of this thesis would not have been possible. I would also like to acknowledge the help of the following organisations: the Tourism Bureau for allowing me to attend the three-week tour guide training course and for providing me with all the training materials; the Tour Guide Association for their advice; and the Language Training & Testing Center for allowing me to develop and pilot the TG test of my own. I am indebted to the following universities and institutions for kindly arranging for me to administer the test: Chung Hsing University, Tam Kang University, Shi Jian University, Xin Wu College and the Language Training & Testing Center. I am grateful to the many teachers and students who participated in the TG test administration. I wish to thank Ms S. Chu, Ms Hsieh, Ms B. Yen, Ms L. Ludwig, Ms D. Young, Mr. G. L. Guo, and Mr W. Calabretta for their kind help in rating the test tapes and their comments on the final draft of the TG test battery. My gratitude also goes to Prof. Z. L. Huang, Dr. Y. H. Chen, Dr. H. L. Wang, Dr. S. F. Lai, Ms. H. J. Lin, and Mr. C. L. Kao for their perceptive comments on the final draft of the test. I owe particular gratitude to Mr G. L. Guo, who assisted me in data collection, and provided me with valuable comments and advice on the test items. I would like to thank my examiners, Professors Alan Davies and Charles Alderson, for their comprehensive comments on this thesis. Finally, I thank my parents for their unending support and encouragement throughout my education and beyond.

Abbreviations

ACTFL: American Council on the Teaching of Foreign Languages
CEELT: Cambridge Examination in English for Language Teachers
CLA: Communicative Language Ability
CNP: Communication needs processor
EAP: English for Academic Purposes
EFL: English as a Foreign Language
ELTS: English Language Testing Service
EOP: English for Occupational Purposes
ESP: English for Specific Purposes
FLPT: Foreign Language Proficiency Tests
FSI: Foreign Service Institute
IELTS: International English Language Testing System
ILTA: International Language Testing Association
IRT: Item response theory
JCEE: Joint College Entrance Examination
LSP: Languages for Specific Purposes
LSTE: Listening Summary Translation Exam
LTTC: The Language Training & Testing Center
MCQ: Multiple-choice question
OET: Occupational English Test
OPI: Oral Proficiency Interview
SD: Standard deviation
SE: Standard error
SEM: Standard error of measurement
TEEP: Test of English for Educational Purposes
TG Test: Tour Guide Test
TOEFL: Test of English as a Foreign Language
TOEIC: Test of English for International Communication
TSE: Test of Spoken English
TWE: Test of Written English
UCH: Unitary Competence Hypothesis

Contents

Chapter 0: General overview

0.0. Introduction

The purpose of this research was (1) to design and develop a specific-purpose test of English for the selection of Taiwanese tour guides (TG English Test), and (2) to validate the test.

The driving force behind the research was a concern that the measurement of the language ability of a would-be tour guide should be based on a theory of language use. The present TG Test is a general proficiency test and takes the structuralist/psychometric approach, which does not seem to adequately reflect the characteristics of language use and the target language use domain of tour guiding. A specific purpose approach, on the other hand, seems more appropriate and will make inferences of test performance more meaningful in relation to the job demands in the workplace.

Based on the notion of communicative competence, a test battery consisting of Listening, Grammar and Speaking tests was developed and administered to 112 university students as a main trial. Classical test analysis and Rasch analysis were carried out for item analysis. Test validation followed Messick's (1989) framework, and the validation process started at the beginning of test design. In this thesis, the justifications for the design of an ESP TG test are discussed, the stages of test development reported, and the different sources of evidence in support of the test presented. Results of item analysis and information gathered on the validity and usefulness of the test indicated the following:

1. The test battery was broadly acceptable to the test takers and test users.
2. Statistical evidence on the performance of the sub-tests and the test tasks seemed to support the constructs as defined and realised in the tests. However, rater statistics showed the existence of two types of rater behaviour: conservative in the sense that they used a narrow range of marks and extreme

in the sense that they used the extreme ends of the scoring scale.

3. Task types affected item performance and the candidate's test performance.
4. Gender difference on test performance was not statistically significant.

The implication of the results suggests that specific purpose language tests can be an alternative to the general proficiency test in the measurement of the tour guide's ability to use language in the context of tour guiding. By extension, this type of test can be piloted in other specific language use contexts such as a medical professional's ability to interact satisfactorily with foreign patients or the measurement of a military officer's ability to communicate and relay messages.

0.1. The structure of the thesis

This thesis contains seven chapters. The present chapter gives an overview of each of the following chapters. Chapter 1 discusses the notions of communicative language use ability in relation to second language specific-purpose testing. Central concepts in language testing and a brief discussion of the two approaches to item analysis (i.e., classical test theory and Rasch analysis) are also presented. Further, stages relevant in the design and development of a specific-purpose test are also outlined. Chapter 2 discusses problems related to the present TG test in the measurement of a would-be tour guide's language use ability. Research questions and procedures to investigate these questions are discussed. Chapter 3 provides the rationales for the assessment of language use ability in terms of the three language use dimensions: listening, speaking and grammatical competence. A working definition for each of the three has been provided. Criteria ensuring the usefulness of the test are offered and discussed. Chapter 4 reports on test results of the main trial; suggestions for future test revision and improvement are also made. Evidence in support of the validity and the usefulness of the test battery is discussed in Chapter 5. The thesis concludes by considering the implications for future test development and the use of LSP testing in the Taiwanese foreign language use context.

Chapter 1: Language use ability and language testing

In this chapter, I will (1) discuss notions of performance and competence in relation to language use and to second language specific purposes testing (LSP), (2) highlight concepts central to LSP testing, (3) present an overview of the stages in LSP test design and development, which will serve as the guideline in the design and development of the Tour Guide (TG) English Test for Taiwanese tour guides, and (4) briefly discuss the rationales behind two approaches to test analysis: the classical test analysis and Rasch analysis.

1.1. On the meaning of communicative language use

The distinction between competence and performance in language use and language testing has been long debated among language testers. Furthermore, the term *competence* has been used widely and divergently in many different contexts; it does not have any precise meaning any more. In the following section, I will examine the different ways in which competence and performance have been defined, beginning with Chomsky's original concept and ending with the Bachman model of communicative language ability (CLA). Then, I will discuss what role performance plays in relation to the different theoretical models and in language testing. I will not be able to solve the competence/performance debate but I will try to explain, synthesise and offer a perspective on the extent of the distinction that has influenced the language testing discipline in the last four decades.

1.1.1. Chomsky

Chomsky's view of knowing a language is reflected in his distinction between the speaker-hearer's knowledge of the language (i.e., competence) and his/her actual use of language in concrete situations (i.e., performance).

> *Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech community, who knows its language perfectly*

> *and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance.*
> Chomsky, 1965:3

The perfect knowledge refers to the mastery of the abstract system of rules an idealised native speaker has in order to understand and produce well-formed sentences in his/her language. The actual use of the language unaffected by what Chomsky terms "the grammatically irrelevant conditions" is in the domain of linguistic performance. For Chomsky, performance is not the realisation of competence; rather, linguistic competence provides "the basis for actual use of language by a speaker-hearer" (Chomsky, 1965).

Chomsky is concerned with idealisation and not with matters related to language use. Also Chomsky excludes the notion that competence suggests ability (Chomsky, 1980). He draws a distinction between knowing the language, the ability to use the language that one knows and actually using it (Brown, 1996).

In short, Chomsky draws the distinction between competence and performance but he is only interested in the perfect linguistic knowledge an ideal native speaker possesses. Chomsky differs from other scholars in two respects. First, he believes that the characteristic feature of human language is not the ability to communicate but the complex syntax in any human language. Second, Chomsky sees the primary function of language as the vehicle of cognitive growth that differentiates human beings from other species.

1.1.2. Hymes

Hymes broadens the concept of competence. For Hymes, Chomsky's conception of competence is far too narrow; the notion of competence and performance does not account for the fact that the one thing we know about language is how to use it appropriately. Thus Hymes differentiates between linguistic and communicative competence and linguistic and communicative performance:

1. *(underlying) competence v. (actual) performance:*
2. *(underlying) grammatical competence v. (underlying) models/rules of performance.* (Hymes, 1972:280)

In his model of language competence, Hymes includes a new type of component: the ability/competence for use, referring to an individual's potential to realise a *possible*, *feasible* and *appropriate* speech act. In Hymes' words, competence is dependent on both "(tacit) knowledge and (ability for) use". Knowledge is a straightforward concept referring to grammatical and sociolinguistic rules. It could be equated with a mental state, associated with a particular cognitive structure. Such a state is distinguished from the capacity/ability to do something with the knowledge people possess. Ability for use refers to a process rather than the mental state (Taylor, 1988). To better understand the concept, McNamara (1996) suggests language testers consider a range of cognitive, affective and volitive factors involved in language use situations. However, these factors have not been properly addressed.

Hymes' contribution is influential. He clarifies the domain of performance and isolates some systematic features of rules governing language use. But, in a subtle way, Hymes has extended Chomsky's notion of competence as "tacit knowledge of grammar" to a knowledge that reflects aspects of communication, i.e., possibility, feasibility, appropriateness and whether people actually use particular utterances. Competence now conveys the notion of ability as well as a social dimension. To complicate the matter further, Hymes uses the term "differential competence" (1971:7), referring to differences among individuals and introducing a relative dimension; thus the term has moved away from what for Chomsky, is an absolute notion, a property of the individual, not allowing any meaningful comparison. Hymes is suggesting that different people have different competencies and there is a social dimension to language ability as well as use.

More recent models of communicative competence are all adaptations of Hymes' model. I will briefly discuss two such frameworks of communicative competence that investigate performance in the second language: the work of Canale and Swain (Section 1.1.3) and

that of Bachman and Bachman & Palmer (Section 1.1.6).

1.1.3. Canale and Swain

Canale and Swain's model is related to Hymes' notion of communicative competence. They maintain that in addition to grammatical knowledge, the domain of language knowledge should include sociolinguistic competence and strategic competence (Canale and Swain, 1980). Discourse competence was subsequently added (Canale, 1983). Grammatical competence consists of knowledge of lexical items and of rules of morphology, syntax, sentence-grammar semantics and phonology. Sociolinguistic competence refers to knowledge of sociocultural rules of use. Strategic competence refers to coping strategies in case of communication breakdowns. Discourse competence is related to the mastery of combination of grammar and meaning to achieve a coherent and cohesive text.

Unlike Hymes, Canale and Swain exclude the ability for use from communicative competence. They place ability for use in the domain of communicative performance defined as "the realisation of the different competencies and their interaction in the actual production and comprehension of utterances." They subsequently define ability for use as "the actual demonstration of this knowledge in *real* second language situations and for *authentic* communication purposes" (Canale and Swain, 1980:6). They argue that there is no theory that adequately explains the ability for use; its inclusion would imply "communicative deficits", that is, inadequate communicative competence.

Canale and Swain's model is essentially behaviour-based as their definition of communicative performance refers only to the actual use of language (McNamara, 1996) and some criticisms of this have been voiced. First, Canale and Swain exclude ability for use from communicative competence, a concept with a focus on the relationship and interaction between "regularities in grammatical competence and regularities in sociolinguistic competence" (p.8). This notion fails to account for an individual's

potential to use language to fulfil certain social functions. Second, Canale and Swain do not explain how the components interact with one another to bring about performance. Lyons (1996) suggests that acquisition of competence is partly dependent upon performance. Spolsky (1989) also notes that he would prefer a model of performance that includes some form of knowledge since performance models presuppose competence models. Canale (1983) later distinguishes the actual communicative performance from the knowledge underlying it and thus aligns himself more with Hymes' position.

Finally, the model is based on theoretical not empirical work (Cziko, 1984). Bachman and Palmer (1982) tested empirically a hypothesised framework of language competence; results indicated one general and two specific trait factors - grammatical/pragmatic competence and sociolinguistic competence. Harley et al. (1990) investigated whether grammatical, discourse and sociolinguistic competence could be distinguished empirically, and found that factor analysis failed to confirm their hypothesis of a three-trait structure of language proficiency.

Despite the criticisms, the Canale and Swain model had a significant impact in language testing in that it lends support to communicative tests after a long period of linguistic-psychometric testing and a short period of general proficiency, unitary testing. Further, the framework includes sub-components of communicative competence and broadens the scope of language testing.

1.1.4. The Proficiency Movement

The requirement of performance is the key feature of proficiency-based testing such as the ACTFL Oral Proficiency Interview (OPI). This type of testing has its origin in an approach to testing foreign language proficiency in some US government agencies such as the Foreign Service Institute (FSI) and CIA (Spolsky, 1995). The FSI type of oral interview and evaluation scale was later adapted by ACTFL/ETS in the 80s to include receptive and productive dimensions of language use. The aim of such a direct and

behavioural approach to testing is to measure the learner's ability to communicate in a foreign language.

The interview type of language testing views language proficiency as "experientially" based on typical language use required for communication as judged by experts (Omaggio, 1983). The three interrelated criteria for judging language proficiency are function, context and accuracy. Grammaticality of utterances is the crucial factor in determining the proficiency level. The theoretical question of what it means to know a language is implicitly assumed by the test tasks and the rating scale, which become the de facto theory of language use ability. Shohamy (1996) sums up the procedures of test development using tasks and the rating scale:

- *a purpose and a context for the test are defined*
- *a sample performance that represents the purpose and context of the test is identified*
- *a task which elicits the performance is constructed*
- *the task is performed by a test taker in a simulated situation*
- *a language sample is obtained*
- *the language sample is assessed by means of a rating scale which provides a description of what it means to know a language.*

Shohamy, 1996: 145

The implication of the proficiency-based approach is that the functions and purposes of a language have been predetermined. The functional definitions are then organised in a hierarchical scale that ascribes to the candidate a particular degree of language proficiency. An operational definition of knowing a language is thus provided.

With the popularity of the oral interview technique, a number of notions have become popular: test tasks, directness, authenticity, rating scales, performance and proficiency. These notions in testing coincide with some of the notions put forward in the communicative language teaching framework and a demand of second language proficiency in the workplace and in educational contexts in the 1980s. Due to a growing awareness that language is communication, the test taker is required to demonstrate his/her communicative skills in the foreign language. Test items no longer focus on linguistic features but on activities that reflect real-life communicative demands. An

indirect measure of the candidate's linguistic proficiency is not considered a sufficient means to measure his/her language use ability because such an approach fails to recognise the full context of language use. People use language for communicative purposes; in a testing situation, the test taker will be required to deal with test material that corresponds to his/her "normal communicative activities" (Widdowson, 1978). Test performance becomes one assessment criterion and is interpreted on the basis of a rating table of several levels in which degrees of language proficiency are described.

It seems that the distinction between competence and performance has become operational. Instead of struggling with complicated models with many components and variables, performance and the rating scale have become the language that everyone can understand. In such a way of assessing language proficiency, the observable is acceptable but the covert mental processes are disregarded.

The "Proficiency Movement" has had a number of criticisms. First, its construct validity has been severely questioned by some researchers (Savignon, 1985; Bachman, 1988; Lantolf & Frawley, 1988; Raffaldini, 1988; Messick, 1994; Shohamy, 1996). The concern is specifically about the fact that the ACTFL Guidelines provide no empirical evidence on how the scale description is calibrated (Salaberry, 2000).

Second, the Guidelines use generic educated native speaker competence as a reference point to measure the proficiency of non-native speakers (McNamara, 1996). However, empirical evidence shows variation among native speakers' performance depending on their professional training and educational backgrounds (Oller & Conrad, 1971; Lopes, 1992; McNamara, 1996).

Third, the proficiency guidelines confound language ability with test methods in test design (Bachman, 1988). Bachman notes that in a performance test such as the oral interview, the modality (i.e., productive) and channel (aural and oral) of the ability (speaking) match the modality and channel of the test method. Therefore, it is difficult to

clearly distinguish ability from test method. In other words, it is difficult to tell whether a particular rating can be interpreted as an indication of the language ability in question or an indicator of an individual ability to perform well under a particular condition. In two studies (Bachman & Palmer 1981& 1982), the researchers found that the factor loading for the oral interview test method was consistently higher than factors associated with the traits being measured.

A further criticism is the claim that oral proficiency interview does not evaluate important aspects of the learner's communicative ability such as sociolinguistic and discourse competencies (Raffaldini, 1988). Compared to van Ek's Threshold Level descriptions of the communicative ability, the majority of functions described in the ACTFL Guidelines only refer to the exchange of factual information and intellectual attitudes. Other types of communicative functions such as socialising, suasion, functions expressing emotional or moral attitudes are either ignored or mentioned here and there at various levels in the guidelines (Raffaldini, 1988). Because the range of functions presented is limited, the Guidelines can only provide limited information on the sociolinguistic and discourse competencies of the learner.

The final criticism is of proficiency as defined by the Guidelines (Lantolf& Frawley, 1988). The proficiency of the learner has been defined according to the functions and contexts specified in the Guidelines. The construct of language ability is dependent on the Guidelines. For example, if a learner can order a meal, s/he may be a '1', if s/he can conduct a casual conversation on familiar topics, s/he may be a '2'. But language ability cannot be assessed solely on information questioning as noted in the Guidelines; rather, it involves creating meaning. Proficiency of the language learner means s/he actively creates and transforms a situation/world through linguistic means to another person/people. It involves interaction. The performance of the individual and what we can infer from it is the central focus, not the functions and purposes of the language specified in the Guidelines.

Included in the list of criticisms are concerns that learners are rated according to the Guidelines in which artificial contexts are promoted but the interactional skills required in real world communication are limited. The roles played by the examiners are highly limited as well. A further concern is the notion of *circular logic* pointed out by Bachman and Savignon (1986) in that the Guidelines are confounded with their test instrument, the OPI. OPI delineates the Guidelines and it in turn represents what is tested in the OPI. To avoid this circular logic, the Guidelines have to be verified by other measurement instruments such as the use of self-assessment (Chalhoub-Deville, 1997).

1.1.5. The Unitary Competence Hypothesis (UCH)

In the late 1970s, John Oller (1979, 1980) advanced his notion of language as a unitary factor rather than a divisible construct. Oller's UCH has both theoretical and empirical justifications. His theoretical justification derives from his postulation of an internalised grammar of expectancy, a capacity which underlies all language performance and enables us to draw on our linguistic knowledge and world knowledge to generate expectancies about the language we are processing (Oller, 1979). The empirical evidence comes from factor analysis studies which suggest the existence of a general factor that underlies all test batteries. In other words, there is empirical evidence for the existence of a general ability with language.

According to Oller, the grammar of expectancy is activated in circumstances which require the learner to process the language under normal contextual constraints. The language ability is demonstrated through pragmatic tests such as cloze and dictation which meet the two naturalness criteria; first, they require the learner to use normal knowledge of contextual constraints on language sequences; second, they require comprehension of the meaningful sequences of the language elements in relation to extralinguistic contexts.

Oller was not specific about whether his hypothesis was a model of competence or

performance (Shohamy, 1996). However, Hymes' concept of ability for use, defined as the rules of performance, seems to be reflected in UCH as Oller writes that his object of interest is "language as it is used for communicative purposes" and that his notion of expectancy is introduced as "a key to understanding the nature of psychologically real processes that underlie language use." McNamara (1996) notes that Oller's notion of *pragmatic naturalness criteria* represents the first major attempt to propose a model of performance in a language testing context. Two aspects of the pragmatic naturalness criteria relate to performance:

> 1. *It must require the processing of temporal sequences of elements in the language constrained by the normal meaningful relationships of such elements in discourse.*
> 2. *It must require the performer of the task to relate the sequences of elements to extralinguistic context via... pragmatic mappings.*
>
> (Oller, 1979: 263)

The first requirement is for naturalness and real-time processing. It relates to language use and an inexplicit psycholinguistic processing mode. The second requirement is about the integration of linguistic and extralinguistic knowledge, that is, the real-world knowledge which points to different aspects of cognitive organisations (McNamara, 1996). In language testing contexts, highly contextualised language tasks that replicate real-life situations conform to the constraints proposed above. They require the test taker to process meaning through contextualised language in real time. Thus performance tests could be referred to as pragmatic tests (Wesche, 1985).

The UCH has received considerable attention, much of it critical not of the sociolinguistic or psycholinguistic aspects of the model but of the notion of a unitary factor, which is not well supported statistically (Vollmer & Sang, 1983). Bachman and Palmer (1982) demonstrated the existence of a higher general factor plus two trait factors which they called grammatical and sociolinguistic competence. Later Oller retracted his claim for the g-factor (Oller, 1984).

The current position is that the strong version of Oller's UCH has been discredited.

Language ability is conceptualised as composed of a general factor in a less prominent position. Linked to this factor are several group factors as demonstrated in Bachman and Palmer (1981 & 1982).

1.1.6. Bachman (1990), Bachman & Palmer (1996)

Bachman (1990)

Bachman's model of communicative language ability (CLA) is an elaboration of Canale and Swain's framework. Bachman distinguishes three components in the model:

1. language competence
2. strategic competence
3. psychophysiological mechanisms

According to the model, language competence involves mainly in organisational and pragmatic competence. Organisational competence consists of grammatical competence and textual competence, whereas pragmatic competence consists of illocutionary and sociolinguistic competencies. Strategic competence involves an ability in assessing and planning to determine the most effective means to achieve a communication goal. Psychophysiological mechanisms refer to the mode and channel in language use in which competence is implemented.

Bachman and Palmer (1996)

Bachman and Palmer (1996) offer a working model of language use specifically for language testing purposes. This model describes (1) the interaction among characteristics of the individual such as his/her language ability, personal characteristics, topical knowledge and affective schemata and (2) how these interact with characteristics of the testing context and test tasks.

Language ability comprises language knowledge and strategic competence. Language knowledge, information specific to language use stored in our memory, includes

organisational knowledge and pragmatic knowledge. Organisational knowledge, including grammatical knowledge and textual knowledge, enables the language user to create and interpret grammatically correct utterances/sentences and to combine these to form texts cohesively, rhetorically or conversationally organised. Pragmatic knowledge, which includes functional knowledge and sociolinguistic knowledge, enables the language user to relate utterances, sentences and texts to communicative purposes and features of the language use setting. Strategic competence, consisting of metacognitive strategies, refers to processes that enable the language user to engage in *goal setting*, *assessment* and *planning*.

Language use is further affected by individual characteristics such as age, sex and the native language. In a testing situation, the test taker's topical knowledge (i.e., knowledge schemata) and affective schemata influence the ways the test taker uses the language and processes the test tasks. These characteristics need to be taken into account for in test design.

Significance of the CLA model

The CLA model reorganises components of language knowledge on the basis of empirical studies (Bachman and Palmer, 1982). For example, cohesion and coherence formerly brought together in discourse competence are redistributed. Cohesion becomes part of textual competence; along with grammatical competence, the two represent abilities involved in "controlling the formal structure of language for producing or recognising grammatically correct sentences, comprehending their propositional content, and ordering them to form texts" (Bachman, 1990: 87). Coherence is realised through pragmatic competence (i.e., through illocutionary and sociolinguistic competencies).

Part of Hymes' "ability for use" is explicitly included in the model under strategic competence, which is defined as "general ability, which enables an individual to make the most effective use of available abilities in carrying out a given task" (Bachman, 1990: 106). Strategic competence is different from language competence; it is more of an

individual's capacity for performance rather than his/her knowledge, and also seems to include affective and volitional factors discussed by Hymes but the extent is not clear.

The role of non-cognitive factors (that is, affective and volitional factors) is explicitly addressed in Bachman and Palmer (1996). Affective schemata are defined as the "affective or emotional correlates of topical knowledge" (Bachman & Palmer, 1996: 65). When combined with the characteristics of a particular communicative task, they determine "the language user's affective response to the task and can either facilitate or limit the flexibility with which s/he responds in a given context" (Bachman & Palmer, 1996: 65). In other words, affective schemata influence language users' willingness to attempt a task and flexibility in adapting their language use in various settings.

The significance of the CLA model is the notion of interaction in language performance and in assessment, lacking in the Canale and Swain framework. For Bachman, interaction is a psycholinguistic and cognitive issue. The interaction of the candidate and his/her social context in a testing context is reflected in terms of his/her ability to handle test content and his/her engagement in the test process.

So far the model is of pivotal importance. First, the model goes beyond Canale and Swain's framework. The model is more coherent and comprehensive in that it tries to be consistent with a wide range of theories of language. It also tries to address important issues concerning theories of language and language use. Second, the model is subject to empirical validation, e.g. the redistribution of cohesion and coherence of discourse competence into organisational and pragmatic competencies. It provides a working rationale for language test developers. Third, the model views the competence/performance relationship from an interactional perspective, in which language knowledge, process, contexts and strategies all play a role in language performance.

To sum up, Chomsky makes the competence/performance distinction; but his concern lies

in the linguistic knowledge of an idealised native speaker. The actual use of the language (i.e., instances of performance) is not the realisation of competence and is therefore irrelevant to our understanding of competence.

Hymes distinguishes between underlying competence and underlying performance and the actual performance. He adds a new component, the ability for use, to his model of language competence. According to Hymes, competence refers not only to linguistic knowledge but also includes sociolinguistic rules. Hymes further sees competence as relative and differing among individuals.

Canale and Swain expand Hymes' model to include yet another component, strategic competence, to deal with communication breakdowns. In their model, language competence includes grammatical knowledge, sociolinguistic competence and strategic competence. Ability for use is in the domain of communicative performance. Canale and Swain's model is essentially behaviour-based.

Bachman's model elaborates Canale and Swain's model. The ability to use language involves essentially three components: language competence, strategic competence and psychophysiological mechanisms. Language competence consists of organisational and pragmatic competence. Bachmanand Palmer's model further includes the topical knowledge and affective schemata, a non-cognitive factor. Language use is seen as the interaction of language ability, topical knowledge and affective schemata, mediated by metacognitive strategies so that situationally appropriate language production is made possible. With reference to language testing, language ability is realised as the interaction of the characteristics of language use context and test taker characteristics. Performance on language tests thus varies according to the test taker's ability and test method (i.e., contexts).

The OPI has a different view of language ability. The definition of language proficiency is based on how the construct is operationalised in the rating scales. Proficiency means

achievement. The test taker's ability to function accurately in a specific context is expressed on a single global rating scale. Levels of performance in a language test are determined by the test task and the rating scale. Associated with such an approach are the concepts of direct testing and authenticity of tasks (Sections 1.2.2. and 1.3.3.).

Language proficiency according to Oller consists of a single unitary ability which underlies all language performance and enables the language user to draw on his/her linguistic as well as world knowledge to generate expectancies about the language s/he is processing. The language ability is demonstrated through pragmatic tests such as dictation and cloze.

The meaning of language ability has been redefined and broadened by the different models of language ability. Presently, the multi-componential and interactional view of language ability has provided a theoretical basis for language testers. The distinction between knowledge and performance suggests that a test not only measures what the test taker knows about the language and how to use it but also to what extent s/he is able to demonstrate this knowledge in a meaningful, communicative and appropriate manner across different contexts. In relation to language testing, language competence may be assessed in terms of the test content designed in such a way that it requires the test taker to vary his/her use of the language through tasks in different contexts and different complexity. Inference thus can be made on the basis of language performance from the language domain of interest.

1.2. Basic considerations in test evaluation

In this section, central concepts related to language testing and test analysis such as types of language tests and validation criteria will be discussed. Issues related to consequential effects of language testing such as test practicality and test fairness will also be examined.

1. 2.1. Purpose of language testing and test use

Language tests serve three major functions: administrative, instructional and research (Jacobs *et al.*, 1981; Henning, 1987; Davies, 1990; Brown, 1996).

Sometimes teachers and administrators need to make decisions based on test information for the purpose of selection. Proficiency tests are especially designed for such decision-making. The focus of a proficiency test is usually on the general skills or knowledge prerequisite to entry into (or exemption from) some institution or work place or selection for more advanced institutions/courses or job obligations. In terms of placement decisions, teachers group students of similar ability levels so they will be better able to tackle the problems or learning points appropriate to each level.

Testing can also provide feedback on the syllabus and improve the effectiveness of teaching. It may be used as an external criterion measure for the evaluation of a programme, course or materials. At the classroom level, teachers may be interested in the strengths and weaknesses of each student in terms of instructional objectives for purposes of correcting students' deficiencies. Language testing can further provide valuable information in our understanding of language and language use. In second language acquisition (SLA) research, language tests can be an elicitation device of interlanguage samples to further our understanding of second language development (Douglas, 1998).

Five test uses are generally recognised: placement, achievement, proficiency, aptitude and diagnostic (Davies, 1990). A placement test attempts to find for each student the appropriate level of instruction in a language programme. The content of the test is usually drawn from the learning points/skills in all levels in a particular language programme. An achievement test is concerned with what the students have learned in a course. It is often used at the end of a course. The content of the test is sampled from the syllabus. A proficiency test measures general ability or skill. Test results may not be informative of students' progress of learning because the purpose of a proficiency test is "to prove not to improve" and there is "no control over previous learning" (Davies, 1990;

Shohamy, 1990).

An aptitude test does not have a specific syllabus to draw on for its content. Like a proficiency test, it attempts to predict the learner's future achievement. Unlike the proficiency test that predicts the test taker's ability in language for a certain purpose, an aptitude test attempts to predict the learner's general ability in language (Davies, 1990).

Diagnostic tests attempt to find out the strengths and weaknesses of the learner or what skill she knows or does not know. Diagnostic tests provide valuable information to teachers and students themselves on their strengths and weaknesses in their learned abilities. Without this specific information, teachers may not be able to address students' weak areas properly (Davies *et al.*, 1999).

1.2.2. Types of tests

In the following section, I will discuss several other categories of language tests: objective vs. subjective tests, direct vs. indirect tests, discrete-point vs. integrative tests and criterion-referenced vs. norm-referenced tests. There will be much overlap among these different categories.

Objective vs. subjective tests

These two types of tests are distinguished on the basis of scoring methods. An objective test is scored against a scoring key or a set of established and acceptable responses. The scorer does not have to have particular knowledge or training in the content of assessment. An example is a test of multiple-choice questions. A subjective test, on the other hand, requires the insight and expertise of the scorer. Opinion or judgement is required in the rating of a subjective test. An example of a subjective test is an oral interview in which the rater is either experienced in eliciting responses from the test takers or has gone through training sessions. Tests like cloze lie near the middle of the objective - subjective continuum (Oller, 1979).

Direct vs. indirect tests

A direct test requires the test taker to perform using the skills the test constructor wishes to assess. An indirect test is one in which inferences about one kind of performance are made through related performance of another kind. In an indirect test, language tasks are abstracted from real-life situations; thus they are different from normal language use. In a direct test, language tasks should be as authentic as possible.

There has been increasing concern voiced that tests should be developed directly to reflect the construct they are supposed to measure (Hughes, 1989; McNamara, 1996). More direct testing has been the result, and this has a number of attractions. First, it is easy to construct a direct test if we are clear about what abilities to measure, and the interpretation of a test taker's performance is straightforward. Second, direct testing has high face validity. Finally, direct testing is more likely to bring about beneficial washback effect. Frederiksen and Collins (1989) view a direct test as a systemically more valid measurement than an indirect test because instructions that improve performance in such a measure are likely to improve the domain area of performance. However, Messick (1994) suggests that a test consisting of briefer structured tasks such as multiple-choice questions or short answers is equally valid in the performance assessment of competencies. Test validity, particularly content and construct validity, is central (Section 1.2.4.)

Finally, tests are only samples (Davies, 1986); in any testing situation, it is not possible to recreate the environment that the test takers are going to function in. Furthermore, examinees are aware that they are in a test and that the tasks cannot really be authentic (Hughes, 1989). Therefore, in a strict sense, all language tests are indirect tests.

Discrete-point vs. integrative tests

J. B. Carroll (1961) originated this distinction between discrete-point and integrative tests. Discrete-point test items assess one and only one language feature at a time.

Traditionally, discrete-point items are used to assess learners' grammatical or vocabulary knowledge. In a pure discrete-point test, the points being measured are determined in part by contrastive analysis of the differences between the language being tested and the native language of the examinee (Lado, 1960). Critics of such an approach feel that the method provides little information on the learner's ability to use the language in a real-world context. They also think that it is difficult to determine which points are being assessed because it is difficult to write a pure discrete-point item. Item selection is yet another problem because we are not sure of the representativeness of the items sampled in a given test (Spolsky, 1968). A discrete-point test is always an indirect test.

An integrative test taps a broader range of language abilities concurrently. It is claimed to have a greater value in assessing overall language proficiency (Carroll, 1961). An integrative test asks the test taker to combine many elements in language to complete a task; for example, listening while making notes or completing a cloze passage. Integrative tests tend to be on the direct testing side of the continuum.

Criterion-referenced vs. norm-referenced tests

Criterion-referencing and norm-referencing are two approaches in the judgement of language test performance. In a criterion-referenced approach, the test has to match teaching objectives or pre-specified criterion behaviour. The purpose is to classify people according to their ability in performing pre-determined sets of tasks. Test results are more descriptive in terms of student mastery of the content areas. The criterion skills or behaviour to be assessed have to be determined before test construction. Then test specifications are written and a cut-off score is set (Davidson et al., 1985). Usually a criterion-referenced approach is to evaluate whether a learner has met the instructional objectives. The challenge when constructing a criterion-referenced test lies in determining which objectives to assess, how to construct items for these objectives, how to administer these items and what constitutes attainment (Hambleton & Eignor, 1978; Brown, 1986).

A norm-referenced test, on the other hand, is a test that evaluates ability against a standard of mean or normative performance of a group. It implies standardisation through prior administration to a large sample of examinees (Henning, 1987). Norm-referenced tests are different from criterion-referenced tests in a number of respects. First, a norm-referenced test must be administered to a large number of people from a target population. Second, acceptable standards of attainment can only be determined after the test has been administered. The standards are determined by referring to the mean score of other examinees from the same population. Test items at different levels of difficulty are included as an attempt to achieve a "normal" (i.e., Gaussian) distribution of scores. If a large percentage of learners pass or fail the test or get very similar scores, the test will be revised or discarded (Henning, 1987).

Norm-referenced tests also have strengths and weaknesses. One of the strengths is the degree of confidence we can place in such a test as the statistical measures of reliability and validity are provided. Also, as the standards have been determined against the performance of other students, these standards are claimed by some writers to be fairer and less arbitrary than is a criterion-referenced test (Henning, 1987). Finally, more comparative information is provided on the relative ability of the student within the entire target population.

One problem associated with a norm-referenced test is the issue of validity; the test is only valid with the population on which it has been normed. The test has to be renormed if the characteristics of the population change. Second, norm-referenced tests relate a candidate's performance only to that of other candidates. Test users are not told explicitly and directly what the examinees are capable of doing in the language being measured.

1.2.3. Reliability

Reliability concerns a test's precision as a measuring instrument. Reliability has many definitions but its main characteristic is consistency (Henning, 1987; Bachman, 1990;

Alderson et al., 1995). Two types of reliability have been identified: stability reliability and equivalence reliability (Ingram, 1977). Stability reliability asks whether a test will yield the same result if administered to the same group of examinees the second time; equivalence reliability examines if two measurement instruments are equivalent. When subjective scoring is involved, rater reliability is computed. Rater reliability concerns the consistency of the rater judgements between pairings of ratings (i.e., inter-rater reliability) and sets of scores by the same rater on different occasions (i.e., intra-rater reliability). These two types of rater reliability are similar in concepts to equivalence reliability and stability reliability.

Identifying the different sources of error variance, estimating the degree to which the error variance is affecting the source and determining the degree of reliability are central to reliability estimation. Three factors have been found to affect the reliability of the test score: the test factor, the situation factor and the learner factor (Henning, 1987; Cohen, 1994). Each of these factors is a potential source of measurement error. Test factors include considerations of the test instrument itself and ratings and may contribute to a greater degree of internal consistency, i.e., the likelihood that the performance of one item will be consistent with performance of another item.

Considerations of rating reliability include the nature of the scoring key, the training of scorers and the number of scorers. With the popularity of performance testing, inter-rater and intra-rater reliability have become the focus of attention. Sophisticated statistical procedures using generalizability theory and Rasch analysis have been employed to take into account the nature of the task being rated and the persons doing the rating to determine rater reliability (Section 1.4.2. for Rasch analysis).

Situation factors include regulatory fluctuations and fluctuations in the environment of the test administration (Henning, 1987). Factors like the interaction between the examiner and the examinee and how instructions are presented to the candidates will affect test reliability. To minimise regulatory effects on a test, training sessions for test

administrators and standardising test procedures have been suggested (Henning, 1987). Environmental inconsistencies will also introduce measurement error, and need to be minimised. Standardisation of the testing environment is one way to reduce environmental inconsistencies.

Learner factors include transient factors and stable factors (Cohen, 1994). Transient factors include test-takers' psychological and physiological states at the time of the test. Stable factors include changes like maturation, further learning or forgetting which bring about fluctuations in true score.

There are different methods of calculating the reliability estimate of a test. The use of a particular method depends on the nature of the test and the threat to present reliability measured and the ease of computation. Each offers a conservative estimate of the reliability of the test score. The most common methods are test-retest methods, parallel forms method, the internal consistency estimates and the measurement of inter-rater reliability. Each of these estimates concerns different aspects of test consistency. The test-retest reliability is appropriate for measuring the stability of a test over time. Parallel-form reliability looks at the stability of test scores over different population samples. The internal consistency estimate, appropriate for estimating the reliability with only one form and one test administration, examines item homogeneity.

To improve reliability, test developers will want to have clearer instructions, more and better items, provide a less anxiety-provoking testing environment and motivate the examinees to complete the tasks. When inconsistent assessment patterns arise, test constructors have to determine if the inconsistency is a result of a learner factor such as a personal reaction to a given task, the measuring instrument or a combination of several factors.

1.2.4. Validity

Validity refers to "the appropriateness of a given test or any of its component parts as a

measure of what it is purported to measure" (Henning, 1987). The concept of validity can be approached from different perspectives and the relationship between these perspectives is interpreted in a number of different ways. According to the 1974 *APA Standards*, validity has three inter-related aspects – criterion, construct and content validity. Each will be discussed below.

Content validity

Content validity is determined by systematically checking the adequacy and representativeness with which the test samples the objectives and/or areas being assessed. Anastasi provides some useful guidelines for establishing content validity:

> 1. *The behaviour domain to be tested must be systematically analysed to make certain that all major aspects are covered by the test items, and in the correct proportions.*
> 2. *The domain under consideration should be fully described in advance, rather than being defined after the test has been prepared.*
> 3. *Content validity depends on the relevance of the individual's test responses to the behaviour area under consideration, rather than on the apparent relevance of item content.*
>
> Anastasi, 1982: 132

Establishing content validity is difficult and problematic (Carroll, 1961; Henning, 1987; Bachman, 1990). The problems primarily come from the difficulties in defining language and language ability. A second problem is the difficulty in defining the domain of language use from which the sample is to be taken and on which interpretation of test result is to be based. A further problem is the operationalisation of real-life behaviour in a test, especially when calculation is required in the test methodology (Weir, 1990). A final problem concerns the variability of opinions on constructing test specifications as to what is being measured in an item (Carroll, 1981).

Because we lack an adequate theory of language use, *a priori* attempts to clarify the construct of language proficiency are necessary. The procedures take us back to Anastasi's Step 1 of the guidelines. The more a test simulates the dimensions of the observable language use in the area to be assessed, the more likely the test is to have

content validity. Content coverage and content relevance (i.e., authenticity) are thus at issue. Both concern the extent to which the selection of test tasks is representative of the domain language tasks which a test should be a sample of (Bachman & Palmer, 1996). There is also a need to look closely at test specifications to ensure that what is to be tested has been described adequately. Expert judgement is required. A second way to examine the content validity is to have a small sample of candidates introspect on the internal processes that are taking place while they are completing the tasks (Faraday, 1982; Cohen, 1985). This procedure would complement expert judgements on test items.

Criterion-related validity

Criterion-related validity is a feature of empirical test validation. There are two types of criterion-related validity: concurrent validity and predictive validity. Concurrent validity concerns the correlation of test scores with another measure at the same time whereas predictive validity concerns the correlation of test scores with some future criterion of performance (Davies, 1977; Henning, 1987).

A validity coefficient is usually obtained and represents the strength of relationship with the external criterion measure. To conduct concurrent validity, the external criterion instrument used is usually a recognised, reputable test of the same ability in the same population. It is administered within days of the administration of the test to be validated (Henning, 1987). For predictive validity, correlational studies of the test administered on two occasions are performed.

We have to be extremely careful in selecting a criterion measure. Validity will be underestimated if the criterion measure itself lacks validity because such validity is dependent to a certain degree on the construct validity of the external criterion measure.

Construct validity

The establishment of construct validity has become an essential part of test construction in recent years. The term first appeared in the 1954 *Technical Recommendations for*

Psychological Tests and Diagnostic Techniques (APA) (Moss, 1992). At the time, construct validation only involved indirect validating procedures; that is, the gathering of data and the testing of hypotheses.

Construct validity is about if a test measures what it is supposed to measure. In other words, test designers are trying to find out if the scores of an assessment instrument permit inferences about an underlying trait. Messick (1975) argues that content and criterion considerations provide relevant but insufficient evidence about the validity of the inference made from a test. He contends that all measurement should be "construct-inferenced." Thus construct validity is at the centre of validity inquiry.

In construct validation, the hypothesised relationship between scores and abilities has to be empirically tested. It requires both logical analysis (i.e., theoretical and operational definition of the constructs) and empirical investigation. For example, we can hypothesise that the ability to read may involve several sub-skills like inferring and guessing. Empirical investigation has to prove the existence of the distinct abilities.

In recent years, considerable attention has been given to the social consequences of assessment in construct validation, in particular, the impact of the test, the value implications of the test use and the effects of instructional changes brought about by the introduction of a new test (i.e., the washback effects).

Language teachers and testers working in the communicative framework usually attempt to assess those of the learner's skills judged to be relevant to his/her present or future needs. The more closely a test reflects the needs and the instruction that precede the test, the greater the degree of construct validity the test will demonstrate. At other times, language teaching adopts the approach of the test at the end of a course. A test can be a very powerful instrument for effecting change in the language curriculum as can been seen in the Sri Lanka Project (Alderson & Wall, 1992). However, Alderson & Wall (1993) suggest that the quality of the washback effect (either positive or negative) may be

independent of the quality of the test. They recommend a set of guidelines and highlight the importance of defining the scope of washback and stating explicitly the washback hypotheses being used for evaluating the systemic validity of an assessment instrument.

Although the concept of validity has been discussed in terms of different types, psychometricians like Cronbach (1988, 1989) and Messick (1988; 1989) have come to view validity as a unitary concept which requires multiple types of evidence to support specific inferences made from a test score (i.e. convergent validation). Messick sees validity as "an integrated 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 others modes of assessment" (Messick, 1989: 5).

Face validity

A final possible form of validity is face validity. Face validity refers to whether a test looks as if it is measuring what it is supposed to measure (Cohen, 1994). Face validity has been discounted by some test researchers (Lado, 1960; Ingram, 1977; Bachman & Palmer, 1981). Davies (1990) notes that the importance of face validity is in its "public relations." The value of face validity lies in the judgement of the test takers and other lay people concerned, e.g., the schools, the administrators or the workplace. Stevenson (1985) argued that content and construct validity should not be sacrificed at the expense of an increasing lay acceptance of a particular form of a test.

Because of the derogatory overtones of the term *face validity*, Bachman (1990) offers the term *test appeal* to stress its effect of acceptability to test takers and test users. Judgement of face validity should refer to the following (Bachman, 1990):

1. The respondents' perceptions of any bias in test content
2. Their understanding of the nature of the task that they are being requested to complete and
3. Their awareness of the nature of their actual performance on the test as a whole and on any particular sub-tests.

These suggestions call for verbal report measures (Cohen, 1984; Anderson, 1991). Verbal report measures can yield empirical data that provide information concerning how test takers perceive tests and how they actually deal with them in testing situations.

Components of validity inquiry

Five sets of categories have been used to guide validity inquiry; they include the traditional construct-content-criterion categories, Messick's facets of validity inquiry, Cronbach's perspectives on validity argument, Frederiksen and Collins' principles of systemically valid testing and Linn, Baker and Dunbar's validation criteria for complex, performance-based assessments. Apart from the traditional psychometric validity criteria, the other guidelines offer validity criteria in the context of educational measurement and performance testing which will be relevant in the validation of the Tour Guide English Test to be discussed in Chapter 2. I will start with Messick's facets of validity inquiry (1989).

Messick(1989) uses two sets of categories for validity inquiry. One category focuses on the functions of testing; it distinguishes between test interpretation and test use. The other category, focusing on the justifications of testing, distinguishes between appraisals of evidence and consequences. Figure 1-1 represents Messick matrix of validity.

Figure 1-1: Messick's Progressive Matrix (1989)

	Test Interpretation	Test Use
Evidential Basis	construct Validity (CV)	CV + Relevance/ Utilities (R/U)
Consequential Basis	CV + Value Implications (VI)	CV + R/U + VI + Social consequences

Messick notes that it is a progressive matrix with construct validity appearing in every cell and highlights construct validity as the integrating force for validity inquiry. Hence, if we want to use a test score for a particular purpose, we must justify this by considering both construct validity and value implications as well as the relevance of the particular

use of the test and the social consequences of using the test score. In short, in order to justify the interpretation of a test score, we must collect evidence supporting the construct validity in relation to test score interpretation and consider the value implications attached to this interpretation. In order to use the score purposefully, evidence to prove that the abilities measured are relevant to the candidates' effectiveness as language users has to be collected. Finally, consequences of decisions made on the basis of the test score have to be examined.

Cronbach's perspective on validity

Cronbach suggests that validity inquiry must link concepts, evidence, social and personal consequences and values" (Cronbach, 1988). He organises his discussion of validity into five categories: functional, political, operationist, economic and explanatory perspectives.

The *functional* perspective has to do with whether a practice has "appropriate consequences for individuals in institutions" (Cronbach, 1988: 6). The *political* perspective refers to decisions on the fairness of a test and to the responsibility of test experts to communicate sensibly to stakeholders so as to improve the basis for these decisions. The *operationist* perspective involves two aspects of analysis. First, it involves the matching of the test content to a domain of performance. The second aspect is the analysis of the fit between the domain of performance and the purpose of the test. The *economic* perspective involves questions about the empirical relationship between the test and some criteria that test designers consider important. It also involves questions about the success of placement or selection based on test information generalised from one context to other contexts. Lastly, the *explanatory* perspective calls for explicitly stating theoretical rationales and challenging them deliberately. The testing of rival hypotheses includes offering alternative explanations of observed behaviour and gathering convergent and discriminant evidence. For Cronbach, this is an important component of validity inquiry, particularly that of construct validation.

Both Messick's and Cronbach's works reflect the centrality of construct validity and the

importance of considerations of the consequences of evaluation in any test. Two aspects of validity that other educators failed to include are included in their schemes: an interpretative validity component and a consequential validity component (Moss, 1992).

In recent performance assessment, some authors have noted that the practice of validation still tends to rely on the traditional sources of evidence, that is, evidence about content representativeness, internal consistency, reliability and correlation with other measurement. These types of evidence are not sufficient for the evaluation of performance assessment, and each of the authors to be discussed below has proposed additional criteria concerning meaningfulness and directness raised by proponents of performance assessment (Section 1.2.8.). First, I consider Frederiksen and Collins' principles of systemically valid testing; second, the work of Linn, Baker and Dunbar.

Frederiksen and Collins' principles for systemic test validity

Frederiksen and Collins (1989) propose a set of principles for the design of a systemically valid testing system. These principles relate to components of the testing system, standards and methods for fostering improvement on the test. Components of the testing system include *set of tasks*, *primary traits for each task* and *subprocess*, *a library of exemplars* and *a training system for scoring tests*. Standards include *directness*, *scope*, *reliability* and *transparency*. Methods for fostering improvement on the test include *practice in self-assessment*, *repeated testing*, *feedback on test performance* and *multiple levels of success*.

Linn, Baker and Dunbar's criteria for test validation

Linn, Baker and Dunbar (1991) raise the concern that traditional approaches to validity may cause performance-based tests to compare poorly to multiple-choice tests. They suggest expanded criteria that would help to "clarify the kinds of information that alternative forms of assessment offer and can thereby help to establish complementary roles for conventional and alternative approaches" (p. 16). Their criteria are listed below:

Consequences: the intended and unintended effects of tests

Fairness: refers to the equitable access to resources that include the learners' family background, exposure to and motivation on the tasks; performance rating that may reflect rater biases; and sources of irrelevant difficulty like prior knowledge.
Transfer and generalisability: extent to which results can be generalised across tasks and raters and degree of generalisability from specific tasks to a broader domain of achievement.
Cognitive complexity: the extent to which the assessment places emphasis on "problem solving, comprehension, critical thinking, reasoning, and metacognitive processes" (p. 19) evidenced by the analysis of the tasks, students' familiarity with the problems and the ways in which learners attempt to solve the problems.
Content quality: the extent to which the content of the assessment is consistent with the current understanding of the field. It includes systematic judgement of the quality of the tasks written by subject matter experts and the way in which learners interpret the content.
Content coverage: the extent to which the test covers the content domain determined by experts.
Meaningfulness: the extent to which the test presents learners with meaningful problems which provide valuable educational experiences.
Cost and efficiency: the extent to which costs are kept at acceptable levels during the process of test development.

All these authors have added categories relevant to the consequential component of validity. All the work points toward the notion that performance assessment is likely to produce more valid interpretations but performance assessment does not yield as good a result as a multiple-choice test in terms of internal consistency, efficiency and comparability. To deal with those concerns, additional criteria like directness, meaningfulness and cognitive complexity are offered and acceptable levels should be achieved for particular assessment purposes.

What is lacking or implicit in these works is the specification of purposes and contexts

for the assessment and the specification of the general aspect of performance to be evaluated (Moss, 1992). The evaluation of the fit among the purpose, context and assessment-based interpretation and use is central to construct validity and the notion of generalisability and possibly replicability.

1.2.5. Relationship between reliability and validity

Reliability is characterised as consistency among independent observations that are intended to be interchangeable (e.g., consistency among tests or evaluations or tasks etc.). The basis of it is objectivity (Lado, 1969). A consistent test will give meaningful results whereas inconsistent tests yield random results (Davies, 1977). However, reliability is not sufficient in itself. Reliability coefficients can still be high even if there is a mismatch between the test and its purpose. For example, if we use TOEFL to measure people's mathematical ability, consistency from one part of the test to another part is still high but we will be trying to measure a totally different trait from what TOEFL is designed for. Therefore, we need to examine the validity of a test as well. Test validity is increased if the test is truer to the target use situation. If the validity of a test is sacrificed to increase reliability, we end up with a test that is a reliable measure of something other than what we want to measure.

Robinson (1979) identifies three areas of difference between less standardised forms of assessment such as performance assessment and objective tests. They are the amount of language produced by the candidates, the types of abilities tested and the norms of language use established. In an objective test, candidates may not be required to produce any language at all; the ability to recognise the appropriate forms is sufficient. In a performance test, the candidates are required to produce language; their ability to produce the kind of language required for successful communication is a crucial factor in assessment. In an objective test, the candidate performance is usually judged against a set of predetermined answer keys, whereas in a performance test, the norm of the language is derived from the candidate's own use of the language. The candidate is assessed on the

basis of the percentage of correct responses as well as the quality of language produced. These differences refer to the discrete-point and holistic continuum of language and they also lead to a tension between reliability and validity. Adding more bits of the language in a test increases test reliability; but the test is made more remote from realistic language use, thus decreasing test validity. An attempt to increase test objectivity/reliability may lead language testers to a restrictive view of what it is that they are measuring.

There are also problems of generalisability across tasks in performance testing. Task reliability, defined as consistency of performance across tasks intended to measure the same ability, is more difficult to achieve (Moss, 1994). Recent validation research takes account of other criteria such as authenticity, directness and cognitive complexity in test design; therefore it sanctions a relatively low level of reliability (Linn *et al.*, 1991). As emphasis has shifted to a concern with the construct and predictive validity of a test, more attention has been given to qualitative analysis of the constructs of the test and whether the test provides predictive information on candidates' language ability in their future language use situation. Statistical analysis of test questions is less emphasised, and a more moderate level of reliability may be considered acceptable if the criteria are met.

But simply because performance-oriented tests that measure language ability cannot reach such a high level of reliability as an objective test can achieve, it does not mean reliance on reliability impedes validity in performance tests. We still need to be certain that test items are reliable before we consider the validity of a test. Reliability is a necessary but not sufficient condition for validity. What matters for reliability and validity is the "adequacy of plan, analysis and sampling of content, and relationship of test to purpose" (Davies, 1977).

1.2.6. Practicality

Issues of practicality should be taken into account in test design and development. A reliable and valid test is of little use if it is not practical. Practicality involves issues like

economy, ease of administration, scoring and result interpretation. The longer it takes to construct, administer and score a test, the more costly the test is. Davies (1990) suggests that a test should be as simple as possible. It is also desirable to make the test as short as possible. The duration of a test may affect its successful operation in ways such as fatigue, the availability of administrators and test rooms etc. During the development process, the test should limit its requirement on the people, time and material. However, test reliability and validity should not be sacrificed because of these practical constraints. A pencil-and-paper test may be an efficient and inexpensive means to measure learners' language ability but its validity in making inferences on candidates' ability to use language is uncertain. On the other hand, it may be costly to develop a test that characterises real-life language use contexts but it is likely to be more valid in terms of assessing language in communication. When developing this type of test, greater attention has to be given to the development of efficient data collection design and scoring design.

1.2.7. Domain of language use

Different situations entail different language use. The area in which particular language behaviour takes place is the language use domain. A clear and well-defined scope of language use helps the construction of a useful test.

Two notions are essential in the design, development and use of language tests: *language use tasks* and *target language use domain* (Bachman & Palmer, 1996). A language use task is "an activity that involves individuals in using language for the purpose of achieving a particular goal or objective in a particular situation or setting;" a target language use domain is "a set of specific tasks and their attendant settings that the test taker is likely to encounter outside of the test itself and that require language use" (Bachman & Palmer, 1996:59).

Douglas and Selinker (1985) propose the concept of discourse domain to further

distinguish the idea of context, a collection of external features from the internal interpretation of these features in the individual. For the test taker, discourse domains establish what the external context is and serve as an input to goal-setting and planning processes in which different components required to deal with the given situation are brought into play. The notion of discourse domain allows the test taker to make sense of the test tasks either by making reference to an existing discourse domain in his/her world knowledge or creating a temporary one so that performance is possible. Test scores are then used to make inferences about the candidate's ability to perform in the particular language use domain of interest.

Bachman and Palmer (1996: 102) suggest the following steps in identifying language use tasks in a language use domain:

1. *identify the stakeholders who are familiar with relevant language use situations, who can help identify the relevant domain and tasks*
2. *identify or develop procedures for gathering information about tasks*
3. *gather information on the domain and tasks in collaboration with stakeholders*
4. *analyze the tasks in terms of their task characteristics, and*
5. *make an initial grouping of tasks into categories of tasks with similar characteristics*

These steps indicate that consultation and collaboration with expert informants, field observations and analysis of language use for task construction are necessary in the design and planning stage of test construction.

1.2.8. Performance assessment

The purpose of this section is to (1) examine different definitions of performance, and (2) discuss criteria for the design of a performance test.

Definitions of performance testing

The term performance testing in second language contexts seems to be self-explanatory, but there have been different perceptions among people who have used it. Those who are familiar with Chomsky's competence/performance dichotomy may interpret performance

in his sense. In the Chomskyan view, performance refers to the use of the underlying knowledge that one needs to perform. In a strict sense, only performance is measurable.

The second interpretation of performance is behavioural. This type of second language assessment has its origin in non-language contexts. McNamara (1996) calls it the work sample approach. Researchers interested in this type of assessment try to account for communicative behaviour and to explain how the test taker uses both his/her knowledge of the language and knowledge on how to achieve a goal (i.e., communicative skills). In this behavioural perspective, competence is closely tied to performance and is demonstrated through actual overt behaviour. To be judged competent, a test taker would have to demonstrate his/her knowledge of appropriate behaviour. OPI is an example. The drawback of the behavioural view is that it deals only with overt manifestations of behaviour without trying to account for the behaviour. It does not provide an adequate explanation of language ability in the communicative framework.

The third interpretation of performance partly stems from the advent of theories of communicative competence and partly from the behavioural perspective. McNamara (1996) terms it the *cognitive-psycholinguistic* approach and Chapalle (1998) names it the *interactionalist* approach. Such an approach assumes that a competent communicator has the necessary underlying knowledge and strategies that will enable him/her to carry out a communicative task in an appropriate manner in a given context. To judge how competent a test taker is, s/he will have to demonstrate this knowledge in language use situations. A performance test in the third sense focuses less on performances but more on how much of the underlying knowledge is revealed through performance which is only a vehicle of the assessment target. Consistency or variability of performance affect score meaning. Generalisability and replicability of the observed tasks to universe tasks relevant to the knowledge/skill domain are at issue (Chapter 5). It is on these lines that I will construct the TG Test to assess a would-be tour guide's ability to use English in the workplace.

Typology of performance tests

Jones distinguishes three types of performance tests: direct assessment, work sample methods and simulation techniques. In direct assessment, a candidate's performance is directly observed in the workplace. Confirmation of his/her ability in a new setting depends on the candidate's performance on the job during an initial period. An example would be an apprentice teacher being observed in the classroom. The trainee is given provisional admission to the workplace and is evaluated over a probationary period (McNamara, 1996). Second language proficiency assessment is usually used to screen or select. But it is not common to admit people to the workplace on a probationary basis before the assessment of language ability takes place. Also, Jones points out that language behaviour is complex; it is necessary to make observations over a relatively long period of time before one is satisfied that an adequate sample has been obtained (Jones, 1985). For these reasons, direct language assessment may not be a feasible alternative.

In the work sample type, assessment takes place in the workplace but the tasks are controlled to achieve standardisation of the assessment. An example would be prospective teachers teaching a mini-class and responding to student questions. The principal feature that distinguishes work sample method and direct assessment is the degree of manipulation of tasks on the part of the examiner. The work sample method is more controlled, whereas in direct assessment there is no manipulation.

Simulation techniques create simulated settings and tasks in such a way that they represent salient features of the real-life context. A certain degree of abstraction is required in simulations. Performance on the tasks is used to predict performance on similar real-life tasks. Jones (1985) considers simulation as the most feasible approach in language testing. Role-play is a frequently used simulation technique; the test taker is asked to assume a particular role in a given situation so that language functions of interest can be elicited.

Second language performance tests

Language performance tests arose in response to a need for selection among foreign students wishing to study either in the U.S. or the U.K. in the 60s (McNamara, 1996). J. B. Carroll recommended a performance component in the language tests for selection of foreign students. He recommended what he called an "integrative approach" to complement the then popular "discrete structure point approach" (Carroll, 1961 [1972]). The integrative approach stresses two points. First, the test is independent of the learner's first language. Second, integrative testing comes in the context of English for Academic Purposes (EAP) testing; therefore, assessment cannot be based on what the learner knows prior to the test but on future activities the learner will have to face (See Section 1.3.).

Carroll's implicit distinction between achievement and proficiency is made explicit by Davie (1977), who argues that a proficiency test is concerned with control and prediction of future performance. The focus is again on academic study. Davies' notion of control suggests skilled execution in performance (McNamara, 1996). In addition, his notion of predictive validity leads to a greater specification of the test context or criterion in which performance is to be predicted. The use of test specifications, which include sociolinguistic component of language use, becomes one feature in performance testing or communicative testing.

In summary, the type of testing recommended by Carroll and Davies focuses on two features: prediction of performance in a real-world context and integration of linguistic and non-linguistic sub-skills in performance.

Types of second language performance tests

McNamara (1996) proposes two types of performance tests: strong and weak. Inferences made from test results are based on the two different hypotheses. In a strong performance test, tasks represent real-life tasks and performance is evaluated on real-world criteria. Aspects of language may or may not be assessed. Criteria reflecting aspects of language use ability are subsumed under a larger part of the criteria set for assessing performance.

Language, in the strong sense, is just the medium of the performance because the target of assessment is the performance of the tasks. Simulation is not possible because such a test requires a full integration of language knowledge and ability for use. Jones (1985) proposed language performance test of this kind. The problems with the strong performance test are that testers are faced with the problem of sampling and the effect on the candidates when they are being observed for an extended period of time not to mention the expense and the testing procedures involved in such an assessment method.

In the weak sense, language performance is the focus of assessment. Candidates may be required to perform either tasks resembling real-world tasks or artificially constructed test tasks. The purpose of performance is to elicit a language sample so that second language proficiency and skills of execution can be assessed. Test results are reported in language-related terms such as "*limited working proficiency*"; a candidate has *speaking vocabulary sufficient to respond simply with some circumlocutions; accent, though often quite faulty, is intelligible"* (S-2 in ACTFL/ILR). Most oral proficiency tests are weak performance tests. The practical advantage of the weak performance tests is that they are easier and less expensive to administer. Also, by focusing on the language sample elicited from the test, considerations of capacities other than language knowledge to fulfil the tasks successfully are minimised, thus avoiding political issues such as test fairness and objectivity in judgement which can affect the test's acceptability to the public.

Messick (1994) also distinguishes two types of performance assessment: product-based and construct-based performance assessment. The product-oriented approach focuses on performance as the target and vehicle of assessment. The concern is the quality of the performance (i.e., the product). The quality of judgement is the key issue in validation. In such an approach, inferences are not to be made about the competence of the test takers. Diving is an example.

A construct-based approach begins by delineating the knowledge, skills or other attributes to be assessed. Then behaviour associated with these is defined. Replicability and

generalisability are the two central concerns since they establish the boundaries of the meaning of the scores. Apart from the three traditional validation criteria, this approach further includes content coverage, transfer and generalisability. In regard to task types, both contextualisation and decontextualisation of problems are appropriate. Complex skills or their component sub-skills, if well-defined, are acceptable. A second language test, according to Messick, is a performance assessment of a construct in which inferences about a candidate's ability are made on the basis of his/her performance on different item types to measure the proposed trait.

Criteria for a good performance test

Apart from the traditional three criteria for a good test (i.e., validity, reliability and practicality), evaluation criteria and how performance is judged are two other important aspects to consider when designing a performance test. First, the language tester has to elicit a rateable sample and quantify it into some type of useful score for interpretation. Then, evaluation guidelines usually in the form of a rating scale are developed with reference to the construct measured. Stiggins suggested three steps for developing the description of a performance test:

> 1. *conduct systematic observations of practitioner work sample and identify the relevant skills (e.g., conduct a job analysis),*
> 2. *involve experienced practitioners in an in-depth discussion of the knowledge and skills that form the basis of their profession, and/or*
> 3. *generate potential lists of relevant skills*
>
> Stiggins, 1981: 6

It is not possible to list all the requisite skills involved in a task; therefore, it is important to sample the skills and make a generalised judgement on that basis. To achieve this, an extended needs analysis has to be conducted; this will provide a detailed description of the specified context, the test tasks, the specific conditions under which the tasks will be performed, and the criteria against which the performance will be judged.

Finally two validity issues associated with performance assessment need to be considered: authenticity and directness. Frederiksen and Collins (1989) propose directness as one important "standard" in performance testing. To them, directness refers

to direct assessment of knowledge and skills. To put it simply, if we are interested in how well the candidate speaks, we measure his/her speaking because the very act of speaking provides us information on the candidate's ability to speak. It is easier to assess productive skills. To assess receptive skills such as listening and reading, the candidate has to demonstrate that s/he has performed the skill successfully. However, it seems not possible to measure constructs of language knowledge or skills directly. They can only be inferred from test performance.

Authenticity refers to task relevance and meaningfulness. Both the language tester and the test taker are aware that in a testing situation, the tasks cannot really be authentic. But it is best to make the tasks as relevant to and representative of the construct as possible (Section 1.3.2.).

1.2.9. Test fairness

In this section, factors affecting fairness in language testing will be discussed. The purpose is not to prescribe judgements about the ethical position of language testers but to summarise some issues confronting people in this profession.

Test fairness has been a major concern in the assessment of language ability. Edgeworth's proposal to investigate the fairness of tests by using Theory of Errors indicated scholars' scepticism over a test's precision of measurement on the candidate's ability in terms of a number (1888, in Spolsky, 1996). In recent years, the objectivity of a test and its empirical validation have no longer been thought to guarantee the fairness of a test. Instead, language testers investigate the value of subjective judgement, sources of test bias and the impact of a test to ensure test fairness.

To this end, researchers have conducted test bias studies, which attempt to identify and reduce the effects of confounding variables on test scores. Test-internal/external bias studies are the two common procedures. In recent research, method bias (Shohamy,

1997) and consequence bias (Messick, 1996) have been included as well.

Judgement in psychometric terms has been central in establishing test validity and detecting test bias. A potential pitfall in using statistical analysis is the choice of a suitable external criterion for the measurement of ability. The validity of the control measurement used as a benchmark for comparing different groups is difficult to ascertain. Thus, correlation is not a sufficient basis to determine validity; the results of test-external bias detection have to be interpreted with caution. To detect a test's external bias, Spurling and Ilyin (1985) suggest the use of judgement to determine if the differences in test performance are results of measurement error or real differences in candidates' ability. The discrepancies could be a result of learner variables (Spurling & Ilyin, 1985) or language distance (Chen & Henning, 1985; Brown and Iwashita, 1996). The nature of differences could be detected by test bias studies.

Test-internal bias studies are not without problems. Test items may be biased against a particular group. The test may be internally consistent but measuring the wrong underlying trait.

Aspects of a test like the item type, discourse type, characteristics of test tasks and background knowledge can affect test scores and result in bias against certain test takers (Bachman, 1990). Item types may affect test scores because different item types are being processed differently by test takers (Shohamy, 1984). Test methods may be considered unfair unless they are proven to be relevant to the test construct.

Studies have shown that cultural content in a test may also be a potential source of bias. Chen and Henning (1985) found that some items in a multiple-choice vocabulary test appeared to be biased in favour of one linguistic and cultural group. Studies done by Alderson and other researchers in the context of ESP testing investigated whether background knowledge was a bias or part of the ability measured (Alderson & Urquhart, 1985; Koh, 1985; Tan, 1990; Clapham, 1996). The results in general indicate that

students' performance appears to be affected by their prior background knowledge as much as their linguistic proficiency (Section 1.3.5.).

Test purpose can also be a factor affecting test fairness. Tests may be used for different purposes from those they are intended for, and this may be considered unfair (Shohamy, 1983, 1996; Hawthorne, 1996). One example would be a public test used to bring about syllabus changes without changes in other educational components like teacher training and/or curriculum change. In such a case, the impact of the test is negative because the consequence of the test could result in the narrowing of the curriculum and teaching to what the test covers.

Traditionally, language testers have discussed washback as either a good or bad force in teaching and learning. Alderson and Wall (1993: 116) note that a test's failure to bring about beneficial washback may be due to problems "which exist within society, education and schools, that might prevent washback from appearing." In other words, the negative effects of testing may be due to forces in schools, educational circles or even the wider society that are beyond the influence of language testers. Thus, to ensure test fairness, one option is to involve stakeholders in test development (Rea-Dickins, 1997). Stakeholders can occupy various roles; for example, language testers, teachers, administrators, sponsoring bodies, subject experts, the government and test takers. Their participation is not limited to expressing their views, concerns or complaints but to equipping language testers with valuable information that they can later take appropriate action on and thus promote greater fairness in the testing process.

The final point to be mentioned concerns ethics in language testing. Davies (1997) argues for professionalisation of this field because it provides a "contract for the profession and the individual with the public" (Davies, 1997). He offers a code of practice in language testing, which has been adopted by ILTA (Davies, 1999 & 2000a). He also urges language testers to be realistic about what kinds of test consequences they will be accountable for.

In short, achieving greater test fairness has been a concern in this field. Language testers are responsible for developing a valid and useful test and where possible to bring about beneficial washback effect. But they cannot be held responsible for all possible social consequences. To achieve a greater degree of test fairness, language testers have to professionalise the field, democratise the testing process by including all stakeholders, minimise test method effects and consider issues of test consequence.

1.3. Specific purpose language testing

Testing languages for specific purpose (LSP) refers to a types of language testing in which "the test content and test methods are derived from an analysis of a specific purpose language use situation" (Douglas, 2000:19). LSP testing suffered from a lack of theoretical justifications in its early days. However, as our understanding of language use ability broadens, LSP testing can be viewed as a type of communicative test which requires an interaction of the language knowledge, background knowledge and the language use context. In the case of LSP testing, the language use context of interest plays a more prominent role in the provision of contextual cues to engage the test taker in a "discourse domain" that enables the test taker to understand the context and create meaning within the context (Douglas & Selinker 1985). Specificity, particularly background knowledge, is at issue (Section 1.3.7.).

In this section, I will first briefly describe the development of teaching English for specific purposes (ESP), which has a direct impact on ESP testing. Second, I will discuss the implications of the communicative competence on LSP testing. Then attractions of LSP testing and features associated with an LSP test will be discussed and presented. Two recent LSP tests in occupational contexts will be described. Finally, issues associated with LSP test design will be discussed.

1.3.1. Development of ESP

The driving force behind the popularity of English for specific purposes (ESP) is practical rather than theoretical. With the rapid increase of importance of English after the Second World War, people needed English for clearly defined reasons such as reading technological texts or conducting business (Hutchinson & Waters, 1987). As early as the 1920s, there was an awareness that different professional fields called for different language use; but the ESP movement only came into full existence in the 60s, and by the 70s, ESP was recognised as a separate area of activity from general English (Widdowson, 1983; Hutchinson & Waters, 1987).

The awareness of learning/teaching for specific purposes also coincided with the development of communicative language teaching, which led ESP course designers to attempt to base their materials and teaching on texts or activities tailored to suit learners' linguistic needs in the real world.

In the early stages of ESP, researchers focused on register analysis (Hutchinson & Waters, 1987). Register analysis often took the form of frequency count (Barber, 1962) or clausal analysis (Huddleston, 1971). The assumption was that each subject area constituted a specific register. The aim was to identify the grammatical and lexical features of the different registers to form the basis of teaching materials that would be more relevant to learner needs. However, Hutchinson and Waters (1987) point out that although scientific English does favour certain forms than others, those forms can also be found in general English. Swales (1985) comments that statistical analysis of linguistic forms has little *explanatory force*. Furthermore, as Widdowson (1979) notes, the fact that a particular text type, e.g., scientific text, has a high proportion of some syntactic structures and low occurrence of others does not reveal anything about that type of discourse as a whole.

In the 70s, linguists such as Widdowson and Selinker, Trimble and Lackstrom began to apply rhetorical analysis to see how sentences combine to produce meaning. The focus

was again on scientific and technical English texts. The concern was to identify the organisational patterns and the linguistic structures in texts that form particular discourse: these textual patterns then formed the basis of an ESP syllabus. There were, however, criticisms of this approach too, one being that Widdowson never clearly explained the relationship between the communicative purposes and the organisational units (Swales, 1985). The second criticism concerns the universality of discourse; if a particular specialist area of discourse is the same among the various linguistic speech communities (for example, the world-wide scientific community), the discourse type should not be unfamiliar to the speakers. Hence, learners do not need to be taught how a particular type of text is constructed in English because they are already familiar with the rhetorical organisation of the text. What they need to be taught is the English language (Swales, 1985).

Parallel to development in the analysis of discourse types of academic texts, there was a growing awareness of learners' needs. ESP course designers began to carry out needs analysis on learners' future linguistic requirements; they started out by identifying the target situations, then carryied out an analysis of the linguistic features in the situations, which later formed the syllabus (Van Ek 1979; Wilkins 1976). The best known model of such an analysis is set out by Munby (1978) in his *Communicative Syllabus Design*.

The Munby Model presents detailed profiles of learner needs in terms of the purposes of communication, the communicative settings, the means of communication, language skills, functions and structures required (i.e., the communication needs processor (CNP)). The CNP was very influential and was used as the basis for the ELTS test (Carroll, 1981). Weir (1983) also used the parameters in CNP in the identification of learner needs for the Test of English for Educational Purposes (TEEP).

For a number of years, Munby-style needs analysis was central to ESP test design. Criticisms soon appeared. The first is that the Munby model reduces language use to a list of skills and sub-skills with no apparent connections and no explanation of how these

skills are realised in language use (Davies, 1981; Widdowson, 1983). Second, the model is impractical (Mead, 1982); the Munby needs analysis will produce a huge list of skills, most of which are not convertible into test items. Third, because there is no indication of the relative importance of the different needs and skills recorded, principled sampling of the testable needs and skills is not possible (Alderson, 1988a). Fourth, the listing of the communicative key is not complete and is not empirically verified (Davies, 1981; Skehan, 1984), and some skills required in real life are not included (Dudley-Evans, 1980; Alderson, 1988b). Finally, the model works with hypothetical students only. It does not take account of real learners, e.g., their existing skills, their wishes, their language learning experience and their background knowledge and their capacity in solving learning difficulties. The model is basically linguistic and sociolinguistic. It does not take into account the psycholinguistic elements in learning a language (Hutchinson & Waters, 1987; Alderson, 1988b). The model, in the extreme case, can only provide for one learner in one specified testing situation (Skehan, 1984).

Hutchinson and Waters claim that the era of needs analysis is over, but this is not so in the case of ESP testing. The Munby model may have been severely criticised but it was not intended as the basis for test specifications in the first place. The detailed and explicit nature of the Munby model helps us understand what an adequate needs analysis should be (Clapham, 1996). The Munby Model of needs analysis may be too detailed and limited in scope, but it does not mean that something of this kind is not necessary.

1.3.2. Assessment in LSP

ESP teaching methodology and the concept of communicative competence have had effects on LSP testing methodology.

Implications of communicative competence on LSP testing

The theoretical distinction between communicative competence and performance suggests that communicative tests not only measure what the learner knows about the

language and how to use it but also to what extent the learner is able to demonstrate this knowledge meaningfully and communicatively in a given language use situation. This distinction has important implications in the design of LSP tests:

(1) Communicative competence is dynamic, relative and context-dependent (Hymes, 1967; Savignon, 1983; Douglas, 2000). An individual's ability to use language involves the interaction of competence and performance as well as the language use domain. Since there are individual differences, language use ability is understood in terms of degrees of proficiency. Thus, in a test, we are measuring relative degrees of language ability.

(2) It is desirable that performance should be measured in a communicative context. An indirect pencil-and-paper test will not necessarily give a valid indication of the candidate's actual language skills. Candidates' success in handling particular communicative situations will be better indicated by the use of a performance-oriented test. In LSP testing, inferred competence is the focus of assessment; the emphasis is on the test taker's ability to manipulate language to achieve communicative goals in a variety of specified contexts. In this regard, a cross-contextual measurement requiring multiple and varied contextualised tasks seems desirable. A modest correlation between contexts is indicative of the test's generalisability.

(3) Related to the concept of contextualisation is "authenticity" and "interactiveness." Widdowson (1979) distinguishes two senses of authenticity. In the first sense, authenticity is synonymous with being natural and genuine, whereas in the second authenticity refers to the interpretation process in which the learner engages in order to make sense of the text. The input may not be authentic as in the first sense; nevertheless the process is authentic to the learner. For Messick (1994) authenticity refers to the "complete construct representation." To achieve an authentic assessment, Messick recommends an assessment combining contextualised

performance tasks and decontextualised items, since very often the learner's experience of language learning is a mixture of realistic activities and the learning of decomposed skills. Such experience of learning should be reflected in the test. Bachman (1990) distinguishes two approaches to defining authenticity: the real-life (RL) approach and the interactional approach (IA), defined as the interaction between the test taker, the test task and the testing context. In IA, a theoretical framework listing the factors affecting test performance is used to construct tests and the framework should include features of language use relevant to the interpretation and the use of the test scores. IA is now the most widely accepted approach to defining authenticity.

Interactiveness is defined as "the extent and type of involvement of the test taker's individual characteristics in accomplishing a test task" (Bachman & Palmer, 1996:25). In other words, the degree to which a given test task engages the test taker's areas of language knowledge, topical knowledge, affective schemata and metacognitive strategies is referred to as task interactiveness, but it is not possible to list all the strategies used by the language user in the target situation. It is, however, possible to take account of language knowledge in the language use domain tasks.

In sum, authenticity has to do with the relevance of the test task to the target language use situations. It is related to the traditional notion of content validity (Bachman & Palmer, 1996). Interactiveness, on the other hand, is dependent on how the construct of an LSP test is defined and what the test taker characteristics are. Therefore, authenticity, interactiveness and construct validity are related, and levels of authenticity and interactiveness acceptable in each test are relative.

(4) The final concept related to communicative competence is the notion of context, which usually includes linguistic contexts and situational contexts, the focus of discussion below.

In attempting to define context, Hymes (1967) suggests the mnemonic SPEAKING in the examination of different situational factors: setting, participants, ends, act sequence, key, instrumentalities, norms and genres. These features indicate the complex interaction of physical, social and psychological factors involved in a context. Performance is based on how the test taker interprets and reacts to the contextual cues present in a given language use situation. Therefore, provision of contextual cues in the test material helps the test taker engage in the appropriate discourse domain, essential for test performance to be more easily interpretable with respect to his/her language use ability.

Implications of ESP teaching methodology for LSP testing

ESP/LSP teaching methodology has direct impact on ESP/LSP testing methodology. First, an LSP test aims at a particular language use context. People taking a test often have a clear idea of why they are taking it. Second, needs analysis is a necessary step in test development; LSP test developers are aware of the test purpose but are sometimes not familiar with the specialist knowledge required. In order to develop a valid test, they carry out a needs analysis on the nature of the tasks and the particular language knowledge associated with those tasks. To achieve this, test developers need to involve subject informants in test development. Third, test input has to be related to the candidates' special needs and the future language use contexts. Because a given LSP test measures the test taker's language performance in a particular context of language use, testing material has to be authentic. Finally, the candidate's level of language proficiency is inferred from his/her understanding and production of appropriate language samples from pre-specified tasks.

1.3.3. Features of LSP testing

Criterion-referencing, directness, performance-orientation, authenticity and the involvement of subjective judgement are features common to many LSP tests (Section 1.3.6).

Criterion-referencing

LSP tests tend to be criterion-referenced. Candidate performance is compared against pre-determined groups of communicative behaviour, used as reference points. Candidates are then assessed on whether or not (or how well) they can perform the sets of pre-specified tasks.

Directness and authenticity

In an LSP test, candidates are asked to use specific skills in specific contexts. In addition, test tasks often reflect the characteristics of the activities candidates are likely to do with language in the real world; therefore, tasks and language skills are sampled and extrapolated from the target language use situations.

Performance-orientation

LSP tests often require some language performance to make the interpretation of the candidate's language ability easier. In language use the whole is bigger than the parts. It is not altogether convincing to infer language ability from tests of discrete points of language. However, depending on the test purpose and the construct defined, performance may not always be necessary.

Subjective judgement

Because performance is a necessary element in LSP tests, judgement of language production are inevitably subjective. Rater training and establishment of a reliable marking scheme are central issues (Section 3.5.).

1.3.4. Attractions of LSP testing

One of the attractions of LSP testing is its high face validity. At least to the test takers, an ESP test appears to measure their ability to use the language in their study or at work.

The second attraction is its diagnostic value. The language requirement in the candidate's

intended workplace or academic institution is often specified in an LSP test; language performance is then evaluated against a set of clearly defined criteria. The test profile matches the communicative needs of the course or workplace requirement for both administrative and pedagogic reasons.

Finally, LSP tests allow for a more exact description of the candidate's desired behaviour in the target language use domain which is restricted, well defined and relevant to the purposes of the learner's language use. Instead of sampling from the whole target language, the test developer can side-step a whole range of problems and concentrate on the language problems associated with a particular context of language use.

1.3.5. Aspects of LSP test design

Alderson (1988b) mentioned three aspects of ESP test design relevant to LSP test design: test content, test method and test validation. As far as test method is concerned, an LSP test tends to be direct and subjective; the focus of measurement is the test taker's ability to use language in specified target situations. Language performance is required; test content and tasks tend to be authentic and representative of the specific language use context and its communicative demands. In terms of test validation, issues such as construct representativeness, generalisability and test impact are considered in addition to the traditional type of validity inquiry. I will now discuss each of the three aspects of LSP test design. Examples of two recent ESP tests in occupational contexts will then be given.

Test content and content selection

Test content provides the specific language use contexts for appropriate discourse to take place. Content analysis is conducted to help the design and selection of test tasks. For the test to be useful and meaningful, present practice in LSP test design favours a correspondence of characteristics between test tasks and domain language use tasks (Douglas, 2000). A test specification (or rubric) consisting of the objectives, the procedures for responding, the structure of the test, the time allotment and the scoring

procedures, is prepared as the basis of test development. Further, a description of the types of language knowledge associated with the given test tasks is usually presented. Procedures such as job analysis, workplace observation, literature review, analysis of texts and/or types of interactions from the target use contexts and consultation with test users and subject experts have been suggested as necessary steps for a valid needs analysis (McNamara, 1996).

Test method

Directness, criterion-referencing, performance-orientation, and the use of subjective judgement are common characteristics in LSP test design (Section 1.3.3.). These characteristics are usually reflected in LSP test methods. The common way of achieving these is to reflect the characteristics of the target language use contexts and simulate the criterion performance in the test tasks. LSP test tasks tend to measure combinatory skills; that is, in a specific test task, candidates may be required to listen and write/summarise or listen, read and write, etc. Test tasks are authentic in that they reflect the types of tasks the test taker may have to deal with or the experience the test takers may have had in completing such tasks. Test items are not limited to a single type but include a variety of item types. Finally, emphasis is placed on the communicative contexts and the measurement of the candidates' language performance.

Test validation

Test validation is concerned with whether the test score is reliable and whether the interpretation and use made of the test score are valid. According to Messick (1989), validity is "a unitary concept", requiring the consideration of different types of evidence in support of the use of the test results. Validity inquiry now usually includes examination of evidential and consequential bases of test use. The entire process is viewed as construct validation which begins at the stage of test design.

Weir (1988) distinguishes between *a priori* and *a posteriori* construct validation. *A priori* construct validation concerns whether the construct of the test at the stage of test

content selection is based on a theory of language and language use. *A posteriori* test construct validation concerns how the construct of the test is empirically and statistically validated. Usually the test is externally validated against a measurement criterion. A serious limitation of concurrent validity is that it only considers the extent to which measures of the same ability tend to agree; it does not tell us anything about the ability we are measuring (Section 1.2.9).

A posteriori validation alone may not be appropriate in LSP tests because each test tackles a particular language use domain, thus making it difficult to find an appropriate measurement criterion. Also, candidates are assessed on both their language ability and other factors that lead to successful completion of the tasks. To examine the claims an LSP test makes, *a priori* validation is necessary and equally important.

In LSP tests, the two concerns in *a posteriori* construct validation are whether the performance demonstrated is the true ability and how objective the raters and the rating scales are. Statistical procedures using IRT Rasch analysis seek to reflect learner ability and the characteristics of the raters and the rating scales(Section 1.4.). Other empirical validation procedures are also necessary; such procedures include collecting feedback from the examinees and the test users, and qualitative item analysis. Research indicates that paying attention to test takers' comments significantly improves the test (Kenyon & Stansfield, 1991).

In short, validation of an LSP test requires a unified inquiry; information gathered in support of the validity of the test should include evidence of the test's construct validity, the consequences and value implications of the test use (Chapter 5). The validation process should start at the stage of test design.

1.3.6. Examples of two ESP tests

In this section, description of two ESP tests in two occupational contexts (EOP) are presented. They are the Occupational English Test (OET) and the Japanese Test for Tour Guides in Australia. Examples of other LOP tests, to name just a few, include the Cambridge Examination in English for Language Teachers (CEELT), the Listening Summary Translation Exam (LSTE) in Spanish and Minnan (a southern Chinese dialect) (Stansfield, 1990; Scott *et. al.*, 1996; Stansfield *et al.*, 2000), and TOEIC (Test of English for International Communication).

Occupational English Test (OET), Australia

OET is a high-stakes ESP performance test for health professionals who want to live and work in Australia. The test was developed in the late 1980s and consists of a speaking and a writing sub-test for all health personnel and a listening and a reading sub-test on profession-specific content (McNamara, 1996). Linguistic ability as well as familiarity with Australian clinical practice are considered essential for successful performance in OET.

OET adopts an empirical approach to ensure content validity. A needs analysis investigating target language use tasks, linguistic features and their relations with the target language use domain was carried out to ensure face and content validity. Subject informants were consulted to establish the construct of the test and to ensure the representativeness of test tasks. For *a posteriori* validation, statistical analyses on item fitness and rater consistency were carried out.

OET is a performance test in the *strong* sense (Section 1.2.8.). This views performance as an interaction of linguistic ability and other non-linguistic and contextual factors like background knowledge, personality traits and communicative strategies used in the target situation (McNamara, 1996). Adequate language proficiency is a necessary but not sufficient condition for successful performance. The target for assessment is performance of the task itself. To measure candidate ability, OET tasks represent target language use,

and performance is judged on real-world criteria. However, McNamara (1996) notes that a pure strong performance test is not easy to construct because of practical difficulties. Also test developers face the problem of weighting linguistic and non-linguistic factors in the assessment. To take account of this, the assessment criteria of OET are defined communicatively on *a priori* grounds with a component of overall impression weighted more heavily. These criteria are later empirically validated by the use of Rasch Analysis (McNamara, 1996).

Japanese Language Test for Tour Guides, Australia

The Japanese Test for Tour Guides is a face-to-face interview with six phases, lasting about 30 minutes (Brown, 1994 & 1995). Like OET, the test is a strong performance test. There two purposes: (1) to indicate to the employer the candidates' language ability through an optional certification process and (2) to select candidates for the TAFE (Technical and Further Education) Japanese tour-guide training courses. Data from various sources are collected for test development. They include direct observations of tours and taping the interactions for later analysis, interviews with expert informants and test users, and literature review (Brown, 1994).

Both linguistic skill and task fulfilment are assessed (Brown, 1994). One difference from other LSP tests is the use of non-native speakers and professionals from the tour guiding industry as raters (Brown, 1995).

To conclude, the two tests described above share some common features:

(1) The needs of the target population , the test users and the target language use contexts are clearly defined;
(2) test takers are assessed directly in either simulated or real-life tasks;
(3) test validation includes both qualitative and quantitative approaches;
(4) the validation process starts at the very beginning of test development; and
(5) the candidates' performance is reported in a profile with clear descriptions of what they can do.

Implications for specific purpose test design and development are:

(1) The purpose of a specific purpose test should be very clearly defined.
(2) The provision of specific language use contexts is essential in engaging the test taker in the appropriate discourse domain.
(3) A needs analysis on the nature of the workplace and the language context is essential for establishing the test's construct validity. This means the inclusion of lay test users and subject experts in test development.
(4) Candidates are measured precisely on the skills necessary to be successful in the workplace.
(5) Qualitative and quantitative approach to test validation are equally important.
(6) Score reports should include descriptive statements of what a candidate can do.

1.3.7. Some issues in LSP testing

The movement of LSP testing, in particular ESP testing, fits neatly with the concept of ESP teaching. LSP tests seem to be potentially able to give a reliable indication of whether the candidate is proficient enough to carry out tasks that will be required. Although specific purpose language testing seems to be a logical consequence of ESP teaching principles, it remains somewhat problematic. Three frequent questions are: (1) the criteria for assessment, (2) specificity, and (3) its practicality (Hutchinson & Waters, 1987; Alderson, 1988; Fulcher, 1999).

Criteria for assessment

As already noted, LSP tests are primarily criterion-referenced and performance-based. One advantage of criterion-referencing is that behaviour criteria are pre-specified and may be known to both language users and candidates. Test scores are interpreted in terms of a profile which might encourage the test user to be explicit about what they want and expect the candidate to be able to do with language (Spolsky, 1990). However, criterion-

referencing is not without practical difficulties. It has been argued that situations suitable for criterion-referencing are not as common as many people believe (Alderson, 1981; Skehan, 1984). Moreover, it is difficult to specify what language-related requirements are needed to perform a particular task and what linguistic demands are essential for successful performance (Alderson, 1981; Porter, 1994).

Relating test performance to external criteria that can be applied to real-life tasks is difficult. It is not certain on what basis one should group particular language features or skills and what evidence one needs for such grouping. At the moment, rating tables similar to the ACTFL scale are used to provide a guideline for the grouping of tasks. Similarly, hypothetical hierarchies of skills postulated by language researchers such as Richards' 1984 listening taxonomy have been popular. The problem with the use of the ACTFL guidelines is that descriptors make specific references to native speakers, whose performances are assumed to be homogeneous. However, studies of native-speaker performances on TEEP and IELTS indicate that native-speaker performance, though better than non-native speakers, are not uniform and far from perfect (Weir, 1988; McNamara, 1996; Clapham & Alderson, 1997). The problem with the use of a skills taxonomy is that there has been no empirical evidence in support of the skill hierarchy (Dunkel, 1991).

The final question raised about the assessment criteria is how far the language tasks can predict the candidate's later performance. Research on EAP has shown that language proficiency only plays a moderate role in students' ultimate academic success (Bailey, 1985; Light, *et al.*, 1987; Criper and Davies, 1988; Ferguson & White, 1994). Other research has recognised the influence of non-linguistic factors such as affective variables and personality traits on the candidate's overall performance (Criper & Davies, 1988). In sum, there are many other variables that influence the candidates' ultimate performance in a given field.

Specificity

The specificity of ESP testing has been a major area of discussion. Since the early 1980s, research has been carried out on the specificity of ESP tests, in particular on the effect of background knowledge on test performance. Most research investigated the effect of background knowledge on ESL reading comprehension but some research examined the candidates' specialist knowledge in listening and speaking.

Alderson and Urquhart conducted three studies on the effect of student academic discipline on test performance (Alderson & Urquhart, 1985). The findings showed contradictory results in student scores on the different test modules. They concluded that prior knowledge had some effect on test scores, but the effect was not consistent and they called for future research to take account of linguistic proficiency and other factors.

Studies by Koh (1985), Hale (1988), Hock (1990) and Tan (1990) indicated that background knowledge helped reading comprehension but the test taker's lack of specialist knowledge may be compensated by his/her linguistic proficiency. In the study by Levine & Hause (1985), background knowledge helped students with advanced language proficiency more. The work by Clapham (1996) on the effect of background knowledge on reading comprehension suggests that language proficiency seems to be as important as background knowledge in the comprehension of reading texts and that science students performed better with science-based texts than other students but as well as the humanities students on humanities-based texts. However she notes that background knowledge is not easily assessed and that the effect of background knowledge depends on students' proficiency level and the specificity of the reading passage.

Jensen and Hansen (1995) compared the effects of prior knowledge on the listening performance of university students. Their findings were that linguistic proficiency has a stronger effect in listening; the effect of background knowledge was only significant in the case of technical passages. However, studies by Long (1990) showed that background

knowledge plays a crucial role in the listening comprehension process. Similar results were found in the Schimidt-Rinehart (1994) study in that topical familiarity improves listening score.

A study by Smith (1989) on some international teaching assistants (ITAs) showed ITAs as a group did not differ significantly in performance if tested in their specific field of study rather than in a general topic area. However, when individual ITA performance was looked at, 4 out of the 38 ITAs would have passed and 4 others would have failed if they had taken the field-specific test. Smith argues that the study results cannot be used to challenge the continued use of the more general "SPEAK" test to predict ITAs' oral proficiency in the classroom. Similar results were obtained by Douglas and Selinker (1993). Finally, results of a study designed to investigate the effect of context on test performance suggest the context-specific test is a better predictor than the general-purpose test (Douglas & Selinker,1992).

Studies using bias analysis have yielded inconclusive results. Henning (1990) investigated specialisation item bias in proficiency/placement tests. The results showed that the effect of systematic specialisation bias was not significant.

Four points emerge from these studies:

(1) Language proficiency plays an important part in test performance.

(2) Background knowledge seems to be relevant to test performance. However, candidates with higher proficiency level may manage to achieve a higher score without background knowledge.

(3) In listening comprehension, familiarity of the topic seems to improve listening scores.

(4) Statistically, performance in an ESP test is not significantly different from that in a general language test, but individual differences in the two types of test performance do exist.

The implications are that candidate performance is influenced by both linguistic and non-

linguistic criteria. One requirement of LSP testing is to provide predictive information on the candidate's language ability in specific language use domains; test tasks have to reflect the language use areas and the content that the situation embodies. However, emphasis should be placed on the measurement of the test taker's ability to use language through the target language use contexts to achieve communication goals.

Practicality

Two practical issues relating to LSP testing have received attention. The first one concerns the provision of more detailed information on candidate performance; the second one is the expense of LSP testing.

Spolsky (1990) proposes that test results should be "full and detailed" in the form of a profile. There are two reasons for this detailed profile: one is for the tester to provide sufficient, accurate and interpretable information on the language ability for the purpose of the test, the other reason is for the test users to be careful in interpreting test scores. Unfortunately, most test users do not usually look at component scores; rather they rely only on the global score (Clapham & Alderson, 1997).

At present, there is a tendency to base EAP testing on a common-core approach (Alderson & Clapham, 1992). It is essentially much more economical to look for common ground between different subject areas. But there are problems with this approach. First, whereas research has indicated the existence of a common-core in the area of study skills, more research is needed to establish what should be included in the core and what should be left to subject specific work (Dudley-Evans, 1988). Second, while candidates are being assessed on their study skills and language use ability, there is no conclusive evidence that the skills will transfer.

EOP testing, on the other hand, attempts to assess how effectively the candidates can use the language in the workplace; content and test task selection is therefore crucial. In occupational testing, the work sample approach seems common (Section 1.2.8.). In test

design, test tasks will represent real world tasks that candidates are going to carry out. Performance is the target of assessment. The work sample approach is expensive and time consuming. There are also practical constraints associated with this approach: degrees of representativeness of test tasks, the nature of performance in a specific purpose test and the effect of background knowledge on test performance; all of which affect the validity of the test. The representativeness of tasks can be enhanced by steps like needs analysis, workplace observation and consultation with subject informants. But further research is necessary on the nature of performance in LSP tests.

Performance, according to the cognitive-psycholinguistic/interactionalist approach, is influenced by the interaction of language use contexts and personal characteristics. To more accurately measure language ability, this approach requires the language tester to consider and specify relevant aspects of language knowledge and contexts. Language ability is a relative concept and is context-dependent in that the performance is a sample of a similar language use to the target language use context.

This approach has its difficulty to overcome as well. As a language tester takes the construct and the context into consideration of the test design, s/he has to specify the language knowledge and processes required in the given context; s/he also has to elicit from the context the defined constructs in test performance. Replicability is another issue to consider in test validation. Presently, test validation using this approach takes in judgmental, empirical and consequential justifications of the given test use (Section 1.2.4.).

In sum, LSP testing is theoretically motivated. However, there are practical problems to overcome. Perhaps the problem is that language use and language performance are complex. It is impossible to explicitly state what the criteria are except in a very small number of tightly defined contexts such as air-traffic English. A more realistic approach may be to establish the major language dimensions of language use in any given target situation, and clarify the important language constructs involved. This requires a

framework describing areas of language knowledge relevant to the target language use situations, and the interaction of language knowledge with topical knowledge, affective schemata, and the strategies involved in a testing situation.

1.4. Test analysis

There are two parts to this section: (1) to explain the two approaches used to report test results of the main trial, namely, classical test analysis and the one-parameter IRT Rasch measurement model, and (2) to describe the two Rasch programmes used for test analysis, namely, *Quest* (Adams & Khoo, 1993) and *Facets* Version 3.2 (Linacre, 1999). Assumptions made in each measurement theory will be discussed and implications of each in test result interpretation will be offered. The purpose is to present essential notions in the two theoretical frameworks for the ease of test interpretations in Section 4.2.; some concepts may overlap with concepts discussed in Section 1.2.

The classical and Rasch approaches to the analysis of test data both have their roots in theories of testing developed within the context of psychometric research. Classical measurement is concerned with test analysis that relies heavily on the correlation coefficient as a statistical procedure. The Rasch model, on the other hand, looks at the probability or improbability of scoring individual items and person response patterns on the basis of the overall pattern of responses in a test. There are important differences between the two approaches, but it may be best to look at them as partially overlapping rather than as rival frameworks for quantifying test data.

1.4.1. Classical measurement

In classical test theory, a person's test score is assumed to comprise his/her true score and an error component. The true score is viewed as remaining constant. The error element refers to any non-systematic variation observed in a person's score from one administration to another or one test form to another. Errors of measurement, with a

mean of zero, are independent of the true score and occur randomly (Guilford & Fruchter, 1978).

Some further concepts related to classical test analysis and test interpretations are test reliability, item analysis, and test validity (Also see Sections 1.2.3. & 1.2.4.)

Test reliability

Reliability concerns the consistency of measurement. The reliability index (i.e., reliability coefficient) of a test is intended to indicate the proportion of the sample's observed score variance due to the sample's true score variance (Guilford & Fruchter, 1978). This assumes that the variance of scores obtained for a group of people on a given test is equal to the variance of their true scores plus the variance due to random error. A perfectly reliable test would have a reliability coefficient of one, which would mean the differences of observed scores reflected the differences of true scores and error played no part.

Procedures for estimating test reliability include the following: (1) test-retest reliability, (2) parallel-forms reliability and (3) internal consistency reliability (See Section 1.2.3.). When scores on a test are rated by two or more judges, an inter-rater reliability is computed (Henning, 1987). The use of a given type of reliability estimate depends on the type of instrument and the purpose for which statistics will be used. The test-retest method is more appropriate for a heterogeneous test, while the split-half method, one variety of internal consistency reliability, is meaningful for power tests but not for speed tests (Guilford & Fruchter, 1978). If a test consists of homogeneous items, internal consistency estimate is the appropriate method.

Reliability, as Guilford and Fruchter (1978: 408) point out, is a stability estimate for a given measurement, not for the measuring instrument itself. They write:

> *It can rarely be said of any instrument, whether a test or some other device, that the reliability of the device is of a certain value, usually in the form of a coefficient of correlation. One should speak of the reliability of a certain*

instrument applied to a certain population under certain conditions.

Another way to look at the consistency of a test set is to compute the standard error of measurement (SEM), which is the standard distribution of measurement errors. The purpose of SEM is to estimate an average of the distribution error deviations across all test takers in a given test administration. Like the reliability coefficient, SEM is reported as a single global statistic for the entire group of test takers and the percentages are based on a normal distribution (Henning, 1987).

Item analysis

The major goal of item analysis is to achieve test reliability and validity. Item parameters in the classical approach are defined in terms of item difficulty level and item discrimination index; item difficulty is defined as the proportion of correct responses and is indicated by an item facility value and item discrimination indicates the degree to which each item differentiates between high- and low-scoring test takers. The importance of these two statistics is that they help the test developer to distinguish between better and poorer items and that they are used in the selection of items to produce tests with certain properties to guarantee reliability. Guidelines on the computation of the two statistics and the level of difficulty and discrimination are provided in many textbooks (Henning, 1987; Brown, 1996) and will not be discussed here.

There are a number of general points regarding classical item difficulty and item discrimination. The difficulty of an item vary according to whether the group is of high or low level for the trait being measured; thus the item facility value not only indicates the difficulty level of an item but also the group tested. Item difficulty is very much sample-dependent and the facility values only remain stable for groups of a similar level. Similarly, item discrimination obtained by separating high and low groups is affected by the distribution of the trait measured within the sample and varies from one sample to another. Another commonly used item discrimination index, the point biserial correlation, has also been found to vary systematically with item difficulty and with the group tested (Wright & Stone, 1979; Baker, 1989).

Finally three assumptions are made for the statistics discussed. First, people who fail to answer correctly do not have a sufficient level in the trait being measured. Second, test takers are given sufficient time to complete the test. Lastly, the effect of correct guessing is assumed to be minimal (Guilford 1954).

Validity

Test validity refers to the degree to which test scores predict the trait measured and includes face validity, content validity, construct validity, predictive validity and concurrent validity (See Section 1.2.4.). There are different approaches to test validation but the main concern of this section is with statistical procedures.

In the classical approach, test validity can be reported empirically in terms of the following:

(1) Validity coefficient: The validity coefficient is indicated by the correlation of the score and an external test criterion (Guilford, 1954). It is a type of correlation coefficient usually using the Pearson product-moment coefficient and is based on the assumption that there are some common factors shared between the test and the criterion being measured. Once a/the validity coefficient is available, the correlation between reliability and validity can then be computed. Concurrent validity and predictive validity can also be reported with some form of correlation coefficient. The resultant correlation coefficients are respectively called concurrent validity and predictive validity.

(2) Discrimination values of the items: Item discrimination indicates how well an item separates high- from lower-level groups. It also expresses a relationship between the item and the criterion being measured. The higher the discrimination values are, the greater the correlation is with the criterion (Guilford & Fruchter, 1978).

(3) Construct validity: Construct validity can be examined empirically. The purpose of construct validation is to provide evidence that the underlying traits being measured are valid. Unlike concurrent validity or predictive validity, construct validity does not

have any one particular validity coefficient associated with it because it is not mainly a matter of numbers (Henning, 1987). Often construct validation involves a series of tests under different pre-specified conditions. The following approaches have been suggested (Guilford, 1954):

a. group differential approach
b. correlation among tests measuring the same variable
c. correlation among items or sub-scales of a test

Procedures for each of the method can be found in many testing textbooks such as Henning, 1987 and Brown, 1996.

One final approach for determining construct validity is through the use of factor analysis. The procedure is similar to (b): correlation among tests measuring the same variable. It involves simultaneous comparison of correlations among two or more traits under two or more methodological conditions (Guilford, 1954; Campbell & Fiske, 1959). However, there are a number of practical problems related to the use of the classical factor analysis model.

(1) The model is based on the normal distribution. Item difficulty level and test difficulty level should be of moderate difficulty so the distribution of scores is symmetrical. Corrective measures have to be taken if the distribution is skewed.
(2) Homogeneous samples are required and the size of sample needs to be larger than 200 (Guilford, 1954).
(3) The classical model does not fit dichotomous variables because common factor variables are conceived as being continuous (Blais & Laurier, 1995).

There are also a number of problems related to the interpretations of factor loading. First, the first factor tends to account for much of the variation in the data. The second problem concerns the difficulty of interpretation when there is little or no variation in ability level in the data set (Gustaffson, 1977).

Traditional test scoring and scaling

This section concerns assumptions of person scores on tests of dichotomously scored items reported as number correct either in percentage or derived terms. The use of number correct scores implicitly assumes that (1) the test is uni-dimensional, (2) that the items discriminate equally and (3) that each item represents an equivalent unit of measurement and forms the basis for an equal interval scale (Baker, 1989).

For purposes of reporting, interpreting and comparing scores, classical test theory has been using various types of derived scales other than the one provided by simple counts of correct responses. For example, standard scores or "z-scores" are obtained by expressing raw scores in terms of standard deviation units above or below the group mean (Hatch & Lazaraton, 1991).

Another commonly used normalised scale is the standard T score. The mean of the T score is set at 50 and the standard deviation at 10 (Hatch & Lazaraton, 1991). The stanine scale with a mean of 5 and a standard deviation of 1.96 is yet another common scale but is less sensitive and discriminating than the T score (Henning, 1987). The main justification for the use of a normalised scale is that the normal curve has many convenient properties. But the normalised scale distorts the shape of the original distribution into a normal distribution.

On the whole, one advantage of a normalised scale is that it makes comparison of the relative standing of persons easier. But the use of a standardised scale has been criticised as one form of score manipulation (Brown, 1980; Parkinson, 2000).

Sample dependence

Classical test statistics depend heavily on the representativeness of the sample and its size. In the case of a small sample of 20 – 30 people, small sample statistics are required to adjust sampling distributions departing from normality (Guilford & Fruchter, 1978).

General remarks

Classical test theory has contributed greatly to the design, construction, pre-testing, scoring, statistical analysis and interpretation of conventional tests. However, it has been acknowledged that classical measurement theory has certain limitations. The first one is the assumption of normal distribution of populations. The normal law of distribution may roughly describe some phenomena like gambling but may not work well for psychological, social or educational events (Parkinson, 2000). With the use of psychometric methods, the outcomes of any investigation are normalised. The next problem is in deciding which reliability coefficient correctly indicates the proportion of true score variance when a number of such coefficients are obtained using different procedures (Thorndike, 1982). Third, the test scoring and scale are based on procedures involving reference to particular groups or item sets; thus, interpretation of scores requires knowledge of the characteristics of the given sample and items. In terms of sampling, the classical test theory is very much sample dependent in item selection and estimation of reliability.

1.4.2. Rasch measurement

An alternative approach to item analysis and interpretation of item and test data is *item response theory* (IRT). It refers to three families of analytical procedures identified as the one-parameter (the Rasch Model), the two-parameter and the three-parameter logistic models. These models differ in complexity depending on the number of item parameters incorporated. The one-parameter model takes account only of item difficulty; the two-parameter model also has a parameter for item discrimination; the three-parameter model attempts to take account of the possibility of guessing. All these models involve a person ability parameter. Thus the one-parameter model involves a total of two parameters: item difficulty and person ability. The focus of the following discussion is on the one-parameter Rasch model because it is readily applicable to small-scale research. However, the central concept of any IRT model is that a relationship is specified between the observable performance of test items and the unobservable characteristics assumed to

underlie the performance. These characteristics are conceived as forming a continuum represented by a numerical scale. A person's standing can then be estimated using his/her responses to suitable items. Items measuring a given trait are also seen as being located on the same scale (Wright & Stone, 1979; Wright & Masters, 1982).

The Rasch model

Since its introduction by Georg Rasch, the Rasch model has been greatly expanded and is now a generic term for a family of models (Andrich, 1978; Wright & Masters, 1980; Linacre, 1991). The Rasch model is a measurement model; that is, it is based on mathematical formulation. The persons and items are graded for ability and difficulty, and judged according to the probability or likelihood of their response patterns on the basis of the observed person ability and item difficulty. The criterion for a successful analysis is that the data fit the model. The degree of fit of the data can be calculated and the acceptability of the measures determined.

Assumptions of the Rasch model

The Rasch model is characterised by the concept of *specific objectivity*, i.e., the separability of person ability and item difficulty parameters in a test (Wright & Masters, 1982). Three properties are related to specific objectivity. First, person ability and item difficulty parameters are independent of each other but they combine to give the probability of getting a correct response. Second, the method of combination is *additive*, meaning that the item and person parameters can be expressed on a linear scale. Third, specific objectivity assumes that the estimate of any parameter is dependent on all the responses but is not affected by any unexpected value of them. In other words, the responses are the *sufficient statistic* for the estimation of the parameter. Any unexpectedly high or low score does not inflate or reduce a test taker's ability estimate, but will affect the validity of the estimates and would cause misfit of the data (Wright & Masters, 1982). While item difficulty varies, the discrimination power is assumed to be the same for all items.

Two further assumptions are required in the Rasch model: local independence and uni-dimensionality. Local independence refers to the assumption that an individual's response to an item is independent of his/her response to any other item. Uni-dimensionality refers to a single underlying measurement dimension. The trait measured, whether simple or complex, have to be approximately the same for all the items in the test. In other words, this assumption requires the items to be relatively homogeneous.

Ability and difficulty estimates

Central to the Rasch model is the advantage that item difficulty and person ability can be placed on the same scale. Thus, via this scale (an equal interval scale) the connection between items and persons can be made. This can be derived by estimating the chances of a test taker achieving a certain score on an item of a given difficulty from the data. The ability and difficulty estimates are represented by numerical values in terms of "log odds units" or *logits*. The mean value for both ability and difficulty is arbitrarily set at zero logits (Linacre, 1999). Items with an above-average difficulty will have a positive sign and those below will be negative in sign. Candidate ability is related to item difficulty. A test taker with ability at 0 logits will have 50% chance of scoring an item of average difficulty correctly. It should be pointed out that according to the Rasch model, item difficulty is an intrinsic property of items, conceived of as being independent of the abilities of any given group of test takers. Similarly, person ability is seen not as an index to a particular test but as representing the ability level of the measured trait (Baker, 1989).

Evaluation of fit

The item and ability estimates are meaningful only when there is a satisfactory fit between the model and the data. The notion of fit concerns the degree of match between the model's expectation of the item-person relationship and the actual set of response data. The measure of fit can be expressed either in terms of a mean square residual or a *t*-value (i.e., the standardised mean square). When the data fit the model, the mean square has an expected value of one and the *t*-value has a mean near zero and a standard deviation close to one. Different Rasch-based programmes suggest slightly different

ranges of acceptable mean square values for item/person fit; Quest suggests 0.75 to 1.3, whereas Facets suggests 0.7/0.8 to 1.2/1.3. Values greater than the upper limit indicate significant misfit – that is, lack of predictability. Values less than the suggested lower limit indicates significant overfit – that is lack of variability. Fit values expressed in *t*-distribution will also be positive and negative according to whether the observed value show greater variation (positive t-value) or less variation (negative t-value). Acceptable t-values are suggested to fall within the range of +2/3 to –2/-3 standard deviations. Values greater than +2/3 indicate significant misfit; values less than –2/-3 show significant overfit. Generally items/persons identified as showing significant lack of fit are termed misfitting items/persons. Their response patterns do not conform to the expected patterns according to the model. These items and person responses need further examination.

When an item is described as misfitting, it indicates that the item may be flawed in some way or that it may not tap the same ability as the other items. It needs to be investigated because the quality of the test and ultimately the validity of any decisions made based on the test score may be affected.

The lack of fit for an individual indicates that the model is an inappropriate means of describing the behaviour of that person on the set of items. If most candidates in a group have responded largely in accordance with the model's expectations, a few instances of misfit can be attributed to anomalous test-taking behaviour such as fatigue, guessing or lack of interest. However, if a significant level of person misfit happens, according to the model, the instrument may not be an appropriate means to measure the trait defined. The test will need revision to reduce the number. McNamara (1996) suggests 2% of the candidates showing person misfit as the maximum limit. However, Hambleton and Swaminathan (1985) noted that sample size could have great influence on tests of fit of the chi-square type such as the Rasch model. Second, fit values in well-controlled data such as MCQ responses are more central than those obtained from free-form responses (Linacre & Wright, 1992). Also, as Rasch (1980) points out, person ability estimate is

usually less accurately measured than item difficulty estimate; an item can be calibrated on an indefinite number of people but the ability of a person can only be measured on the limited items in the given instrument. When interpreting person fit, the nature of the data and the size of the sample may have to be considered.

Test reliability and validity

Test reliability is concerned with consistency of measurements. In a norm-referenced test, reliable measurement is indicated by the tendency to rank order the test takers in the same way on repeated measurement of the test (Henning, 1987). High variance of score distribution or great distance between scores of a test is less likely to be reliable. The traditional true-score approach defines reliability as the proportion of variability in observed scores attributable to the variability in true scores. Reliability estimates depend on the observed sample variance and a test error variance. There are various ways to derive a reliability estimate (Section 1.2.3.).

In the Rasch model, a third factor is considered, namely, the extent to which the items work together to define a variable (Wright & Masters, 1982). Error variance is considered to have two parts: a modelled error variance and the consistency of the items as one variable, and is estimated by the overall fit mean square for the test and the sample. Thus, the Rasch reliability estimate deals with three components: the mean square, the modelled error variance and the observed score. The modelled error (S.E.) tells us how precisely we are able to estimate a person's ability when the items are internally consistent. Reliability is defined as the "proportion of the observed sample variance not due to measurement error with which the test separates the persons" (Linacre, 1999). It is a correlation reporting the repeatability of a local combination of the test and the sample. Two statistics are of interest in a reliability report: (1) reliability estimate, referring to a measurement of the person ability/item difficulty with measurement error removed and (2) person separation, which is the ratio of the sample error-adjusted SD to the average measure of SE. A person measure with reliability 0.98 and a person separation of 7.32 will be interpreted as the candidates being well separated

in ability, by about seven times more than the measurement of error. If the sample is normally distributed, there are about 10 measurably different levels of ability in this sample ({(4x7.32+1)/3} and the replicability of the measurement under the same condition is 0.98.

Test validity in classical test theory is divided into internal validity and external validity. In the Rasch approach, test validity is estimated in terms of the fit statistics of each item to the model independent of the sample distribution (Wright & Masters, 1982). The mean squares are standardised into fit statistics with an expected mean of near zero and SD near one. If the fit statistics of an item are acceptable then the item calibration is valid. The internal consistency of person response patterns can be examined by way of the fit statistics as well. If the fit statistics of a person's performance are acceptable, then their measure is valid; however, perfect fit to the model rarely happens (Smith, 1987).

Checks on model assumptions

The fit statistics provide information as to whether the data satisfy the assumptions required by the model. There are various procedures, in addition, to check on specific assumptions.

One assumption of the Rasch model on the item parameter is that items do not vary in discrimination. Sometimes, it has been suggested that in application of the Rasch model the item-score correlations such as biserials or point biserials be computed to determine the extent of variation in discrimination. Most Rasch-based programmes have this function built in.

The assumption of unidimensionality requires that the items are homogeneous in a data set. Thus, the aim in checking the unidimensionality of the data set is to detect possible sources of heterogeneity. Factor analysis has often been suggested (Baker, 1989).

Bejar (1980) suggested plotting pairs of item difficulty estimates obtained from the

complete data set against those obtained from subsets of the data. However, Spurling (1987) considered the comparison of ability estimates obtained for the same persons using the complete data set and the subset to be more appropriate since the test-based and subset-based item estimates were based essentially on the same information and could not be expected to depart from unidimensionality.

How to check whether the assumption of local independence is violated is another important issue. One method is to test items in different contexts and then examine the effect of this on the item parameter (Yen, 1980).

1.4.3. Comparison of classical measurement and Rasch measurement

Classical measurement and Rasch measurement are overlapping psychometric methods. They therefore share some of the assumptions underlying the estimates of internal consistency of classical test theory, including local independence, unidimensionality, and speededness. Speededness refers to the time given to complete the test being sufficient; it means the test given has to be a power test. The Rasch model further assumes that the difficulty/ability is a true interval scale and that the differences between the observed scores and expected scores are normally distributed (Linacre, 1999).

The difference between the two approaches lies largely in the degree of dependence on the sample. Classical test theory is very much group-dependent; the ability of a given sample affects test statistics and, ultimately, the interpretations and the use of test results. The Rasch approach looks at the probability of a person scoring an item on the basis of the overall response patterns. Item difficulty is conceived as being independent of the sample. Similarly, person ability is independent of item difficulty. Rasch measurement has attracted some criticisms (Goldstein, 1979; Brennan, 2000). It has to be stressed that one approach to test analysis and interpretation is not better than the other. The best way is to look at both measurement theories as examining test reliability and validity from different angles.

Finally, the assumptions underlying any measurement theory will never be completely satisfied. The choice of a measurement procedure is a matter of selecting a coherent approach to measuring data. The concern here is that test results can be validated and thus interpreted meaningfully for future use and research.

1.4.4. Rasch models and test data

In this section, I will outline four of the six Rasch models commonly used in the analysis of language test data. They are the Rasch dichotomous model, the partial credit model, the rating scale model and the many-facet rating scale model. I will also give a brief description of the rationales underlying the two computer programmes *Quest* and *Facets* used in the main trial test analysis. The purpose is to provide basic information needed to understand test interpretation from the Rasch perspective.

1.4.4.1. The family of models

The dichotomous model

The dichotomous model is the basic Rasch model introduced by Georg Rasch in the 1950's (Linacre, 1999). It deals with dichotomously scored test data, i.e., correct or incorrect responses. Dichotomous items are thought of as one-step items. If this one step is completed, the candidate scores 1 on the item. If it is not completed, the person scores 0. The probability of the person scoring 1 is dependent on his/her ability and the difficulty of the item. The dichotomous model is expressed as

$$P_{nik} = B_n - D_i$$

Where P_{nik} is the probability of Person n responding in Category k in Item i

B_n is the ability of Person n and

D_i is the difficulty of Item i.

The dichotomous model is the basic Rasch model. A detailed description of the model and its estimation procedure is available in Wright & Stone (1979).

The partial credit model

The partial credit model is an extension of the basic dichotomous model developed by Masters (1982). It handles data from items scored using partial credit scoring. A partial credit item of two steps has performance levels of 0, 1, and 2. The probability of a person scoring 1 on this item is identical to that in the basic dichotomous model. However, since there are more than one level of performance in this item, the first step is not the only step in this item. The second step from level 1 to level 2 can be taken only if the first step from level 0 to level 1 has been completed. The person's probability of completing step 2 depends on his/her ability and the difficulty of this step. The probability of a person scoring each step is estimated independently. The partial credit model can be expressed as

$$P_{nik} = B_n - D_i - F_{ik}$$

Where P_{nik} is the probability of Person n responding in Category k in Item i

B_n is the ability of Person n

D_i is the difficulty of Item i and

F_{ik} is the difficulty of scoring Category k in Item i.

The central concept of the partial credit model is that the step structure of a partial credit item may vary from item to item and from test to test.

The rating scale model

The rating scale model was developed by Andrich (1978). It is also an extension of the basic Rasch model. The rating scale model handles scores derived from rating scales in general. The points in a rating scale are considered to be ordered. A person's probability of choosing 2 on a 4-point scale (0 – 3) is interpreted as that of choosing 1 over 0 (i.e., first step taken) and also 2 over 1 (i.e., the second step taken) but failing to choose 3 over 2. The relative difficulty of the rating points should not vary from item to item. The

rating scale model is a simplified partial credit model (Wright & Masters, 1982). The model can be expressed as

$$P_{nik} = B_n - D_i - F_k$$

Where P_{nik} is the probability of Person n responding in Category k in Item i

B_n is the ability of Person n

D_i is the difficulty of Item i and

F_k is the difficulty of scoring Category k.

The many-facets Rasch model

The facets model was developed by Linacre (1989) who extended the partial credit model to include the rater aspect. Also, Linacre (1994) showed that it is possible to include in further aspects such as task difficulty, judge interaction etc. to the model. Each aspect in the rating situation is called a facet in the computer programme *Facets* , which Linacre developed (1987). The facets model is an expansion of the partial credit model (Linacre *et al.,* 1997). The basic Rasch model has been expanded to include parameters describing the partial credit steps associated with an item. The facets model further expands the partial credit model to include raters in the measurement process. According to the model, four factors dominate the rating awarded to a candidate: the ability of the candidate, the difficulty of the task performed, the severity of the rater and the way in which each rater applies the rating scale. The basic facets model can be expressed as

$$P_{nijk} = B_n - D_i - C_j - F_k \quad (1)$$

Where P_{nijk} the probability of Person n achieving a score k from a rater j

B_n is the ability of the person

D_i is the difficulty of the item i

C_j is the severity of the rater j and

F_k is the difficulty of the step up from category k-1 to category k.

This model is for a rating scale that is the same for all raters on all items (Linacre, 1997). Three more models are available to explain different rating situations. They are:

$$P_{nijk} = B_n - D_i - C_j - F_{jk} \quad (2)$$

Where F_{jk} is the difficulty of the step from category k-1 to category k for rater j and other parameters are defined as before.

This model is for a rating process in which each rater uses his/her own interpretation of the rating scale.

(3) is an item-scale model in which each item is constructed with its own rating scale:

$P_{nijk} = B_n - D_i - C_j - F_{ik}$ (3)

Where

F_{ik} is the difficulty of the step from category k-1 to category k of the scale for item i.

The last one is a many-facet model used in the computer programme FACETS to explain the interaction of persons, tasks, items and raters (Linacre, 1999). It is expressed as

$$P_{nmijk} = B_n - A_m - D_i - C_j - F_k \quad (4)$$

Where P_{nmijk} the probability of Person n achieving a score k from a rater j

B_n is the ability of the person

A_m is the challenge of the task m

D_i is the difficulty of the item i

C_j is the severity of the rater j and

F_k is the difficulty of the step up from category k-1 to category k.

All of the formulations discussed are mathematical equations of probability. These equations are prescriptive rather than descriptive in the measurement of item difficulty, person ability or rater severity in order to provide a common scale to make comparison easier. Properties of the original Rasch dichotomous model are also shared by the four related models, namely "specific objectivity" (i.e., separable item and person parameters) and the assumption of unidimensionality (Section 4.2.2.).

1.4.4.2. Quest and Facets

Quest (Adams & Khoo, 1993) and *Facets* Version 3.2 (Linacre, 1999) are two computer programmes on the Rasch measurement. Both deal with dichotomous and polytomous test data but *Facets* further includes judge mediation. Available analyses of the two programmes include the basic Rasch model, the rating scale model and the partial credit model. *Facets* offers an extended partial credit model to include rater behaviour. Available computer output regarding item, person and rater is listed in Table 1-2 below. Test results of the TG main trial will be presented and discussed in Chapter 4.

Table 1-2: Computer output of two Rasch programmes

Quest used to analyse:	*Facets* Version 3.2. used to analyse:
Listening and Grammar Tests	Speaking Test
Data type: dichotomous	Data type: polytomous
1. Item estimates and its summary 2. Item analysis 3. Person estimates and its summary 4. Item/Ability scale 5. Item and person fit 6. Kidmap	1. Data summary report 2. Iteration report 3. Unexpected responses report 4. Measurable data summary 5. All facet vertical ruler, i.e. all facet scale 6. Individual facet summary report 7. Candidate measurement report 8. Rater report 9. Item report 10. Catgeory report (i.e., rating categories) 11. Scale structure report, probability curves and expected score ogive

1.5. Summary

This chapter has four parts: discussion of the frameworks of language and language use, concepts in language testing, theoretical justification of LSP testing and a brief outline of the two statistical procedures used for item analysis.

Recent advances in the understanding of communicative language use and specific purpose language testing have provided language testers with theoretical justifications for the design and use of LSP tests. Candidates in a test of communicative skills not only

have to show they possess the necessary language knowledge and the knowledge of appropriate use, they also have to demonstrate this knowledge in a situationally appropriate and meaningful manner. This has particular implications in LSP test design: interaction of language knowledge and language use context, relative competence of language use, performance-orientation and the notions of task authenticity, interactiveness, and direct testing.

ESP teaching methodology also contributes important notions to LSP test design: a clearly identified test purpose, investigation of the language demand to be faced in the candidate's future workplace, and interpretation of language ability in terms of the test performance. However, there are practical constraints concerning the development of an LSP test; it is difficult to specify with certainty what language-related requirements and processes are necessary to fulfil a task, and what linguistic demands are essential for successful performance. Equally uncertain is the role of non-linguistic factors in performance and if test performance can transfer to real life.

Chapter 2: Research design

This research has two objectives: development and validation of an ESP Tour Guide English Test (TG Test) for the selection of Taiwanese tour guides.

The current TG Test is a general proficiency test, which has some inadequacies as a measurement instrument of the language ability of potential tour guides. An ESP test, on the other hand, has several apparent strong points (Section 1.3.). In this proposal, the reasons for the use of LSP testing methodology in the design of the TG English Test will be explained in Section 2.1. In Section 2.2., the two research questions will be presented; they are (1) whether the new test is valid and (2) whether the new test is practical and acceptable to test users. The value implications of the ESP TG test will be examined as well. Procedures for answering the two questions are presented and discussed in Sections 2.3 – 2.4. Expected outcomes will be listed in Section 2.5.

2.1. Statement of the problem

This section comments on the current Tour Guide English Test with particular reference to its drawbacks. Recommendations for improvement will be made.

2.1.1. Description of the present Tour Guide Test

The Tour Guide English Proficiency Test (TG Test) is a sub-test of the annual Tour Guide Examination held by the Taiwanese Tourism Bureau. The foreign-language test is offered in English, French, Spanish, German, Japanese, Korean and some other Asian languages. The Language Training & Testing Center (LTTC) in Taipei is responsible for the development and administration of the English, French, Spanish, German and Japanese tests. Up to now, the TG language tests have been a general foreign language proficiency test (FLPT) consisting of two parts: a paper-and-pencil test of multiple-choice questions on Listening and Use & Usage, and a tape-mediated oral test. Item types are described below with a focus on the TG English Test.

Listening Test:

There are three parts to the Listening test. Part 1 consists of 15 short questions or statements. Candidates are asked to choose the most appropriate response to the question asked. Part 2 has 15 short statements. Candidates are to choose the best response that describes/paraphrases the statement heard. Part 3 has 20 mini-dialogues or statements. Candidates are required to choose the best response from the four options provided.

Usage Test:

The test has 80 questions on language "use or usage". However, test items mainly measure language usage as the test content is focused on linguistic features such as tense/aspect, parts of speech, and use of functional and grammatical words.

Oral Test:

The oral test has five parts:

Part 1: Reading. Candidates are asked to read five sentences of various length and phonetic difficulty out loud in 45 seconds.

Part 2: Translation. Candidates are to read five Chinese sentences silently and then give the English translation. Parts of the English translation are given as prompts. Candidates are given 1 minute to complete.

Part 3: Answering questions. This part contains 10 questions of increasing difficulty. After hearing the questions, candidates are given 15 seconds to answer each of the first five questions and 30 seconds to answer each of the remaining five questions. Questions are only spoken once.

Part 4: Oral mini-essay. Candidates are given one and a half minutes to prepare a given topic in which the candidates are required to give their opinions, and then record their response in one and a half minutes.

Part 5: Picture description. The candidates are shown a picture depicting an event for one and a half minutes and then record their answers in one and a half minutes. Wh-questions intended as prompts are provided to help candidates prepare their answers.

2.1.2. Drawbacks of the present Tour Guide Test

The general proficiency approach seems to be inadequate for the following reasons:

(1) The language use ability in relation to the future language use contexts may not be adequately reflected. General proficiency tests tend to measure the test taker's linguistic capacity to cope with undefined language use situations. In a general proficiency approach, we are measuring the test taker's language ability to solve future language problems without any formulaic solution. A specific-purpose test, on the other hand, assesses restricted competence in a particular language use setting. The contexts and the test tasks should reflect the knowledge/skill and the communicative demands present in the target language use domain.

(2) The language construct seems to be narrowly defined in the FLPT. Tour guiding involves many areas of language knowledge. The traits being measured in the present TG Test seem to be limited to grammatical knowledge only, which is inappropriately narrow. Grammatical knowledge enables the language user to control the formal structure of language in order to produce or comprehend grammatically acceptable utterances. But being grammatically competent is not a sufficient condition for a competent language user. S/he also has to be able to use the language knowledge as a source to create or interpret meanings as well as express his/her intentions appropriate to the particular language use setting. Test tasks should reflect the characteristic language use of the target language use situations.

The TG test is intended to make inferences about the language ability of would-be tour guides; test design has to consider the language demands in this particular language use context. The present test does not seem to measure the full scope of the language knowledge in tour guiding. Identifying all communicative language characteristics involved may not be possible. But it is possible to identify areas of language knowledge involved in tour guiding through a thorough needs analysis and reflect these characteristics in test tasks to make the interpretation more appropriate to the language use demands in tour guiding.

(3) The present TG test does not generally reflect the communicative demands faced by a tour guide. Authenticity and interactiveness are at issue.

Authenticity

Authenticity has to do with content validity (Section 1.3.). It refers to the extent to which test tasks correspond to tasks in the target language use domain. Authenticity can have three meanings: being realistic, communicative and operational (Strevens, 1988).

The present TG Test is not realistic with respect to test input and output. The input of the test is short and de-contexualised; the references to persons, objects and actions are unknown to the participants. Regarding response output, in real life language use situations, candidates are seldom presented with three or four options and their understanding is signalled by selecting the correct choice intended by the tester. Equally rare are instances in which a tour guide is asked to interpret a partially completed English sentence or compose an oral essay, remotely related to his or her experience. Instead, people signal understanding by speech or writing. In addition, in multiple choice questions, what is intended by the language tester to be the best possible answer may not be perceived as possible by the candidates, as people bring in their own topical knowledge, world view and personality into each test situation.

The present TG test method is not communicative in that it requires limited interaction and performance on the part of the test taker. Moreover, the context provided does not allow for appropriate language use to take place. Interaction involves intention of expressions, expressing content of a message, likely expectations of the participants and accordingly, modifications of the content/expressions. It involves receptive and productive skills. The present TG test only involves interaction in a stimulus-response manner.

The present test does not seem to provide a context for appropriate language use to take place. Language use takes place in a context. Appropriate linguistic forms vary accordingly. A competent language user must be able to handle appropriateness in terms of the language use setting and the linguistic context.

S/he must be able to manipulate language in wide enough contexts. The present test has not been able to provide such contexts.

With respect to methods of eliciting test performance, there is little productive candidate performance so there is no direct way of telling how well the candidate can function as a competent tour guide. In other words, the present test does not translate language use ability in ways and contexts corresponding to real life language use settings.

Finally, the present test is not operational. The candidates are not given an opportunity to fulfil the tasks in an active manner. Rather, they are required to sit back and choose the one possible answer throughout one test sitting in the paper-and-pencil part of the test. With respect to the speaking test, the types of communicative demands generally faced by a tour guide such as making an announcement or interpreting for the clients are absent. For the test to be operational, the candidate has to be treated as an expert and be allowed to talk about typical subject matters in their future workplace as freely as possible.

Interactiveness

Interactiveness refers to the degree to which test tasks engage the candidates' area of language ability. Test tasks in the present test do not seem to engage the candidates in the appropriate discourse domain and the use of strategic competence, which makes the interpretation of test performance less meaningful. Most items measure grammatical competence; successful completion of the items is largely dependent on a process of elimination and knowledge of lexical signification and propositional content expressed by syntactical structures in the items. Pragmatic knowledge, i.e., knowledge about social rules and appropriateness, is largely ignored. The candidates have not been given minimum opportunities to identify, assess the test tasks and plan accordingly for appropriate response.

(4) The last criticism of the present test is about item selection, test method and score report. The present test takes a structuralist/psychometric approach that emphasises reliability and item analysis. Items selected are to a large extent

determined by technical criteria such as difficulty and discrimination indices rather than some other social criteria derived by a needs analysis and/or other means of empirical language use analyses. There is nothing fatally wrong with technical solutions to item selection but there is no reason why other criteria could not be introduced at an early stage of test development to enhance face and content validity of the test.

In terms of test method, Multiple-choice questions (MCQ) have been found to lead to gender bias in favour of male candidates (Murphy, 1980; Hellekant, 1994)). Inclusion of varied task types could rectify the problem.

Regarding score reports, the written part of the test is reported in terms of a numerical value indicating the relative standing of the candidate compared with other candidates in the group. It does not indicate what the test taker can or cannot do with the language skills measured in the particular test. The score would be more meaningful if inferences could be made on the basis of test performance.

To sum up, the present TG Test needs to be improved in the following ways: provide a clear definition of language use ability in tour guiding, specify the language use situations, and take account of task authenticity and interactiveness.

2.1.3. Suggestions for improvement

The problems stated in the previous section suggest that the general proficiency approach and the multiple-choice format may not be adequate in measuring and predicting a would-be tour guide's language use ability. Therefore, I would like to make the following recommendations to improve the test.

(1) Designing a new TG test on the basis of the theory of communicative language use (Hymes, 1972; Canale & Swain, 1980: Bachman, 1990): Tour guiding is primarily about communication. A successful tour guide tries to convey meaning in an appropriate manner. The focus of the assessment should be on the test

taker's ability to use or manipulate the language to achieve communication goals. In recent conceptualisation of language use, two sets of factors interact with each other. They are the test taker's characteristics such as his/her language ability, topical knowledge and the affective schemata on the one hand, and the language use contexts on the other, thus accounting for individual differences in performance (Section 1.1.6.). This can be translated into a language testing situation in which test performance is the interaction of the test taker's person characteristics and the language use contexts in the test mediated by the test taker's strategic competence.

(2) Considering the use of LSP testing: Future language use contexts are commonly reflected in a specific-purpose test. Performance criteria are then built into simulated tasks so that inferences about the candidate's language use ability can be made with more confidence. However, such behaviour criteria are often overlooked in a general proficiency approach.

Two major studies on EAP tests were carried out to investigate the relationship between a student's language test score and his/her final academic success. The ELTS validation study indicated a predictive correlation of about 0.3, while an IELTS predictive validation study yielded a correlation of 0.39 (Criper & Davies, 1988; Ferguson & White, 1994). These findings suggest a "weakly positive relationship between language proficiency and academic outcome" which is a satisfactory result because other factors like personality characteristics, affective schemata and the candidates' world view/topical knowledge also play a part in language use.

Research in EOP is scarce. But EOP testing has some apparent attractions: precise definition of language use in the specific language use contexts, directness in test method, criterion-referencing and performance-orientation (Section 1.3.). These attractions are closely linked to the interpretation of test results and the candidate's future job performance.

Tour guiding requires some specific language use behaviour. Awareness of the situational contexts and the appropriate handling of the language use contexts

will enhance the candidate's performance as a tour guide. In order to more adequately measure the test taker's language use ability as a tour guide, a specific purpose language test seems a better choice. However, this option to language testing should not be associated with "laziness" on the part of the language tester; rather, LSP tests should be seen as "richer proficiency tests" as suggested by Davies (2000b).

(3) Inclusion of variety of item types: In a large-scale test, the multiple-choice test format has been the default format for ease of scoring. Research findings have indicated that test methods have an effect on candidate performance and ultimately the test result (Murphy, 1980; Bachman & Palmer, 1982; Hellekant, 1994). Inclusion of a variety of item types may safeguard against test method effects. Moreover, different types of interaction (e.g. greeting, explaining, presenting factual information etc.) in tour guiding, which require different language knowledge and strategies, have been observed. Therefore, a variety of item types reflecting the salient features in tour guiding would reflect the test construct.

A third reason for the inclusion of varied item types is that it seems to provide authenticity in terms of the test taker's language learning experience. Also, in real life communication, the test taker is rarely presented with four alternatives from which one choice is to be made to signal understanding and production.

The final reason is that a decision on the use of an ESP test assumes the tasks are constructed to reflect the characteristics of target language use. In occupational or academic contexts, the candidate's language use ability is manifested through different activities and tasks. One common practice in ESP testing is simulation of the language use contexts and language tasks, which often result in the inclusion of different item/task types.

(4) Involvement of test users and subject informants in test development: ESP tests often list the performance criteria required of the test takers. People responsible for the workplace usually have a better-articulated view on the nature of the workplace and its language demands than the test designers; their expectations of

the candidates' performance may be different from those of the language experts. Thus their involvement in the test development helps specify test tasks which may help increase the construct validity of the test.

In short, to improve the present TG Test, I proposed to base the test on the theory of communicative language use and specific purpose language testing to measure the test takers' language ability in tour guiding. This implies the assessment of the test takers' ability to demonstrate their language knowledge through the targeted language use contexts and tasks. Test users, particularly experts from the tour guiding industry, were involved in test development and test validation.

2.2. Research aim

There were two aims in this thesis: (1) the design, and development an ESP TG test to measure prospective tour guides' language use ability; (2) the examination of the validity of the new test and some consequences of its use. The new TG test took a cognitive-psycholinguistic approach to test design; the candidate's ability was inferred from his/her test performance on various tasks designed to measure the construct realised in three language use dimensions: ability to listen, speak and to use the grammatical knowledge meaningfully and appropriately. Performance was defined as the candidate's demonstration of his/her ability to use the language related to tour guiding through tasks in specified language use contexts. Language production from the test was the basis of evaluation and score interpretation. Messick's (1989) facets, which include evidential and consequential examinations of the test and test use, was used as the framework of the TG test validation. On the basis of the information gathered, the validity of the new TG test will be discussed; implications for specific-purpose testing in the Taiwan foreign language use context and future directions will be suggested.

2.2.1. Research questions

In this thesis, the following two questions are investigated.

Research question 1: Is the new Tour Guide English Test empirically valid?

This question considers the evidential bases of test interpretation and test use, which includes the traditional content, construct and criterion-related test validity. Procedures for construction validation of the TG test are described in Section 2.3.

Construct validity:

Construct validity is now the most important consideration in test evaluation. The construct of the TG Test goes beyond grammatical competence and is manifested primarily in aural and oral skills in tour guiding. To examine the construct validity of the three TG sub-tests, I used the following approaches.

(1) Qualitative examination of content coverage and representativeness: Expert informants were consulted by way of a questionnaire. Student feedback was also collected.
(2) The goodness of fit of the tests and the rating scale were investigated with the use of Rasch analysis.
(3) Factor analysis was performed to examine the homogeneity of the test tasks.

Content analysis:

Content relevance and coverage were examined. Language experts and subject experts were asked to check test specifications and test content for domain behaviours measured and tasks sampled.

Correlation studies with external criteria:

Correlational analyses against the TOEFL score and the end-of-year grade were conducted to examine the relationship of the TG test with two established criteria.

Item investigation:

Item stability and the consistency of rating were examined.

Research question 2:

2a. Is the new Tour Guide English Test appropriate, fair and meaningful?

2b. How practical is the TG test? What are the value implications of the use of this test?

Performance testing in Taiwan is not a new concept. In some professions such as teaching, medicine and pilot training, student candidates are required to perform either in a simulated setting or the target setting to demonstrate their ability for certification. In educational contexts, students demonstrate their knowledge of a subject by completing tasks designed to measure their mastery of knowledge in a given subject. Language performance tests, however, are rare. In large-scale language tests, indirect and objective tests are usually developed and administered for fairness and efficiency. Some large-scale tests have a performance element such as the Writing part of the Joint College Entrance Examination (JCEE) for high school students and the oral test of the FLPT for civil servants. But the two tests are essentially composed of multiple-choice questions. Strictly speaking, the JCEE writing component is not a performance test in that the writing task requires each candidate to write a composition about 100 words in English with cues provided. It does not reflect academic writing nor does it allow students to actively create meaning within a discourse domain. The FLPT Oral Test has a number of drawbacks discussed earlier.

There is a growing awareness of measuring the general public's language use ability to cope with foreign language use contexts in Taiwan; hence the LTTC's ongoing project of assessing the learner's language proficiency commissioned by Ministry of Education. But, there are practical constraints on LSP tests; they are time-consuming to develop, and they can be expensive to take.

In this study, the appropriateness, fairness and meaningfulness of the new TG test, its practicality, and its likely effect on language learning/teaching were investigated. To examine the appropriateness, fairness and meaningfulness of the TG Test, expert judgements were collected. Experts included one senior administrator in the Tour

Guide Association, three EFL lecturers (two in Tourism Industry; one in Foreign Languages), and one language tester.

To examine the practicality of the test and the value implications of the test use, the following were investigated:

(a) Cost and efficiency

- human resources, e.g., the test writer, raters, monitors and clerical support
- material resources, e.g., test space, equipment, test material and security
- the time allowed, i.e., scheduling of test development and the length of the test

(b) some likely short-term washback effects

- teacher's and students' awareness of this test method
- students' understanding of their language use ability with reference to the test

2.3. The Tour Guide English Test

In this section, procedures in developing and administering the test, methods used to validate the test and the time line for each step in test development and validation are outlined.

2.3.1. The development process

The test development process included the following stages.

Identifying the test purpose, the test takers and test users

The Tour Guide English Test is used as a screening test to control entry to the workplace. The purpose of the test is to make decisions on an individual's language ability to work as a tour guide. The test has been a general proficiency test up to now. For reasons discussed in Section 2.1, an LSP approach is considered to provide better predictive information on the test taker's ability to cope in his/her future job. Therefore, I wished to develop and pilot an ESP TG test.

Three groups of people might be interested in the test results: the test takers, the administrators in the Tour Guide Association, and the test developers. Administrators in the Tour Guide Association were interested in test results for licensing, training and supervision purposes. Candidates were concerned about their performance and the test developers were concerned about the validity and reliability of the test and whether the new test yielded information on the test taker's language ability as a tour guide.

At this stage, resources available for this research were identified as well. This project was not sponsored by the Government; therefore the research was under-funded, and resources and help available were limited. The LTTC and four universities and colleges agreed to provide administrative/clerical support and the space for testing. The Tourism Bureau allowed me to attend the three-week tour guide training course; they also provided me with the training materials for future needs analysis. The Tour Guide Association recommended two senior practising tour guides to be the expert informants. In addition, five tour guides agreed to be observed and a number of colleges in Taiwan encouraged students to participate in the test.

A form to gather candidate bio-data was designed (Appendix 2-1). It includes gender, age, educational and professional background, length of exposure to English, and length of stay in English-speaking countries. A self-assessment of their English language use ability in listening and speaking skills was given to the candidates at the

time of registration (Appendix 2-2). This was for the test developer to find out about individual candidates' perception of their language ability for future research.

Conducting the needs analysis

Needs analysis is the basis of test design and test content development. In order to develop a valid performance TG Test, a needs analysis of the nature of the workplace and the linguistic demands of tour guiding is essential in grouping tasks into categories of tasks. Two steps in the identification of the language use tasks in the target language use domain of tour guiding were identified:

(1) Identifying expert informants who are familiar with the language use settings and who can help identify and select relevant language use situations and tasks: The expert informants consulted were Mr. Liu, Secretary General, from the Tour Guide Association and Mr. Kuo, a senior practising tour guide and a tour guide trainer. In addition, Ms. Zuo from the Tourism Bureau explained the government's expectations of a tour guide: good command of a foreign language and good knowledge of Taiwan so that the tour guide can tell his/her clients about the country. However, she was not available to comment on the test and test content.

(2) Gathering information on language use domain and tasks. There were three procedures: (1) attending the three-week tour guide training course to familiarise myself with the nature of the work place and the job; (2) interviews with the expert informants to gain insight into the work demand; and (3) workplace observations to gather task types in tour guiding.

Writing up test specifications

Test specifications of the Listening, Speaking and Grammar tests were developed and included test objectives, general description of the language functions, language elements, test tasks, sample items, characteristics of test input and output, time allotment and scoring methods. In addition, testing environment and the test format, length of the tests, and test instructions were described in the test specifications.

Rating scales and scoring methods were developed in this stage. Selection and training of raters were arranged (See Section 3.5.). The TG rating scale consisted of

three parts of six levels: an overall proficiency band scale (which was later discarded, See Section 3.5.), a holistic rating table for listening ability and an analytic rating scale for speaking ability.

Both objective and subjective scoring were used with criteria for correctness provided. The Grammar and Listening tests used dichotomous scoring. In the Speaking test, the candidates' oral ability was evaluated against the descriptors in the rating table. Test results were reported in three scores: Listening, Grammar and Speaking with scores converted to their equivalent band levels.

Three groups of raters from different professional and language backgrounds participated in the training and rating: three native-speaking EFL teachers, three Chinese EFL teachers and two senior tour guides. A one-day training session was arranged (Section 3.5.). Extra sessions and help were provided. Information on raters was collected as well (Appendix 2-3).

Constructing the TG test tasks

Test tasks based on the item specifications were created. Test materials were either taken from authentic texts or scripted for this test purpose. Six native speakers of English (including three EFL teachers), three university EFL lecturers and the expert informants reviewed the pilot version of the test.

Piloting the test and rewriting test questions

Fifteen subjects participated in the pilot test and commented on the test and the test questions. Revision was carried out on the basis of item statistics and test taker comments.

Administering and monitoring the main trial

112 subjects participated in the main trial. The testing conditions as given in the test specifications were observed.

Scoring the test and reporting test results

In this stage, rating behaviour was examined. Test results in the form of a profile were reported to the teachers of the participating students.

Test validation

Test reliability and validity were examined. Test results were analysed with the use of classical test theory and Rasch analysis. Inter-rater reliability was performed. Expert judgements on test usefulness were collected. Test validation following Messick's (1989) framework of test validation was carried out.

Implications and recommendations

Test results were discussed and the areas needing improvement suggested. Implications for the use of specific purposes testing in Taiwan were also suggested.

2.3.2. Test administration

In this section, I will describe the test environment, test instructions and the candidates.

Test environment:

The TG test was piloted and administered in language laboratories in various schools. Candidates were given a seat number. Test booklets were placed on the desk. For the Listening test and the Speaking test, candidates were required to put on the headsets to listen and answer the questions. For the Grammar test, answer sheets were provided.

Each test administration lasted for about 2.5 hours. The order of testing was Grammar, followed by Listening and Speaking, with a 10-minute break between the Listening and Speaking tests. At the end of the test, candidates were asked to fill out the questionnaire on the test.

I worked with one EFL teacher as monitors to ensure the guidelines listed in the test specifications were followed. Any abnormalities such as environmental disturbances, clarity of the test instructions, quality of the recording, etc. that might affect test results were noted down.

Test instructions:

Sample items were given to the candidates before the test. To minimise confusion during the test, monitors explained the procedures in Chinese before the actual test. Candidates could ask questions related to testing procedures. Instructions on how to do different parts of the test were read on the tape; they were also printed in the test booklets.

Candidates:

The Tourism Bureau regulation on the eligibility of the candidates was observed. That is, the candidates have to be over 20 years of age with a tertiary education degree or an equivalent diploma.

Candidates were asked to fill out a registration from to establish a population profile. They were asked to fill out a self-assessment form in their own time before the test and gave it back to me. A questionnaire to be filled in by the test takers was handed out after the test (Appendix 2-4).

2.4. Scheduling of the Tour Guide English Test

The following is the time line I follow for the test development and test validation.

Activity	Place and suggested deadline	
	Taipei	Edinburgh
Identify expert informants	31/11/98	
Observe tours	15/12/98	
Identify and group language use domain tasks		15/01/99
Write up Test Specifications		31/01/99
Prepare questionnaires, etc. on pilot test		31/01/99
Write pilot version of the test		15/02/99
test tasks		Dec/98
rating scales		Jan/99
test rubric		Feb/99
Consult expert informants		01/03/99
Prepare test materials for pilot test	08/04/99	
Pilot test & hand out questionnaires for comments (students)	12/04/99	
Final version of test	20/05/99	
analysis of test results	16/04/99	
test revision	18/05/99	
prepare for the final test version	20/05/99	
Send out self-assessment to candidates	15/05/99	
Administer final version of test	21/05/99	
Hand out questionnaires for comments (on test, test method, practicality, acceptability, etc. – test users, students)	21/06/99	
Scoring	15/07/99	
Test validation		31/08/00

2.5. Expected outcomes

By the end of the research project, I hoped to achieve the following objectives.

(1) To produce an acceptable and useful TG test and a rating scale, along with the respective test specifications and test results for examination by other testing professionals.
(2) To show that an ESP test can be an option in the measurement and the prediction of a would-be tour guide's language ability and perhaps by extension the language use ability of candidates in other specific language use contexts.
(3) To show that collaboration among language experts, test users and subject experts can better ensure test content and validity and its usefulness.
(4) To record the various constraints encountered in ESP test development and test administration so final recommendations could be made with respect to future test development or revision.
(5) To provide some learning experience for students and teachers so they are aware of such a test method. It was hoped that the type of language use ability assessed would provide some insight into their understanding of language learning and teaching.

Chapter 3: Design and development of the Tour Guide English Test

3.0. Introduction

This chapter seeks (1) to provide theoretical justifications for the assessment of the ability to listen, speak and use grammatical knowledge accurately, appropriately and meaningfully, (2) to offer a working definition of these three abilities for test construction, (3) to describe the rating scale for the judgement of test performance and rater training, (4) to discuss criteria to ensure the usefulness of the TG test, and (5) to describe the TG test itself.

Traditionally the knowledge and production of language elements such as sounds, intonations, stress, words, arrangement of words, and their linguistic and textual meanings have been measured in terms of separate skills and systems: listening, grammar, reading, speaking, writing, and grammar and vocabulary. It was assumed that the degree of mastery of these linguistic elements varied; therefore, each of the elements constituted a variable to be tested (Lado, 1961). Language ability meant good mastery of language systems and skills; language forms were the focus of measurement, and the learner's language ability was estimated on the basis of the correctness of responses on the skills measured. For example, a listening test measured if the learner understood what was read to him/her; a speaking test measured how accurately the learner produced sounds and sentences; understanding the linguistic meanings would be the focus of a reading test; a grammar test would test control of the language problems/differences between the learner's native language and the target language; the learner's writing ability was defined in terms of his ability to produce the language accurately. Because discrete language elements were tested, results could be reported with a score indicating a percentage of correctness. Scores were then compared against a known norm (See Section 1.4.).

According to the CLA framework (See Section 1.1.6.), language knowledge consists of organisational knowledge as well as pragmatic knowledge. The learner is viewed as being able to relate linguistic signals to produce meanings appropriately in different

language use contexts. Further, language ability consists of factors other than the learner's language knowledge. The learner's world knowledge, affective factors and the strategies s/he uses have contributing effects on his/her language performance. In other words, the learner is no longer treated as a black box; instead, s/he actively brings in other factors in language use contexts for communicative purposes. Language performance is further affected by variables arising in different situations. They include the sort of input the learner gets (e.g., cognitive load), and the relations with the participants in an interaction. A test in the communicative paradigm has to consider all the factors that could affect real-life language performance and try to approximate the most salient features of language use in a testing situation. Because language use involves variables other than the language system, tests tend to consist of tasks with a purpose and a context to help the test taker achieve a goal and to provide him/her with fresh starts.

3.1. Practical reasons for tests of Listening, Speaking and Grammar

There are two practical reasons for separate tests of listening, speaking and grammar.

(1) Separate assessment of language ability in different skills has been widely accepted. Change in testing procedures would cause test takers unnecessary anxiety and suspicion of the validity of the test. The present TG test measures language use abilities in listening, speaking and grammatical knowledge; decision was made to maintain the present test format.

(2) The best way to measure the test taker's ability of any skill is to get him/her actually to do something with the use of that skill. Tour guiding involves primarily face-to-face interactions. Good listening and speaking skills are essential, and a good knowledge of language use and usage can help communication more effectively. Therefore, the TG Test will include the types of skills and knowledge most necessary in the workplace, namely, listening and speaking. A grammar test is included to explicitly measure the test taker's ability to use grammatical knowledge with accuracy, appropriateness and meaningfulness. The candidate's ability in reading and writing are not assessed in the TG test because these two skills are less central in the workplace.

Current understanding of what is involved in listening, speaking and the use of grammar helps one to construct tests with more confidence. In Sections 3.2 – 3.4, brief reviews of research on listening, speaking and grammar will be discussed to gain insight into the processes involved to derive a working definition of the three constructs.

3.2. The Tour Guide Listening Test

Recent research on first and second language acquisition has shown that listening ability plays an important role in language acquisition and development, interpersonal relations and academic success. (Brown and Yule, 1983; Faerch and Kasper, 1986; Feyten, 1991; Dunkel, 1991). In a review of some corporations' perceptions of the importance of listening in corporate settings, listening has been perceived as the most necessary and important communication skill for entry-level employment, job success, general career competence and managerial competency etc. (Dunkel, 1991). These findings confirm that listening is an important component in language use ability.

Although there exists no consensus on the unitary or divisible nature of language proficiency, a growing body of evidence from tests of second/foreign language listening comprehension seems to suggest that listening is a distinct skill from measures of other language skills (Buck, 1992; Dandonoli & Henning, 1990; Lund, 1991) and indicates the necessity of a listening component in the assessment of communicative language use ability.

Presently, there is no general agreement on the components of listening comprehension, the factors affecting the success and failure of comprehension, and the best techniques for assessing this construct (Dunkel, Henning & Chaudron, 1993). Before presenting my own attempt to develop a valid TG Listening Test, in the following section, I will

(a) briefly examine the nature of the second/foreign language listening comprehension construct,

(b) operationally define a listening comprehension construct to fit the particular language use domain, and

(c) identify critical aspects of listening comprehension to be addressed in the Tour Guide Listening test.

3.2.1. The nature of foreign/second language listening comprehension

Various models of L1/L2 comprehension have been proposed to capture the essence of listening comprehension. Wolvin (1990, in Dunkel 1991) described 12 models of first language listening which involve auditory and visual elements of reception, perception, discrimination and response as well as cognitive and affective elements. Faerch and Kasper (1986), making reference to Jarvella and Nelson's (1982) psycholinguistic model which assumes that comprehension is usually "a product of several cognitive subsystems working together in a harmonious way", state that the emphasis within first language comprehension is on "higher-level processes" of meaning reconstruction. In Nagle and Sanders' (1986) L2 comprehension model, comprehension and learning are viewed as interrelated and interdependent but distinctive cognitive phenomena. They distinguish between automatic and controlled decoding processes which interact with and affect implicit and explicit linguistic knowledge as well as other types of non-linguistic knowledge. The researchers postulate that "comprehension becomes more efficient as knowledge increases, processes become automatic and experience confirms the reliability of the learner's decoding, inferring and predicting" (p.22). In sum, as Rost (1990) points out, "although some models of verbal understanding have been attempted, they are for the most part broad descriptions of linguistic and pragmatic competence or narrow descriptions of verbal processes" (p.6). They cannot offer much help in operationalising precisely what listening involves in a given test .

Despite the perceived inadequacies of models of listening comprehension, researchers agree on a number of points with respect to the processes of listening comprehension, skills required for successful listening and factors affecting a learner's listening ability. These help test developers to make more accurate judgements on whether or not they are succeeding in their assessment techniques. They also provide the basis for the TG listening specifications.

Listening as an interactive process

A general consensus among testing researchers with regard to listening comprehension is that it involves an interaction between linguistic codes and the learner's ability to understand and make inferences on the basis of his/her linguistic knowledge, world knowledge and his/her personal experience etc. (Anderson & Lynch, 1988; Bernhardt, 1991; Lund, 1991; Richards, 1989). There is a noticeable move away from notions of listening as auditory discrimination to contextualised listening comprehension (Brindley, 1998; Davies, 1978; Rost, 1990; Weir, 1990). The use of a much more complex and interactive model of listening implies that tests of listening comprehension should reflect the learner's ability to understand authentic discourse in contexts for particular communicative purposes. Therefore in a listening test items should reflect a focus on an understanding of meaning and extracting information from the text rather than a focus on linguistic decoding of the passage.

Listening skills

A number of scholars have provided useful taxonomies of listening comprehension skills as well as listener tasks and functions related to these skills (Richards, 1985; Lund, 1990; Munby, 1978; Rost, 1990). These skills are arranged in a hierarchy from lower order, involving understanding utterances, to higher order, involving inferencing and critical evaluation. For example, Richards' 1985 taxonomy lists 33 microskills of L2 participatory (or conversational) listening and 18 skills involved in non-participatory academic listening. Although empirical evidence supporting those taxonomies is scant, these skills serve to highlight aspects of the ability the listener needs to demonstrate if she is to function as a skilful listener. Very often, these skills appear in test specifications according to their hypothesised level of difficulty. The postulated developmental taxonomy is further used to form the basis of levels of listening ability which appear in proficiency rating scales such as the American Council on the Teaching of Foreign Languages (ACTFL). Until and unless further empirical investigation proves a mis-ordering of the hierarchy of skills, the taxonomy is likely to remain generally popular.

Factors affecting listening comprehension

In a review of some 115 studies in listening comprehension, Rubin (1994) proposes five factors believed to affect listening comprehension. They are: (1) text characteristics, (2) interlocutor characteristics, (3) task characteristics, (4) listener characteristics and (5) cognitive operations.

Text characteristics

Text characteristics refer to variations associated with a listening passage/text or its visual support. This includes (a) lexical and syntactic aspects of a text, (b) the learner's perception of a given listening passage in terms of the text's acoustic variables such as speech rate, hesitations/pauses, sandhi-variations, stress and rhythmic patterns, and (c) text type and the amount of visual support for the text. Research has not yet offered conclusive findings on whether a native-speaker's syntactic modifications improve comprehension. However, four variables do seem to influence comprehensibility: the learner's proficiency level, type of input by the native speaker, type of the text and the amount of background knowledge required (Chiang & Dunkel, 1992; Rubin, 1994).

Text type has been noted as a factor affecting listening comprehension as well. Shohamy and Inbar (1991) consider the comprehensibility of three types of texts: a news broadcast, a lecturette and a consultative dialogue. Results indicate that the news was the most difficult, followed by the lecturette, with the dialogue being the least difficult text type. The researchers suggest that a listening test should include a number of texts which form a sample of the range of genres on the oral/literate continuum of listening (Shohamy & Inbar, 1991).

Interlocutor characteristics

In a study of gender bias and perceived speaker expertness in 98 university ESL students' listening recall, Markham (1988) reported that

(a) students recalled more from the non-expert male speaker than from the female non-expert,

(b) advanced students recalled more from the male expert than from the female expert, and

(c) performance was better when the presentation was given by a female expert than by a female non-expert.

The findings suggest that gender bias may influence ESL students' recall of orally presented material, which will be considered in test design.

Task characteristics

Task characteristics refer to variations in the purpose of listening and associated tasks. Brown and Yule (1983) identify two listening purposes along a continuum of listening aspects: interactional and transactional. Listening in interactional discourse means being an active participant in collaborative discourse. In such types of listening, the listener's ability to display understanding and signs of participation in expected ways is very important. However, this type of task is primarily social and it is artificial for the test taker to listen to.

Transactional purposes of listening refer to instances of listening in which the listener does not interact with a speaker. Clarification strategies cannot be enacted directly and listener understanding is not normally displayed (Rost, 1990). Transactional use of language is primarily for communicating information. Understanding of the purpose of a given listening text facilitates comprehension.

Shohamy and Inbar (1991) consider how type of question influences success in L2 listening comprehension tests. They find that subjects perform better on questions answerable by referring to local cues in the text than on those answerable by referring to global cues. They conclude that it is more difficult to generalise, infer and synthesise information than to look for data-specific information. In the study, they also report that students who respond to global questions are also able to respond to local questions but not vice versa. Lund (1991) looks at how task types affects learners' ability to remember more main ideas or details. Results suggest a significant task effect on (1) the recall of task-specific propositions, (2) the proportion of macropropositions to micropropositions recalled and (3) the number of distortions in recall.

The results of the studies do not, however, indicate that one task type is superior to another. Depending on the objectives for listening, test takers seem to modify their behaviour according to the task they are given. Valid listening comprehension tests should, therefore, attempt to include different task types.

Listener characteristics

Listener characteristics such as proficiency level, background knowledge, memory, affect, age, and gender can have considerable impact on listening comprehension. Language proficiency affects listening ability to some extent. Schmidt-Rinehart (1994) finds a consistent increase in comprehension scores when learners' overall language proficiency increases. It is not clear what roles grammatical knowledge and pragmatic knowledge play at different proficiency levels. But learners use linguistic knowledge to decode texts where there is little background information available. Learners at lower levels of proficiency also tend to apply their linguistic knowledge to understand the listening passages (Teng, 1999).

Background (or schematic) knowledge is an equally important facilitator in listening comprehension (Widdowson, 1983; Anderson & Lynch, 1988). Brown and Yule (1983) describe schematic knowledge as "organised background knowledge which leads us to expect or predict aspects in our interpretation of discourse" (p. 248). Anderson and Lynch refer to schematic knowledge as one of the "information sources in comprehension" (1988:13). They suggest that the lack of such information impedes comprehension. Empirical studies, exploring the relationship between prior knowledge and listening comprehension, have shown background knowledge improves listening comprehension (Chiang & Dunkel,1992; Long, 1990; Markham & Latham, 1987; Schmidt-Rinehart, 1994). Long (1989) even differentiates and highlights the critical role played by background knowledge (i.e., content schemata) and textual knowledge in L2 listening comprehension. It is not yet known how background knowledge interacts with proficiency level but research indicates that advanced learners tend to rely on the use of their schematic knowledge more than lower level proficiency learners (Section 1.3.7.).

Affect also influences listening comprehension. In a study examining the listening strategies of beginning students of Japanese, Fujita (1984, in Rubin, 1994) reported self-confidence as one of the major factors affecting their listening. In another study on the relationship between apprehension in a second language, listening comprehension and language competence, Aneiro (1989, in Rubin, 1994) found that apprehension was significantly related to lower L2 language proficiency. Madsen, (1982) in a test-anxiety study, concludes that tests that evoke a high level of anxiety are less valid measures of students' performance when these students are susceptible to debilitating anxiety. However, Bailey's diary study (1983) suggests that there is an optimal and moderate level and that it is the level of anxiety the individual is experiencing at the moment that determines whether the anxiety is facilitating or debilitating.

The relationship between memory and listening comprehension has been investigated. Dunkel *et al.* (1989) consider the influence of short-term memory on encoding lecture material in English. They find that subjects (both NSs and NNSs) who have good short-term memory recognise significantly more concept information and detailed information than subjects who have low short-term memory. In addition, they find that NSs recognise significantly more of the lecture concepts and detail than do NNSs.

Learner characteristics influence language performance, and are therefore one aspect to consider in test design. An understanding of the test takers' characteristics is relevant to decisions in the TG test design.

Cognitive operations

Currently researchers favour an interactive and parallel model of L2 listening comprehension. For L1 listeners, there is extensive evidence for real time interactive language processing in which lexical, structural and interpretative knowledge sources communicate and interact during processing (Rubin, 1994). Research findings in L2 listening are contradictory. Lund (1991) provides evidence for a reliance on top-down processing in L2 listening. Van Patten (1988) finds that paying attention to form interferes with learners' comprehension of content. O'Malley *et al.* (1989) find that "effective listeners seemed to be listening for larger chunks, shifting their attention to

individual words only when there was a breakdown in comprehension." In sum, studies seem to indicate a delicate interaction between top-down and bottom-up processing.

Listening strategies have also been much researched. Two types of strategies have been proposed: cognitive and meta-cognitive strategies. Cognitive strategies involve solving learning problems by considering how to store and retrieve information, whereas meta-cognitive strategies involve planning, monitoring and evaluating comprehension (Rubin, 1994). Murphy (1985, in Rubin, 1994), working with intermediate-level ESL university students, reports that more proficient listeners most frequently use "wide distribution" strategies and the less proficient listeners most frequently use "text heavy" strategies. "Wide distribution" refers to an open and flexible use of strategies, and "text heavy" refers to a dependence on the text and a consistent use of paraphrase. In a study on how the use of certain strategies correlates with proficiency level, Rost and Ross (1991) report that the use of global queries is found among beginning-level students. Less proficient students tend to ask for repetition, rephrasing or simplification. But more advanced students use forward inference (i.e., asking a question using information already given in the story) and continuation signals (backchannel communication). Effective listeners are also reported to use more self-monitoring, elaboration and inferencing (O'Malley, Chamot & Kupper, 1989).

The various empirical findings about the aspects affecting listening provide some insights to our understanding of how L2 listening comprehension is processed. Yet there still appears to be no universally accepted view of the L2 listening comprehension construct. It is necessary to articulate the nature of this construct so that the language tester can be more accurate in his/her judgements of the validity of the test instruments and techniques. The lack of argument about this will pose difficulty for the assessment of L2 listening comprehension and consequently test validity. To solve the problem, at present, it seems best to provide (1) a tentative working definition of the construct to be measured and (2) a framework of L2 listening comprehension that will adequately explain the interrelationship of the components for the listening test.

3.2.2. Framework of listening comprehension in tour guiding

Effective listening involves more than just comprehension. It includes elements of reasoning abilities and requires modalities in addition to the aural modality. Since there is no universal definition of listening for operational guidelines to be established, the five factors discussed will be used to delimit a listening comprehension construct for the design of the TG Listening Test.

Two broad facets within comprehension seem to be involved: person ability and text/task difficulty. The interplay of these facets is critical for deriving a construct of listening comprehension and our assessment of a candidate's listening performance. Unless an incremental relationship is to be found between person ability and task difficulty in terms of response type, listening comprehension is hardly a unitary measurement construct. The construct should be defined with regard to different tasks and test takers for different assessment purposes. This, in my opinion, is in line with Bachman and Palmer's (1996) framework of language use and language test performance. Table 3-1 presents the listening framework to be used in the design of the Tour Guide Listening Test.

Table 3-1: TG listening ability framework

1. Purpose: to measure the candidate's listening ability
2. Construct: listening ability is to be defined as the ability to successfully complete the different tasks for different purposes.
3. Characteristics of text/task: to be discussed in Sections 3.2.3 & 3.2.4.
    a. text type
    b. response format and item types
    c. types of skills (i.e., macro- or micro-skills needed)
    d. sample items
    e. difficulty level: contextual support, rate of speech, lexical/syntactic complexity as well as the type of output for tasks are considered.
    f. scoring method
3. Characteristics of the test takers: Including
    a. descriptions of personal characteristics

b. knowledge/skills/abilities required for comprehension

c. performance required

4. Cognitive abilities required for comprehension (e.g., whether the test taker needs to use bottom-up/top-down processing; whether he is asked to identify, interpret, analyse, evaluate or synthesise information)
5. Characterising listening ability (i.e., outcome of assessment in terms of a rating scale to be discussed in Section 3.5.)

3.2.3. Specifications of the Listening Test

The aspects suggested in Table 3-1 were used in the preparation of the Listening test specifications and the development of the test. The following are the Listening test specifications.

Listening test specifications

1. **Objectives**

The Listening Test is to measure a prospective tour guide's ability to understand spoken English. The focus is on the candidate's ability to understand different types of incoming information, his/her ability to process the information and his/her ability to identify and interpret the information to complete test tasks.

2. **Skills to be sampled**

- Macro-skills
  - a. listen for specific information
  - b. understand gist of message
  - c. recall specific information
  - d. information transfer
  - e. following directions/instructions
  - f. comprehending details
  - g. remembering vocabulary and numbers

- Micro-skills
  - a. ability to recognise functions of stress and intonation to understand information structure of an utterance

b. ability to distinguish word boundaries

c. ability to understand reduced forms of words and/or elliptical forms of sentences

d. ability to detect key words

e. ability to recognise sentence constituents

f. ability to understand meanings expressed in different grammatical forms

g. ability to identify and reconstruct topics and coherent structures from ongoing discourse

h. ability to retain chunks of language of different lengths for a short period of time

i. ability to process speech at different rates and in different accents

j. ability to adjust listening strategies to different listening purposes

3. **Time allowed**: Approximately 40 minutes
4. **Number of items**: 45 questions
5. **Test tasks**: Each listening text with its task(s) measures a number of macro-skills. There are 6 listening passages; each about 2 - 4 minutes long. Candidates will be given time to read through the questions once and time to answer the questions. They may also be asked to answer questions while listening. Taking notes is allowed.
6. **Item types**: The test should include the following question types.

a. Identification and labelling	7 questions
b. Matching	6 questions
c. Information transfer	11 questions
d. True or false	4 questions
e. Sentence completion	12 questions
f.. Short answers	5 questions

7. **Text type**: Test texts include conversation, monologue, and public announcements. The length of texts varies from 2 minutes to 4 minutes. Texts are mainly transactional in nature.

8. **Authenticity of texts**: Texts types used will reflect common speech types in tour guiding and texts are either genuine samples of the target language use domain or texts modified from information brochures or current EFL textbooks. All of the texts have been modified and re-recorded to suit the testing purpose. Effort has been made to recapture the authenticity of the original texts.
9. **Voices and accent**: The listening texts and questions will be spoken in American English by two American EFL teachers (one male and one female) whose accent has been perceived as Standard American English.
10. **Sample items**: See *Item specifications for Listening Test* below.
11. **Test format**: Candidates are asked to listen to a tape of recorded material and are given a test booklet that contains tasks for them to carry out. They should be able to complete the tasks in about 40 minutes. The recording includes all test instructions, test questions, and where necessary, pauses and repetitions for the candidates to complete the tasks in time. Candidates should listen to the recording individually through a headphone, but if this facility is not available, they should listen to the recording through a tape-recorder.
12. **Scoring method and marking**: All items should be objectively markable. A comprehensive mark scheme is provided. Candidates are awarded one point for each question correctly answered. Maximum mark is 45. Candidates have to get 23 points to pass the test. For the marking scheme, please see Appendix 3-1.
13. **Degree of skill assessed**: The candidate's listening proficiency is assessed according to the rating scale on listening proficiency.

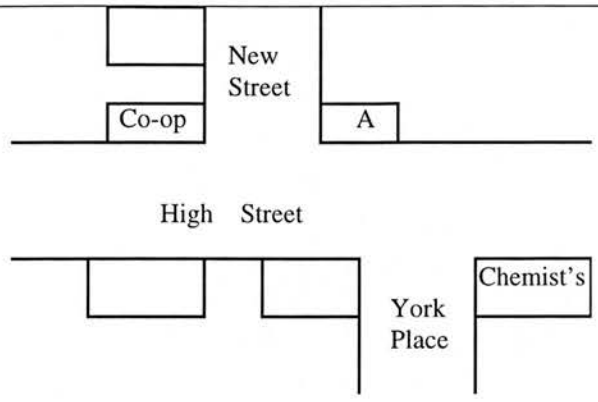
Listening Test Item Specifications

1. Language skills to be sampled: See Point 2 in Listening Test Specifications above.

2. Sample items

   a: Identification and labelling: Candidates will listen to a passage. They will identify and label the positions of objects, people or places in the picture provided.

   Example: Listen to conversation and indicate on the map the different shops mentioned.



A: Newsagent

B: Bakery

C: Post Office

b: Matching: Candidates are asked to match a set of events with time, people, etc. or a set of short conversations/commentaries with different purposes or tones.

Example: Listen to a radio commentary about the opera singer, Maria Callas and match the events mentioned with the year.

1923 __F__	A: Attracted world-wide attention
1937 _____	B: Concert tour
1947 _____	C: Voice lost steadiness
1955 _____	D: Last performance
1959 _____	E: First international debut
1965 _____	F: Born in New York
1973 _____	G: Studied in Athens
1977 _____	H: She died.

c: Information transfer: Candidates are asked to listen to a passage and fill in the missing information according to the speaker.

Example: You will hear a tour guide talking about a Scottish town. Listen and fill in the information which summarises what the tour guide says.

1. Location: _____
2. Built in: _____
3. By: _____
4. Opening time of the castle: _____
5. Present population: _____

d: True or false: Candidates are asked to listen to a passage and decide if the statements they hear are true or false.

Example: You will hear a talk on the six National Parks in Taiwan. Listen to the passage and questions asked and decide if each question is true or false according to the passage you hear. Write T for a true statement and **F** for a false statement.

(The candidates hear)

Kenting National Park is the first national park in Taiwan.
(and they write **T** because Kenting National Park **is** the first national park)
(0) _T_ _ _

e: Completing the sentences: Candidates hear a passage and complete sentences asked.
Example: According to *Feng Sui*, the best kind of room is _ _ _ _ _.

f.: Short answers: Candidates are asked to listen to a passage and answer the questions in less than 4 words.
Example: You will hear a talk on the six national parks in Taiwan. Listen and answer the following questions. (You may answer in Chinese.)

(The candidates hear)
Which is the first national park?
(The candidates write)
(0) _Kenting National Park or 墾丁國家公園

3. Stimulus input: Candidates hear test material from an audio-tape either through a headphone in a language lab or a tape-recorder where this facility is not available. The recording includes test instructions, test questions in the target language and where necessary pauses and repetitions for candidates to carry out the tasks. Test instructions will be printed out on the test booklet as well.

4. Expected response: Candidates are given a test book containing tasks and test instructions. Candidates are required to write their answers in the test book. Types of responses include selection and short answers (1 - 4 words).

3.2.4. Development of the Listening Test

The following is a list of criteria used in the Listening test development.

Material selection

Listening passages were selected on the basis of the following:

1. Text length: The duration of texts should not be too great. A tour guide usually speaks for 2 to 5 minutes about a place, a person or an event, to explain and to clarify any misunderstandings with his/her clients. A long listening passage did not seem to be necessary for this particular test purpose; it would also impose memory load and cause listener fatigue.
2. Information organisation: The presentation of the sequence of events was a consideration. A good tour guide tries to be explicit, organised and informative

when providing information. The genre is narrative most of the time. In the listening test, the texts used should be mainly narrative. Information contained should be as organised and explicit as possible.

3. Explicitness of information: There were three major considerations in mind.
   (a) Texts should contain the necessary information to complete the test tasks so that the candidates were not required to process too much irrelevant information.
   (b) Some texts should contain a few other redundant facts as well so that the more proficient candidates could benefit from the extra information provided in the text.
   (c) Texts should be spoken in simple English and easy for the candidates to understand.
4. Processing load: The amount of information to be processed and the time available to process the information were the two concerns in text selection. The amount of information to be processed has to do with the explicitness of information. As for the time available to process the information, a little slower than normal rate of speaking was suggested when recording the texts. In addition, to minimise the processing load, the topic of each text was included in test instructions. Candidates were given time to pre-view the test questions and, for some tasks, extra time was given to complete the tasks.
5. Familiarity of topics: most text topics were related to typical topics in guiding tours.
6. Text orientation: Passages are primarily transactional in nature.
7. Text type: The inclusion of different text types in a listening test enhances its construct validity (Shohamy and Inbar, 1991). Therefore, the listening text types included conversation, announcement and mini-lectures etc.

Item writing: Four criteria were considered in constructing the listening test items.

1. Orientations of questions: Tour guiding is mainly information-oriented. Therefore, the tasks are primarily information-oriented. That is, questions concentrate on the factual content of the listening texts. Questions such as "What is the speaker's tone/attitude?" are subjective and were avoided.

2. Types of response: Candidates were treated as participants in the test who were asked to comprehend and analyse listening passages through a variety of task types such as matching, filling in the grid, completing a sentence etc.
3. Question types: Listening comprehension was the focus of assessment. Questions requiring lengthy language production were avoided. Instead, candidates were asked to complete tasks that did not require much language production. Some examples were labeling pictures, filling in the grids, matching and short answers.
4. Item checking: Items were checked by 6 native speakers (NS) of English for grammaticality and suitability/difficulty. Three were EFL teachers and three had no EFL teaching experience. Items were further checked by Mr. G. L. Kuo and one of his colleagues from the (Taiwanese) Tour Guide Association. Three Taiwanese professors of English, Professor Z. L. Huang, Professor Y. H. Chen, and Professor H. L. Wang also helped check the test. Five volunteer Taiwanese postgraduate students in Edinburgh pre-tested the test for item difficulty, time allotment and clarity of test instructions and test questions.
5. For the final version of the test, please see Section 3.7.

3.3. The Tour Guide Speaking Test

3.3.1. Why include a speaking test?

The ability to speak a foreign language has been valued as one "highly prized language skill" (Lado, 1961), and the assessment of the oral ability has been included in many of the large-scale proficiency tests like TEOFL and IELTS. The assessment of the speaking ability centred on the linguistic system of the target language, but current practice in language teaching views language as for communication; oral ability plays a direct role in getting meanings across. The assessment of oral ability is often one component in the measurement of communicative language use ability. With regard to tour guiding, a tour guide' oral ability is central to the job; therefore a speaking component should be included in the test battery.

3.3.2. The nature of speaking

Studies of discourse analysis suggest that spoken and written language differ in manner of production and in forms (Brown and Yule, 1983); and speaking is very different from writing.

Speaking is also a complex skill. It is essential for the language tester to know what speaking is and consists of so that the construct can be defined accordingly for a given test purpose. In the remainder of this section, a very brief account of what speaking is composed of will be presented, starting with an examination of the role of the speaker as an information processor, followed by a discussion of a model of speech production by Bock and Levelt (1993) and ending with a discussion related to speech functions, skills required in speech production and speaker strategies to facilitate speech.

The role of the speaker as information processor

Levelt (1989) conceptualises the intentional use of speech as involving a set of highly complex information-processing skills. Several components represent the psycholinguistic processes that generate fluent speech. The *Conceptualizer* generates a preverbal message, a conceptual structure including the speaker's communicative intention, utterance planning and monitoring his/her production and the manner of production. In the *Conceptualizing* phase, the speaker has to have access to two types of knowledge in order to encode the message: procedural knowledge and declarative knowledge. Procedural knowledge refers to the information currently accessible to the speaker and is stored in *Working Memory*. Declarative knowledge, available only in long-term memory, refers to the speaker's world knowledge and the situational knowledge of present discourse context.

The message generated in the *Conceptualizer* then goes to the next processing component, *Formulator*. The *Formulator* translates the conceptual structure into a linguistic structure. Two encoding procedures are activated: Grammatical encoding and Phonological encoding. Grammatical encoding refers to the procedures for accessing lemmas, lexical items in their meaning and sense, and to the procedures for

syntax building. Phonological encoding refers to the building of an articulatory plan, which is carried out in the *Articulator*; the product is the overt speech.

The final two components are related to internal and overt self-monitoring of the speech. *Audition* enables the speaker to listen to his/her own speech as well as the speech of his/her interlocutors. The *Speech-Comprehension System* allows access to information on forms and lemmas in the lexicon in order to recognise the words and to retrieve their meanings. The *Speech-Comprehension System* also allows the speaker to attend to his/her own internal speech.

According to Levelt (1989), each of the components is an autonomous specialist, capable of transforming its characteristic input into its characteristic output without interference or feedback from other components. There is a distinction between controlled and automatic processing applied to these components. Conceptualizing and monitoring are controlled activities requiring the speaker's continuing attention. Formulating and articulating processes are assumed to be automatic. They are speedy and reflex-like requiring little attention and can proceed in parallel.

Processes of speech production

In this part, the foci of discussion are the component of Grammatical encoding, the part that creates the skeleton of an utterance, and theories of monitoring which accounts for how errors happen and how they are detected. Grammatical encoding is responsible for both the selection of appropriate lexical concepts (i.e., lemmas) and the assembly of a syntactic framework. Grammatical encoding contrasts with phonological encoding which comprises the assembly of sound forms and the generation of intonation. The product of the two is a specification of an utterance. The actual speech is materialised in the articulating phase.

Bock and Levelt (1993) provide a model of language (specifically speech) production in which four levels of processing are conceptualised: the message level, the functional level, the positional level and the phonological level. The message captures features of the speaker's intended meaning and it also provides the raw

material for the processes of grammatical encoding into two sub-sets: functional and positional. The primary components of functional processing are lexical selection and function assignment. Lexical selection involves the identification of lexical concepts suitable for conveying the speaker's meaning, whereas function assignment involves the assignment of grammatical roles or syntactic functions. Positional processing is responsible for "constituent assembly" and inflection, that is, the creation of ordered set of word slots and morphological slots. Finally in phonological processing, the phonological structures of the utterance are spelled out in terms of phonological segments of the words and the prosody of larger chunks of phrases.

Errors are generated during the processing. When a non-targeted lemma is activated, an error occurs, and it is, according to Bock and Levelt, the result of incorrect lexical selection. Three major types of errors have been suggested: substitutions, blends and exchanges (Bock & Levelt, 1993). Substitution refers to a lexical error in which a non-target concept is activated rather than the intended one; such errors are usually word association errors; blending refers to the phonological merging of two words which are near synonyms; exchange involves words and phrases of the same category.

Errors are detected by way of self-monitoring (Levelt, 1989); the speaker monitors his/her speech delivery and necessary repairs are made accordingly. However, not all errors are detected. Two theories of monitoring have been proposed for error detection: editor theories, and connectionist theories (Levelt, 1989; MacKay, 1987).

The major feature of editor theories is that speech production results are fed back through a device external to the production system. The device is called an editor or a monitor. The editor can check "in-between" results at the different levels of processing. The editor at each level incorporates the knowledge of the component it monitors. Motley et al. (1982) propose a more restricted device of monitoring which suggests the editor cannot inspect all intermediary output in the generation of speech but only the pre-articulatory output. Editing follows phonological encoding. The editor can intercept troublesome output before it becomes articulated. According to the Editor theories, there is a distinction between self-monitoring and other-monitoring, but both are highly dependent on context and task.

According to the connectionist theories, there are no external mechanisms involved in the control of the speech to the speech production apparatus. Self-control is due to the inherent feedback being operative in the generation of speech. There is no editor checking the speech production. MacKay's (1987) node structure theory explains how error is detected in speech generation. According to this theory, the network for language has several layers of mental nodes: layers of "propositional nodes", "conceptual nodes", "lexical nodes", "syllable nodes", "phonological nodes" and "feature nodes", all of which are shared by a production system and a comprehension system. An activated node primes all nodes connected to it, which in turn prime other nodes connected to them. The most primed node becomes the activated node. Errors happen because the wrong nodes become activated. An error is detected through backward priming which activates all the nodes connected to the original node. Thus the error can be detected almost immediately.

It is difficult to disconfirm either theory. The node-structure theory is more economical as it equates the network for the production and the understanding of language, whereas the monitor theory suffers from reduplication of knowledge. However, there is evidence suggesting interference between speech production and speech perception (Levelt, 1989). Also, error detection can be substantially delayed (Levelt, 1983). Furthermore, the speaker may have reasons to bother or not to bother about a language feature s/he considers to be an error. The mechanisms operating in self-monitoring seem more plausible in explaining error detection.

Types of errors

Errors detected according to Levelt's (1989) model of speech processing are either lexical or phonological in nature. However, research in second language acquisition suggests two types of errors that learners are not often aware of but that affect performance: interlingual and intralingual errors (Larsen-Freeman & Long, 1991). Interlingual errors are errors due to interference of the learner's native language. Intralingual errors include overgeneralisation, simplification or reduction , communication-based errors and teacher/textbook-induced errors observed in all learners regardless of their native languages. In a testing situation, the learner may not

be aware of the types of errors s/he makes. Nevertheless, these errors happen and they affect the test taker's test performance.

Functions of speech

Spoken language, like written language, serves to fulfil different functions. Brown and Yule (1983) distinguish two basic language functions: transactional and interactional. The transactional use of the language is to transmit information; the concern is whether the message is clearly conveyed. Interactional language, on the other hand, is for social interactions. The purpose is to maintain social relationships. Some characteristics of interactional talk are constant shifts of topics and a great deal of agreement on them (Brown & Yule, 1983). The participants usually end up feeling comfortable in an instance of interaction. In transactional use, some interactional elements may be embedded but many social interactions may contain very little transactional content.

Types of speech

There are three broad speech types: long vs. short turns, planned vs. unplanned speech and reciprocal vs. non-reciprocal speeches. Brown & Yule (1983a) make a distinction between "long" and "short" turns. Short turns consist of only one or two utterances, often tend to be interactional and are often used in conversations to achieve the purpose of being friendly, hospitable, comforting or maintaining the face of the participating speakers. The use of a particular short turn depends on the context. Native speakers usually have a rich repertoire of such turns.

A long turn, on the other hand, consists of a string of utterances which may last as long as a lecture and is considerably more demanding than what is required of a speaker in a short turn. When the speaker takes the floor for a long turn, s/he must give a coherent sequence of utterances to make the listeners understand what s/he is trying to say. S/he usually knows the nature and the content of the talk beforehand. The ability to construct a long turn is acquired; it needs adequate models, practice and feedback (Brown & Yule, 1983a) and the ability varies among speakers.

A second distinction of speech types is that of planned and unplanned discourse (Ochs, 1979). The latter refers to discourse that does not have any "forethought and organisational preparation," whereas planned discourse has been "thought out and organised prior to its expression" (Ochs, 1979:55). Ochs points out that the distinction constitutes a continuum and emphasises that most daily communication falls somewhere in the middle of the continuum. Ochs also notes that some types of discourse are more planned than other types. Conversations are usually not planned but a lecture is.

Linguistic differences have been observed in the use of the planned and unplanned speech types. In unplanned discourse, Ochs (1979) finds that speakers tend to (1) rely more on the immediate context to help them convey their message, (2) make use of syntactic structures that tend to emerge early during acquisition, and (3) make extensive use of repetition and word replacement. Danielewicz (1984) also finds that there are differences in the unplanned spoken texts from speakers at a dinner table and their planned spoken texts. The differences are both global (e.g., the way evidence is used to construct an argument) and specific (e.g., the complexity of clause and sentence construction) in nature.

A final distinction in speech types refers to whether there is immediate reaction from the interactants, as in a conversation, or not, as in a news broadcast. This distinction of speech type affects the forms of language to be used. For example, the speaker in a face-to-face interaction has to pay attention to his/her listeners and adapts his/her message according to his/her listeners' reaction. S/he will use certain devices to ensure his/her message is understood. These devices will be discussed in *Processing constraints*.

Characteristics of spoken language

Two types of spoken language have been identified: the form of language occurring in conversational speech and the form of spoken language highly influenced by written language. The discussion will be concentrated on the former. Brown and Yule (1983b) give an account of the features of the spoken language summarised below.

(1) The syntax of spoken language is less structured than that of written language

(2) The speaker is typically less explicit than the writer. In speech, clauses are connected by *and*, *but*, *then,* whereas in written language rhetorical organisers of larger stretches of discourse such as *firstly*, *more important than* and *in conclusion* are used to mark relationships between clauses.

(3) In written language, rather heavily premodified noun phrases are quite common, but they are rare in spoken language. In addition, in spoken texts, short chunks of speech are structured in a way that only one predicate is attached to a referent at a time, for example: *it's a biggish cat + tabby + with torn ears*, or in :*old man McArthur + he was a wee chap + oh very small + and eh a beard + and he was pretty stooped.*

(4) Topic-comment structure, as in *the cats + did you let them out?* are quite common in spoken language, but in written language, sentences are generally structured in subject-predicate form.

(5) Passive constructions are relatively infrequent in informal speech. Instead, active constructions with indeterminate group agents are more noticeable as in: *Oh everything they do in Edinburgh + they do it far too slowly.*

(6) The speaker may rely on paralinguistic cues to supply a referent when s/he talks about the immediate environment: (looking at the rain) *frightful isn't it.*

(7) In conversation, the speaker may replace or refine expressions as s/he goes along: *this man + this chap she was going out with*

(8) A good deal of rather generalised vocabulary is used in informal speech: *a lot of*, *got*, *do*, *thing*, *nice*, *stuff*, *place* and *things like that.*

(9) In spoken language, the same syntactic form tends to be repeated, for example: *I look at fire extinguishers + I look at fire exits + I look at what gangways are available + I look at electric cables what + are they properly earthed + are they properly covered*

(10) The speaker may use a large number of fillers such as *well*, *erm*, *I think*, *you know*, *if you see what I mean*, *of course*, *and so on.*

In short, the features of spoken language are quite different from those of written language. Constrained by time, the speaker's ability to plan and organise the message is greatly affected. As a result, syntax tends to be simpler and the vocabulary used tends to be general and non-specific. Finally, information contained in an utterance is less densely packed and easier to understand.

Speaking skills

Speaking involves two basic skills: motor-perceptive skills and interaction skills (Bygate, 1987). Motor-perceptive skills refer to how well the learner perceives, recalls and articulates in the correct order sounds and structures of the language. The skill is context-free. Interaction skills refer to making decisions about communication and making use of basic motor-perceptive skills. This primarily involves strategies and tactics to achieve communication such as what to say, how to say it according to the speaker's intention while the desired relations with the interactants is maintained. Two types of abilities are central to the use of interaction skills: the ability to organise information into patterns (i.e., routines) and the ability to negotiate meanings in case of communication problems.

Routines are "conventional ways of presenting information" (Bygate, 1987:23); they are part of the speaker's schemata. Depending on the type and mode of speech required, the speaker will decide which pattern/convention s/he is going to use. Bygate (1987) identifies two types of routines: information routines and interaction routines.

Information routines can be further classified as expository or evaluative (Brown & Yule, 1983a). Expository routines involve ways of presenting factual information and evaluative routines refer to conventions in the drawing of conclusions requiring reasoning. Information routines can be difficult and complex, but ability to handle a large chunk of information orally is central to both first- and second-language use.

Interaction routines are turn-taking routines in particular social/physical contexts, e.g., greeting, at the service counter. These types of routines are not structured in an

ordered manner; learners in an unfamiliar situation can make mistakes and may sound brusque, rude or disorganised (Bygate, 1987).

Native-speakers build up an extensive repertoire of routines which is a product of their familiarity with particular types of communication and reflect different categories of knowledge. Very often routines become stock patterns, which might provide a baseline in the assessment of a learner's oral performance.

Apart from the ability to present information appropriately, the speaker in addition needs to know how to solve communication problems. This requires negotiation skills, a general category that includes the way participants in an interaction signal understanding, their decisions on taking turns, and their control of the topic.

When people are taking part in a conversation, they tend to choose a level of explicitness and detail which they think appropriate for the participants to understand. Normally they aim for a sufficient level of understanding according to Grice's (1975:45) co-operative principle: "Make your contribution just as informative as required."

Perfectly explicit communication is unattainable. Too much explicitness gives the listener too much information to process. But the speaker may appear to be arrogant, pretentious or unco-operative if the message is not explicit enough. To find the right level, the speaker often has to predict what his/her listener knows. Common negotiation strategies for ensuring understanding include paraphrase, use of metaphor and use of vocabulary with varying degrees of precision or repetition (Bygate, 1987).

Processing constraints

Time, task difficulty and cognitive difficulty have observable effects on oral performance. Time pressure affects the speaker's ability to plan and organise the message; speakers tend to (1) use devices to facilitate production and (2) try to compensate for the difficulties arising out of time pressure. Bygate (1987) suggests four ways in which the speaker facilitates the production of speech:

(a) simplifying structure

(b) use of ellipsis:

(c) use of formulaic expression and

(d) use of fillers and hesitation devices

Simplification often involves a tendency to connect part of sentence by the use of co-ordination instead of subordination, to repeat the same sentence structure, and to avoid complex noun phrases. Ellipsis refers to the omission of parts of a sentence. A great deal of background knowledge on the part of the listener is assumed by the speaker, which is expected in most conversational situations. The use of formulaic expressions contributes to oral fluency in routine situations; the speaker does not have to monitor his/her choice of words, nor does s/he have to construct new utterances when s/he needs to vary his/her use of expressions on each fresh occasion. The final set of strategies is used to create more time for the speaker to find the right word or organise his ideas.

Task and cognitive difficulty are also conditions that can affect the speaker, who will feel more comfortable if the information s/he is about to give is familiar to him/her and the information provides its own structure. For example, it is easier to give an account of a series of events the speaker witnessed than to provide a justification for the occurrence of these events.

Contexts of speaking

The place and nature of an interaction, the number of people involved in it, and their status relative to the speaker also affect the speaker's performance. It is easier for the speaker if the listeners are his/her peers or junior to him/her. It is also easier for him/her to talk to one listener than many. With regard to the place/situation of an interaction, it is easier for the speaker if s/he is speaking in a familiar and private environment than in a large hall. It is easier to use an informal rather than a formal style of speech. It is also easier for the speaker if the information s/he has is familiar so s/he understands it thoroughly. Finally, it is easier if the information the speaker is about to give has its own structure (such as a sequence of events) so the language is externally supported by the requirements of the task (Brown & Yule, 1983a).

Summary

To speak communicatively and competently in a foreign language, the learner must have good knowledge of the spoken language, background knowledge and ability to use the knowledge/skill in an appropriate manner to achieve communicative goals in a specific context. To measure how well the test taker speaks a foreign language, it is necessary to get him/her to actually say something. To do this, s/he must act on (1) the knowledge of the target language, (2) knowledge of language use in different contexts and (3) skilful demonstration of the knowledge. There are constraints and conditions related to the production of speech: time limitations, types of speech required, planning, and presence or absence of listeners among others. As a result, the intended speech may not go according to the ideal plan and errors may arise. In test design, it is highly desirable to incorporate these constraints and conditions in test tasks to approximate real life communicative demands.

3.3.3. Speaking in second language testing

In the 1960's, tests of speaking ability were based on the ability to discriminate and produce words and structures of a language in the stress and intonation patterns of that language (Lado, 1961). In technologically advanced places, these tests were often conducted in language laboratories. The assessment of oral language ability according to the "Proficiency Movement" framework is based on the impression of the test taker's performance in specific contexts (Section 1.1.4.). The test often takes the form of an oral interview in which one or two interlocutors ask the examinee questions on pre-arranged topics. While the interlocutor pays attention to different aspects of the oral language such as pronunciation, fluency, accuracy and vocabulary etc., an overall score is assigned against a rating scale.

The current view of spoken language is that it is normally used to fulfil some communicative purpose. The communicative oral trait to be measured is usually defined for a given test purpose, its main components and their relationship identified because a well-defined construct helps language testers better understand what is being measured, and what variables affect this measurement.

Hymes (1974), in an attempt to identify and isolate the variables that interact in authentic language use, arrived at a model of authentic language performance involving linguistic, non-linguistic, social and psychological variables operating in constant interaction. Labov (1972) pointed out that oral language varies according to the task and circumstances in which the task is produced. Research findings in learner language indicate that interlanguage has both systematic and variable elements (Ellis, 1985). A learner will use one rule on one occasion and another on a different one (Tarone, 1983). This variability, occurring in all areas of the oral language, is caused by the task and context in which the oral language is produced (Labov, 1976). Tarone's (1983) research findings lead her to propose that different interlanguage styles fall along a continuum from vernacular (unattended speech) to careful style (attended speech); in her view (ibid), the vernacular is the most stable and consistent style, independent of L1 and L2 influences. She suggests learner language should be observed in the vernacular. Similarly, according to Ellis (1985), the vernacular is more natural than planned discourse because acquiring a second language means acquiring the ability to communicate spontaneously. Thus the assessment of oral ability will have to include tasks that elicit unmonitored spontaneous speech as well as tasks that require planned discourse depending on the purpose and context of the given task.

3.3.4. Framework of communicative speech events

In this section, a framework of speech events for the design of the TG Speaking test tasks is attempted. In accordance with the discussion in Sections 3.3.2. and 3.3.3., the framework includes the following components.

1. Purpose of assessment
2. Speech functions: whether interactional or transactional
3. Speech type: whether the speech is a long or short turn; planned or unplanned; formal or informal; reciprocal or non-reciprocal
4. Participants and role relationship: familiar or unfamiliar; large or small group and the relative status to the speaker
5. Topic: e.g. narrating, giving instructions or explaining etc.
6. Skills required: information or interaction routines

7. Setting: whether the speech is conducted indoor or outdoor
8. Gender of interlocutor:

3.3.5. Test methods

Earlier tests of oral proficiency were non-communicative; test takers were required to repeat words and sentences, giving answers to pattern questions and substitution drills. With the growing awareness of what speaking is and involves, and an emphasis on communicative language use, such tests were viewed as unauthentic. Testers began to treat the test taker as an active participant in meaning negotiation; they began to develop tests which could be considered communicative and authentic and which would better reflect psycholinguistic processes in speech production.

Nowadays an oral test requires the test taker to produce a real language sample, i.e., the kind of language spoken among real people for a purpose. Canale and Swain (1980) call for tests that require learners to perform their knowledge in meaningful and communicative contexts. Jones (1975) proposes tests that would provide information on functional language ability in tasks approximating real world communicative tasks. Morrow (1977) recommends giving test takers the opportunity for spontaneous use of the language in authentic settings and activities that the candidate recognises as relevant. Clark (1975) calls for direct tests whose format and procedures duplicate the setting and procedure of real language use. One example is the OPI.

In recent years, researchers have raised a number of questions as to the validity of OPI (Section 1.1.4.). First, they have questioned whether a test that measures a single type of oral interaction, i.e., conversation, can provide a valid sample of other types of interactions (Shohamy, 1983; Kramsch, 1986; van Lier, 1989). There are also questions regarding the effect of a number of contextual variables which influence oral production; variables such as the role relationship, personality and gender of testers and test takers, the purpose of the test and the setting tend to affect oral output and test scores (Shohamy, 1988; Brown, 1995; Chalhoub-Deville, 1995a and b; Turner, 1998).

It has been felt that there is a need to control these variables by conducting oral tests in a more uniform way to ensure reliability and validity without compromising their communicative features (Shohamy, 1994). Semi-direct oral tests have been developed in this direction. The two well-known examples are ETS's Test of Spoken English (TSE) and SOPI by the Center for Applied Linguistics. In this type of test, test takers respond to pre-recorded and visual tasks. Their proficiency levels are then evaluated by trained raters against a rating scale. Semi-direct tests are uniform tests because all test takers are required to perform similar language tasks that attempt to include a variety of communicative characteristics and to elicit a wide enough range of oral interactions and discourse strategies.

There are differences between a direct and a semi-direct test. A semi-direct test is based on one-way monologue, whereas in a direct test, there is two-way communication. In terms of structure, in a direct test a question is followed by an answer which is then followed by another question, whilst in a semi-direct test, instead of question-answer-question, the test taker is asked to perform different unconnected tasks. With regard to genre, an interview resembling a conversation takes place in a direct test, but in a semi-direct test, test takers are usually required to describe, report or give a monologue on a given topic. Studies comparing direct and semi-direct tests indicate (1) differences in the discourse used by test takers in the two types, (2) the semi-direct test is perceived as more difficult, and (3) the face-to-face interview is preferred, but if an alternative approach is required, test takers consider themselves adequately tested through semi-direct means (Clark, 1988; Shohamy, 1988). Findings suggest that the live interview is the preferable option, but that in circumstances in which such direct testing is not feasible, semi-direct testing is perceived to be valid.

3.3.6. Design of the Tour Guide Speaking Test

Speaking is difficult, complex and involves the interaction of the test taker's language knowledge and contextual cues present at the time. The quality of speech production may thus vary according to the test taker's language ability, the nature and complexity

of the tasks, and his/her executive ability. To develop the Speaking tasks, the following guidelines were suggested:

Aim: The test is designed to measure the skill and efficiency of would-be tour guides' oral communication.

Test construct: Spoken language fulfils different functions depending on the communicative purpose and contexts. Test takers were, therefore, required to produce speech samples appropriate to different occasions in tour guiding. In addition, contextual variables such as function/purpose, audience type, gender and status, topic, and formality are considered to play an important role in the test taker's speech production; therefore, they were incorporated into test tasks.

Sampling the oral language use domain: Ideally, the Tour Guide Speaking Test should consist of all speech interactions and contextual variables related to tour guiding. A test, however, can only tap a sample of the oral trait; therefore, language functions were restricted to those most salient in tour guiding. Test takers were required to produce both planned and unplanned speech.

Suggested components of speaking: See Section 3.3.4.

Test method: For this particular test administration, semi-direct testing was used because there was a lack of trained interlocutors to conduct the interviews. One disadvantage of semi-direct testing is the lack of real-life interaction. There is no way to follow up the candidate's responses. The advantage is the uniformity of elicitation procedures and economy.

Test takers listened to a pre-recorded question tape and recorded their speech onto a response tape which was later evaluated by two trained raters against a rating scale.

Charaterising speaking ability: Test takers were evaluated against a rating scale (See Section 3.5.)

3.3.7. Specifications of the Speaking Test

1. Test objectives

The test is to measure the test taker's oral ability as a tour guide. Candidates are expected to demonstrate that they can understand aural input and that they can respond effectively and appropriately.

2. Communicative tasks to be assessed:

a. introducing yourself to a group of people
b. interpreting and translating signs, notices and advertisements
c. providing information on Taiwan to an interested tourist group and answering enquiries
d. explaining and providing a tour itinerary
e. justifying a course of action

3. Language functions to be sampled:

a. Greeting,
b. Describing location, duration and features,
c. Comparing and contrasting,
d. Giving instructions,
e. Giving advice,
f. Giving opinions,
g. Requesting and reminding,
h. Explaining,
i. Presenting facts,
j. Presenting temporal sequence,
k. Sympathising and persuading, and
l. Explaining a course of action

4. Authenticity: authenticity is considered in terms of task type. Tasks which have been observed as common in tour guiding are designed to elicit different speech functions and types (See Points 5 and 6 below).

5. Speech functions to be assessed: interactional and transactional

6. Speech types: both long and short turns, planned and unplanned speech

7. Mode of communication: expository and evaluative routines in both transactional and interactional contexts

8. Time allotment: about 30 minutes

9. Number of tasks: 5

10. Test method: Semi-direct, tape-mediated. Candidates sit in a language lab. They listen to a test tape and record their response on an answer tape. The test tape

contains test instructions, test questions, pauses and repetitions necessary for candidates to carry out the test. A female NS American EFL teacher will do the recording.

11. Task types and sample items: Five tasks will appear in the test.

Task 1: Warming up

Candidates are asked to talk about themselves: their name, work, schooling, hobbies etc.

Task 2: Interpretation and Translation

Candidates will be given 10 short signs or sentences in Chinese. They are asked to translate those signs/sentences into English.
Example: Candidates read " "
Candidates say "*Please take off your caps/hats and please be quiet while inside the building.*"

Task 3: Answering short questions

Candidates will answer 10 questions they hear.
Example: Candidates hear "*In the western world, 7 is considered a lucky number and 13 is thought to be very unlucky. In Taiwan, what's the lucky number? The unlucky number? Why are they considered lucky and unlucky?*"

Task 4: Making an announcement

Candidates will be given a tour itinerary and asked to make an appropriate announcement.

Task 5: Solving a problem

A somewhat difficult but typical situation in which a decision has to be made will be presented to the test taker in which s/he is asked to decide a course of action and justify his/her action.

12. Scoring method: Each candidate will be evaluated by two trained raters. Task 1 in the test will not be evaluated

13. Proficiency levels: Refer to *Tour Guide English Language Oral Proficiency Levels* (See Section 3.7.). In addition, symbols such as up (↑) and down (↓) may be used at each of the five levels. For example, at Level 3: *Competent*, 3↑or 3↓ can be used by the raters when they think the candidate is better than competent but not quite good to be assigned to the next higher level or the opposite.

14. Raters and training sessions: Each candidate is evaluated by two raters drawn from the following pool of persons:

- three native EFL teachers (NS),
- three non-native EFL teachers (NNS),
- two tour guide trainers (NNS-tg)

There is a chief assessor who organises the training and rating sessions. Training sessions are held after the pilot test for one day. Extra sessions will be provided if required. Tapes containing different levels of proficiency speech samples are provided. Raters listen to the tape, evaluate the speech samples and discuss any differences/disagreements they may have.

15. Marking of the test: each candidate is marked by 2 raters with one of the following combinations:

(a) NS : NNS
(b) NNS : NNS-tg
(c) NS : NNS-tg

In case of a discrepancy in marking, the chief assessor or an EFL teacher evaluates the tape and give the final rating. A discrepancy refers to no overlapping of the two ratings of a particular candidate. For example, if Rater A gives a 3↑ (*Competent*) and Rater B gives a 3↓, the candidate gets a 3 (*Competent*). If the candidate gets 3↓ from Rater A and 2↑ (*Satisfactory*) from Rater B, the candidate also gets a 3 (*Competent*). However, if the candidate gets a 3↑ from Rater A and 2↓ (*Satisfactory*) from Rater B, a third rating is required.

3.3.8. Development of the Speaking Test

The Speaking Test consisted of five different tasks. The use of different tasks was to give the test taker fresh starts and to elicit different language use samples. The following principles were used in task development:

- There is a purpose for each of the tasks.
- The tasks elicit both short exchanges and extended speech.
- For the tasks that attempt to elicit extended speech, there should be a graded scale of support in the information to be conveyed for the test takers, which ranges from a lot of external support such as providing the visual input for the speech sample to be elicited to much less external support such as giving your reasons for undertaking a course of action.
- The input stimuli should be controlled. Each of the tasks dictates what the test taker has to talk about. The tasks selected are considered typical in tour guiding.

Tasks were then checked by six native speakers of English (including 3 EFL teachers) and two expert informants for difficulty and content relevance. In addition, three

applied linguists checked the test difficulty. The test was also pre-tested for difficulty with native speakers of Mandarin before the pilot test. The procedures are similar to those reported in Section 3.1.4. The final version of the Speaking test is presented in Section 3.7. Table 3-2 displays the speech events of the tasks.

Table 3-2: Speaking Test speech events

Task 1: Warming up	Purpose: Ability to introduce oneself Speech function: interactional Speech type: planned, unreciprocal, informal Role relationship: to clients Skill required: interactional routines Setting: indoor/outdoor
Task 2: Translation & Interpretation	Purpose: Ability to translate signs or interpret to the clients Speech function: transactional Speech type: unplanned, unreciprocal, formal & informal Role relationship: to clients Skill required: information routines Setting: indoor/outdoor
Task 3: Answering short questions	Purpose: Ability to answer client's short questions appropriately Speech function: transactional Speech type: unplanned, unreciprocal, quite formal Role relationship: to clients Skill required: mainly information routines, Setting: indoor/outdoor
Task 4: Making an announcement	Purpose: Ability to make an announcement to a group of tourists Speech function: transactional Speech type: planned, formal, unreciprocal Role relationship: to clients Skill required: mainly information routines Setting: indoor/outdoor
Task 5: Solving a problem	Purpose: Ability to state a problem and explain the course of action to be followed Speech function: transactional Speech type: planned, formal, unreciprocal Role relationship: to clients Skill required: information routines Setting: mostly indoor

3.4. The Tour Guide Grammar Test

3.4.1. Why include a Grammar Test

Mastery of linguistic form has traditionally been part of language assessment. The grammar component has been included in many of the national examinations and commercially available tests such as the JCEE, FLPT and TOEFL. Presently, trends in foreign language teaching seem to encourage a practice of focusing on form which is expected to improve learner performance (Sharwood Smith, 1993; Williams, 1995).

This trend seems to suggest recognition of the fact that assessment of grammatical proficiency will remain a component in language teaching and learning. Moreover, tests can provide some learning experience and diagnostic information about the test taker's areas of strengths and weakness. Finally, test users and test takers usually expect a grammar component in a language test, and this may enhance the face validity of the TG test battery. Therefore, a grammar test was included.

3.4.2. The construct to be measured

In the communicative competence framework, grammatical competence is only a part of a single competency for effective language use. The ability to understand and produce grammatical utterances is a very restricted ability; more important is the ability to understand and produce utterances appropriate to the language use context. From the communicative perspective, when language testers are trying to measure the learner's grammatical competence, they are in fact measuring the first three of the five dimensions of language production suggested by Long (1990): *grammatical accuracy*, *appropriacy*, *adequacy*, *truth value* and *morality*. An utterance may be grammatical but it may still fail to achieve its goal in a given situation because it may not be appropriate in that language use context. It is not easy to assess *truth value* and *morality* in a grammar test nor is it necessary. But language forms, their social meanings and the appropriateness of use can be measured provided the contexts are clearly defined.

Traditional grammar tests measured the learner's grammatical ability at the sentence level; correctness was the main concern. Within the communicative competence framework, however, grammatical ability is viewed as a communicative device that is functionally motivated (Halliday, 1985). It is also an essential component to appropriate language use (Hymes, 1972). Therefore, a communicative grammar test should attempt to assess linguistic forms, their meanings and their appropriate use.

Rea Dickins (1991) proposes a three-level definition of communicative grammar: form, meaning and use, which serves as a working definition of the construct which the TG Grammar test seeks to measure.

Grammar as form reflects a view that grammar consists of a system of rules. A closer examination of structural rules, however, often reveals a mismatch between structures and their associated meanings in everyday use; for example, the prescribed use of *some* and *any* in pedagogical grammar books and the use of present tense in a sports commentary. This is because linguistic meanings are often further conditioned by lexis, contexts and conventions governing language use in any speech community. In short, an individual's grammatical competence goes beyond grammar at the sentence-level to include a dimension relating to grammar in use. The assessment of the learner's grammatical knowledge has to integrate form and meaning to produce sociolinguistically appropriate use in a given context. A grammar test should at least attempt to measure grammatical knowledge at three levels: forms, meanings and their use.

3.4.3. Measuring grammar as forms and meanings

Since grammar is viewed as an integration of form and meaning, items requiring syntactic decoding alone, as shown in Example (1), are not sufficient as a means to predict the learner's ability to use grammar.

Example (1):

In the following question, choose A, B, C or D to complete the sentence or question.

Who _____ your teacher?
A: be B: is C: are D: were

Example (2) demonstrates some integration of syntax and semantics of the linguistic description.

Example (2):

In the following question, choose A, B, C or D to complete the sentence or question.

Jane _____ before crossing the road but she didn't.
A: must look B: may look C: would have looked D: should have looked

Items which follow a discrete-point approach, inevitably are also decontextualised. They may not be a good instrument for the measurement of grammatical competence as the selection of the correct answer could be done by a mere process of elimination or guessing. An alternative approach would be to produce the required grammatical forms appropriate to the context as shown in Examples (3).

Example (3a):
Read the passage carefully and supply the correct verb form according to the meaning of the passage.

If you look ahead of you, you _____ (see) the Presidential Office Building...

Example (3b):
Finish the second sentence so it has the same meaning as the sentence before.

Be careful or you'll hurt yourself.
If _____.

Examples (2) and (3) require attention to more than linguistic well-formedness. They require learners to relate form to the contexts and meaning. However, the item types in those three examples are still concerned with grammatical accuracy. Grammar in the communicative framework is seen as a way to convey meaning where meanings are created within contexts (Rea Dickins, 1991). Correct grammatical forms are selected on the basis of the meanings they are able to express. If items are constructed as an end in themselves, they can only tell us that some structures are possible and may occur; they do not reveal to the users of the contexts and the purposes of their occurrences. In a grammar test, the contexts and functions which items attempt to perform in a text are equally important. Example (4) is an illustration of testing grammar at a higher level of analysis with respect to the context and purpose of the occurrence of grammatical structures.

Example (4):

A word is missing in each of the blanks in the following newspaper article. Read the article and fill in the missing words. Each blank requires ***one word*** *only. The first blank has been done for you as an example.*

The ruins of Taiwan city are located in An-ping, Tainan City. Built by the Dutch in 1624, An-ping Fort was the first to be built in Taiwan. The fort was first known _____ Fort Orange. Later, the name was changed to Fort Zeelandia. The fort was built with red bricks brought _____ Indonesia....

To some extent, such an item type, though focusing on specific discrete grammar points, provides context and forces the learner to use his/her linguistic knowledge for communicative purposes. In other words, users of the text have to produce grammatical elements determined by their functions within the sequence of linguistic events in the text. To do this successfully, the test takers have to analyse the linguistic context and event, and make decisions about the ways grammar functions to signal meaning rather than a mechanical recall of the right grammatical elements.

So far, item types discussed do not assess the level of use. The reasons are: (1) The questions are all decontextualised; (2) There is no authentic communicative focus; (3) Test takers are controlled in the completion of items. To compel attention to use in a grammar test, grammar tasks should be specified in terms of a realistic situation and a focus on the exchange of information as Example (5) illustrates.

Example (5):
A situation is given to you. Read the situation and write two possible responses if you were in that situation.

You are buying some "Thank You" cards in a stationer's for the office. You want a receipt. You ask the shop assistant for it. What would you say?

Response 1: _____
Response 2: _____

Such an approach provides a context and reasons for engaging in the activities for the test takers. It also provides the opportunity for the test taker to create his/her own message and produce grammatical response appropriate to the context. In this type of task, there is an integration of form, meaning and use.

3.4.4. Item types in a grammar test

Depending on the purpose of a test, assessment of grammar as structure and semantics through discrete-point items can be as valid if these items bring in meanings and contexts. Rea Dickins (1991) suggests that items aimed at assessment of grammatical forms and their associated meanings through a discrete-point approach tap different aspects of grammar than items with a focus on communicative grammar. The latter has been suggested to activate a network of integrated competencies on syntax,

semantics and pragmatics that functions as a means for successful communication. It therefore seems logical to (1) include various item types from discrete-point items to items that assess more global language use in a grammar test and (2) tasks that require conveyance of meaning through the choice of appropriate linguistic forms.

3.4.5. Test design and specifications of the Grammar Test

Aim: The test was designed to measure a would-be tour guide's grammatical knowledge and his/her ability to use the knowledge appropriately in the context of tour guiding.

Test construct: For this particular test purpose, to be grammatically competent, the candidate has to

(1) have knowledge of English usage and

(2) be able to use the forms meaningfully and appropriately in different linguistic and situational contexts.

In other words, grammatical competence refers to an understanding of language forms, their meanings and the ability to integrate these forms and meanings for appropriate use.

Test specifications:

1. **Test objective**

The test is to measure a prospective tour guide's grammatical and pragmatic knowledge of the English language. The focus is on (1) the formal structure of English, (2) English vocabulary with special attention to those features/vocabulary typical in tour guiding and (3) the candidate's ability to identify characteristics of transactional and interactional language use.

2. **Language elements to be measured**: Test items attempt to assess accuracy, appropriateness and the candidate's ability to select form according to discourse context. For detail, see *Grammar Test Item Specifications*.

3. **Time allowed**: 40 minutes

4. **Number of items**: 65

5. **Test tasks**: Test tasks require identification and selection of appropriate answers and language production.

6. **Item types**: The test should include the following question types.

Multiple-choice questions (12 questions)	Q1 - 12
Verb form (10 questions)	Q13- 22
Sentence transformation (15 questions)	Q23 - 37
Fill in the blanks (13 questions)	Q38 - 50
Complete the conversation (10 questions)	Q51 - 60

7. **Sample item**: See *Grammar Test Item Specifications.*

8. **Scoring method**: All items should be objectively markable. A comprehensive answer key will be provided. Candidates will be awarded 1 point for each response correctly answered*. Full score is 65 points. Candidates need to score 26 points to pass the Grammar Test. (*For Questions 56 - 60, each item requires two responses from the candidates; each response is treated as one item and is worth of 1 point.)

9. **Marking scheme**: Appendix 3-1.

Grammar Test Item Specifications

1. **Language system elements to be sampled**:

Nouns, pronouns
Verbs
Adjectives
Adverbs
Articles, possessives, demonstratives
Prepositions
Infinitives and gerunds
Conjunctions
Modal verbs
Adverbials
Negative: not, too…to
Conditionals: if, wish, unless
Comparison
Tense: present, past & future
Aspect: present progressive, past progressive, present perfect
Active and passive
Word order
Clauses: relative, subordinate
Reported speech

2. **Sample items**:

A. Multiple-choice questions:

For Questions 1 - 12, read the question and choose the best option that completes the sentence/question.

Example: *What time ____ it now?* Ans.: B

A: be
B: is
C: are
D: were

B. Verb forms: For Questions 13 - 22, two passages are given. Read the passages and

supply the correct verb forms according to the meaning of the passages.

Example:

Located on the outskirts of Taipei, Shrlin is the largest of Taipei's 16 administrative districts, _(cover)__ a total area of 64.87 square kilometres. The first Chinese settlers in this area __(be)__ mostly from Zhangzhou in mainland China. They __(settle)__ around the Jrshan Rock for its defensive advantages and __(establish)__ a thriving commercial centre.

C. Sentence transformation: For Questions 23 - 37, rewrite the second sentence in such a way so that it has the same meaning as the sentence printed before it. Write your answers on the answer sheet provided.

Example 1: *Be careful, or you'll hurt yourself.*

If _____ you'll hurt yourself.

Example 2: *I wasn't comfortable. Then I had the painful tooth extracted.*

Until I _____.

D: Fill in the blanks: For Questions 38 - 50, some words in the passage have been left out. Read the passage and supply each of the gaps with one suitable word.

Example:

A: *Grandfather, who was Matsu?*

B: *Well, Matsu is the Goddess of the Sea. According to legend, Matsu was ___1__ in China during the Sung Dynasty in Fujian province. ___2__ was a daughter of a fisherman. Matsu ___3__ supernatural powers as a young girl. One day, she __4___ asleep. In a dream, she saw ___5__ father and brothers drowning in __6___ sea. In the dream, she lifted a brother in __7___ of her arms and caught her __8__ shirt in her teeth. Just as she __9___ about to bring them back ashore, her mother called __10__ name. She opened her mouth to __11___ and lost grip of her father's shirt. __12___ she woke up, her face was wet __13___ sweat and her low garments were soaking __14___. Her brothers were miraculously saved __15___ her father drowned.*

E: Complete the conversation: There are two parts to Complete the conversation. For questions 51 - 55, one side of a conversation is provided. Read the conversation and complete the other half of it according to the suggested cues.

Example: Angela: I mustn't forget to collect the tickets.

You: **Offer to do it** _____.

Angela: Will you? Thank you. That's very kind of you.

For Questions 56 - 60, a situation is given. Read the situation and write two most appropriate responses if you were in that situation.

Example: *You are buying some "Thank You" cards at a stationer's for the office. You want a receipt. You ask the shop assistant for it. What would you say?*

Response 1: _____

Response 2: _____

3. **Language input**: Test instructions and items are written in the target language. They are printed in an A4 size test booklet. Candidates are required to read and then either select or produce the answers.

4. **Expected response**:

Questions 1 - 12: Candidates are required to select the answer from 4 possible answers
Questions 13 - 22: Candidates are required to produce the correct verb forms ranging from 1 - 2 words
Questions 23 - 37: Candidates are required to complete a sentence according to the cues.
Questions 38 - 50: Candidates are required to write one word per item.
Questions 51 - 60: Candidates are required to write 1 - 2 sentences per item according to the cues.

5. **Recommended time allotment**:

Multiple-choice questions: approximately 4 minutes
Verb forms: approximately 3 minutes
Sentence transformation: approximately 15 minutes
Fill in the blanks: approximately 8 minutes
Complete the conversation: approximately 10 minutes

3.4.6. Development of the Grammar Test

There were two stages to the development of test tasks: material selection, item writing/revision.

Material selection: Language elements were sampled from high school English textbooks and EFL tourism textbooks. Texts were selected and modified from various sources such as newspapers, tourist pamphlets, government publications related to tourism, EFL textbooks and language sampled from field observation.

Item writing: Questions were constructed according to pre-specified item types listed in the Grammar test specification.

1. Orientations of questions: Grammar questions focused on the following three areas:
    - Ability to identify and produce linguistically well-formed sentences
    - Ability to produce parallel structures with similar meaning and function
    - Ability to produce grammatical sentences as appropriate to the contexts.
2. Types of responses: The candidates were asked to either identify or produce their own language samples according to the types of tasks.
3. Question types: Types of questions included MCQs, providing the correct verb forms, filling in the blanks, sentence transformation and providing appropriate responses.
4. Item checking: Six native speakers of English including three NS EFL teachers checked the items for difficulty and suitability. The tour guide informants and three applied linguists further checked content appropriacy. Five Taiwanese postgraduate students volunteered to take the test in order to check item difficulty, time allotment and clarity of test instructions from the test taker's point of view.

Item revision: In the development stage, items were revised whenever there was feedback from the experts. A major revision was made after the pilot test.
Final version: See Section 3.7.

3.5. Development of the rating scale

3.5.1. The use of the Tour Guide rating scale

There are two purposes in the use of the Tour Guide rating scale: to describe levels of performance to the test takers and the test users, and to provide a guideline for the raters. In the rating scale, language abilities are organised to form a hierarchy; the descriptions explain what is meant by a given level and help test users to interpret test

results more easily. The rating scale also acts as a common standard for the raters to ensure reliability and validity. For some scales such as the ACTFL Guidelines, the descriptors serve a third purpose: to provide guidelines for test construction, but this is not applicable to the TG scale.

3.5.2. Development of the Tour Guide rating scale

This section reports how descriptions for the Tour Guide rating scale were arrived at. Problems associated with score interpretation are discussed.

Components of the scale

The Tour Guide rating scale includes three broad band scales: an overall general proficiency band scale, and two sub-skills scales on Listening and Speaking. The general proficiency scale was later abandoned (Section 4.5.). The Speaking scale consists of a general scale and a scale on each of the four categories: *fluency*, *pronunciation*, *grammatical accuracy & vocabulary range*, and *communicative effectiveness*. *Grammatical accuracy & vocabulary range* also accounts for scores on the Grammar Test. Six levels are assigned in each of the scales with a number value attached to each of the levels: (0) *No foreign language use ability*, (1) *Below standard*, (2) *Satisfactory*, (3) *Competent*, (4) *Good,* and (5) *Distinction*. The general scale was meant to serve as a quick indication of the test taker's overall language use ability. The descriptors in Listening and Speaking provide indication of ability in specific areas of interest.

The levels are modified from ESU Framework (Carroll & West, 1989). In spite of some doubts raised by language testing researchers (Fulcher, 1987; Matthews, 1990), the Framework offers a comprehensive series of descriptions of language performance and can be readily revised for any given examination. Given the limited time for TG test development, the availability of the Framework seemed to be the best option at the time. The scale's validity was empirically validated with the use of the Facets statistics purchase (Section 4.2.3.).

Steps in modification of the descriptors

The TG proficiency levels were modified according to the following four steps.

(1) ESU Bands 3 - 8 were selected to form the basis of the draft descriptors. The nine-point scale was not used because the concern was to differentiate more able language users from less able ones; people whose language ability is either below or above the range of language ability of interest were not our concern. It was decided to use a 5-level scale originally. Level 0 was later added because some of the students did not attempt any questions. It was not clear whether these candidates failed to understand the questions or they were not interested in doing the test; therefore, *(0) No foreign language use ability* was assigned to any one who failed to respond.

(2) The band levels were then modified against real language samples from practising tour guides recorded during workplace observation. University EFL lecturers, experienced EFL teachers and raters were asked to rate the samples against the rating scale and adjustments were made.

(3) Two additional EFL teachers and two native-speakers of English were asked to rate the samples against the scale. This was to check the clarity of the descriptors and to obtain further comments on the scale.

(4) The scale with the language sample was sent to the expert informant for further comments. Further adjustments were made in the training session.

The final version of the Rating Scale is presented in Section 3.7.

Difficulties encountered in modification of the scale

Several difficulties were faced when modifying the descriptors. The first was to decide whether to keep the Tour Guide scale as a general proficiency scale or to change the descriptions to fit the contexts of tour guiding. It was decided to keep the scale as a general proficiency scale for two reasons. First, the language domain of interest is tour guiding but language functions and features observed did not seem to deviate very much from those observed in any other language use contexts; they were just more common in tour guiding. Second, the language data obtained were not considered to be extensive and representative enough to derive a scale with specific relevance to language use in tour guiding.

The second problem was the progression of proficiency through the levels, which includes consistency of differences between levels. It was felt that a consistent progression through levels was necessary for the users because it seems intuitive to interpret levels of proficiency as an interval scale. However, there seems to be no clear indication of a linear progression of language development. Unless empirically verified, the descriptors are only what language testers think is happening in communicative situations (Fulcher, 1987). Alderson's (1997) solution of careful choice of quantifiers was followed. Raters during the training session were asked to agree on their understanding of the different quantifiers used in the descriptions. But the problems of linear progression and the validity of the descriptors still remain. It is hoped that with the use of statistical methods we will gain insight to the reliability of the oral scale. (Section 4.2.3.).

Relating scores to scales

It was decided to relate raw scores on the Speaking Test to the Speaking scale. For the Listening and Grammar tests, scores were converted into band levels. The pass/fail and a cut-off score for each test was suggested and decided after statistical information on the test items was available (Section 4.4). As there are no other comparable tests to relate the TG test scores to, it was decided to use expert judgements and statistical information to arrive at the level for a pass. It was also decided to report scores as well as their equivalent band levels because scores alone tends to give a false impression of the accuracy of test results. Bands, on the other hand, give a range of possible scores on test behaviour and enable us to interpret test results with more confidence.

The scale was not empirically validated before its use; when interpreting the test scores, test difficulty, test method, test content and the performance of other test takers were considered as well (Section 4.5.).

3.5.3. Rater training

Eight volunteer raters were trained to evaluate the candidates' performance: three NS EFL teachers, three NNS EFL teachers and two tour guides. Four of them are also

experienced raters. The rating scale with a marking sheet and six speech samples were given to the raters a week before the training for them to familiarise themselves with the rating procedure. The speech samples consisted of recordings of tour guides. Speech samples from the test were not available at the time of Training Session A. We had planned to obtain some student samples after the pilot test for the raters to evaluate together.

Two training sessions were arranged: one before the pilot test (Session A) and one after (Session B). Session A lasted for 3 hours and Session B, 1.5 hours. One of the tour guide informants was away on an assignment therefore he was not able to evaluate the tapes. He asked to be removed from the rater list. In Session A, the seven remaining raters first went through the rating scale to clarify any misunderstanding or misinterpretation of the descriptions. Then they compared their ratings on the sample speeches and discussed the reasons they gave a particular rating based on the descriptions and scales. In the last part of the training session, raters evaluated five more speech samples together. Finally, raters agreed on the use of analytic rating. Final rating was to be the mean of the four sub-categories of the rating scale.

Session B was arranged after the pilot test. Because of a technical fault, only two tapes were recorded. It was decided to ask the raters to evaluate the two speech samples together. Two EFL teachers (Raters II and LL) were not able to attend because of teaching obligations. A separate session was arranged for these two.

In general, the following observations were noted:

1. Experienced raters favoured holistic rating; inexperienced raters preferred analytic rating.
2. Experienced EFL teachers (more than 10 years) were more lenient than less experienced EFL teachers (1-3 years).
3. Raters assigned the same overall rating but their ratings differed in the four sub-scales.
4. The tour guide rater seemed to follow the descriptions in the scale more accurately.

For (1), we agreed to use analytic rating. Rater severity was a more pressing problem. Six speech samples of different language ability were given to the two raters with less teaching experience as an attempt to provide them with more exposure. In one extra session, the three of us listened to each sample, discussed the descriptions in the sub-scales and decided on one most appropriate rating. By the end of the session, we seemed to agree on the ratings. However, according to the Rasch results, one of the raters (Rater JJ) was too severe (Section 4.3.). Rater LL, the NS rater, was the second severest. It is probable that because she did not participate in Session B, she was not able to discuss her ratings with other raters. If the TG test was to be implemented on a national basis, then the presence of all raters in all training sessions would be required. Since all the raters were volunteers, arrangements had to be made to suit their convenience.

3.6. Criteria to ensure the Tour Guide test usefulness

In this section, measures taken to maximise the validity and usefulness of the TG test are presented and discussed.

Test specifications describe the contexts that can provide evidence about the test taker's ability and provide judgmental guidelines for the mapping of the observed behaviour to inferred ability. However, this is not necessarily direct evidence about language competence. Language testers and test users may have to interpret the observed behaviour (i.e., test performance) in the light of other supporting evidence about how the content or tasks interact with the test taker and how the test taker or test user see the tasks. In other words, language testers need evidence that is coherent, principled and useful for the test purpose, and tests should be designed and developed to capture such evidence.

Bachman & Palmer (1996) propose six qualities for the evaluation of test usefulness to assist test design and development, which I used as the criteria to ensure the usefulness of the TG Test. These qualities are reliability, construct validity, authenticity, interactiveness, impact and practicality. Sources of possible invalidity and facets affecting validity were identified, and a minimally acceptable level in each

of these criteria was set to ensure the TG test would be useful. Table 3-3 lists these sources and facets and the suggested acceptable standards and procedures to maximise the usefulness of the TG test.

Table 3-3: Criteria to ensure TG test usefulness

Test Quality	**Sources of possible invalidity & facets affecting validity**	**Standards and procedures to maximise usefulness**
Reliability: The following procedures will be used to collect empirical evidence: 1. Administer a pilot test to a sample of 15 students and calculate item facility values and discrimination indices. 2. Speech samples rated by raters and calculate inter-rater reliability. 3. Model-fitting, i.e., item fit statistics	1. Setting: The test will be taken in language labs with partitions between seats to minimise interference, particularly during the Speaking Test. 2. Test input: Language input is in English. 3. Format of response: Generally, short responses are required. Grammar Part V requires a complete phrase/sentence. 4. Internal consistency of items: A pilot test will be administered to collect data on items. Item revision will follow. 5. Scoring: Tests will be marked by EFL teachers with answer key. 6. Rating behaviour: Training sessions and examples of different ability levels will be provided to reduce inconsistency.	Reliability estimate should be high.
Construct Validity:	1. The language domain selected does not fit the area of target language use. 2. Constructs are not defined appropriately. 3. Tasks designed do not engage the test taker in the language area measured. To minimise the above, field observations, interviews with senior tour guides will be conducted. Further, expert informants will be consulted during test design and development.	To examine if the test measures the constructs proposed, the following are considered: 1. Content/task relevance and representative-ness: high 2. Internal relationship of the tasks: high 3. Score stability: high 4. Relationship with an external criterion: moderate
Authenticity	1. Characteristics of tasks do not correspond to descriptions of the characteristics of target language use 2. Task characteristics such as input, response etc., do not match those described in test specifications	There should be a complete match between critical features in the target language use areas and test tasks. Tests and their test specifications should be reviewed by field experts, EFL teachers and test takers.
Interactiveness	1. Test tasks do not engage the test takers in areas specified in the test specifications. 2. Involvement of factors other than language ability, background knowledge, & the strategic competence	Language ability should be high. Background knowledge and affective schemata levels should be moderate. Questionnaires will be given to the test takers and expert informants for their comments on the

		test content and test tasks.
Impact	1. Wrong decisions made because of incorrect interpretation of score. 2. The test takers do not fully understand the test purpose.	The impact of the test is high as the TG test is a relatively high-stakes test. To reduce (2), the test takers are informed of the purpose of the test and are given sample test items for preview. To prevent (1), experts consulted to make decisions on the cut-off score and test performance is to be reported in terms of a band level.
Practicality		Resources used will be monitored. They include human resources, and material resources

3.7. The Tour Guide English Test battery and the rating scale

This section presents the complete versions of the TG test battery, the rating scale, and in the case of the Listening and the Speaking tests, the scripts. The rating scales were modified from the ESU Framework; levels and descriptions in the framework have been adjusted so that the ability levels for the present test purpose could be more clearly reflected.

Form Code: TGE L001

Name: _____ **Registration Number:** _____

The Tour Guide English Language Listening Test

Test Booklet

Time allowed: About 40 minutes

Number of questions: 45

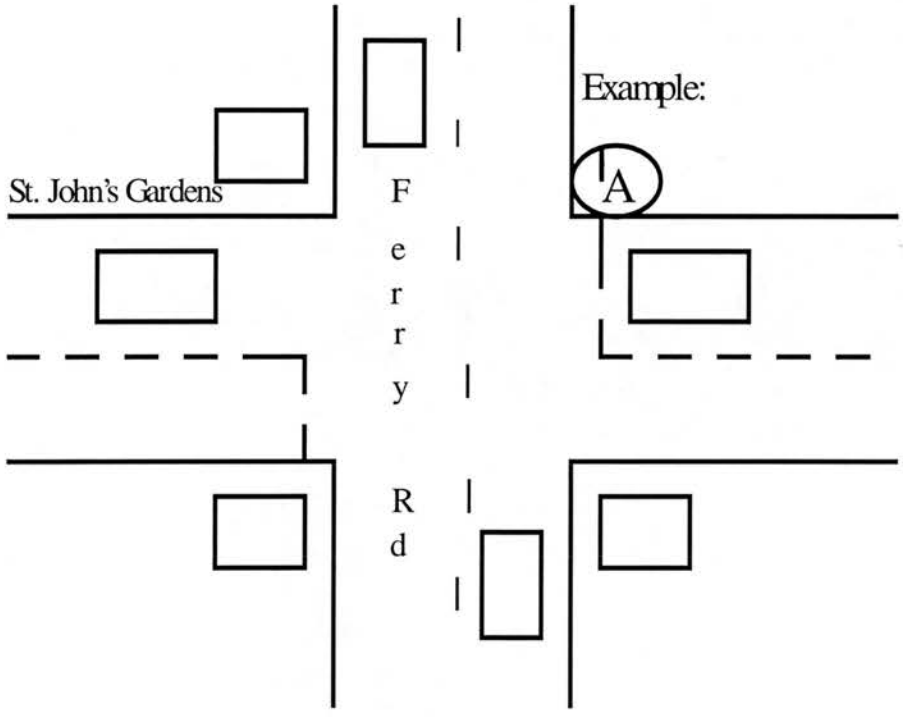
General instructions:

There are 6 listening passages with 45 questions altogether in 9 tasks. The tasks are printed in this test booklet. In each of the tasks there are several questions. Listen to each passage and then do the task. You will be given time to read through the questions before you listen to each passage. After you listen, you will also be given time to write your answers. You may take notes while you are listening to the passages. Write your notes on the paper provided.

(The test continues on the next page.)

Task 1 (Listening Passage 1)

For Questions 1 - 3, listen to the conversation between a policewoman and the driver and indicate the positions of the Volkswagen, the Honda Civic and the witness before the accident in the appropriate boxes in the picture below. The "SLOW" sign has been indicated for you as an example.



Example:

A: "SLOW" sign

(1) B: Volkswagen

(2) C: Honda Civic

(3) D: the witness

(The test continues on the next page.)

Task 2 (Listening Passage 2)

For Questions 4 - 11, listen to the story of Agatha Christie and fill in the missing information.

Example: Agatha Christie is known as *the Queen of Crime*

Section A: Personal data

Date of birth:	(4) _ _ _ _ _ _ _ _ _ _
Place of birth:	Torquay, England
Education:	At home, by her mother School in Paris
Dates of marriages:	(5) _ _ _ _ _ _ _ _ _ _ (6) _ _ _ _ _ _ _ _ _ _
Names of her husbands:	Archibald Christie Max Mallowan
Date of death:	(7) _ _ _ _ _ _ _ _ _ _

Section B: Some important dates in her life

Date	What happened/ What she did
1914 - 1918	worked in a hospital
(8) _ _ _ _ _ _ _ _ _ _	disappeared from her home and was found in Yorkshire suffering from amnesia.

(The test continues on the next page.)

Section C: Four of her books mentioned in the talk

Titles	Dates published
The Mysterious Affair at Styles	(9) _ _ _ _ _ _ _ _ _ _
The Murder of Roger Achroyd	1926
(10) _ _ _ _ _ _ _ _ _ _ _ _ _ _ _	1934
And Then There Were None	(11) _ _ _ _ _ _ _ _ _ _

(The test continues on the next page.)

Tasks 3a and 3b (Listening Passage 3)

Task 3a

For Questions 12 - 15, decide whether each statement you hear is True or False. Write **T** for true and **F** for false. The first one has been done for you as an example.

Example: _T_

(12) _____

(13) _____

(14) _____

(15) _____

Task 3b

For Questions 16 - 21, match the colours used in Chinese opera in Box 甲 with their associated meanings in Box 乙. Not all meanings will be used. The first one has been done for you as an example.

Box 甲

Example: _A_ Red		
_____ (16) Green	_____ (17) Black	_____ (18) White
_____ (19) Purple	_____ (20) Yellow	_____ (21) Blue

Box 乙

A: Righteousness	B: Pride	C: Reliability	D: Honesty
E: Cunning	F: Danger	G: Cheerfulness	H: Insincerity
I: Faithfulness	J: Greed	K: Anger	

(The test continues on the next page.)

Task 4 (Listening Passage 4)

For Questions 22 - 26, listen to the talk on traditional architecture in Taiwan and then answer each of the questions you hear in 4 words at most. The first one is an example.

Example: Southern China

(22) _____

(23) _____

(24) _____

(25) _____

(26) _____

(The test continues on the next page.)

Tasks 5a and 5b (Listening Passage 5)

Task 5a

Listen to these announcements at Edinburgh Railway Station and fill in the details of the train on the grid. The first one has been done for you as an example.

	Destination	**Departing from**	**Platform**	**Time of departure**
Ex.	Glasgow (Queen Street)	Edinburgh	14	11:30
(27)	Edinburgh	Glasgow (Queen Street)	_____	11:20
(28)	London (King's Cross)	Edinburgh	19	_____
(29)	Plymouth	Edinburgh	10	_____

Task 5b

Listen again to the announcements and complete the sentences in Questions 30 - 32.

(30) The 11:30 shuttle service to Glasgow Queen Street offers _____ and light refreshments.

(31) Mr. Jonathan Brown should go to _____ to meet his mother.

(32) The first-class accommodation on the train to London King's Cross can be found _____.

(The test continues on the next page.)

Tasks 6a and 6b (Listening Passage 6)

Task 6a

Listen to the tour guide talk about famous places on the Royal Mile in Edinburgh, Scotland's capital. Put the correct letter next to the places mentioned in the box. The first one has been done for you as an example. The tour starts from the castle and is heading east to Holyroodhouse Palace.

(map of the Royal Mile)

This stretch of road is called the Royal Mile.

Example: _C_ Writers' Museum

(33) _____ St. Giles Cathedral

(34) _____ Museum of Childhood

(35) _____ John Knox House

(36) _____ People's Story Museum

(The test continues on the next page.)

Task 6b

Listen to the talk again and fill in the missing information. The first one has been done for you as an example.

Example:

Length of the Royal Mile: *a mile and 200 yards*

(37) Edinburgh Castle: Opens _ _ _ _ _ _ _ _ _ _ _ _ _ _ _.

(38) The Writers' Museum: Admission is _ _ _ _ _ _ _ _ _ _ _ _ _ _ _.

(39) Oldest parts of St. Giles: the _ _ _ _ _ _ _ _ _ _ _ century.

(40) John Knox House is now a _ _ _ _ _ _ _ _ _ _ _ _ _.

(41) The Museum of Childhood: Described as "the _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ place in the world."

(42) The word GATE in Canongate means _ _ _ _ _ _ _ _ _ _ _ _ _ _ _.

(43) The Tolbooth is Canongate used to be a _ _ _ _ _ _ _ _ _ _ _ _ _ _ _.

(44) Holyroodhouse Palace: Official residence of the Queen who usually comes in the month of _ _ _ _ _ _ _ _ _ _ _ _ _ _ _.

(45) Holyroodhouse Palace: Now there is an _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ of water colours.

-The End -

The TG Listening Test

(Tape script - listening)

The Tour Guide English Language Listening Test

General instructions:

There are 6 listening passages with 45 questions altogether in 9 tasks. The tasks are printed in your test booklet. In each task there are several questions. Listen to each passage and then do the task. You will be given time to read through the questions before you listen to each passage. After you listen, you will also be given time to write your answers. You may take notes while you are listening to the passages. Write your notes on the paper provided. Do you have any questions? If you do, please raise your hand and an assistant will come and help you. ...(pause 5 seconds)... If you don't have any more questions, let us begin the test ...(pause 2 seconds). Now turn to Task number 1 in your test booklet and listen. (5 seconds)

Listening Passage 1

Listen to the conversation between a policewoman and the driver of a car which has been involved in a road accident. The policewoman is asking the driver for details of the accident. Listen and mark the positions of the Volkswagen, the Honda Civic and the witness before the accident in the appropriate boxes in the picture in Task number 1. The location of the SLOW sign has been indicated for you as an example. Now look at the picture for 10 seconds...(Pause 10 seconds.)..... *Now listen carefully to the conversation.*

Policewoman: And what's your name, sir?

Drummond: Drummond. John Drummond.

Policewoman: Do you live around here, Mr. Drummond?

Drummond: Yes, I do. I live on this road, actually. Ferry Road. Number 18.

Policewoman: So your address is 18, Ferry Road. And are you the driver of this Volkswagen?

Drummond: Yes, I am. And just look at this....

Policewoman: The Honda Civic is in pretty bad shape too.

Drummond: The idiot drove too fast!

Policewoman: Can I see your driver's licence, sir?

Drummond: Yes. Here you are.

Policewoman: How fast were you driving?

Drummond: I was under the speed limit. I know this road very well. I go up and down it every day.

Policewoman: I see. Well, exactly what happened?

Drummond: I was driving up Ferry Road. I slowed down a little as I was passing the junction with St. John's Gardens.

Policewoman: Did you see the other car?

Drummond: Yes. I saw it and I thought it was going to slow down. There's a "SLOW" sign on St. John's Gardens before the junction with Ferry Road so I had the right of way. But the Honda came right across the road and ran into my right side. He didn't slow down at all.

Policewoman: Did you try to brake?

Drummond: Yes, I tried but I didn't have a chance.

Policewoman: Do you have any witnesses?

Drummond: Yes, the lady who lives in the corner house on this side of the road was just coming out. She saw everything.

(3 seconds)

Now indicate the positions of the two cars and the witness just before the accident by writing down B, C and D in the appropriate boxes in the picture. You will have 20 seconds to complete this task.

(20 seconds)

Now, let's move on to Listening Passage 2.

(text modified from *Developing Strategies*, 9)

Listening Passage 2

Now listen to a talk on Agatha Christie, who is known throughout the world as the Queen of Crime. Her 77 novels and books of stories have been translated into every major language in the world and her sales are calculated in tens of millions. Listen carefully to the talk. While you are listening, complete the table marked Task Number 2 in your test book. The first one has been done for you as an example. Before we begin, read through Sections A, B and C for 30 seconds... (Pause 30 seconds.) *Now listen.*

Agatha Christie was born in 1891 in Torquay, England. She was educated by her mother at home and she also went to school in Paris. In 1914 she married Archibald Christie. During the First World War she worked in a hospital and began to write detective stories. In 1920 her first book, the *Mysterious Affair at Styles* was published. In 1926 her most famous book, *The Murder of Roger Achroyd* was published and in the same year she disappeared from her home. She was finally found in Yorkshire, suffering from amnesia. She wrote many books, some of which were made into plays, for example *And Then There Were None*, published in 1940 and some of which were made into films, for example, *Murder on the Orient Express*, published in 1934. In 1928 she divorced her husband and two years later she married Max Mallowan, an archaeologist. She accompanied him on some of his expeditions abroad and one or two of her books were based on this experience. She died at the age of 85 in 1976.

(3 seconds)

Now, you will be given 30 more seconds to complete this task.

(30 seconds)

Now, let's move on to Listening Passage 3.

(text modified from *Focus Listening*, Unit 14)

Listening Passage 3

Listen to the talk on facial symbolism in Chinese opera and then answer the questions in Tasks 3a and 3b. You may take notes during the talk. Before we begin, read through the questions in Tasks 3a and 3b for 20 seconds...(Pause 20 seconds.)..... *Now let's begin.*

Facial Symbolism in Chinese Opera

I'd like to talk about facial symbolism which I think is one of the most fascinating elements in Chinese opera. The actors use a variety of colours to tell the audience about the characters they are playing. A little understanding of facial symbolism will certainly help us more deeply appreciate Chinese opera.

Making up faces in Chinese opera is a very specialised skill. It is also an art. The actors and actresses use colourful paints to paint a variety of facial symbols and lines on their faces. Different facial symbols stand for different characteristics, social positions, and ages.

Basically there are four types of symbols: Zheng lian" *whole face*, San kwai wa" *three parts*, Hua san kwai wa" *complicated three parts* and Sui lian" *most complicated.* Each of the types requires different designs on the forehead, the sides of the nose, the cheeks, eyebrows and mouth. They also depict four types of characters in Chinese Opera: furious and angry people, villains, ghosts and fairies, and historical figures. In general, the less colourful and complicated the faces are, the higher the characters' positions. The more colourful and complicated the faces, the lower in social position and the more cheerful the characters are in the play.

(3 seconds)

Now, let's stop here and do Task number 3a. In this task, put ***T*** *if the statement you hear is true according to the talk and* ***F*** *if it is false. The first one is an example. You hear "in Chinese opera, actors use different colours to tell the audience about the characters they play." The statement you just heard is true therefore you write* ***T*** *in the space provided. (2 seconds) Now let us begin.*

Question Number 12, according to the talk, making up faces in Chinese opera requires training and ability. (Pause 3 seconds.)

Question Number 13, different facial symbolism in Chinese opera represents different types of characters. (Pause 3 seconds.)

Question Number 14, in Chinese opera, a character with simpler drawings on the face means he or she is lower in position. (Pause 3 seconds.)

Question Number 15, a cheerful character has more colours on his or her face. (Pause 3 seconds)

Now listen to the second part of the talk. You may take notes while you are listening. (Pause 2 seconds.)

The colours used to draw facial symbols stand for the personality of the characters. For example, red means loyalty and righteousness. *Kuan Yu* in Chinese opera is depicted with a red face. A green face means an angry man. Fierce, unpleasant and dangerous people have blue faces. A black face signifies openness and honesty. A yellow face means the character is cunning but a purple face indicates the character is a reliable person. Fairies usually have gold and silver faces. A white face like the character *Tsau Tsau* means insincerity and deception, but a white face with a square or a reverse U-shaped black spot on the nose means the character is a clown in the play.

I have just given you a very general description of facial symbolism in Chinese opera. There are many other varieties and exceptions that I haven't mentioned. In order to understand more about the facial symbolism, you simply have to see a Chinese opera or two and observe the characters in the play. You'll find out how interesting it is. (3 seconds)

Now let's do Questions 16 - 21 in Task number 3b. Each of the facial colours in Chinese opera has a particular significance. Match the meaning with the colour it stands for. You will not use all the choices in Box 乙. The first one has been done for you as an example. You have 45 seconds to complete this task.

(45 seconds) *Now, let's move on to Listening Passage 4.*

(modified from promotional brochure by National Fu Hsin Dramatic Arts Academy)

Listening Passage 4

Listen to the mini lecture on Taiwan's traditional architecture. After you hear the talk, answer Questions 22 - 26 in Task number 4. You may take notes during the talk. Now, listen carefully.

The beauty of Taiwan's traditional architecture: Part 1, the roof

The origin of Taiwan's traditional architecture comes from southern China. Apart from the ideology of architecture brought by the first mainland settlers, Taiwanese architecture has added elements in a style that reflects the thinking in *The Book of Changes* and some popular beliefs prevalent in our everyday life. These elements may be best realised in the work of the roof and the framework of a building.

The roof is the most exquisite and most important part of traditional Taiwanese architecture. The four most distinguishing features of the roof are: the Horse Back, the Swallow Tail, the Talisman and the Cylinder-shaped Brick.

Horse Back refers to the two ends of the roof and is usually in the shape of one of the five elements in *The Book of Changes*, namely Metal, Wood, Water, Fire and Earth.

Swallow Tail means that both ends of the roof raise up like the tail of a swallow. It symbolises the social status of the house owner.

Talisman refers to the pottery figure set on the roof. It is there to ward off evil spirits.

Cylinder-shaped Brick refers to the glazed semi-cylinder shaped tile on the slope of the roof. It is again used to protect the house.

The roof is always the last part to be completed. Upon its completion, a solemn ceremony is held to give thanks to Master Lu Ban, patron of Construction. The ceremony is usually in the early morning on an auspicious day.

(Pause 3 seconds)

Now listen and answer questions 22 - 26 in Task number 4. For each question, you must answer in ***no more than 4 words****. The first one is done for you as an example.*

Example: Where does traditional Taiwanese architecture originate?

The answer to this question is southern China. Therefore you write <u>*southern China*</u> *in the space provided. Now let us begin.*

Question Number 22. In which two parts of a building do we see elements of popular belief and thinking in the Book of Changes realised? (Pause 7 seconds)

Question Number 23. Which type of roof represents the five elements? (Pause 7 seconds)

Question Number 24. What does a Swallow Tail roof tell us about the owner of a house? (Pause 7 seconds)

Question Number 25. According to the talk, what is the Talisman used for? (Pause 7 seconds)

Question Number 26. At what time during the day is the ceremony to Master Lu Ban usually held? (Pause 7 seconds)

Now, let's move on to Listening Passage 5.

(text modified from a pamphlet on Taiwanese architectural features)

Listening Passage 5

Listen to the railway announcements in Edinburgh Railway Station and fill in the details of the train on the grid. Then listen again and complete the sentences in Questions 30 - 32. Before we begin, read through the questions for 20 seconds... (Pause 20 seconds.)... *Now, listen and fill in the missing information in Task number 5a.*

Platform 14 for the 11: 30 Scot Railways shuttle service to Glasgow Queen Street calling at Haymarket, Linlithgow, Falkirk High and Glasgow Queen Street. Service of drinks and light refreshments is available. Platform 14 for the 11: 30 Scot Railways shuttle service to Glasgow Queen Street.

This is a railway customer call for the person by the name of Mr. Jonathan Brown. Mr. Jonathan Brown. Meeting his mum please go to Thomas Cook which is opposite the main concourse. Mr. Jonathan Brown meeting his mother please go to Thomas Cook which is situated at the main concourse. Mr. Jonathan Brown.

The 11:20 Scot Railway terminating service from Glasgow Queen Street is approaching Platform 14.

Platform 19 for the 11:30 Great North-Eastern Railway service to London King's Cross calling at Newcastle, York, Peterborough, and London King's Cross. Hot food facilities are available. First class accommodation is situated at the front. Platform 19 for the 11:30 Great North-Eastern Railway service to London King's Cross.

Platform 10 for the 11:25 Virgin Train service to Plymouth calling at Birmingham New Street, Cheltenham Spa, Gloucester, Bristol Parkway, Bristol Temple Meads, Taunton, Torquay, Totnes and Plymouth. Hot food facilities are available. First class accommodation is situated at the front. Platform 10 for the 11:25 Virgin Trains service to Plymouth.

(Pause 3 seconds.)

Now listen again. While you're listening, complete the sentences in Questions 30 to 32.

(Play Announcements again.)

(Pause 5 seconds)
Now, let's move on to Listening Passage 6.

(text modified from 27/2/99 recording in Edinburgh Railway Station)

Listening Passage 6 ***(Edinburgh Tour 28/2/99 2:30pm (the script has been modified.))*

Listen to the tour guide talk about famous places on the Royal Mile in Edinburgh, Scotland's capital. While you are listening, put the correct letter next to the places mentioned in the box. Then listen again and fill in the missing information about each place. The tour starts from the castle and is heading east towards Holyroodhouse Palace. Now, before you listen, read through the questions for 30 seconds.....(Pause 30 seconds)..... *Now listen.*

Now, we're just beginning the ..... Royal Mile now. It's called the Royal Mile because it has the Castle at the one end and the Palace of Holyroodhouse at the other. It's actually an old Scots mile..... a mile and 200 yards.

We're leaving the castle esplanade now. Edinburgh Castle is a working castle. It opens all year round. The castle is on top of an extinct volcano. The oldest part of the castle dates back to the eleventh century. That's where Queen Margaret's Chapel is. The chapel is also the oldest building in the castle.

Now just alongside the left hand side of the bus is a little lane. Lady Stair's Close it's called and that's the way through to the Writers' Museum.... and it's free, no charge, open from Monday to Saturday. The museum holds the artefacts of three of the most famous writers: Sir Walter Scott, Robert Louis Stevenson and Robert Burns. In the Writers' Museum. Open Monday to Saturday and, as I said, the entrance is free.

Those of you upstairs, ahead of you on your right hand side is the Crown Spire of St. Giles Cathedral, a famous landmark on the Royal Mile. Those of you sitting downstairs, you'll see St. Giles Cathedral coming on your right hand side. It's a very large church. St. Giles is the high kirk of the Church of Scotland. Parts of this building date back to the twelfth century. And the cathedral goes hand in hand with the history of Edinburgh, the history of Scotland in fact.

Directly coming ahead of us on the right hand side is my favourite museum, the Museum of Childhood. The museum's been described as "the noisiest place in the world". And it is one of the most frequently visited attractions in Scotland. All the toys you, your parents, your grandparents played with could be found here. Just passing on your right hand side, the Museum of Childhood. Open Monday to Saturday and the entrance is free. It's worth a visit.

Almost directly across the street is John Knox House, which is a museum now. John Knox was a famous Protestant reformer and had much to do with the reformation of the Church in Scotland in the sixteenth century. The house has hand-painted ceilings and it's entered by forestairs which was once a common architectural feature in the Royal Mile. But there are only a few surviving examples now.

The place where we are now was at one time the country side. The area is called the Canongate.

"Gate" is the Scots word for "walk". This is where the canons, or monks used to walk up and down from Holyrood Abbey to St. Giles Cathedral. Outside Edinburgh it was a separate burgh or town and had its own town council. And indeed it had its own jail. Prisoners were housed in the Tolbooth. The Tolbooth nowadays is the People's Story Museum. It's open Monday to Saturday and it shows how ordinary people, not kings and queens but how ordinary people would have lived from the eighteenth century onwards up to about the nineteen fifties.

Now those staircases on your left hand side are the entrance to the People's Story Museum. Open Monday to Saturday. The entrance is .... free.

Now the next stop is the Holyroodhouse Palace. The Queen usually visits during the month of July, during which time she stays at the palace. The palace is a private residence here in Edinburgh. It's the official residence in Scotland. When the queen is not here, it's open to the public. A very interesting place to visit. Currently there's an exhibition of water colours from the private collection of Prince Albert and Queen Victoria. This is where we stop for anyone wishing to visit Holyroodhouse Palace.

(Pause 5 seconds.)

Now listen again and while you're listening, complete the description of the places in task number 6b.

(Play the talk again.)

(Pause 5 seconds)

This is the end of the listening test. Thank you for your participation. Now, please remain seated until the test assistant tells you to leave. Test assistant, please collect the test booklets. Thank you.
(Pause 3 seconds.)

This is the end of the recording.

Form Code: TGE S001

Name: _____

Registration Number: _____

The Tour Guide English Language Speaking Test

Time allowed: 30 minutes

General instructions:

There are five parts to this speaking test:

1. Warming Up,
2. Translation and Interpretation,
3. Answering Short Questions,
4. Making an Announcement and
5. Solving a Problem.

Part I, Warming Up, will not be assessed. You will hear questions on a tape. Your responses will be recorded onto another audio-tape which has been provided. They will be the basis of our judgement of your spoken language ability. The time allowed in each part will be indicated. Try to speak as much as you can. There is plenty of time for you to answer each question.

Part I: Warming Up (30 seconds)

In this part of the test, could you tell me your name, where you're from, where you're studying or working, what you do in your spare time and what you would like to be doing two years from now. You have **30 seconds** to complete this part.

Name:
Nationality:
Current work:
Interests:
Future plan:
Other information:

Part II: Interpretation and Translation (2.5 minutes)

There are 10 signs and sentences in this part of the test. Please interpret these signs and sentences in such a way that your interpretation is similar to the meaning given in Chinese. You have **2.5 minutes**.

1. 本大廈全面禁煙。
2. 請輕聲說話。
3. 節目進行中，請勿閃光拍照、攝影。
4. 出清存貨，三折起。
5. 請靠右走。
6. 我們在故宮博物院約停留三個小時。
7. 忠烈祠的衛兵交接儀式每一小時舉行一次；下一次在十點舉行。
8. 基隆市 中山公園的觀音雕像是全台灣最高的一座；約25公尺高。
9. 謝謝各位參加本公司"台南市區遊",希望有機會能再爲各位服務。
10. 下車時，請不要忘記各位的隨身物品。

故宮博物院：National Palace Museum
忠烈祠：Martyrs' Shrine
衛兵交接儀式：Changing of the Guard
台南市區遊：Tainan City tour

Part III: Answering Short Questions (4.5 minutes)

In this part, you will hear 10 short questions. You need to answer each question. For Questions 1 - 4, you will have 15 seconds to answer each question. For Questions 5 - 8, you will have 30 seconds and for questions 9 - 10, 45 seconds.

Question 1:	15 seconds	Question 6:	30 seconds
Question 2:	15 seconds	Question 7:	30 seconds
Question 3:	15 seconds	Question 8:	30 seconds
Question 4:	15 seconds	Question 9:	45 seconds
Question 5:	30 seconds	Question 10:	45 seconds

Part VI: Making an Announcement (6 minutes)

You are taking a group of foreign medical professionals for a day trip to Danshuei. You have to explain the itinerary for the day. Read the following itinerary, then try to present the information as if you were talking to a group of real people. You have **3 minutes** to prepare and then **3 minutes** to record your response. Please try to use as much of the information provided for you as possible. Draft paper is provided.

Itinerary for Danshuei (淡水)

Time	**Place to see**
9:30am	Taipei →Mangrove Conservation Area → Oxford College and Mackay's Old Residence
12:30pm	Lunch
1:30pm	Chingshuei Temple & Fuyou Temple → Fort San Domingo → Danshuei Ferry Pier to watch sunset
6:pm	Dinner

1. Depart 9:30am and return 8:30pm. (Return time is approximate.)
2. Transportation: Taipei to Danshuei by coach
   Danshuei to Taipei by MRT from Danshuei Station
3. Buffet lunch at *Old Danshuei* Seafood Restaurant. Dinner at *Danshuei Pier* Beer Pub (with live music).
4. Mineral water on coach

Mangrove Conservation Area：紅樹林保護區

Oxford College：牛津學堂

Mackay's Old Residence：馬偕故居

Chingshuei Temple：清水祖師廟

Fuyou Temple：福祐宮

Fort San Domingo：紅毛城

Danshuei Ferry Pier：渡船碼頭

Part V: Solving a Problem (8 minutes)

In this part of the test, you are going to explain to your clients concerning a change in the tour itinerary. You are a tour guide and you are taking a group of 15 foreign visitors on a 5-day round-the-island tour. You are scheduled to visit Hualien tomorrow. Unfortunately, a typhoon is approaching and it is going to hit Hualien tomorrow. For safety concerns, your company has decided to cancel the trip to Hualien. Instead, you and the group are to spend the day in Taipei and then go to the next destination, Kengting. To compensate the clients for their loss, your company is going to give a gift voucher of NT$2,000 to each of the clients. You are going to explain the change of itinerary and the company's course of action to your clients. You have **4 minutes** to prepare your explanation and then **4 minutes** to record your response. You may use the prompts as the basis of your preparation. You may also add your own ideas. Draft paper is provided.

- The End -

(Speaking test script)

The Tour Guide English Language Speaking Test

General instructions:

Hello. You will hear me asking you questions throughout this test. There are five parts to this test: Warming up, Interpretation and Translation, Answering short questions, Making an announcement and Solving a problem. Part I, Warming Up will not be assessed. You will hear questions on a tape. Your responses will be recorded onto another tape. The time permitted for each part of the test is indicated in the test booklet. This will give you a rough idea of how much time you need to prepare your answers and respond to the questions. Try to speak as much as you can. You will have plenty of time to answer each question. Your oral language skills will be judged on the basis of your responses to the questions. To indicate the end of a part and the beginning of the next part, you will hear a tone like this (**Tone**). *Now, do you have any questions? If you do, please raise your hand and an assistant will come and help you. During the test, you may not ask any more questions. (**Pause 5 seconds**) If you don't have any more questions, turn to Part I: Warming Up and let us now begin.*

Part I: Warming Up

In this part of the test, could you tell me your name, where you're from, where you're studying or working, what you do in your spare time and what you would like to be doing two years from now? You have 30 seconds to complete this part. You may now begin.

(30 seconds) (Tone)

Part II: Interpretation and Translation

Now let's move on to Part II: Interpretation and Translation. There are 10 signs and sentences in this part of the test. Please interpret these signs and sentences in such a way that your interpretation is similar to the meaning given in Chinese. You have two and a half minutes. You may now begin.

(after 2.5 minutes) (Tone)

Part 3: Answering short questions

Now let us begin Part 3: Answering short questions. In Part III, you will hear 10 short questions. You need to answer each question. For Questions 1 - 4, you will have 15 seconds. For Questions 5 - 8, you will have 30 seconds, and for Questions 9 - 10, 45 seconds. Ready? ***(3 seconds)***

Question number 1. How do you say THANK YOU and GOOD-BYE in Mandarin? (15")

Question number 2. What does TAIPEI mean? (15")

Question number 3. Apart from English, what other languages do you speak? How well do you speak them? (15")

Question number 4. How many major political parties are there in Taiwan? What are their names? (15")

Question number 5. Is it safe to go out at night in Taiwan? Why or why not? Please explain to me. (30")

Question number 6. I'd like to visit Taiwan, when's the best time of the year to come? When is it not a good time to come? (30")

Question number 7. Why do lots of people ride scooters in Taiwan? (30")

Question number 8. What does a typical Taiwanese breakfast consist of? (30")

Question number 9. In your opinion, what is the most worthwhile place to visit in Taiwan? Why? (45")

Question number 10. What's the most important holiday in Taiwan? How is it celebrated? (45 seconds)

*(**Tone**)*

Now, let's move on to Part IV: Making an announcement

Part IV: Making an announcement

You are taking a group of foreign medical professionals for a day trip to Danshuei. You have to explain the itinerary for the day. Read the following itinerary, then try to present the information as if you were talking to a group of real people. You have ***3 minutes*** *to prepare and then* ***3 minutes*** *to record your response. Please try to use*

as much of the information provided for you as possible. Draft paper is provided for you. You may start your preparation now.

(after 3 minutes)

Now record your response. Don't forget the usual greetings. Ready? You may start now.

(After 3 minutes) (Tone)

Good. Now let's move on to Part V: Solving a problem.

Part V: Solving a problem. In this part of the test, you are going to explain to your clients concerning a change in the tour itinerary. You are taking a group of 15 foreign visitors on a 5-day round-the-island tour. You are scheduled to visit Hualien tomorrow. Unfortunately, a typhoon is approaching and it is going to hit Hualien tomorrow. For safety concerns, your company has decided to cancel the trip to Hualien. Instead, you and the group are to spend the day in Taipei and then go to the next destination, Kengting. To compensate the clients for their loss, your company is going to give a gift voucher of NT$2,000 to each of the clients. You are going to explain the change of itinerary and the company's course of action to your clients. You have 4 minutes to prepare your explanation and then 4 minutes to record your response.

Now, read the prompts again. You have 4 minutes to prepare your explanation. You may use the prompts as the basis of your preparation. You may also add your own ideas. Draft paper is provided. You may start your preparation now.

(After 4 minutes)

Now record your response. You have 4 minutes.

(After 4 minutes) ***(Tone)***

Stop recording. (3 seconds)

This is the end of the speaking test. Thank you for your participation. Please remain seated until the test assistant tells you to leave. (3 seconds)

This is the end of the recording.

Form Code: TGE G001

Name: _____

Registration Number: _____

The Tour Guide English Language Grammar Test

Time allowed: 40 minutes

Number of questions: 65

General instructions:

There are five parts to this test. Instructions for each part of the test will appear before the test questions along with an example. A recommended time to do each part of the test will also be given.

Write your answers on the answer sheet provided. Do not write on the test paper!

Part I: Multiple choice questions

Recommended time: 4 minutes

Instructions:

For Questions 1 - 12, choose A, B, C or D to complete the sentence in each question.

Write your answers on the answer sheet provided.

Example:

What time _____ it now?

A: be

B: is

C: are

D: were

The best answer to the question above is B: is. Therefore you should choose B.

(The test continues on the next page.)

1. While playing tennis, Alice says, "I _____ the game. I think it's going to suit me."

A: enjoy

B: enjoyed

C: am enjoying

D: have enjoyed

2. John seemed very anxious _____ the event to be a success.

A: of

B: in

C: about

D: for

3. _____, he would have made a serious mistake.

A: Had it not been for his solicitor

B: If it wasn't his solicitor

C: If it hasn't been his solicitor

D: Not for his solicitor

4. In fact, we heard Susie ____ the whole piece from beginning to end on the piano.

A: to play

B: play

C: was playing

D: played

5. _____ usually have great difficulty in getting a job. They need more help from the government.

A: Homeless

B: Homelessness

C: The homeless

D: The homelessness

(The test continues on the next page.)

6. It's _____ raining. How much longer can it go on?

A: already

B: yet

C: still

D: no longer

7. It will be easier to go _____ the bus to get out of the traffic.

A: by

B: in

C: on

D: of

8. I wish the room _____ a bit bigger.

A: is

B: will

C: were

D: would be

9. We saw several weddings today. It _____ today is a good day for getting married.

A: feels

B: seems if

C: sounds like

D: looks as if

10. The news of the president's retirement _____ be announced soon. No one is sure when.

A: will

B: would

C: can

D: could

(The test continues on the next page.)

11. Jane _____ before crossing the road but she didn't.

A: must look

B: may look

C: would have looked

D: should have looked

12. I parked my car outside a shop but _____ I was in the shop, the police towed it away.

A: as

B: when

C: while

D: at the time

(The test continues on the next page.)

Part II: Verb forms

Recommended time: 3 minutes

For questions 13 - 22, read the two passages carefully and supply the correct verb form according to the meaning of the passages. You may need to write one to three words for each question. The first one has been done for you as an example. Write your answers on the answer sheet provided.

Passage 1: *A tour guide is giving a brief history of the Presidential Office Building to his clients. Read the passage and supply the correct verb forms.*

If you _look_ ahead of you, you *(13. see)* the Presidential Office Building, where President Lee *(14. work)*. *(15. Build)* in 1919, originally, the building *(16. call)* the Supreme Office and it *(17. use)* by the Japanese Governor during the Japanese occupation. We *(18. rename)* the building after the defeat of Japan in 1945.... Now, here is a good place for you to take pictures.

Passage 2: *The following is a description of a cottage in a travel brochure. Read the passage and supply the correct forms of the verbs.*

Surrounded by pine trees, with direct access to a sandy beach and a heated pool *(19. face)* the sea, this charming 19th century manor house *(20. retain)* all of its original character. It *(21. situate)* just 4 km from Royan in the small resort of Vaux sur Mer. Beautifully decorated and with fine furnishings, this elegant hotel *(22. have)* a bar, lounge and tennis court, and there is a golf course 4 km away.

(The test continues on the next page.)

Part III: Sentence transformation

Recommended time: 15 minutes

For questions 23 - 37, finish the second sentence so it has the same meaning as the sentence before. Write your answers on the answer sheet provided.

Example 1:

Be careful or you'll hurt yourself.

If *you are not careful* you'll hurt yourself.

Example 2:

I wasn't comfortable. Then I had the painful tooth extracted.

Until I *had the tooth extracted, I was in pain*.

(The test continues on the next page.)

23. It appears that we don't have any film left.

We seem to _____ of film.

24. Emma says something different. It is not the same as she does.

What Emma _____ what she does.

25. Mary will return from her business trip to Dublin very soon.

It won't be _____ from her business trip to Dublin.

26. Many people have seen the show. It shows how popular it is.

The fact that _____ shows how popular it is.

27. For further information, please contact the Tourism Bureau Information Center.

Further information _____ the Tourism Bureau Information Center.

28. Only senior staff members are allowed to use the company car park.

The company car park is _____ senior staff members only.

29. I cannot agree with that statement.

That is the statement _____.

30. "Would you like to have lunch with me?" John said.

John invited _____.

31. My cousin is the most annoying person I've ever met.

I've yet _____.

32. I was surprised at how approachable the new boss is.

I didn't expect _____.

33. I will never lend money to Robert no matter what.

Under no _____.

(The test continues on the next page.)

34. It has been two weeks since anyone saw John.

John _____.

35. Joan didn't attend the meeting. She apologises.

Joan sends _____.

36. This door must be kept closed at all times.

At no time _____.

37. The suitcase was so heavy that Tommy could not lift it.

The suitcase was too _____.

(The test continues on the next page.)

Part IV: Fill in the blanks

Recommended time: 8 minutes

In the newspaper article below (Questions 38 - 50), a word is missing in each of the blanks. Read the article and fill in the missing words. Each blank requires **one word** only. The first blank has been done for you as an example. Write your answers on the answer sheet provided.

Tour of Historical Sites of Taiwan:

Ruins of An-ping Fort

The ruins of Taiwan city are located in An-ping, Tainan City. Built by _the_ Dutch in 1624, An-ping Fort was the first to be built in Taiwan. The fort was first known (38) Fort Orange. Later, the name was changed to Fort Zeelandia. The fort was built with red bricks brought (39) Indonesia. The bricks were mortared with a mixture of sugar syrup, glutinous rice (40) crushed oyster shells. They made a very strong foundation. The square shaped fort is built (41) top of a two-storey platform with lookout towers (42) the four corners giving the place a grand appearance. At the north-west corner of the fort, there used to be (43) city surrounded by a ten-meter high wall. From (44) remains of the wall, we can still see (45) the interior was constructed with wooden beams, and there (46) still traces of metal studs. After Cheng Cheng-kung (47) the Dutch in 1661, the fort was used as Cheng's residence and it was renamed *Wang Cheng*, the City of the Prince. (48) the late 19th century the fort was in a dilapidated (49) and during the Japanese occupation (1895 - 1945), the houses (50) levelled to form steps. A platform and a lighthouse were built on top. Today, only the outer walls survive from the original construction by the Dutch.

(The test continues on the next page.)

Part V: Complete the conversation

Recommended time: 6 minutes

There are two parts to Complete the conversation.

Part Va:

For questions 51 - 55, read the conversation between you and Sandy, an acquaintance of yours. Some of your part has been deliberately left out. Read the conversation carefully and write the sentences numbered 61 - 65 according to the suggested cues. The first one has been done for you as an example. Write your answers on the answer sheet provided.

Sandy: Hello! Isn't it a lovely day!

You: Yes, beautiful!

Sandy: Can you find somewhere to sit? I'm sorry this room is so untidy.

You: **Disagree** No, it isn't. It's fine.

Sandy: Well, I've got this afternoon off. Where shall we go?

You: **(51) Suggest a visit to Taipei Zoo** _____

Sandy: Sorry, I didn't quite catch that.

You: **(52) Repeat what you said** _____

Sandy: Where is that exactly?

You: **(53) Explain** _____

Sandy: All right. That sounds fine. But I must go to the bank first and I've got all these letters to post too.

You: **(54) Offer to help** _____

Sandy: Oh, that's kind of you. Thanks. Well, shall we meet in half an hour then?

You: **(55) Agree and say goodbye** _____

(The test continues on the next page.)

Part Vb:

For questions 56 - 60, a situation is given to you. Read the situation and write two possible responses if you were in that situation. Write your answer on the answer sheet provided.

Example:

You are buying some "Thank You" cards in a stationer's for the office. You want a receipt. You ask the shop assistant for it. What would you say? Write the answers on the answer sheet provided.

(a): ***Could you give me a receipt, please?***

(b): ***Can I have a receipt please?***

(The test continues on the next page.)

56. You are taking some foreign visitors to Keelung tomorrow. The city is well known for its rainfall. You want to remind your guests to bring an umbrella. What would you say?

(56a): _____.

(56b): _____.

57. You are taking a group of visitors to hike the Tsauling Historic Trail. It's a very warm day. You offer an elderly lady some mineral water. What would you say?

(57a): _____.

(57b): _____.

58. You and your foreign clients are visiting the National Palace Museum. You want to show them the well-known jade exhibit, the Jade Cabbage. You are trying to get your clients to follow you. What would you say?

(58a): _____.

(58b): _____.

59. You and your foreign visitors are visiting the Martyrs' Shrine. It is a solemn place. You are reminding your visitors to be quiet and respectful. What would you say?

(59a): _____.

(59b): _____.

60. You want to borrow a client's newspaper so you can check the cinema times. What would you say?

(60a): _____.

(60b): _____.

- The End -

Form Code: TGE G001

Name: _____ Registration Number: _____

The Tour Guide English Language Grammar Test – Answer sheet

Part I: Multiple-choice questions

1. _____ 2. _____ 3. _____ 4. _____ 5. _____

6. _____ 7. _____ 8. _____ 9. _____ 10. _____

11. _____ 12. _____

Part II: Verb forms

13. _____ 14. _____ 15. _____

16. _____ 17. _____ 18. _____

19. _____ 20. _____ 21. _____

22. _____

Part III: Sentence transformation

23. We seem to _____ of film.

24. What Emma _____ what she does.

25. It won't be _____ from her business trip to Dublin.

26. The fact that _____ shows how popular it is.

27. Further information _____ the Tourism Bureau Information Center.

28. The company car park is _____ senior staff members only.

29. That is the statement _____.

30. John invited _____.

31. I've yet _____.

32. I didn't expect _____.

33. Under no _____.

34. John _____.

35. Joan sends _____.

36. At no time _____.

37. The suitcase was too _____.

Part IV: Fill in the blanks

38. _____ 39. _____ 40. _____ 41. _____

42. _____ 43. _____ 44. _____ 45. _____

46. _____ 47. _____ 48. _____ 49. _____

50. _____

Part V: Complete the conversation

51. _____

52. _____

53. _____

54. _____

55. _____

56a: _____

56b: _____

57a: _____

57b: _____

58a: _____

58b: _____

59a: _____

59b: _____

60a: _____

60b: _____

- The End -

Descriptive statements of Tour Guide General Proficiency levels

5. **Distinction**

Uses the language with proficiency approaching that in the candidate's mother tongue (L1). Copes well with demanding and complex language situations. Makes occasional minor lapses in accuracy, fluency, appropriateness and organisation but does not affect communication. Only rare uncertainties in conveying or comprehending the content of the message.

4. **Good**

Uses language effectively in most situations, except the very complex and difficult. A few lapses in accuracy, fluency, appropriateness and organisation, but communication is effective and consistent, with only a few uncertainties in conveying or comprehending the content of the message.

3. **Competent**

Uses the language with confidence in moderately difficult situations. Noticeable lapses in accuracy, fluency, appropriateness and organisation in complex situations. Communication and comprehension are effective on most occasions, but are disrupted when difficulties arise.

2. **Satisfactory**

Uses the language effectively in all familiar and non-pressuring situations. Rather frequent lapses in accuracy, fluency, appropriateness and organisation, but usually succeeds in communicating and comprehending general message.

1. **Below standard**

Uses basic range of language, sufficient for familiar and simple situations. Many lapses in accuracy, fluency, appropriateness and organisation. Language ability restricts continuity of communication and comprehension. Frequent efforts are needed to ensure communicative intention is achieved.

0. **No foreign language use ability:** Unable to use the language even in simple situations

Descriptive statements of Tour Guide Oral Proficiency levels

5. **Distinction**

Handles a full range of oral interaction (i.e., personal, social and business) with confidence. Competence approaches that of the candidate's native language. Relevant message is fully and effectively conveyed with interesting treatment of topic and is well adjusted to the listener's language and knowledge. Message is well organised and sustained. Style is well adjusted to context and purpose. Fluency is good with few false starts and/or hesitations. There is only occasional need for repair. Controls a good range of social/business language with high degree of accuracy in interaction. Has slight traces of L1 accent but pronunciation, intonation and stress patterns assist communication.

4. **Good**

Handles a wide range of oral interaction (i.e., personal, social and business) with good confidence and competence. Message is delivered clearly in an effective manner. Presentation of the spoken text is relevant and appropriate to the listener's language and knowledge of the topic. Speech is clearly organised with suitable sequencing and cohesion. Participates readily in oral interaction but with some lapses in fluency, flexibility and appropriateness. Has a good language repertoire but shows some lapses of linguistic accuracy and linguistic uncertainty. Uses effective coping strategies. Speech features show L1 influence but they do not affect communication.

3. **Competent**

Handles moderate-level oral interaction (i.e., personal, social and business) with good confidence and competence but shows some problems with higher-level interaction (e.g., lengthy talk or discussion). Message is adequately adjusted to listeners' language knowledge of topic and content. Some restriction in participation because of language limitation. Has an adequate mastery of text organisation but shows some uncertainties over appropriateness of style. Some loss of fluency (e.g., false starts and/or hesitations) which hampers full participation in oral interaction. Handles a fair range of language. Has a good grasp of usage and accuracy in spite of some lapses. Speech marked with L1 features but rarely affects communication.

2. **Satisfactory**

Handles simple personal, social and business oral interaction with good confidence and competence. Conveys major points of message but with little subtlety and frequent loss of detail. Shows some difficulties in sustaining conversation/talk. Fairly frequent need for rephrasing and repair. Shows frequent false starts or hesitation. Has a basic mastery of text organisation but a little variation in style. Has inadequate sense of appropriateness to context and purpose. Handles a limited range of language. Needs to search for words and uses circumlocution. Frequent errors in accuracy. Fairly frequent lapses in fluency, but they do not affect basic communication. Speaks with obvious L1 accent which at times impairs communication

.

1. **Below standard**

Can handle simple oral interaction. Generally lacking confidence in communication. Frequent need for repetition and rephrasing. Text organisation is haphazard and shows little variation in style. Frequent need for repair. Participates in structured interaction but very restricted in freer interaction. Shows a narrow range of language with little variety. Frequent errors. Heavy L1 accent. Language limitation is restricted to handling basic facts and opinions. Communication may break down because of shortcomings in usage and pronunciation.

0. No foreign language use ability

Utterance is sporadic. Generally shows no ability to handle oral interaction of any type.

Sub-skill descriptors for TG oral proficiency

I. Fluency

5. **Distinction**

Speech is natural and easy almost as fluent as the candidates' L1. The candidate rarely searches for language and can express himself or herself clearly and coherently most of the time.

4. **Good**

Speech is coherent most of the time. The candidate rarely hesitates before responding but may pause to marshal ideas.

3. **Competent**

Speech is somewhat slow but coherent without pausing unduly.

2. **Standard**

The candidate is able to produce short and fairly coherent speech. S/he needs to pause to search for language.

1. **Below standard**

Speech is short, incoherent with frequent pauses.

0. **No foreign language use ability**

Only able to produce single words.

II. Pronunciation

5. **Distinction**

Able to speak the language in a way that approximates a native-speaker.

4. **Good**

Able to speak the language in a way that approaches native-speaker standard. Speech features show L1 influence.

3. **Competent**

Able to pronounce words that are easily understood by a native-speaker. Use of stress, rhythm and intonation approaches native usage but a foreign accent is evident.

2. **Satisfactory**

Obvious foreign accent but most sounds are comprehensible to a native-speaker.

1. **Below standard**

Candidate's production of sounds is only understood by a sympathetic native-speaker.

0. **No foreign language use ability**

Candidate's production of sounds is unintelligible even to a sympathetic native-speaker.

III. Grammatical accuracy and vocabulary range

5. **Distinction**

Uses a complete range of forms and structures with confidence and competence as would in his/her mother tongue. Performs with high degree of accuracy. No obvious weakness in the use of general vocabulary. The candidate's vocabulary is wide enough so that s/he can express himself or herself without much difficulty in almost all situations.

4. **Good**

Able to use easily and naturally a wide range of structures to express himself or herself clearly and precisely. Performs with a fair degree of accuracy. No obvious weakness in the use of general vocabulary. Sometimes may use circumlocutions. But vocabulary is wide enough to allow them to express themselves without many more difficulties than they do in their mother tongue.

3. **Competent**

Able to produce extended utterances with simple and complex structures. Able to use appropriate forms for different situations most of the time. Performs tasks with reasonable grammatical accuracy though there are noticeable lapses and errors. Able to use everyday vocabulary appropriately and paraphrase where gaps of vocabulary occur.

2. **Satisfactory**

Able to use considerable range of structures to form more complex sentences and has a repertoire of adverbials, connectors and prepositions to do so. Performs communicative tasks with a limited degree of grammatical accuracy. There are frequent grammatical errors. Able to use basic vocabulary relating to candidate's work and personal interests.

1. **Below standard**

Able to produce simple sentences with appropriate word order, gender and case. Able to indicate past, present and future time, singular and plural, positive and negative, questions and requests. Usually very inaccurate in communicative tasks. Candidates control a limited range of vocabulary. Frequently hesitate and search for words. Generally poor level of vocabulary knowledge.

0. **No foreign language use ability**

Unable to produce grammatical sentences. Shows little knowledge of vocabulary.

IV. Communicative effectiveness

5. **Distinction**

Able to understand and respond to language spoken at normal rate by a native speaker with no difficulties in comprehension. Speech style is well adjusted to topic, context and purpose.

4. **Good**

Able to understand and respond to language spoken at normal rate by a native speaker but shows few difficulties in understanding some highly complex speech. Speech style shows occasional lapses in appropriateness and use of vocabulary.

3. **Competent**

Able to understand the gist of the message at normal native speed. Able to understand most details though occasional repetition may be required. They should be able to guess from context. Aware of registers but shows some uncertainties as to register choice and appropriateness to context and purpose of interaction.

2. **Satisfactory**

Able to understand short sentences at normal native speed. Able to understand the gist of the message though may have difficulties with details in lengthy discussions without repetition. Has inadequate sense of register and appropriateness.

1. **Below standard**

Understands a native-speaker when s/he speaks short and simple sentences at a somewhat slower rate than normal speech. Shows little sensitivity to register and appropriateness.

0. **No foreign language use ability**

Unable to understand the target language.

Descriptive statements of TG Listening Proficiency levels

5. **Distinction**

Handles a full range of listening input with confidence and competence that approaches the candidate's L1. Extracts full content of the message with only occasional loss of detail and/or subtlety. Uses a full range of techniques to evaluate, apply or relay message. Sometimes may need repetition, rephrasing or repair. Rare uncertainties over organisation, style or fluency of text. Flexible adjustment in listening strategies for lengthy and detailed discussions or social texts delivered at normal speed. Has a full range of language in own and related areas of interest and is able to compensate for distortions and errors in listening texts.

4. **Good**

Handles a wide range of listening texts with confidence and competence. Extracts majority of message with only occasional loss of detail and/or subtlety. Uses a good range of techniques to evaluate, apply or relay message. Occasional need for repetition, rephrasing or repair. Few uncertainties over organisation, style or fluency of text. Effective adjustment in listening strategies at normal speed but less effective than in L1. Handles a wide range of social and business language and has little difficulty in compensating for distortions and errors in listening texts.

3. **Competent**

Handles moderate levels of listening operations with confidence and competence. Extracts major points of message with noticeable loss of detail and/or subtly. Fairly frequent need for repetition, rephrasing or repair. Adequate ability to handle organisation, style or fluency of text. Uses adequate techniques to store, apply or relate straightforward listening input delivered at normal speed but has problems in initial adjustment to style, accent and speed of delivery. Employs good strategies when listening with full attention. Handles a moderate range of language in general and particular areas of interest which sometimes compensates for distortions and errors in listening texts.

2. **Satisfactory**

Handles simple listening input with confidence and competence. Extracts essential points of message, with great loss of detail and little grasp of subtlety. Frequent need for repetition, rephrasing or repair. Frequent problems with organisation, style and fluency of text. Uses limited range of techniques to store, apply or relate message delivered at normal speed and directed toward him/her. Has a limited range of language in own particular area of interests that occasionally compensates for distortions and errors.

1. **Below standard**

Handles simple listening input with adequate competence and confidence. Can identify the topic of the talk and comprehend the gist of message with little detail. Further comprehension depends on L1 or visual support. Stores basic factual information. Success of applying or relating of message depends on level of comprehension. Constant need for repetition, rephrasing and/or repair. Constant problems with organisation, style or fluency of text. Very limited ability to handle input at normal speed. Requires clear and slow speech directed at him/her. Has a narrow range of language in own particular areas of interests and is not able to compensate for distortions or errors.

0. **No foreign language use ability**

Unable to understand what has been spoken

Rating scale

Form Code: _____ **Rating:** _____

Name: _____ **Registration Number:** _____

	Distinction 5	Good 4	Competent 3	Satisfactory 2	Below standard 1	No ability 0
Fluency	5	4	3	2	1	0
Pronunciation	5	4	3	2	1	0
Accuracy & Vocabulary	5	4	3	2	1	0
Communicative effectiveness	5	4	3	2	1	0

Rater's Comments:

Rater's name: _____ Date: _____

3.8. Summary

This chapter looks at the following areas:

(1) What is and is not involved in the ability to listen, speak and use grammatical knowledge accurately, meaningfully and appropriately. This is then related to an exposition of what is intended to be assessed in each of the TG sub-tests.

(2) A framework to guide the design of each sub-tests for the measurement of the traits proposed.

(3) Criteria and procedures in test and task development.

(4) Development of the rating scale and rater training.

(5) The criteria to ensure TG test usefulness.

(6) Presentation of the entire TG Test battery and the rating scale

In the next chapter, reports on test administration and test results on classical test analysis and Rasch analysis are presented and discussed; areas for improvement based on the statistical results are suggested.

Chapter 4: Test administration and test analysis

This chapter reports on (1) administration of the TG test, (2) test results and test interpretation, (3) test revision, and (4) setting the cut-off score.

4.1. Test administration

The pilot test

The test was piloted on 15 volunteer university students to gather information on item difficulty, the clarity of test instructions, the time allotment, and the appropriateness of the test content. We planned to ask as many tour guides as possible to participate in the pilot test and to gather their feedback for revision. However, the Tour Guide Association advised against this because the tour guides would think they were being tested and would lose their "face" if their performance was poor. Therefore, no tour guides took part in the pilot test.

On the basis of test statistics and student feedback from the pilot test, the following changes were made for the main trial:

(1) adding five more items to the Listening Test and giving a little more time to preview the listening questions,
(2) shortening the Grammar Test by 10 items (i.e., 65 items instead of 75),
(3) changing Part V of the Speaking Test to a more common scenario, namely a change of itinerary because of the weather instead of handling a complaint, and
(4) recording test tapes with American English at a slower than normal rate.

The main trial

The participants:

The main trial was administered over a period of two weeks. 112 volunteer students participated in the test. Table 4-1 provides a summary of the test takers and the participating institutions.

Table 4-1: Summary of participating institutions

School	Number of participants
Chung Hsing Univ.	14
Tam Kang Univ.	60
LTTC	15
Xin Wu College	23

All candidates were briefed and were given the sample items for reference.

Timing and scoring:

Time allotment for each of the sub-tests was: Listening and Grammar Tests 40 minutes; Speaking Test, 30 minutes.

Composition of the test:

The revised Tour Guide English Test battery consisted of three sub-tests: Listening (45 items), Grammar (65 items) and Speaking (five tasks). For details of the tests, please see Chapter 3.

4.2. Analysis of test results

The main trial results of classical test analysis and Rasch analysis are reported in this section. Section 4.2.1. reports on classical test results and Section 4.2.2., Rasch results. In each section, the Listening test results are presented first followed by the Grammar test results and finally, the Speaking test results.

4.2.1. Classical analysis

The following statistics were computed.

- Raw score distribution
- Descriptive statistics of Listening, Grammar and Speaking
- Item analysis – Listening and Grammar
- Correlation of the sub-tests and
- Rater statistics and inter-rater reliability

Raw score distribution

Raw score distributions for the Listening, Grammar and Speaking tests are summarised in Table 4-2, the histograms presented in Figures 4-1a to 4-1c. Listening raw scores ranged from 2 to 38 points, Grammar raw scores from 0 to 50 points, and Speaking scores from S0: *No foreign language ability* to S4: *Good.*

Table 4-2: Score distribution for Listening, Grammar and Speaking tests

Listening score	No. of people	**Grammar score**	No. of people	**Speaking score**	No. of people
0 – 5	13	0 – 5	4	0	23
6 – 10	17	6 – 10	13	1	23
11 – 15	29	11 – 15	23	1+	2
16 – 20	23	16 – 20	25	2 -	1
21 – 25	14	21 – 25	19	2	40
25 – 30	10	26 – 30	19	2+	3
31 – 35	5	31 – 35	5	3	17
36 – 40	1	36 – 40	2	3+	3
		41 – 45	1	4	3
		46 – 50	1		

N=112

Figure 4-1a: Histogram for Listening Test results

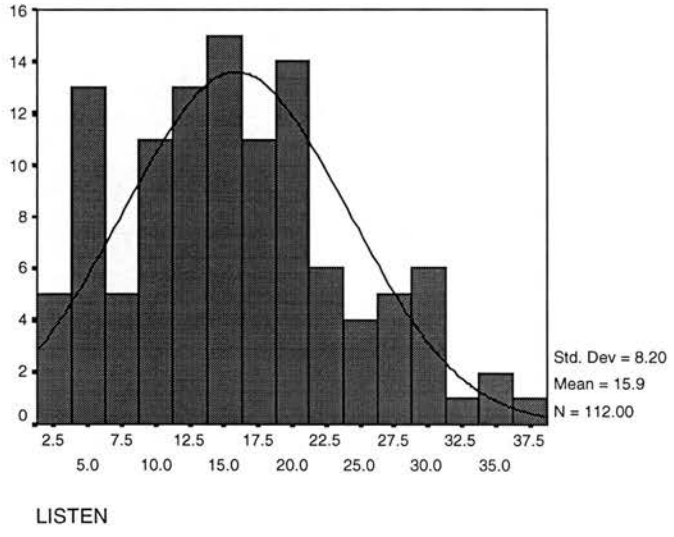


Figure 4-1b: Histogram for Grammar Test results

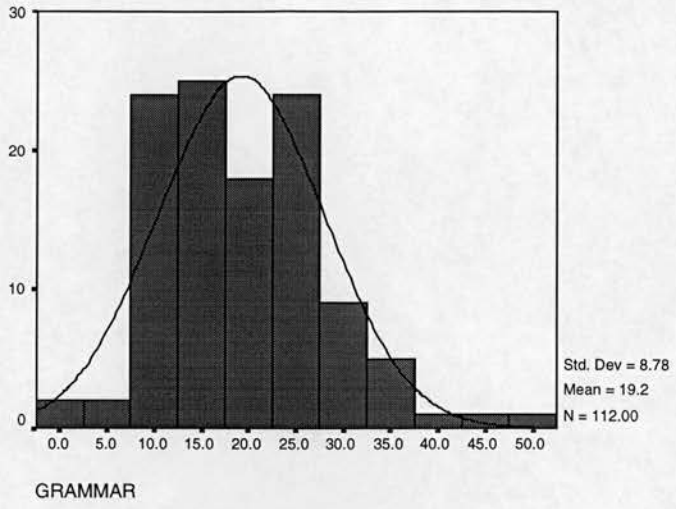
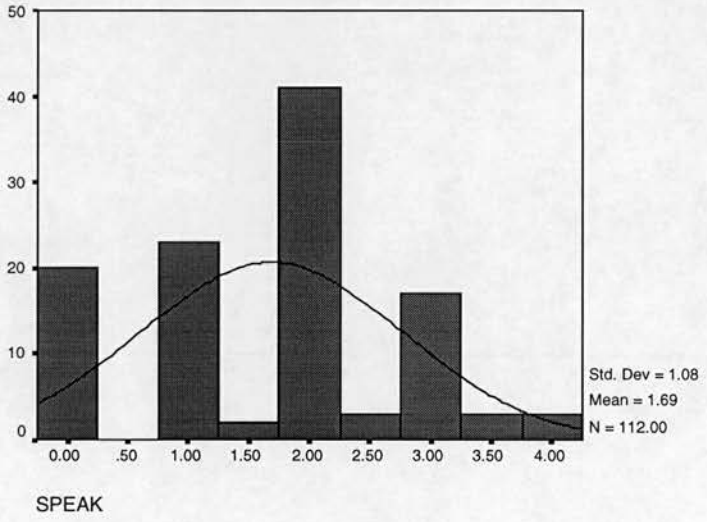


Figure 4-1c: Histogram for Speaking Test results



Descriptive statistics for Listening, Grammar and Speaking

Descriptive statistics for the three sub-tests are summarised in Table 4-3. Listening and Grammar mean scores are low but score ranges are wide. The pass level will be adjusted (Section 4.4.). Individual items and task types are discussed below.

Table 4-3: Descriptive statistics of Main Trial Listening, Grammar & Speaking

	Listening	Grammar	Speaking
Mean score	15.875	19.1518	S1.68: Below standard
Median	15.00	18.50	S2: Satisfactory
Mode	13.00	11.	S2: Satisfactory
Std. Deviation	8.2047	8.7839	1.07
Range	36	50	4
SEM	0.775	0.83	0.101

N=112

Item analysis – Listening and Grammar

Table 4-4 displays the mean item facility values of the Listening and Grammar tests with their standard deviations and reliability coefficients (α estimates).

Table 4-4: Listening and Grammar mean item facility values, SD and α estimates

	Listening	Grammar
Mean	0.35	0.30
SD	0.25	0.20
α	0.91	0.87

The mean item facility values are low, which suggests that the two tests are difficult for this sample of students. However, items seem to work homogeneously as the rather high α estimates indicate.

Individual item facility values and discrimination indices of the Listening and Grammar Tests are listed in Tables 4-5a and 4-5b below. The difficulty range of 0.2 – 0.9 has been suggested. Items falling outside the range were either revised or removed. Henning (1987) suggests a range of 0.33 – 0.67. However, decisions were made to allow some flexibility for the following reasons. First, the inclusion of some easy items may help motivate the test takers. Second, inclusion of easy and difficult items may create a more sensitive and discriminating test. Finally, traditional analysis is very much group-dependent: the ability of the sample chosen influences item facility values which in turn affect test reliability. Some items may have low facility values; nevertheless they could be judged as good items and may be able to separate different ability groups.

Listening test items

Table 4-5a lists the facility values and discrimination indices of the listening items.

Table 4-5a: Facility values and discrimination indices of Listening Test

Task No.	Item No.	Item Facility	Item Discrimination		Task No.	Item No.	Item Facility	Item Discrimination
1	1	0.13	0.27		4	24	0.09	0.27
1	2	0.13	0.3		4	25	0.06	0.19
1	3	0.1	0.19		4	26	0.09	0.27
2	4	0.59	0.38		5a	27	0.58	0.53
2	5	0.19	0.27		5a	28	0.72	0.65
2	6	0.06	0.13		5a	29	0.69	0.65
2	7	0.69	0.14		5b	30	0.2	0.54
2	8	0.13	0.24		5b	31	0.34	0.54
2	9	0.18	0.25		5b	32	0.05	0.08
2	10	0.01	0.03		6a	33	0.57	0.65
2	11	0.18	0.24		6a	34	0.63	0.63
3a	12	0.85	0.14		6a	35	0.64	0.7
3a	13	0.8	0.35		6a	36	0.52	0.56
3a	14	0.51	0.43		6b	37	0.66	0.6
3a	15	0.67	0.16		6b	38	0.46	0.75
3b	16	0.57	0.57		6b	39	0.46	0.76
3b	17	0.49	0.49		6b	40	0.41	0.66
3b	18	0.35	0.65		6b	41	0.14	0.41
3b	19	0.46	0.46		6b	42	0.21	0.46
3b	20	0.3	0.38		6b	43	0.17	0.32
3b	21	0.15	0.25		6b	44	0.41	0.75
4	22	0.0	0.0		6b	45	0.13	0.38
4	23	0.15	0.41					

Overall, the following task types were considered easy for this sample:

- Task 3a, Items 12 – 15, True/False,
- Task 5a, Items 27 – 29, Information transfer - Filling in the train information, and
- Task 6a, Items 33 – 36, Identification & Labelling - Matching.

The following task types were difficult for the candidates:

- Task 1, Items 1 – 3, Identification and Labelling – Locating the whereabouts of the vehicles,
- Task 2, Items 4 – 11, Information transfer – Filling in personal information
- Task 4, Items 22 – 26, Short answers, and
- Task 6b, Items 41 – 45, Information transfer - Filling in the key word.

It seems that this sample of students did well in questions which involve in a process of elimination such as Tasks 3a and 6a but they found information retrieval types of tasks difficult, namely Tasks 1 and 4. Tasks 2 and 5a require the test takers to fill the grids with information they heard. Students did better in Task 5a, writing down train information than Task 2, filling in important personal information. Students may have had more practice in filling in time and numbers rather than dates and other

personal information in the classroom. The poor performance on the last few items (Items 41 – 45) may be caused by fatigue or a loss of interest.

The overall item discrimination power is good except for Item 22 which asks students to indicate the two parts of a traditional Taiwanese house in which the influences of the *Book of Changes* can be detected. The answer is *the roof* and *the framework of the building*. We thought that Item 22 would be an easy question because the speaker explicitly mentioned the two features. About 15 students managed one answer. With hindsight, Item 22 is difficult and should be revised.

Grammar Test items

The Grammar facility values and discrimination indices are displayed in Table 4-5b. The easiest tasks for the students are: Part II (Items 13 – 22): Verb forms, and Part Va (Items 51 – 55): Completing the dialogue. Performance of Part I (Items 1- 12): MCQs seems to be satisfactory.

Part III (Items 23 – 37): Sentence transformation, Part IV (Items 38 – 50): Filling in the missing word and Part Vb (Items 56 – 60): Giving appropriate responses are the three most difficult tasks according to item statistics. The low facility values of Part Vb may be caused by the lack of time; it is not certain if students found this task difficult. Sentence transformation and filling in the missing word are common high school classroom exercises but are generally considered demanding because these two exercises require understanding and production of language other than simple linguistic decoding. Students seemed to perform well on items which involve a process of elimination such as MCQs, or items which require a localised knowledge of the structure such as verb forms, or items with cues provided such as completing the dialogue. Students did not seem to score well on items which require global understanding of the language knowledge such as filling in the missing word and sentence transformation.

Table 4-5b: Item facility values and item discrimination indices of Grammar Test

Part No.	Item No.	Item Facility	Item Discrimination		Part No.	Item No.	Item Facility	Item Discrimination
I	1	0.28	0.05		III	34	0.15	0.21
I	2	0.16	0.11		III	35	0.01	0.03
I	3	0.36	0.22		III	36	0.04	0.05
I	4	0.55	0.27		III	37	0.15	0.19
I	5	0.32	0.14		IV	38	0.23	0.38
I	6	0.86	0.19		IV	39	0.36	0.52
I	7	0.42	-0.21		IV	40	0.39	0.32
I	8	0.31	0.05		IB	41	0.35	0.11
I	9	0.3	0.32		IV	42	0.15	0.24
I	10	0.09	-0.06		IV	43	0.23	0.35
I	11	0.48	0.46		IV	44	0.38	0.32
I	12	0.29	0.35		IV	45	0.36	0.43
II	13	0.56	0.43		IV	46	0.37	0.3
II	14	0.52	0.62		IV	47	0.19	0.25
II	15	0.32	0.19		IV	48	0.52	0.3
II	16	0.43	0.41		IV	49	0.02	-0.05
II	17	0.54	0.59		IV	50	0.36	0.43
II	18	0.83	0.22		Va	51	0.61	0.65
II	19	0.39	0.4		Va	52	0.41	0.43
II	20	0.39	0.45		Va	53	0.28	0.33
II	21	0.1	0.11		Va	54	0.17	0.27
II	22	0.71	0.38		Va	55	0.58	0.56
III	23	0.0	0.0		Vb	56a	0.37	0.7
III	24	0.35	0.46		Vb	56b	0.32	0.65
III	25	0.0	0.0		Vb	57a	0.24	0.54
III	26	0.39	0.3		Vb	57b	0.14	0.41
III	27	0.08	0.16		Vb	58a	0.19	0.35
III	28	0.12	0.32		Vb	58b	0.09	0.19
III	29	0.33	0.3		Vb	59a	0.18	0.24
III	30	0.55	0.13		Vb	59b	0.08	0.24
III	31	0.01	0.03		Vb	60a	0.15	0.41
III	32	0.0	0.0		Vb	60b	0.01	0.3
III	33	0.03	0.08					

Detailed comments on individual Listening and Grammar items are presented below.

Listening

The following Listening items have facility values outside the range of 0.2 – 0.9. Items marked with an asterisk "*" are new items for the main trial.

Table 4-6: Listening items with low facility values

Task Number	Item Number	Facility Value	Item Discrimination
1: Identification & Labelling	1	0.13	0.27
	2	0.13	0.3
	3	0.10	0.19
2: Information transfer/Filling in important dates	5	0.19	0.27
	6	0.06	0.13
	8	0.13	0.24
	9*	0.18	0.25
	10	0.01	0.03
	11	0.18	0.24
3b: Matching	21	0.15	0.25
4: Short answers	22	0.0	0.0
	23	0.15	0.41
	24	0.09	0.27
	25	0.06	0.19
	26	0.09	0.27
5b: Completing the sentence	32	0.05	0.08
6b: Completing the sentence/phrase	41	0.14	0.41
	43*	0.17	0.32
	45*	0.13	0.38

Table 4-6 seems to indicate that task types rather than individual items have an impact on item performance. Individual items also need revision; for example, students failed to associate "Tolbooth" with the answer "prison" in Item 43, a talk on Edinburgh's Royal Mile, despite the fact that it is mentioned in the passage; Item 43 should be discarded as it may contain unfamiliar cultural reference; for Items 21, 32 and 45, it seems that students simply failed to retrieve the correct word, suggesting insufficient vocabulary.

The following Grammar items fall outside the suggested facility range of 0.2 - 0.9.

Table 4-7a: Main trial Grammar items with low facility values

Task No.	Item number		Facility	Dis
I	2	John seemed very anxious ___ the event to be a success. A. of B. in C. about D. for	0.16	0.11
I	10	The news of the president's retirement ___ be announced soon. No one is sure when. A. will B. would C. can D. could	0.09	-0.06
II	21	It (situate) just 4 km from Royan in the small resort of	0.10	0.11
III	23	It appears that we don't have any film left. We seem to _____.	0	0.0
III	25	Mary will return from her business trip to Dublin very soon. It won't be _____ from her business trip to Dublin.	0	0.0
III	27	For further information, please contact the Tourism..... Further information _____ the Tourism	0.08	0.16
III	28	Only senior staff members are allowed to use the .... The company car park is _____ senior staff.....	0.12	0.32
III	31	My cousin is the most annoying person I've ever met. I've yet _____.	0.01	0.03
III	32	I was surprised at how approachable the new boss is. I didn't expect _____.	0	0.0
III	33	I will never lend money to Robert no matter what. Under no _____.	0.03	0.08
III	34.	It has been two weeks since anyone saw John. John _____.	0.15	0.21
III	35.	Joan didn't attend the meeting. She apologises. Joan sends _____.	0.01	0.03
III	36.	This door must be kept closed at all times. At no time _____.	0.04	0.05
III	37.	The suitcase was so heavy that Tommy could not lift it. The suitcase was too _____.	0.15	0.19
IV	42.	...of a two-storey platform with lookout towers ___ the four corners giving the place a grand appearance.	0.15	0.24
IV	47.	After Cheng Cheng-kung ___ the Dutch in 1662, the fort...	0.19	0.25
IV	49.	...the fort was in a dilapidated ___ and during the Japanese occupation...	0.02	-0.05
Va	54.	(Offer help.) _____	0.17	0.27

(Questions 56 - 60 are displayed in Table 4-5b.)

In general, items in Part III: Sentence Transformation and Part IV: Filling in the missing word are difficult. Regarding the individual items listed in Table 4-7a, some

need revision or removal. Question 2 tests knowledge of prepositions. The intended answer is D: *for*, but, C: *about* elicited many responses and needs revision. Question 10 assesses modal verbs, but the stem is ambiguous and two options seem possible. As one expert informant pointed out later, such questions should be avoided in the future (Chapter 5). Finally, some questions in Part III do not seem to be appropriate; for examples, Questions 27, 31, 33 and 36 seem to measure literary style as well and should be discarded.

Table 4-7b lists Items 56 – 60: Giving appropriate responses. Each item described a situation in which the candidates were required to provide two appropriate responses. This task measures the candidate's ability for appropriate use. Most students failed to do this part; their responses are indicated by 9 (=missing data). Therefore, we do not know if this task has truly fulfilled its purpose in the measurement of appropriate use. Taken the testing time and the number of items into consideration, in future test revisions, one response for every situation seems adequate.

Table 4-7b: Item facility values and discrimination indices of Grammar Items 56 – 60

Situations	Facility Value	Item Discrimination
56. You are taking some foreign visitors to Keelung tomorrow. The city is well known for its rainfall. You want to remind your guests to bring an umbrella. What would you say?	56a: 0.37 56b: 0.32	56a: 0.7 56b: 0.65
57. You are taking a group of visitors to hike the Tasuling Historic Trail. It's a very warm day. You offer an elderly lady some mineral water. What would you say?	57a:0.24 57b: 0.14	57a:0.54 57b:0.41
58. You and your foreign clients are visiting the National Palace Museum. You want to show them the well-known jade exhibit, the Jade Cabbage. You are trying to get your clients to follow you. What would you say?	58a: 0.19 58b: 0.09	58a: 0.35 58b: 0.19
59. You and your foreign visitors are visiting the Martyrs' Shrine. It is a solemn place. You are reminding your visitors to be quiet and respectful. What would you say?	59a: 0.18 59b: 0.08	59a: 0.24 59b: 0.24
60. You want to borrow a client's newspaper so you can check the cinema times. What would you say?	60a: 0.15 60b: 0.01	60a: 0.41 60b: 0.3

To conclude, in general, the Listening and Grammar task types seem to affect item difficulty. Overall, tasks with cues provided or tasks involved in process of elimination such as verb forms, matching, MCQs, and true/false performed better than tasks requiring global understanding of the language or text. Such tasks include filling in the missing word, sentence transformation, short answers and completing the

sentence/phrase. Regarding information transfer type of tasks, students did better on filling in times and numbers than filling in personal information.

With regard to individual items, some need revision and some should be discarded. However, as discussed earlier, the performance of a given item seemed to be influenced by its task type; therefore test revision needs to consider both items and task types. In terms of the relationship of different task type in the measurement of the proposed traits, a factor analysis was performed to examine the adequacy of each of the tasks; results will be discussed in Chapter 5. For the time being, revision of the Listening and Grammar tests and test items is based on the consideration of the overall length, the number of tasks/questions, length of the listening passages and item statistics. Details of revision are discussed in Section 4.3., and the revised test presented in Appendix 4-10.

Listening Test

1. The 40-minute testing time seems appropriate and will remain unchanged.
2. Maximum listening text length should be shortened to 3 - 4 minutes.
3. True/False seems easy and can be removed. On the other hand, short answers seems difficult for the test takers and should be made easier.
4. Questions with specific cultural reference such as Item 43 should be discarded.

On the whole, the Listening Test could have shorter passages and fewer questions, e.g., 35 - 40 questions. The testing time could remain unchanged, i.e., about 40 minutes.

Grammar Test

1. The 40-minute testing time is about right. However the number of items needs reducing to 40 – 45 instead of 65.
2. Language elements or structures less common in the profession should be removed from the test, for example, some items in Part III: Sentence Transformation such as Item 36.
3. For Part Vb, one response instead of two seems adequate.

Correlation of the three sub-tests

The correlation of the three tests is listed below in Table 4-8.

Table 4-8: Correlation of Listening, Grammar & Speaking Tests

	Listening	Grammar	Speaking
Listening	---	0.54	0.62
Grammar		---	0.55
Speaking			---

Note: Spearman's correlation, N=112; $p < 0.01$

These *r* values suggest a moderate correlation between the components of the test battery, indicating a definite relationship yet each test seems to measure a distinct trait. This relationship seems in accord with the test rationales discussed in Chapter 3 in that grammatical competence, listening and speaking seem distinct language use dimensions; tasks were designed to measure the proposed constructs accordingly.

Rater statistics and inter-rater reliability

Three types of rater statistics are reported in this section: (1) descriptive statistics of individual raters, (2) mean rating of first rater and his/her second rater and (3) correlation of ratings among raters.

Individual rater descriptive statistics are displayed in Table 4-9a. These statistics are only indicative of the rater behaviour as each rater was assigned to different candidates with varied oral ability; the statistics are less reliable and meaningful. To compare the relative rater severity, the mean ratings of each rater listed in Table 4-9b are examined.

Table 4-9a: Rater statistics

	No. of ratings	Mean rating	S.E.M.	SD
JJ	45	1.03	0.11	0.77
HH	42	1.34	0.12	0.80
BB	25	1.52	0.24	1.21
GG	24	1.64	0.25	1.27
II	20	2.66	0.15	0.69
DD	26	2.29	0.11	0.59
LL	20	0.85	0.22	0.99
KK	22	1.58	0.23	1.07

Table 4-9b: Mean ratings of the judges

Number of persons shared	1st Rater	2nd Rater
9	JJ 1.88	GG 2.5
4	JJ 0.5	KK 1.69
15	JJ 0.65	HH 0.65
6	JJ 0.95	II 2.25
11	JJ 1.13	DD 2.27
14	HH 1.88	II 2.8
9	HH 1.6	DD 2.16
4	HH 1.43	KK 2.37
6	BB 2.7	DD 2.5
18	BB 1.04	LL 0.94
13	GG 1.3	KK 1.17
1	KK 3.25	BB 3

Raters JJ and HH tended to give lower ratings than the second rater; the other raters seem to be fairly compatible in their rating. There are two possible explanations for the low mean ratings of JJ and HH. First, they were not certain about the rating scale levels so they used the three levels they were sure of, namely, 0: *No foreign language use ability*, 1: *Below standard* and 2: *Satisfactory*. This has to do with their training. The second explanation is that they were less experienced and therefore were less able to differentiate levels of candidate performance.

A Spearman correlation was computed to examine how each rater correlated with the others. Table 4-9c shows the correlation matrix for the inter-rater correlation.

Table 4-9c: Inter-rater correlation

	JJ	HH	BB	GG	II	DD	LL	KK	FINAL
JJ	---	0.88*		0.81*	0.67	0.58		0.45	0.71*
HH		---			0.75*	0.55		0.21**	0.85*
BB			---			0.99*	0.87*		0.94*
GG				---				0.94	0.97*
II					---				0.84*
DD						---			0.87*
LL							---		0.91*
KK								---	0.92*
FINAL									---

Note: Spearman's correlation, N=8; * indicates $p<0.01$; **= the raters shared only 4 candidates

On the whole, the raters seemed to be compatible with one another. JJ, HH and GG were inexperienced raters. Rater JJ showed the least compatibility with the other raters. Of the 46 tapes JJ listened, 19 were incompatible with the second rating and required a third rating, which tended to agree with the second rating. Her incompatibility was indicated by the moderate correlation of 0.71 between her rating and the final rating. An examination of her ratings showed that Rater JJ tended to

award S1s or S2s to the candidates while some of them were awarded a 2+ or 3 by the second rater. Rater JJ needs more guidance and monitoring in the future. The overall inter-rater reliability is 0.77 – 0.78, which is not particularly high and suggests a need for further rater training.

4.2.2. Discussion of the classical test analysis results

Classical test analysis results suggest that the Listening and Grammar Tests are difficult for this sample population. Regarding the length of the tests, students were not able to complete the Grammar test within the time allowed, which suggests that the test needs to be shortened. The time allotment for the Listening test seemed to be about right. However, the poor facility values of the last five items seem to suggest students were tired at the end and Listening Test needs shortening as well.

Task types seem to influence candidate performance, which in turn affects item difficulty estimates; some task types are more difficult than others, for example, short answers, filling in personal information, and identification & labelling for the Listening test, and sentence transformation and filling in the missing word for the Grammar test. In future test revision and development, the types of tasks to be included need examination; this, however, does not mean excluding tasks with low facility values, which may be difficult for particular groups of candidates but may not necessarily be bad tasks.

The conclusion to be drawn from the item statistics is that students do not seem to have been exposed to varieties of task types in the classroom. With regard to classroom learning and teaching, students could be offered more varieties of tasks in the classroom, which will involve careful selection of teaching materials and classroom management.

Regarding the Speaking test results, the mean candidate ability is S:1.68 (*Below standard*). However, most candidates fall in S2: *Satisfactory,* which suggests that students understand the general spoken message but their communication needs some improvement. With reference to the learning and teaching of the oral skill in the

classroom, emphasis could be placed on fluency, appropriateness and organisation of the speech.

The inter-rater reliability of 0.77 – 0.78 is rather low and suggests a need for more training and careful rater selection. In terms of rater severity, Raters JJ and HH are more severe than the other six raters. They will need to be given more training and be closely monitored in the future.

4.2.3. Rasch analysis

In this section, Rasch results are summarised and interpreted. The Listening and Grammar Test statistics were computed with the use of the computer programmes Quest (Adams & Khoo, 1992) and Facets Version 3.2 (Linacre, 1999) for the Speaking test.

For the Listening and Grammar tests, the following are presented:

- item difficulty estimates,
- item fit,
- person ability estimates,
- person fit,
- the difficulty/ability scale,
- item analysis and
- the candidate kidmap.

The following persons and items were not included in the person and item calibrations: Candidate 1003 for the Grammar Test, Item 22 of the Listening Test, Items 32 and 35 of the Grammar Test. Rasch measurement cannot estimate persons with perfect or zero scores and items to which responses are either all correct or all incorrect as it is not possible to place such persons or items on the ability/difficulty scale since all that is known in such cases is that (1) the test taker is either too able or not able enough for the set of items administered to them and (2) the items are either too easy or too difficult for the group of people who take the test (Section 1-4). Thus, in this section, the Grammar Test had 111 measurable persons and 63 item calibrations; the Listening Test had 112 measureable persons and 44 item calibrations.

For the Speaking test, the following will be reported:

- candidate performance (i.e., candidate ability estimate),
- rating behaviour (i.e., rater severity) and
- the rating scale statistics.

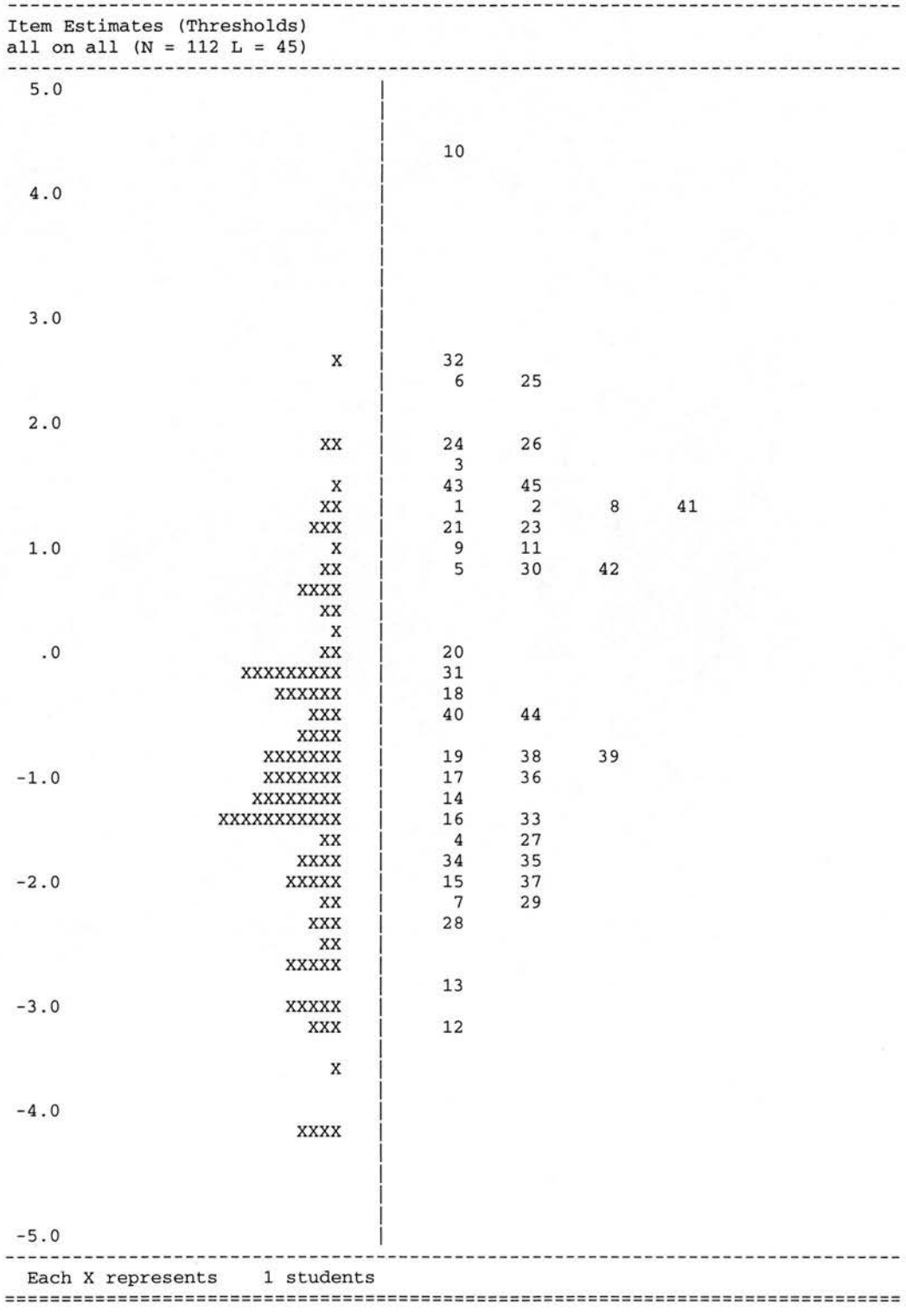
Listening Test

Item difficulty estimates (Thresholds)

Figure 4-2 displays the Listening item difficulty and person ability on the common logit scale. Detailed item estimates are listed in Appendix 4-1. In Figure 4-2, the scale extends from +5 to –5 logits. Each person is represented by an X and is separated from the items on the right by the dotted line in the centre.

Item difficulty ranged from +4.49 to –3.27 logits; Item 10 at +4.49 was the most difficult item, and Item 12 at –3.27 logits, the easiest. Excluding Item 10, item difficulty estimates ranged from +2.56 to –3.27. The mean difficulty estimate is set at zero by the analysis with a standard deviation of 1.72. Standard errors ranged from 1.02 to 0. Excluding the largest S.E. of 11.02 (for Item 10), standard errors ranged from 0.42 to 0.22. The reliability estimate is 0.97.

Figure 4-2: TG Listening item estimates and person ability



Item fit

Item fit is reported in terms of the mean squares (MnSq), the ratio of observed error variance to modelled error variance, and *t*-values (*t*) referring to the standardised weighted mean square. Two types of mean squares and *t*-values are reported: the unweighted (Outfit), sensitive to outliers and the weighted (Infit), sensitive to in-liers (Adams & Khoo, 1993). The mean square has an expected value of one. If the estimated mean square is larger than one, the overall response pattern shows some variation from the modelled response pattern. If the value is less than one, the observed overall response pattern displays uniformity. The observed response pattern seldom displays exact fit to the model's expectation; therefore, MnSq values between 0.75 and 1.3 and *t*-values between ±2 have been suggested to indicate good fit (Adams & Khoo, 1993). The overall Listening item fit statistics are reported in Table 4-10.

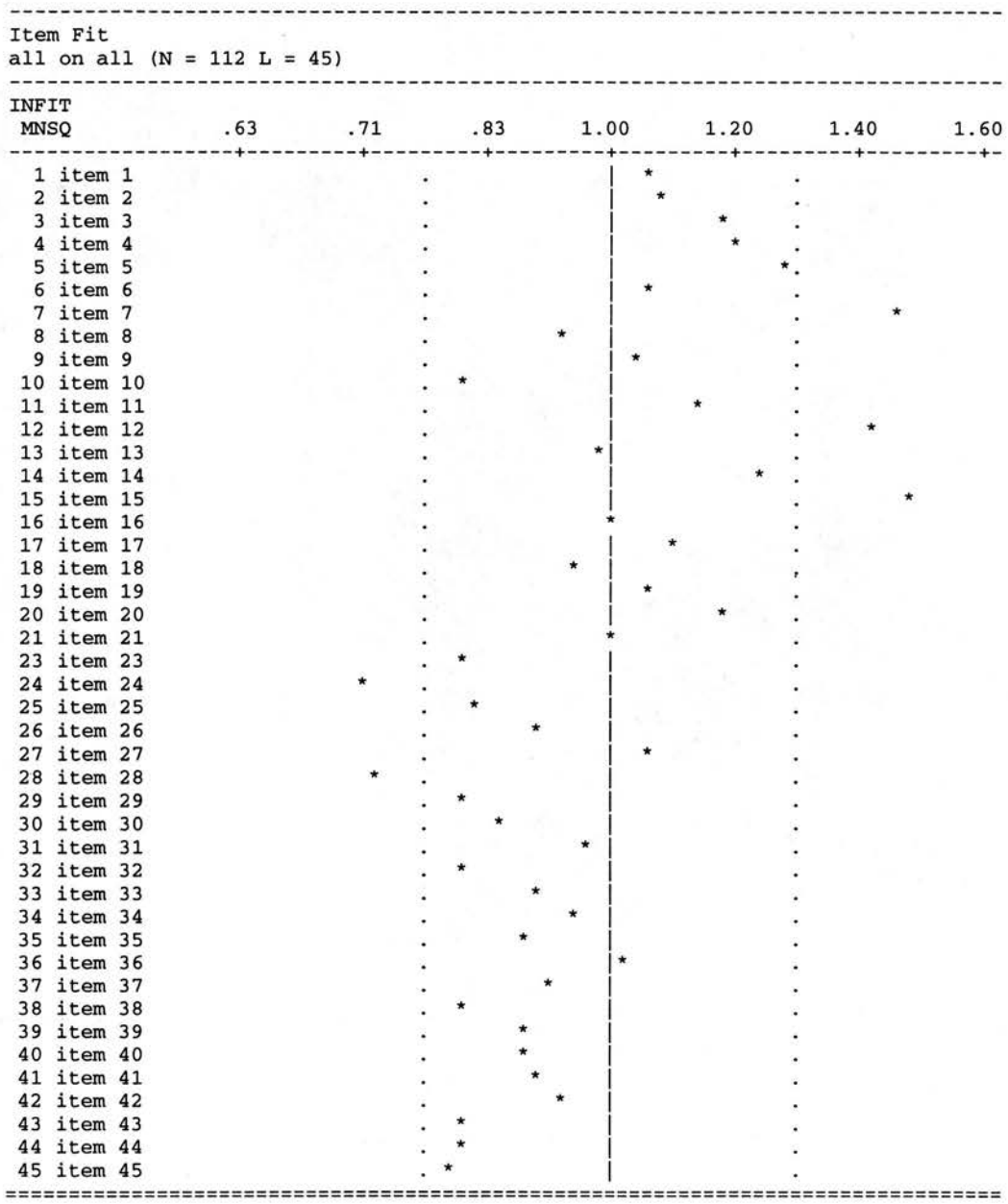
Table 4-10: Listening item fit statistics

Infit Mean Square		Outfit Mean Square	
Mean	0.99	Mean	0.96
SD	0.19	SD	0.46
Infit *t*		Outfit *t*	
Mean	-0.01	Mean	-0.01
SD	1.40	SD	1.27

On the whole, the observed responses fitted the expectations of the model, indicating the listening items seemed to work well together.

Visual representation of individual item fit is presented in Figure 4-3. Detailed fit statistics are given in Appendix 4-1.

Figure 4-3: TG Listening item fit



Listening items ranged from the Infit mean squares of 0.72 – 1.47. Five items fell outside the range of 0.75 – 1.3; they are Items 7, 12, 15, 24 and 28. Table 4-11a displays their MnSq and t-values and Table 4-11b displays these five items.

Table 4-11a: Misfitting Listening items

Item No.	Infit MnSq	Outfit MnSq	Infit-t	Outfit-t	Difficulty
7	1.47	1.63	3.7	2.3	-2.02
12	1.42	2.09	2.1	1.9	-3.27
15	1.47	2.40	3.8	4.7	-1.91
24	0.72	0.29	-1.2	-1.3	1.93
28	0.72	0.73	-2.5	-1.0	-2.30

Items 7 and 15 exhibit 47% variation from the expected value of 1, and Item 12, 42%. The large infit mean square values indicate that the error variance of the observed response patterns deviate from their modelled error variance by more than 40%, which suggests that these items had unexpected response pattern and need examination.

Items 24 and 28, on the other hand, show 18% less variation than the expected value of 1, indicating the actual response patterns are more regular than expected. Except for Item 24 with a difficulty index of 1.93 logits, the other four items seem easy. Item 12, the easiest item, has a difficulty estimate of –3.27; it should be removed. Item 24 is considered a difficult item as the difficulty estimate of 1.93 logits indicat. Its low infit mean square value of 0.72 means that the persons who are expected to score the item have already been stipulated by the model; thus, according to the model, the observed response pattern is a little too predictable. Item 24, on the basis of the item statistics, should be revised.

The five misfitting Listening items are listed in Table 4-11b. Apart from Item 24, the rest are considered easy items. According to the model, these items require revision or removal.

Table 4-11b: Misfitting Listening items

Task 2	7. Date of death (of Agatha Christie): _____
Task 3a	12. Making up faces in the Chinese opera requires training and ability: _____ (true or false)
	15. A cheerful character has more colours on his/her face: _____ (true or false)
Task 4	24. What does a swallow-tailed roof tell us about the owner of a house? _____
Task 5a	28. Departing time of the train to London King's Cross from Edinburgh: _____

Person ability estimates

Visual representation of the person ability estimate is given in Figure 4-2. The more able the person is, the higher s/he is on the logit scale. Detailed person ability estimates are provided in Appendix 4-2. Listening person ability ranged from +2.69 to –4.07 logits, corresponding to raw scores of 38 and 2 points. The mean ability for this sample group is at –1.04 logits (SD=1.39), indicating the overall ability of the group is about one logit lower than their expected ability. The standard errors ranged from 0.38 to 0.76. They are lowest for candidates whose abilities are closer to the centre of the possible ability range (i.e., near 0 logits). The standard errors of the top and bottom scores are 0.5 and 0.76 respectively, reflecting a relative lack of information of the response vectors for persons falling at the extreme ends of the range. The reliability estimate is 0.9.

Person fit

Again, four fit statistics are reported: Infit MnSq, Outfit MnSq, Infit-*t* and Outfit-*t*. These statistics indicate the degree of variation of individual person response pattern according to the overall person response pattern. Mean squares are expected to be one and t-values are expected to be near zero. Acceptable ranges for each of the fit statistics are the same to those of item fit statistics: 0.75 – 1.3 for MnSq values and ±2 for *t*-values. The difference between person fit and item fit is that the squared residuals (i.e., the difference between observed and expected data) are summed over items for a person rather than over persons for an item. These residuals are then converted to t-statistics. T-values in person fit statistics provide a summary of the discrepancy between the person's observed and expected response pattern. The overall person fit statistics are given in Table 4-12.

Table 4-12: Listening person fit statistics

Infit Mean Square		Outfit Mean Square	
Mean	1.00	Mean	0.96
SD	0.23	SD	0.62
Infit *t*		Outfit *t*	
Mean	0.01	Mean	0.12
SD	1.12	SD	0.83

In general, the fit statistics suggest that the overall person response patterns seemed to be in good accord with the model's expectation.

t-values will be used to identify candidates whose observed response patterns deviate markedly from those predicted by the model. *Quest* has suggested an acceptable range of +2 to –2 (Adams & Khoo, 1993). Eight persons, listed in Table 4-13, have shown Infit-t values greater than 2.

Table 4-13: Listening Test misfitting persons

Name	**Estimate**	**Infit MnSq**	**Outfit MnSq**	**Infit-t**	**Outfit-t**
1009	0.86	1.57	1.75	2.55	1.59
3038	-0.62	0.57	0.41	-2.54	-1.59
3045	-1.70	1.75	3.43	3.26	2.26
4001	-1.38	0.63	0.40	-2.29	-1.02
6007	0.41	0.45	0.34	-3.31	-2.32
6026	-0.77	0.6	0.47	-2.33	-1.26
7001	0.71	1.51	1.78	2.29	1.74
7014	0.86	1.77	2.35	3.26	2.49

Their large t-values suggest that their response patterns have caused disturbance to the test as an instrument of measurement and could be indicative of a lack of construct validity of the test. However, if taking all observations including outliers into account, only three candidates departed from their expected ability measures. They are Candidates 3045, 6007 and 7014; their response patterns are listed in Table 4-14. Detailed response patterns are displayed in Appendix 4-7.

Table 4-14: Response patterns of Candidates 3045, 6007 and 7014

Question No.	Candidate 6007	Candidate 7014	Candidate 3045
22	0	0	0
10	0	0	0
32	0	1	0
25	0	1	0
6	0	0	0
24	0	0	0
26	0	1	0
3	0	0	0
43	0	0	0
45	0	0	0
1	0	1	0
2	0	0	0
8	0	1	0
41	0	1	0
21	0	1	0
23	0	1	0
9	0	0	0
11	0	1	0
5	0	0	0
30	1	1	0
42	1	1	0
20	0	1	0
31	1	1	1
18	1	0	0
44	1	1	0
40	1	1	0
19	1	1	1
38	1	1	0
39	1	0	0
17	1	0	1
36	1	1	1
14	1	1	0
33	1	0	1
16	1	1	1
27	1	1	0
4	1	1	1
35	1	0	0
34	1	0	0
15	1	0	1
37	1	1	1
7	1	1	1
29	1	1	0
28	1	1	0
13	1	1	1
12	1	1	0

In the column "Question No.", all items are ranked according to difficulty from the most difficult to the easiest. Candidates are expected to score the easier questions, indicated by 1. Items on top are expected to have fewer 1s whereas in the intermediate zone, a mixture of 1s and 0s is expected. The response patterns of the three candidates are unexpected in that

(1) in the difficult zone in which more 0s than 1s are expected, Candidates 6007 and 3045 failed to score any 1s;

(2) in the intermediate zone in which a mixture of 1s and 0s are expected, Candidates 6007 and 7014 scored more 1s than 0s and Candidate 3045 only scored a few items;

(3) in the easy item zone in which more 1s are expected, Candidate 6007 scored all the items.

The response patterns of the three test takers deviate from the modelled response pattern. Candidate 6007's response pattern is characterised by 0s in the upper half of the difficulty continuum and 1s in the lower half continuum. By the way he scored the items, Candidate 6007 seems a cautious test taker, unwilling to take risks. He only answered items he was sure of. Candidate 7014 managed to score some of the difficult items but failed some of the easy items, which could be caused by carelessness. For Candidate 3045, a mixture of 0s and 1s are observed in the bottom area of easier items where more 1s are expected. Examination of the types of items Candidate 3045 scored showed that he scored mostly on matching and true/false items, suggesting he was only able to cope with some types of tasks.

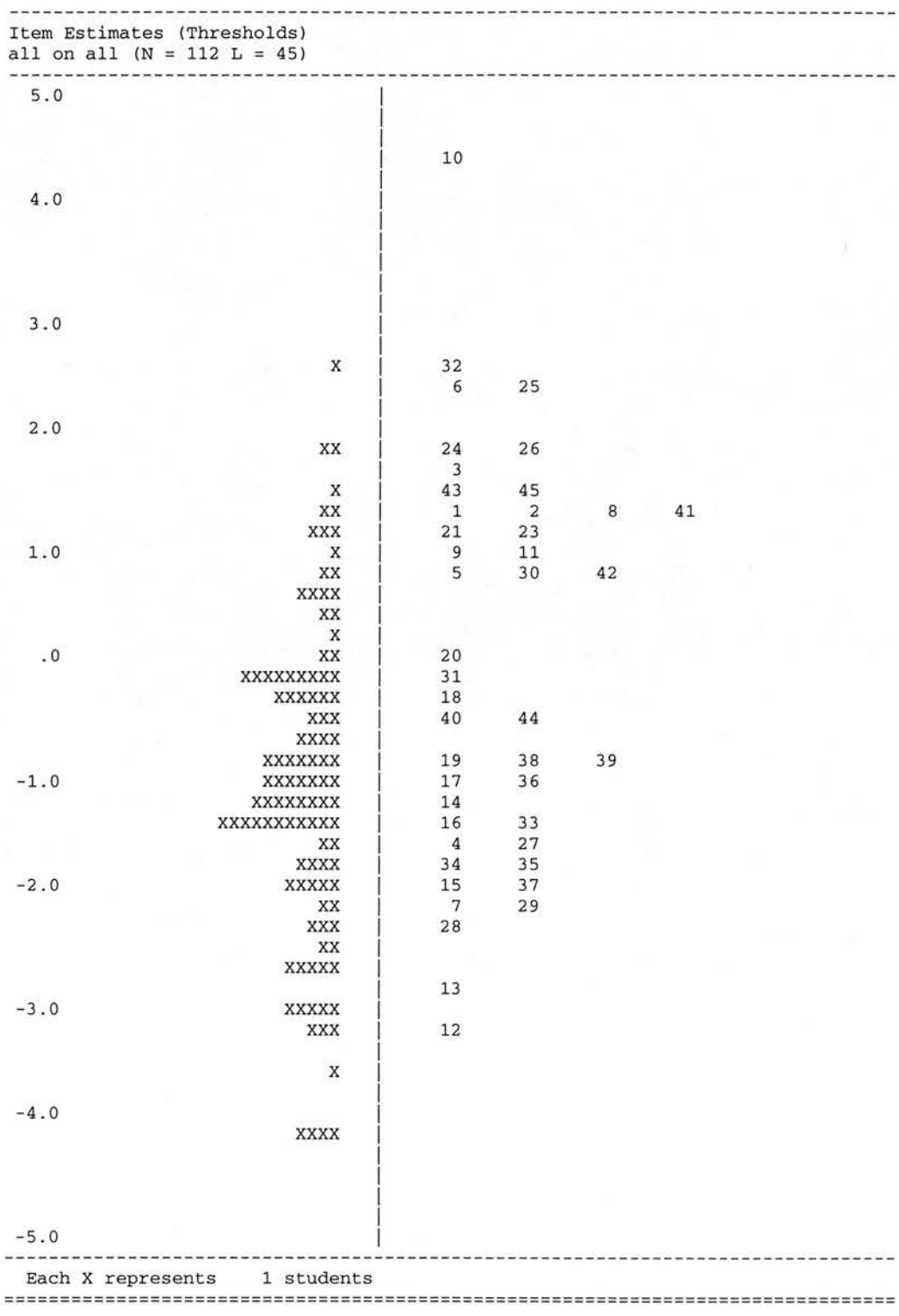
The difficulty/ability scale

Because item difficulty and person ability are all interpreted in terms of logit scores, it is possible to compare the relative standing of item difficulty and person ability on the common logit scale as shown in Figure 4-2 (reproduced on the next page). On the difficulty/ability map, the logit scale ranging from +5 to –5 appears on the left; the dotted line in the centre separates the candidates on the left and the items on the right with each X representing a candidate and each number representing the item numbers. Candidate abilities extended from +2.69 to –4.07 logits, corresponding to raw scores of 38 and 2 points respectively, and item difficulty from +4.49 to –3.27 logits.

There are no candidates who are capable enough for all the items; the most competent candidate stands at +2.69 logits whereas the most difficult item is at +4.49 logits. Item 10, the most difficult item, is separated by 1.8 logtis to the rest of the items. Excluding Item 10, candidate ability matches item difficulty. At the lower end of the scale, approximate five candidates whose abilities fall below item difficulty. There are two possible explanations; first, the listening test is difficult and should be made easier and second, these five candidates are below the ability level the test intended to measure; the test was not for them.

Four items (Items 10, 32, 6 and 25) were difficult and two items (Items 12 and 13) were easy; test revision should attempt to make the four items less difficult and remove the two easy items.

Figure 4-2: TG Listening item estimates and person ability



Item analysis

The results of the Listening item analysis are presented in Appendix 4-3. For each item, the following information is provided: (1) the proportion of correct and incorrect responses, (2) the percentage of correct and incorrect responses, (3) a point biserial correlation and the p-value, (4) the mean ability, (5) difficulty estimate with its associated measurement error and (6) goodness of fit (i.e., the Infit MnSq). Finally, the overall mean test score with its SD and reliability index are given. (1) to (4) refer to information provided by the classical analysis and (5) to (6) provide Rasch item analysis. The focus of discussion is on item discrimination. Three items have been identified to have poor discrimination power. They are listed in Table 4-15a.

Table 4-15a: Listening items with poor discrimination power

Item No.	Pt-biserial	Dif. estimate	Infit MnSq
7	0.19	-2.02	1.47
12	0.17	-3.27	1.42
15	0.14	-1.91	1.47

The low point-biserial correlations suggest that these three items did not discriminate high- from low-level groups. Their large Infit mean squares have also indicated the unusual response patterns of these items. The difficulty estimates of these items further suggest that they are so easy that they do not separate high ability test takers from low ability ones; they should be removed from the Listening test. Table 4-15b displays the three items.

Table 4-15b:

7. Date of death (of Agatha Christie): _____
12. Making up faces in the Chinese opera requires training and ability: _____ (true or false)
15. A cheerful character has more colours on his/her face: _____ (true or false)

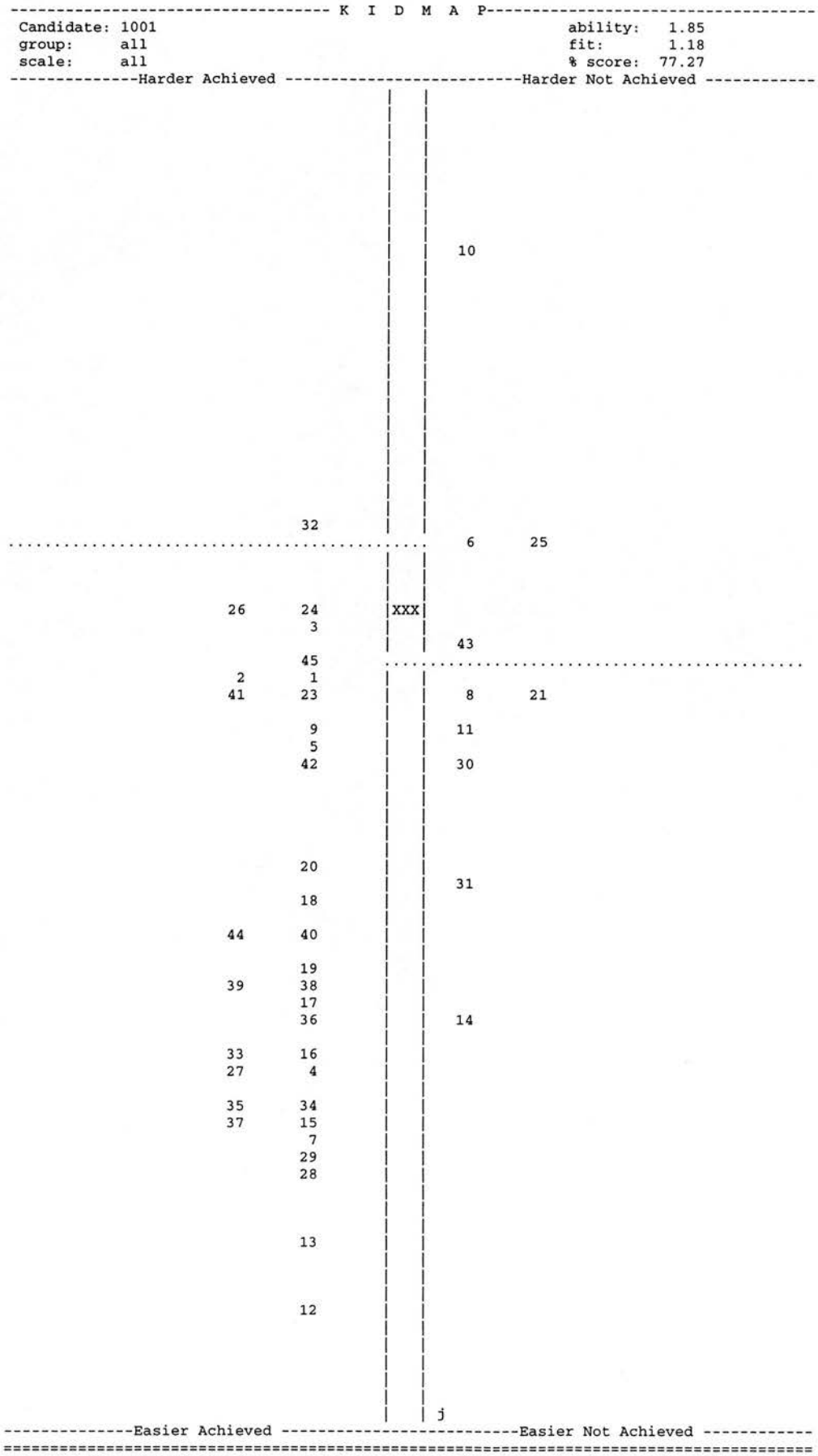
Items 12 and 15 do not seem to require listening to score. Item 7 measures the ability to listen for specific number and is an easy item according to the statistics.

Person ability map

Finally, for each of the candidates, a person ability map (i.e., a Kidmap) is provided. Figure 4-4 displays the kidmap of Candidate 1001 on his performance in the Listening Test as an example.

In Figure 4-4, items are divided into four groups according to the candidate's estimated ability: easy items correctly answered, easy items incorrectly answered, difficulty items correctly answered and difficulty items not achieved by the candidate. The focus should be on those items predicted to be easy for the candidate but who fails to score. The kidmap provides useful information regarding the performance of the candidates in the test and could be used in the classroom for purposes like remedial teaching or as a reference point to assess the student's progress in the future.

Figure 4-4: Kidmap of Candidate 1001 on Listening



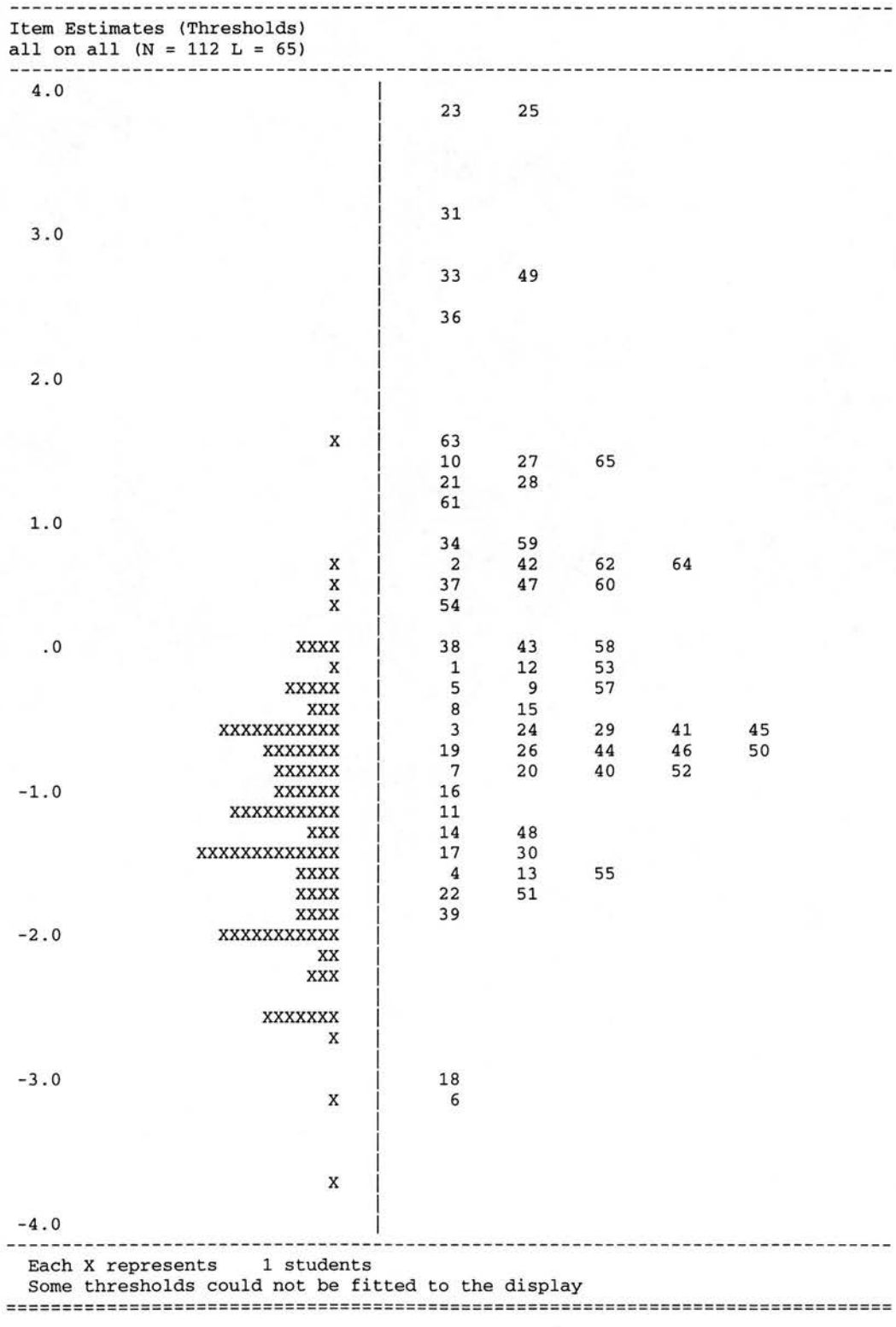
Grammar Test

The Rasch results of the Grammar Test are reported below and include item difficulty estimates, item fit, person ability estimate, person fit, comparison of person ability and item difficult, item analysis and the individual candidate's kidmap.

Item difficulty estimate

Visual representation of Grammar item difficulty estimates is provided in Figure 4-5. Detailed item difficulty estimates are listed in Appendix 4-4. The Grammar item difficulty ranged from +3.80 to –3.17 logits; excluding the top six most difficult items and the bottom two easiest items, the rest fell between +1.48 and –1.80 logits. The mean is set at zero by the analysis and the standard deviation is 1.43. Standard errors ranged from 1.10 to 0.20; excluding the three items with the standard errors of 1.10 and 0.59, the rest fell between 0.36 and 0.20. The reliability of estimate is 0.94.

Figure 4-5: Grammar item difficulty estimates and person ability



Item fit

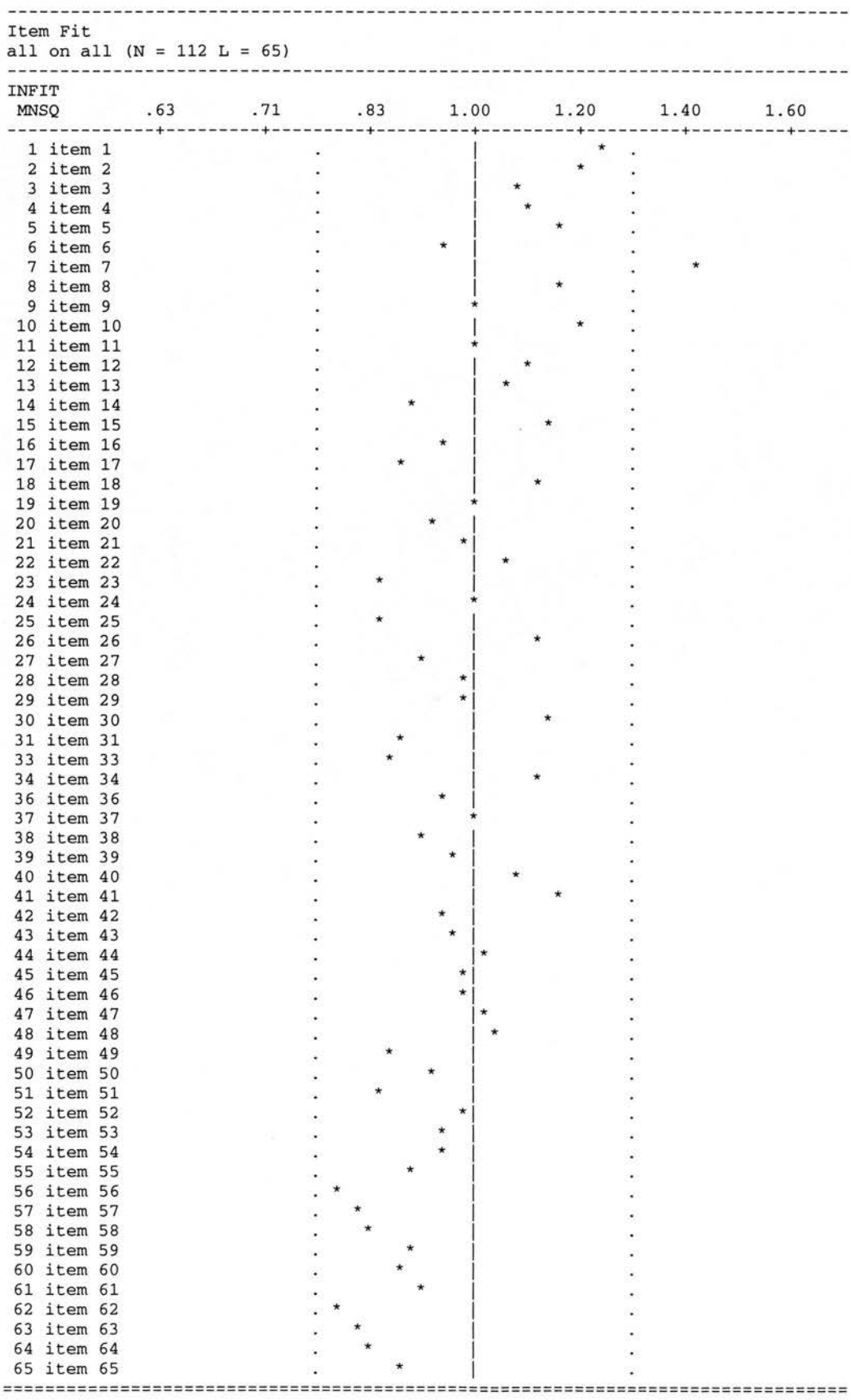
The overall grammar item fit statistics are reported below.

Table 4-16: Item fit statistics of the Grammar Test

Infit Mean Square		Outfit Mean Square	
Mean	0.98	Mean	0.99
SD	0.12	SD	0.41
Infit t		Outfit t	
Mean	0.04	Mean	0.12
SD	1.23	SD	1.47

Overall, Grammar items fitted the modelled expectation. Visual representation of item fit statistics is provided in Figure 4-6. Individual item fit statistics are listed in Appendix 4-4. Except for Item 7, the rest of the Grammar items showed good fit.

Figure 4-6: Grammar item fit



Item 7 has an Infit mean square of 1.42, indicating a 42% variation of the actual responses from the expected responses. In other words, the item has attracted some unexpected responses. According to classical test analysis, Item 7 has a point biserial correlation of –0.12, which suggests that the item has failed to measure the proposed trait and therefore should be discarded. Table 4-17 displays the item and its fit statistics.

Table 4-17: Listening items with negative point biserial correlations

Item No.	Infit MnSq	Outfit MnSq	Infit-t	Outfit-t	Pt. biserial
7	1.42	1.58	5.3	5.0	-0.12
It will be easier to go ___ the bus to get out of the traffic.					
A: by B: in C: on D: of					

47 out of the 112 candidates scored this item. 57 of the 112 candidates chose "A: by," which is considered a possible answer by some native speakers of English and should be revised.

Person ability estimate

The person ability estimates ranged from +1.59 to –3.69 logits, corresponding to raw scores of 50 and 3 points respectively. The mean ability is –1.20 (SD=0.85); in other words, the overall ability level is more than one logit below the expected ability. Individual person ability estimates are provided in Appendix 4-5. The associated standard errors ranged from 0.62 to 0.29; excluding the three extreme low scores, the standard errors ranged from 0.41 to 0.29. The reliability of estimate is 0.85.

Person fit

The overall person fit statistics are given in Table 4-18 below.

Table 4-18: Person fit statistics of Grammar Test

Infit mean Square			Outfit Mean Square		
	Mean	1.01		Mean	0.99
	SD	0.17		SD	0.48
Infit t			Outfit t		
	Mean	0.01		Mean	0.01
	SD	1.01		SD	0.81

The overall candidate response pattern seems to conform to the model-expected pattern.

Individual person fit statistics are summarised in Appendix 4-5. Of the 112 candidates, three persons fell outside the expected Infit t-values of +2 to –2; they are Candidates 1005, 6018 and 7005. Taking outliers into consideration, however, only Candidate 1005 departed from the model-expected ability by 4.89 standard deviations. The fit statistics of the three candidates are given in Table 4-19.

Table 4-19: Misfitting persons for the Grammar Test

Name	Ability estimate	Infit MnSq	Outfit MnSq	Infit-t	Outfit-t
1005	1.5	2.09	4.28	4.18	4.89
6018	-0.72	0.76	0.61	-2.24	-1.26
7005	-0.87	1.35	1.31	2.70	0.95

Table 4-20 lists the response pattern of Candidate 1005. Detailed person response patterns are listed in Appendix 4-8. The items are ranked according to difficulty with the most difficult one on top followed by less difficult items. Items at the bottom are the easiest. Candidate 1005, with an ability estimate at 1.5 logits, scored many of the most difficult items but failed some of the easier items. He is exceptionally good as he attended the American School in Taipei; his failure to score one of the easier items may be caused by boredom or carelessness.

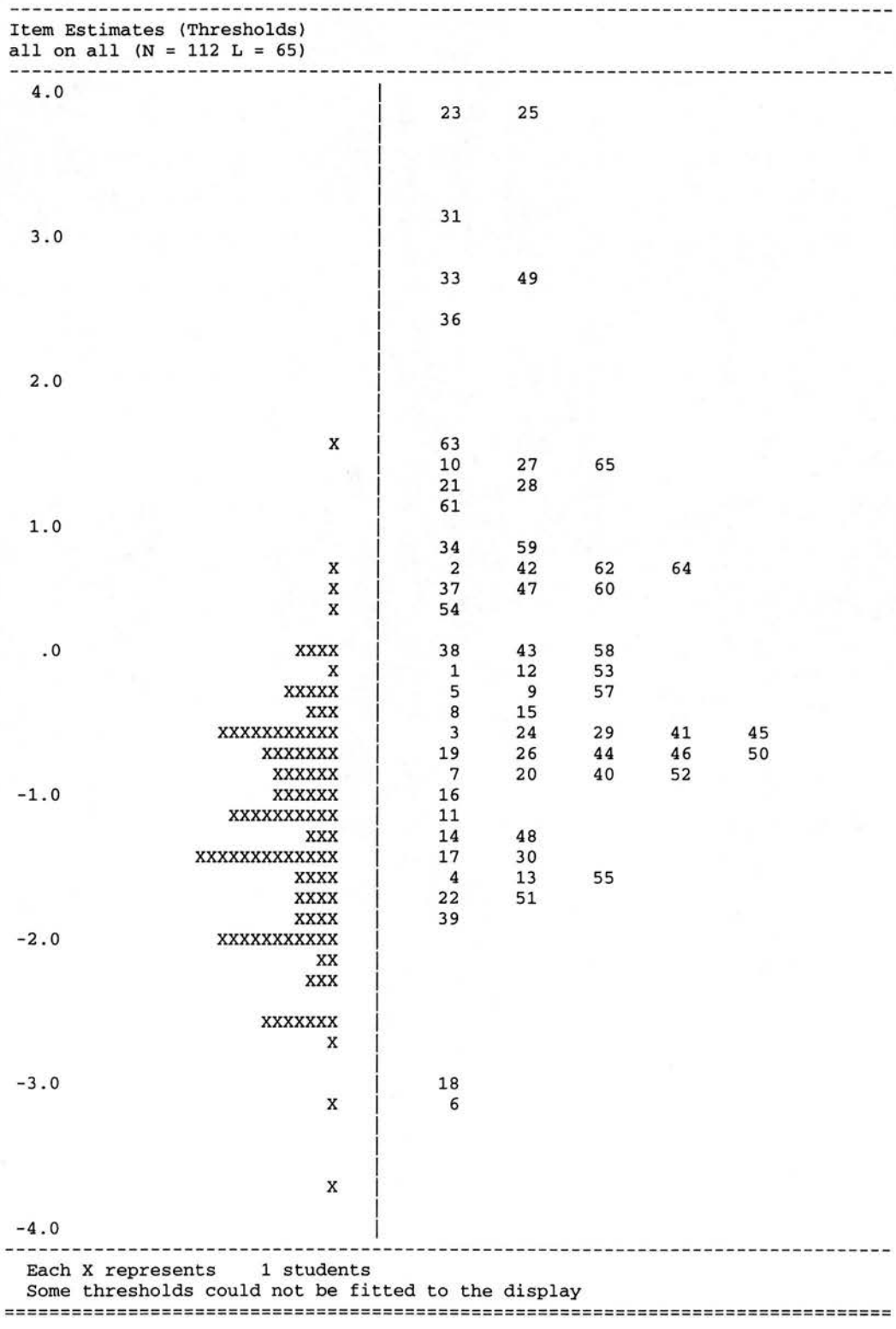
Table 4-20: Response patterns of Candidate 1005 on Grammar Test

Question Number	Candidate 1005
32	0
35	0
23	0
25	0
31	0
33	1
49	0
36	1
63	1
27	1
10	0
65	1
21	1
28	1
61	1
59	1
34	1
64	1
62	1
42	0
2	0
37	0
47	1
60	1
54	1
38	1
43	1
58	1
53	1
1	1
12	1
57	1
5	1
9	0
15	1
8	0
24	1
45	1
41	1
29	0
3	0
50	1
56	1
26	1
46	1
44	1
19	1
52	1
40	1
20	1
7	0
16	1
11	1
48	1
14	1
30	1
17	1
55	1
13	1
4	0
51	1
22	1
39	1
18	1
6	1

The difficulty/ability scale

Figure 4-5 (reproduced on the next page) shows the difficulty/ability scale of the Grammar Test. Candidate ability clusters between 0 to –2.49 logits while most items fall between 1.48 to –1.76 logits, indicating items tend to be more difficult for the ability level of this sample group. Six items are too difficult and two items easy for the candidates. The difficult items are 23, 25, 31, 33, 49,36, and the easy items are Items 6 and 18. In future test revision, these items should either be revised or removed. The ability of all candidates except one is above the item difficulty level when the six difficulty items are excluded. The candidate falling below the difficulty scale of the items is Candidate 1003 who was late and did not do the Grammar Test. In general, candidate abilities fall within the item difficulty estimates.

Figure 4-5: Grammar item difficulty estimates and person ability



Item analysis

Item analysis is provided in Appendix 4-6. This section concentrates on discussions of individual items with poor discrimination power. Table 4-21a lists the statistics of those items and Table 4-21b lists the item stems.

Table 4-21a: Grammar items showing poor discrimination

Item No.	Pt-biserial	Dif. estimate	Infit MnSq
2	-0.01	0.59	1.2
7	-0.12	-0.84	1.42
10	-0.07	1.36	1.19
1	0.04	-0.07	1.24
5	0.15	-0.27	1.16
8	0.16	-0.36	1.16
15	0.18	-3.02	1.12
34	0.16	0.79	1.13
36	0.19	2.37	0.95
41	0.17	-0.45	1.15

The negative point-biserial correlations of the first three items mean that low-ability group scored these items while high-ability group failed them, suggesting that these items did not tap the intended trait. When examining these three items, each item suggests defects not detected earlier. For example, the intended answer for Item 2 is "D:for", however, "C:about" is a more common collocation for the adjective anxious, and thus attracted more candidates. For Items 7 and 10, two answers are possible according to some native speakers of English. These distractors require revision.

The remaining seven items in Table 4-21a indicated a point biserial correlation below the recommended value of 0.2 (Hatch & Lazaraton, 1991). These items were designed to measure the test taker's ability to use different language elements such as the definite article (Item 5), a preposition (Item 41), past participles (Item 15), the passive voice (Items 34 & 36), the second conditional (Item 8), and tense and aspect (Items 1 & 34). Item 36 may be unsuitable in that this structure is not common in spoken texts. Item 1 is difficult because aspect is not an easy concept for the Chinese to fully understand. Moreover, Distractor A:enjoy is considered possible by some native speakers of English. As for Item 8, two answers are considered possible. The

distractors in these two items need revision. The problem with possible answers in some of the MCQ items also shows the difficulty in constructing MCQ questions. The rest of the items do not seem to be particularly problematic according to their fit statistics and expert judgements; they will be retained.

Table 4-21b: Item stems with poor discrimination power

Items with negative point-biserial correlations

2. John seems very anxious _____ the event to be a success.
   A: of B: in C: about D: for

7. It will be easier to go _____ the bus to get out of the traffic.
   A: by B: in C: on D: of

10. The news of the president's retirement _____ be announced soon. No one is sure when.
   A: will B: would C: can D: could

Items with low point-biserial correlations

1. While playing tennis, Alice says, "I _____ the game. I think it's going to suit me."
   A: enjoy B: enjoyed C: am enjoying D: have enjoyed

5. _____ usually have great difficulty in getting a job. They need more help from the government.
   A: Homeless B: Homelessness C: The homeless D: The homelessness

8. I wish the room _____ a bit bigger.
   A: is B: will C: were D: would be

(15. Build) in 1919, the building (call) the Supreme Office and it (use) by the Japanese Governor.....

34. It has been two weeks since anyone saw John.
John _____.

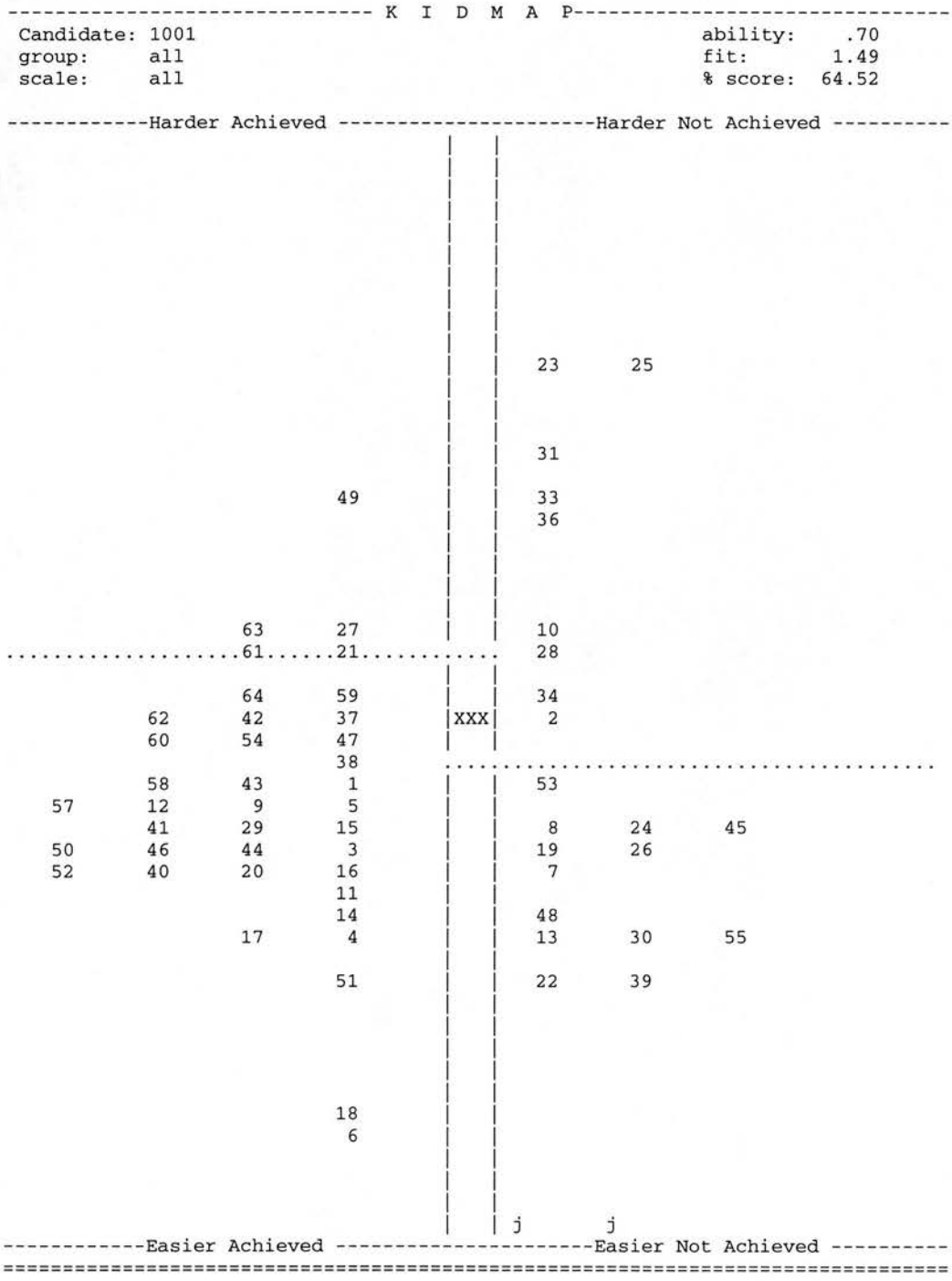
36. This door must be kept closed at all times.
At no time _____.

41. The square shaped fort is built _____ top of a two-storey platform with lookout towers....

Person ability map

Individual kidmaps of the candidates on the Grammar Test are provided. Figure 4-7 shows the performance of Candidate 1001 on the Grammar Test. Items are grouped into four areas according to the candidate's ability: easy items answered correctly, easy items answered incorrectly, difficult items answered correctly and difficult items not achieved by the candidate. Again, one should look at items aimed to be at the candidate's ability level but are answered incorrectly.

Figure 4-7: Kidmap of Candidate 1001 on the Grammar Test



Speaking Test

The analysis of the Speaking Test is discussed in this section. Three types of test results will be presented:

- Candidate performance
- Rater severity and
- The rating scale

Candidate performance reports on candidate oral ability in terms of a logit score; rater severity reports on degrees of harshness in the rating process, and the rating scale reports on how well the rating scale distinguishes between levels. Zero scores were excluded from the estimation as they imply out-of-bounds measures (Section 1.4.).

Candidate performance

Candidate ability ranged from +11.30 to –14.10 logits, corresponding to *S4: Good* and *S0: No Foreign Language Use Ability* respectively. Excluding candidates with zero scores, the ability range fell between +11.30 to –9.40 logits, corresponding to *S4: Good* and *S1: Below Standard.* The ability range is expected to be wide because individual oral ability varies. The average candidate ability is +0.74 logits, indicating the mean candidate ability measure is a little better than the model expected. The standard error is 0.54; excluding those with zero scores, standard errors range from 0.42 – 0.71. The person ability separation index is 7.32; it indicates a good separation among candidate ability measures. The reliability estimate is 0.98. In general, the test as a measurement instrument seems able to separate person ability into different levels.

The candidate measure fit statistics are reported in Table 4-22. The mean square fit statistic is expected to be 1 and the standardised value of the mean square (ZStd) is expected to be near 0. Acceptable mean squares are in the range of 0.7 – 1.3 and of ±3 for ZStd for *Facets*. Interpretation will be based on examination of the mean square because in polytomous data, mean squares are more useful in detecting aberrant response patterns (Linacre, 1997).

Table 4-22: Candidate fit statistics of the Speaking Test

Infit Mean Square		Outfit Mean Square	
Mean	0.7	Mean	0.7
SD	0.9	SD	1.1
Infit ZStd		Outfit ZStd	
Mean	-1.5	Mean	-1.6
SD	2.5	SD	2.5

The mean square of 0.7 indicates 30% less variance in the observed performance pattern than is expected, which indicates better than expected fit to the model and suggests that candidate ability tends to be uniform. However, the lack of performance variation may also be linked to the rating and will be discussed in conjunction with rater severity. The Infit ZStd of –1.5 fell within the acceptable range and suggests the overall internal consistency of candidate performance fitted the modelled expectation.

Fifteen candidates displayed large mean squares indicating variation. Their logit measures and mean square values are given in Table 4-23.

Table 4-23: Speaking Test misfitting persons

Candidate	Measure	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd
1007	1.10	2.8	2	2.6	2
3005	-3.45	2.5	1	2.2	1
3011	5.73	3.3	4	3.0	3
3014	1.84	1.8	1	1.7	1
3026	-9.40	3.8	2	2.1	0
3030	0.84	1.9	5	2.0	5
3032	-9.40	3.8	2	2.1	0
3034	-4.97	1.7	2	1.6	1
3035	-2.19	2.9	2	1.9	0
3044	6.48	2.2	4	2.1	4
3045	0.84	1.9	5	2.0	5
3054	-3.65	2.7	2	2.0	1
4015	1.62	1.5	1	1.6	1
7007	9.13	1.5	1	1.5	1
7010	5.01	2.4	2	2.2	2

These larger than expected mean squares suggest the presence of unmodelled variation. A closer examination of the scores reveals discrepancies between the two ratings, implying that a third rating is required. Table 4-24 lists the ratings of these candidates with their two raters. Individual candidate performance is listed in Appendix 4-9.

Table 4-24: Ratings of the Speaking Test misfitting persons

Candidate	Fluency	Pronunciation	Accuracy&Vocab	Communicative effectiveness	Raters
1007	0 2	0 3	0 2	0 2	JJ KK
3005	2 2	2 1	2 1	2 1	GG KK
3011	4 3	4 2	4 3	3 3	GG KK
3014	3 2	3 2	3 2	3 2	GG KK
3026	0 0	0 0	1 0	0 0	BB LL
3030	2 2	3 3	3 1	2 2	BB LL
3032	1 1	1 0	1 0	1 0	BB LL
3034	1 0	1 0	1 1	1 1	BB LL
3035	2 0	2 1	2 1	2 0	BB LL
3044	2 2	3 2	3 3	3 2	BB LL
3045	2 2	2 3	3 1	2 2	BB LL
3054	0 2	1 2	0 1	1 1	JJ KK
4015	3 3	3 2	3 2	3 3	BB DD
7007	3 3	2 4	2 4	3 3	HH II
7010	2 4	1 4	1 3	2 3	HH II

These discrepancies indicate rater unreliability and a need for more training.

Analysis of rater severity

Rater statistics are reported in Table 4-25. The column "Measure" reports each rater's rating behaviour on the logit scale. Positive values indicate severity and negative values means leniency. The rater reliability estimate is 0.98. The mean Infit MnSq is 0.8, indicating a 20% less than expected variance is observed. Individual rater behaviour is listed under the column "Measure." Raters ranged from 2.29 to –1.53 logits in severity. Rater JJ is the most severe judge and Rater DD, the most lenient. The mean square values ranged from 2.6 to 0.0. Excluding the largest mean square of 2.6, the rest fell between 1.2 to 0.0. According to the model, Rater GG fitted the modelled expectation, which suggests that his ratings were consistent. Raters JJ and HH displayed 20% variation from the modelled expectation, which according to the model, suggests a tendency to use extreme scores.

The low mean squares of Raters KK, DD, II and LL suggest that their ratings agreed with the model so well that there was little additional information available about

their rating behaviour. Their rating is predictable and may be the result of a restricted use of categories. Predictability may be regarded as beneficial according to classical test analysis. In Rasch analysis, however, predictability implies that raters are not supplying much information concerning differences among the candidates.

Rater BB, with a mean square of 2.6, is problematic. The large value suggests a disruptive response pattern. She may have used idiosyncratic response categories. A talk with her revealed that she converted the TG rating scale to the FLPT rating scale, which explained her large un-modelled variance. Further guidance in her rating is required.

On the whole, raters vary in severity. Their rating patterns reveal conservatism (less than 1 MnSq) and extremism (larger than 1.5 MnSq). The presence of four conservative raters may explain the predictability of the overall candidate performance. Further, the number of persons showing variation could be accounted for by the pairing of conservative and radical raters.

The rater fit statistics provide statistical information relevant to future rater training. Conservative raters will be encouraged to use the full range of rating categories. For those who tend to give extreme scores, progression of categories will be highlighted in future training sessions. For this particular test administration, extra training sessions could have been helpful.

Table 4-25: Rater measurement

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	N rater
488	296	1.6	1.61	2.29	.14	1.2	2	1.2	1	1 JJ
572	312	1.8	1.89	.93	.14	1.2	2	1.1	0	2 HH
484	216	2.2	2.14	-.80	.18	1.0	0	0.8	-1	3 GG
304	160	1.9	2.02	.02	.19	2.6	9	3.6	9	4 BB
512	304	1.7	2.12	-.65	.16	0.1	-9	0.1	-9	5 KK
864	416	2.1	2.29	-1.53	.13	0.0	-9	0.0	-9	6 DD
768	320	2.4	2.28	-1.49	.14	0.2	-9	0.2	-9	7 II
224	208	1.1	1.84	1.23	.18	0.5	-6	0.3	-4	8 LL
527.0	279.0	1.9	2.03	.00	.16	0.8	-2.6	0.9	-2.7	Mean (Count: 8)
199.6	75.5	0.4	0.22	1.29	.02	0.8	6.4	1.1	6.1	S.D.

Analysis of the Rating Scale

The rating scale statistics provide information on how well separated each of the rating categories is. The statistics, presented in Table 4-26, are an indication of the validity of the rating scale. "QUALITY CONTROL" provides information on the validity of the rating categories. The column "Avge. Meas." indicates the logit score for a rating assigned to the level (i.e., the category score). The statistics show that each level is well separated by at least 5 logits. The column "MOST PROBABLE" gives the lowest possible logit value at which the category is most probable to be assigned. The values also indicate that each scale point is well separated. Values obtained by the Thurstone approach under "THURSTONE THRESHOLD" further indicates clear separations of the levels. On the basis of the statistics, we might infer that each level seemed to be well understood and applied by the raters.

Table 4-26: Rating scale statistics

DATA				QUALITY CONTROL			STEP		EXPECTATION		MOST	THURSTONE	Cat	Obsd-Expd	Response
Category	Counts		Cum.	Avge	Exp.	OUTFIT	CALIBRATIONS		Measure at		PROBABLE	THRESHOLD	PEAK	Diagnostic	Category
Score	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	at	Prob	Residual	Name
0	88	4%	4%	-10.53	-11.22	.7			(-10.09)		low	low	100%	-4.0	No foreign language ability
1	640	29%	33%	-4.27	-3.94	.5	-9.03	.21	-5.34	-9.01	-9.03	-9.03	94%	-5.0	Below Standard
2	1052	47%	80%	.41	.29	.8	-1.99	.08	.52	-1.97	-1.99	-1.99	86%	6.8	Satisfactory
3	336	15%	95%	4.29	4.24	.9	3.11	.09	5.48	3.09	3.11	3.10	85%	2.8	Competent
4	116	5%	100%	9.18	9.07	.8	7.91	.17	(8.99)	7.92	7.91	7.90	100%	-.6	Good
5															Distinction
									(Mean)		(Modal)		(Median)		

Figure 4-8 illustrates the probability of occurrence of each level. It depicts the probability of a category score (i.e., a level) to be chosen at any ability level represented by the horizontal axis. For example, at the logit scale of +0.0, the probability of getting a 2 is above 50%. The clear peaks and the separation of the scale levels indicate the levels were well separated by the raters. The statistics provide empirical support for the construct validity of the rating scale.

Figure 4-8: Probability Curves of the rating scale

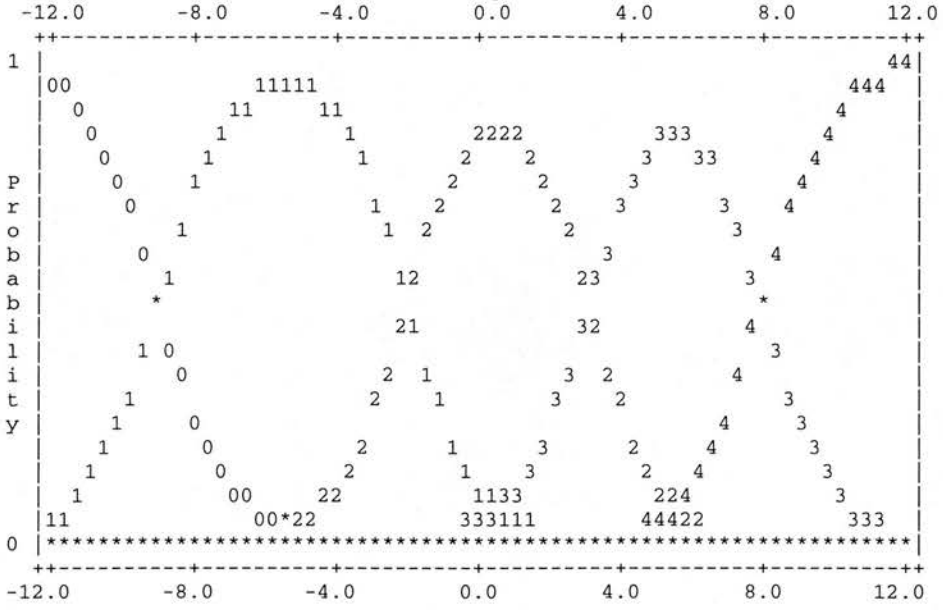
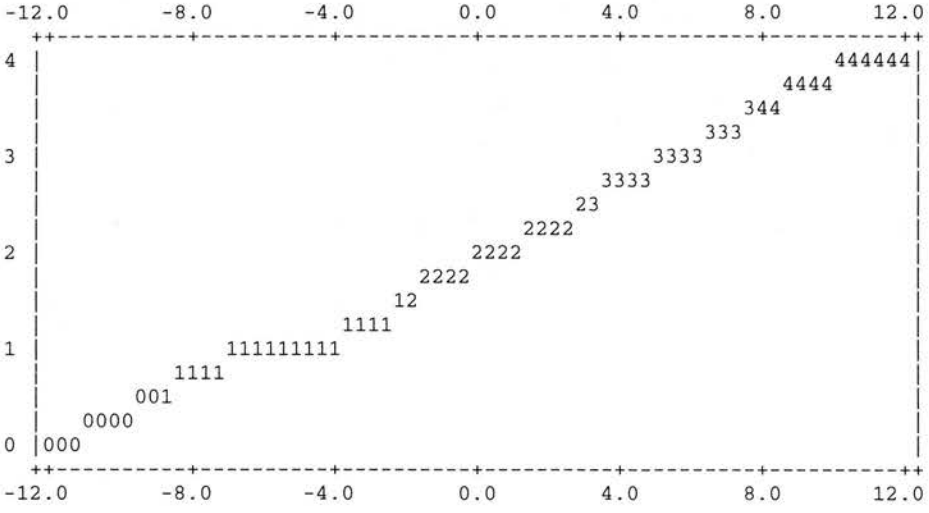


Figure 4-9 shows the expected score ogive. The horizontal axis indicates the ability continuum. The vertical axis displays the category levels of the rating scale. The ogive shows the rating expected for any ability measure relative to the task and rater.

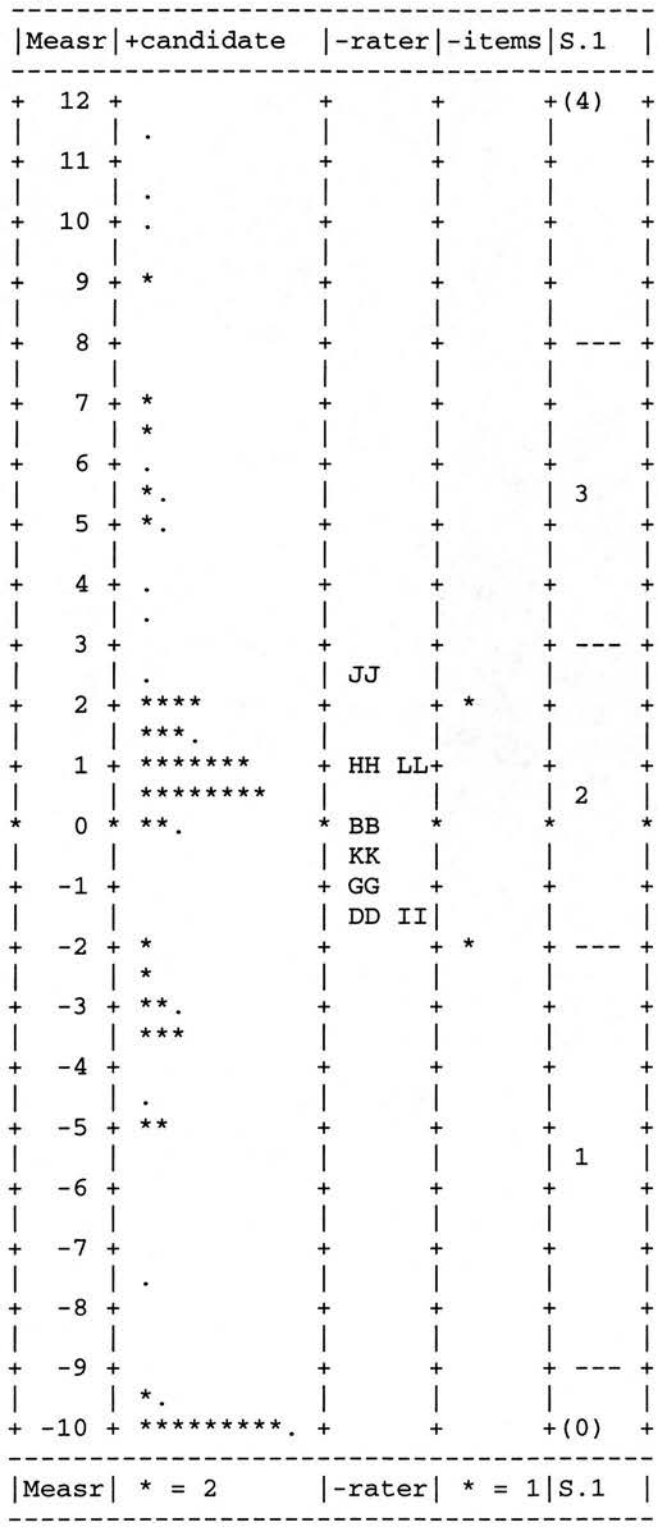
Figure 4-9: Expected score ogive



Finally, Figure 4-10 places the facets measured on a common logit scale. The graph shows the relative standing of the candidates, the raters and the four category levels used in the test on a common logit scale. The column "Measr" refers to the logit scale

ranging from +12 to –10. The candidate ability stretches 22 logit units and the rater severity, 5 logit units. Under the column "items", the two sets of ratings are treated as two distinct items and are separated by a difference of 4 logit units, with the first rating being more severe than the second rating. The difference is very much affected by rater severity because Raters DD and II, the two most lenient raters, were both assigned as the second rater. Finally, the last column is the rating scale with S4 at 12 logits and S0 at –10 logits, indicating the higher the level, the more able the candidate is in terms of the logit score. The modelled level is a little below S2: *Satisfactory*.

Figure 4-10: All Facet Vertical Rulers



4.2.4. Discussion of the Rasch test analysis results

Listening and Grammar Tests

The Rasch results indicate the candidates' ability is a little more than one logit below their modelled ability, which suggests that the Listening and Grammar tests are difficult. Five Listening items showed a large misfit. On closer examination, four of them were identified as easy items and should be discarded because the test takers could have guessed the answers without listening to the passage. The fifth misfitting item was too difficult and should be revised or removed. In terms of item difficulty, four items were difficult for most of the test takers. Revision or removal of these items should be considered. Two items are easy and they should be removed. The reliability estimate is 0.97.

For the Grammar test, three items should be discarded: one misfitting item and two items with negative point-biserial correlations. In terms of item difficulty, six items are too difficult and two items are too easy. The difficult items should be made easier and the two easy items should be discarded. The reliability estimate is 0.94.

In regard to candidate performance in the two tests, five candidates were unable to do any of the items in the Listening test. But all of them were able to do the Grammar test. However, six Grammar items are too difficult for the sample and should be discarded.

Speaking Test

The average person ability was at 0.74 logit, suggesting the overall candidate performance was better than expected. The reliability estimate of 0.98 and the person separation index of 7.32 suggest that candidate ability could well be differentiated by the test and the reliability of the same measurement procedure to candidates from the same population pool is high. The mean MnSq value of 0.7 indicates that candidate ability does not seem to vary much.

The severity of rater varied from 2.29 to –1.53 logtis. Four raters showed conservativism and three raters showed unmodelled variation in their rating. The statistics suggest a need for further rater training. The rating scale statistics indicated

that each level was well-separated, which may suggest the construct validity of the rating scale.

4.3. Test revision

The revision is based on test analysis results and suggestions discussed in Section 4.2., and includes the testing time, the number of items appropriate, the language elements to be assessed, and revision or removal of individual items. The Speaking test, according to the statistics, seems to perform satisfactorily for this test administration; therefore it will not be revised for now.

The following is a summary of the suggestions made in Section 4.2. The revised versions of the Listening and Grammar Tests from the main trial are presented in Appendix 4-10.

Testing Time:

Listening: 40 minutes

Grammar: 45 minutes

Number of items:

Listening: 35 items

Grammar: 40 items

Item revision

In terms of individual item revision, the following items were removed:

Listening: Task 3a,

Grammar: Items 2, 3, 6, 7, 9, 10, 12, 18, 23, 27, 33, 35, 36, 42, 47, and 49.

For Questions 56 – 60, only one response is required.

The following items were revised/replaced:

Listening: Task 2, Task 4, and Task 6b

Grammar: Items 25, and 31.

The following tasks or items were added:

Listening: Item 23

Grammar: Item 18

4.4. Setting the cut-off scores

This section presents discussion of how the cut-off scores were set. Candidate performance in the Listening and Grammar tests was interpreted according to classical test results. This performance was reported in terms of the band levels; raw scores were also reported (Section 3.7.). For this test administration, proficiency level was derived by dividing the raw score range into 6 equal levels. Adjustments such as score standardisation will be made when results of parallel forms are available. The Listening Test had a range of 36 points from 2 to 38 points; the Grammar Test had a range of 50 from 0 to 50 points. Oral ability was based on the oral rating scale. Table 4-27 lists the levels with their equivalent raw score ranges.

Table 4-27: Conversion table of Listening and Grammar raw scores to proficiency levels

Proficiency level	Listening	Grammar
Distinction	36 pts and above	42 pts and above
Good	30 –35	35 – 41
Competent	23 – 29	26 – 34
Satisfactory	16 – 22	17 – 25
Below standard	9 –15	9 – 16
No foreign language ability	0 - 8	0 - 8

For the cut-off score, the minimal acceptable scores of the Listening and Grammar tests had been provisionally set at one SD above the mean raw score and S3 for the Speaking Test. In other words, the suggested cut-off score for Listening is at 24 points ($\bar{X}$=15.78, SD=8.2); 28 points for Grammar Test ($\bar{X}$=19.15, SD=8.78). However, in view of the difficulty of the Grammar test and in order to ensure fairness and allow borderline pass, it was decided to lower the Listening cut-off score by one SEM and two SEMs for the Grammar test. Therefore for the Listening Test, a score of 23 points was considered a pass and a score of 26 points for the Grammar Test was

a pass; both scores are within the band scale of a *Competent* user. For the Speaking test, according to the statistics displayed in Figure 4-10, an appropriate cut-off point is between S2.5 and S3. However, oral ability is considered the most important language use ability for a tour guide; S3 was therefore retained as the minimally acceptable pass level. Table 4-28 lists the minimal scores for each test.

Tabel 4-28: TG cut-off scores

	Listening	Grammar	Speaking
Cut-off raw score	23	26	S3
Equivalent level	Competent	Competent	Competent

Using the suggested cut-off scores, about 11% of the total candidates would pass the test. The pass rate was similar to that of the current pass rate (See Appendix 5-2), which seems to be low. The estimated low pass rate suggests that the test needs to be made easier.

4.5. Summary and conclusion

The main trial TG Test was administered to 112 volunteers. Classical test analysis and Rasch analysis were performed. Results indicate the following:

1. The Listening and Grammar Tests were difficult for the sample of students. The Speaking Test on the other hand was relatively easier. The results suggest that in future test revision, the Listening and Grammar Tests should be made easier. For now, suggestions were made to shorten the tests and revise or remove the statistically unsatisfactory items.
2. Task type seemed to partly account for item performance. For example, tasks such as short-answers, identification and labelling, completing the phrase/sentence, sentence transformation and filling in the missing words had low facility values. Item types such as matching, true/false, MCQs, and verb forms were considered easier. Task type should be considered in future TG test revision and development.
3. Inter-rater reliability and the Rasch rater statistics suggest raters varied in their rating behaviour. Four raters were conservative, two raters tended to award extreme scores and one rater deviated from the guidelines. This

suggests a need for more rater training and guidance. Regarding the rating scale, the Rasch rating scale statistics indicated each level was well separated, which seems to support the construct validity of the scale.

Revision of the Listening and Grammar Tests was discussed. The present testing time of 40 minutes for the Listening and Grammar tests was maintained. But the two tests were shortened by 10 items for the Listening Test (35 instead of 45 items) and 20 items for the Grammar Test (45 instead of 65 items). The cut-off scores were set at 3: *Competent* for each of the sub-tests for this test administration. Further adjustments will be required when statistical information on parallel forms is available. In the next chapter, information gathered in support of the TG test validity and usefulness will be presented and discussed.

Chapter 5: Validation of the Tour Guide English test

5.0. Introduction

This chapter reports on the validation of the TG Test. The framework adopted for the validation is Messick's (1989) progressive matrix (Section 1.2.4.). Test validation is viewed by Messick as one single concept; the different facets are intertwined and not easy to separate, but I shall attempt to disentangle these facets to some extent when presenting my conclusion.

Messick (1989) sees construct validation as a single, central inquiry in practice and theory, but one which has various aspects related to rationale of, evidence for, interpretation and uses of a given test in a specific social context. He proposes a distinction between empirical evidence for construct validation and the consequence of employing the construct, i.e., its *evidential basis* and its *consequential basis.* Similarly, Messick distinguishes construct validation based on analyses of *test interpretation* and *test use*. These four facets form a progressive matrix (Table 5-1) in which efforts to establish the trustworthiness of test interpretation can be made.

Table 5-1: Messick's progressive matrix of construct validation

	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

One implication of this matrix is that meaning and values as well as test interpretation and test use are intertwined in test validation (Messick, 1989). Second, empirically oriented construct interpretation is the integrating power of test interpretation and test use. These implications have been addressed by testing researchers like Bachman (1990). Others have asserted the centrality of construct validation and the importance of its social consequences (Davies, 1990). In this chapter, following Messick's progressive matrix, the four facets will be examined in relation to the use and interpretation of the TG test.

Test validation is a long-term process. Aspects such as construct validity and usefulness may be evaluated within a short time after a test is administered, but aspects such as predictive validity and impact may take a long time to assess. Limited by time, the present study will only examine the aspects outlined in Table 5-2.

Table 5-2: Aspects of TG test validation (after Messick, 1989)

	Test Interpretation	Test Use
Evidential Basis	Construct Validity (CV) • content coverage (i.e., content validity) • internal test structure • relationship between test tasks • consistency of response patterns • relationship with an external criterion	CV + Relevance/Utility (R/U) • appropriateness, • meaningfulness & • fairness
Consequential Basis	CV + Value implications • score consistency and generalisability • gender difference/bias	CV + R/U + VI + Social consequences • practicality • value implications: students' and teachers' reaction to the test

5.1. Evidential bases of the TG test interpretation

At the core of test validation is the examination of the construct underlying the test instruments. Traditionally, construct validity is inferred empirically by way of data gathering and hypothesis testing (Section 1.2.4). Recently, attention has been given to construct-related evidence in support of the test construct, which includes the traditional content validity and criterion-related validity as well as the construct validity. The three form one concept in construct validation and each contributes to the construct validity of the test. Content validity provides judgmental evidence in support of the domain relevance and representativeness of the content of the instruments; expert judgements are required. Criterion-related validity is based on the degree of empirical correlation between test scores and criterion scores. Construct validity is based on an integration of any evidence that provides the interpretation or meaning of the test score. Almost any kind of information about a test can contribute to some understanding of the test construct. Traditionally, the primary emphasis has been placed on the appraisal of the relationships among item scores, between test scores and/or with other measures. Construct validity subsumes content validity and

criterion-related validity. Therefore, to examine the construct validity of the TG tests, the following will be investigated:

(a) content coverage and content representativeness and relevance: expert and student feedback
(b) internal test structure: item fit statistics
(c) relationship between test tasks in the sub-test of Listening and Grammar: factor analysis
(d) consistency of response: person fit statistics
(e) relation with an external criteria: correlation with student final grades and
(f) test reliability

5.1.1. Content coverage and content relevance

The relevance and representativeness of the test content and test tasks are examined in relation to the language use domain of tour guiding. The content of the test was selected on the basis of (1) field observations (2) training materials for would-be tour guides and (3) interviews with practising tour guides (See Chapter 2). Test tasks were later developed accordingly.

To evaluate the relevance of the test content, 5 experts, consisting of 3 university EFL lecturers, 1 tour guide and 1 language tester, checked the following:

(1) test specifications outlining the knowledge and skills relevant to tour guiding obtained from field observations and from textbooks published by the Tourism Bureau for tour guides and
(2) the test instruments

Appendix 5-1 displays their responses. Written comments are listed in Table 5-3. In general, these experts thought items matched with the descriptions of the test specifications, and the test is appropriate and satisfactory as a measuring instrument. The content of the Listening and Grammar tests was selected and modified from (1) the tour guide training texts, (2) from general sources such as EFL teaching materials, brochures etc., or (3) from genuine spoken texts by practising tour guides. Generally, the experts thought the content was relevant. Specific comments on the three tests follow.

Listening Test

The motive for Task 2 was to measure the test taker's ability to listen to and fill in a well-known person's bio-data. Providing biographical information is common when conducting a tour. Originally I decided on a Chinese folk hero but later thought candidates may already have sufficient knowledge of the person, and they may not need to listen to the passage to fill in the important dates in his/her life. Therefore I decided to choose a fairly well known person from a different country, Agatha Christie, so students could also learn about the person. However, one expert did not think Task 2 on *Agatha Christie* appropriate. She suggested an adventure story or a talk on a cultural aspect of a country. This particular comment will be considered in future test revision.

Although one expert suggested MCQs instead of different task types, I wished to keep a variety of test types. One reason is to provide sufficient contextualisation for the test takers (Section 1.3.). The other is to avoid test method effect.

Grammar Test

Experts thought the different parts in the test appropriate. Two considered the test "too difficult" for the students; they suggested shortening the test and restricting the questions to usage only. One expert questioned the validity of Part V as an appropriate task type in the measurement of the candidate's grammatical knowledge. She thought this part was more appropriate as a speaking task. As explained in the grammar test specification (Chapter 3), knowledge of grammar as well as knowledge for appropriate use were the two objectives in the design of the test, and so for the time being I would like to keep Part V. But her comment will be noted in future test development. Concerning the length of the test, I realised there were too many items for the students, particularly in the Grammar Test, and therefore, the number of items was reduced (Section 4.3.).

Speaking Test

All of the experts thought the tasks appropriate. However, one expert thought that for Parts IV and V, some students may have just read the instructions and the context instead of producing their own speech sample. This was also a concern when Part V

was being developed. However, when rating the answer tapes, raters seemed to be able to differentiate those who were reading test instructions from those who were not. For this test administration, this had not been a problem. Regarding Part IV, it seems not possible just to read out an itinerary to imaginary clients because it would sound odd: Students who were not able to do this part were simply silent. But her comment will be considered when developing parallel forms.

Table 5-3: Written comments of the experts

Comments:
1. This is the first time researchers develop the TG test from a linguistic point of view. The test combines language and tour guiding and hopefully will better select people with appropriate language ability.
2. The Listening and Speaking tests are fine but the Grammar test is too difficult for students. I think tour guides will need to be proficient in listening and speaking but grammar is not as important.

In regard to the appropriateness of individual items, the 5 experts have the following comments.

Listening Test

One expert thought Question 12 – 20 (Tasks 3a and 3b) needed revision. She did not think answering these questions would require listening to the entire passage. Task 3a: *True/False* required the candidates to listen to the entire passage before they were able to indicate if the statement heard was true or false. Task 3b: *Matching* required candidates to listen for specific information; students had to listen to the passage on the tape again for the specific information needed to score. I concluded that in order to complete the two tasks, the test takers had to listen to the entire passage attentively; therefore, this comment will be disregarded. However, Task 3a, according to the test statistics, seems too easy and was removed from the revised Listening test (Section 4.3.)

Three experts did not think Questions 22 – 26 (short answers) appropriate as a listening task. One expert considered this task more appropriate for a reading test, but she did not give any reasons nor did she suggest an alternative task type. Short-answers can be difficult but could be made easier by, for example, providing a partially completed general statement for student to complete (Section 4.3.), providing more cues and contexts. As to the appropriateness and legitimacy of short-answers,

an internal relationship between tasks and task performance was examined and suggestions made (See 5.1.2.).

Grammar Test

Experts had reservations about Part III: *Sentence Transformation*. Two of them indicated that some questions were inappropriate; no reasons were given. Classical item analysis indicated that questions in Part III were difficult for this sample. Rasch item fit estimates showed, however, that these questions were within the acceptable range. It may be that both the test takers and the experts were not used to this task type in the test; its use in future test will be considered. For Part IV: *Filling in the missing word*, two experts did not like the choice of some deleted words, but no alternative words were suggested. It seems experts considered Part IV a good grammar task but they differed in the choice of words.

One expert thought Part V: *Giving the appropriate responses* measures oral ability rather than grammatical knowledge. However, as one of the objectives is to measure the ability for appropriate use (See Chapter 3), this task seemed appropriate and will be retained for a second trial. In view of the number of Grammar items and the time given, the number of items in Part III have to be reduced and its appropriateness for the test purpose considered.

Speaking Test

One expert was not sure if the questions in Part III: *Answering questions* were appropriate. But she did not give any reasons, nor did she suggest a more appropriate task. The same expert also thought half of the questions/tasks in the test would be sufficient; again, no reason was given. Another expert pointed out that too much context had been provided in Parts IV and V so that students may have just read the test instructions instead of producing their own language sample. Regarding her comment, I will try to present the context in less complete sentences in future test development.

Overall, the experts had the following comments:

(1) the tests and their content appropriate,

(2) some task types such as short-answers, and giving appropriate responses may not be appropriate for the given test purpose. Some questions in sentence transformation, filling in the missing words and answering questions may need revision and

(3) the number of items in the test battery need reducing and this has been implemented (Section 4.3.).

Student comments

Student comments on the test were also collected. Of the 112 questionnaires distributed, 50 were returned. The questionnaire is presented in Appendix 2-4. Student comments were summarised in Table 5-4.

Table 5-4: Student comments

1. I think the Listening test is:	Difficult	About right	Don't know
	33 (66%)	15 (30%)	2 (4%)
2. I think the Grammar test is:	Difficult	About right	Don't know
	25 (50%)	25 (50%)	---
3. I think the Speaking test is:	Difficult	About right	Don't know
	28 (56%)	21 (42%)	1 (2%)
4. On different task types in the test battery, I like:			
	Yes	No	
Listening – Identification and Labelling	29 (58%)	21 (42%)	
Matching	35 (70%)	15 (30%)	
Information Transfer	23 (46%)	26 (52%)	
True or False	37 (74%)	13 (26%)	
Sentence Completion	25 (50%)	25 (50%)	
Short Answers	31 (62%)	18 (36%)	
Grammar –			
	Yes	No	
Multiple-choice questions	30 (60%)	20 (40%)	
Verb forms	39 (78%)	11 (22%)	
Sentence transformation	31 (62%)	19 (38%)	
Filling in blanks	34 (68%)	16 (32%)	
Completing conversation	33 (66%)	17 (34%)	

Speaking –

	Yes	No
Interpretation & Translation	29 (58%)	21 (42%)
Answering short questions	31 (62%)	19 (38%)
Making an announcement	30 (60%)	20 (40%)
Solving a problem	34 (68%)	16 (32%)

5. On the whole, I like the test.	Yes	No	Don't know
	26 (52%)	12 (24%)	12 (24%)

Written comments:

1. "I think the test is quite suitable for us college students. But since we don't spend as much time as we did in high school. The standard in English of college students is going from bad to worse. So I still think the test is quite difficult for me to do."
2. "If my English is improve, I think I will like the test more and do it well."
3. "I think that these questions are fairly easy. I also think that the amount of questions is just alright."
4. "Time isn't enough to answer so many questions and it's hard for me to listen clearly to all the questions because my English is very poor."
5. "It's a little too hard for me."
6. "I am not getting used to this kind of test because I am slow. I can't response to each question at once. It's nothing to do with English comprehension. It's get to do with your memory."
7. "I am just not in the mood!"
8. "I like the test but it made me feel bad, frustrated."
9. "It's full of challenge. I can understand my English level."
10. "It's too difficult. I have never joined this kind of test before. My spoken English is very poor. I still need to work hard in English."
11. "Before this kind of test, we could read the sample first."
12. "It's so difficult."
13. "Yes, I do like the test but this test requires a little bit professional knowledge. I'm not a tourism major student. I think the only problem is that the speaking test. The time isn't enough for me to answer the questions. Well, maybe I speak too slow or thought too much. So I couldn't manage to answer each question very well. Actually, it's very nice to have this opportunity to take this test. But I think I didn't do it very well. Maybe it is because this is the first time I took the test. I'm not well prepared for this type of test. Just a little bit depressed!"
14. "I like the test. I think the conversation and the usage are all right. I think if the oral test was given by the real people will be better."
15. "Maybe you should offer a clock to let us know about the time. I'm not prepared to be a tour guide so there are a lot of information I don't know. So I think the survey is not useful."
16. "I'm not familiar with those names of places."
17. "I like your test. Best of luck!"

(Comments 18 – 22 are translations of student comments.)

18. The time given for the Grammar test is short and there are far too many proper nouns in the test.
19. The time for the Grammar test is short and there is a lot to read and write.
20. The time for the Grammar test is short. I can't think thoroughly before I answer the questions. x 2
21. I don't know if I like the test. I don't think my English is bad but the test is a bit difficult for me.
22. I hope you will publish this test so more people will know about such test type.

In general, students thought the tests difficult in terms of the questions, the time given and the number of items as well as their language ability. As stated under test purpose in Chapter 3, the test taker's spontaneous reaction to the language input is part of the assessment. Therefore, unlike in other types of tests in which more than enough time is usually given, students were not given plenty of time to think over their answers. Their overall language proficiency may be another reason, and the tendency to avoid failure may also play a part in their performance. A third reason may be their unfamiliarity with the test format and the type of language use measured in the test. Even so, over half of them liked the tests and the task types.

Overall, student comments seem to accord with expert comments regarding the number of items. The entire test lasts for about 2.5 hours. To give the test taker more time to do the test, one solution would be to lengthen the test; the other would be to reduce the tasks and items. Lengthening the test may cause test fatigue and affect test results; reducing the tasks and items was preferred.

Conclusion

Overall the test seemed to be broadly acceptable to the expert judges and students. Some items and tasks need revision; tasks such as short-answers and giving appropriate responses need further studies to establish their validity as a task type in this particular test. The TG test content in general seemed representative of the target language area and the tasks appropriate. The overall test length should remain approximately the same but specific tests may need shortening to allow students more time for the questions.

Some expert comments such as the use of MCQs, reduction of the Speaking tasks, the choice of listening texts, and removal of some test tasks, were disregarded. The comments were invaluable in themselves; yet, each expert seemed to have examined the testing instruments from a slightly different point of view than the one the language constructor had in mind. This suggests that experts may not have necessarily understood the test rationales. The language developer has to consider the extent of the usefulness of these comments.

5.1.2. Internal test structure

Content validity provides judgmental support of the domain relevance and representativeness of the test; it may not present sufficient evidence to sustain inferences made from the test score. Empirical evidence has to be investigated as well. The first piece of empirical evidence to be examined is the item fit, which concerns the validity of the items working together as one variable in the measurement of the underlying trait (Section 1.4.). Table 5-5 lists the mean fit statistics of the TG sub-tests.

Table 5-5: Item fit statistics of the TG sub-tests

	Listening	Grammar	Speaking (Rating scale)
Infit Mean Square	0.99 (SD = 0.19)	0.98 (SD = 0.12)	1.1 (SD= 0.6)
Outfit Mean Square	0.96 (SD = 0.46)	0.99 (SD = 0.41)	1.3 (SD= 0.8)
Infit *t/ZStd*	-0.01 (SD= 1.40)	0.04 (SD = 1.23)	-0.9 (SD= 8.1)
Outfit *t/ZStd*	-0.01 (SD = 1.27)	0.12 (SD = 1.47)	-0.7 (SD= 8.3)
Reliability Estimate	0.97	0.94	1.00

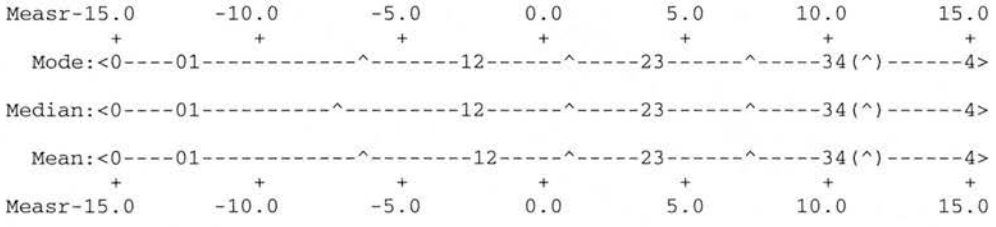
When the data fit the modelled expectation perfectly, the mean square value is expected to be one and the *t*-value is expected to be near zero. Overall, items (ratings in case of the Speaking test) in the sub-tests fitted the model's expectation, indicating that items in each of the tests acted as a cohesive set and measured one single trait. Individual item statistics indicated that apart from the five misfitting Listening items, one misfitting Grammar item and two poor discriminating items (Section 4.2.), the other items were within the acceptable range of fit and displayed satisfactory discrimination power. Thus, we have evidence in favour of the construct validity of the sub-tests.

The construct validity of the rating scale is examined as well. Figure 5-1 presents the rating scale structure bar chart. The "mode" indicates the probable starting and end points of a category level along the measurement scale starting from –12 logis to +12 logits. ^ indicates the location of probability peak of a category level. For example the peak probability of level 1: *Below standard* occurs at about –5 logits.

The "median" indicates the point in which there is a 50% probability of being rated below or above a category. ^ indicates the mid point of each zone. For example at –2 logits, the probability of being awarded a 1 or 2 is at 50%.

The "mean" indicates the expected category score along the logit scale. ^ indicates the location where the category score is the expected score. For example, at a little above 0 logits is the expected score of 2: *Satisfactory*. The bar chart indicates that each level has been well separated, suggesting that raters seemed to have understood the categories and have applied them with clarity. Overall, the data support the validity of the rating scale.

Figure 5-1: Rating scale structure bar chart



5.1.3. Relationship between test tasks

The construct validity can further be established by examining the relationship between test tasks by means of factor analysis. Factor analysis was carried out with the use of SPSS 8.0. Maximum-likelihood was the method used for extracting factors and oblique rotation was used. The reasons for the choice of the two specific methods are:

(1) The maximum-likelihood method is a common factor model. It hypothesises a minimum number of common factors necessary to reproduce the observed correlation. The model does not assume a g-factor and usually starts with one common factor (Kim & Mueller, 1978).

(2) A test of significance has to be carried out before factor extraction to examine sample adequacy. Such procedure helps later result interpretation (Fruchter, 1954).

(3) Oblique rotation is used because the variables (i.e., the tasks) are assumed to be correlated.

Tables 5-6a to 5-6b display the results. Table 5-6a shows that two factors were extracted for the Listening test. All tasks were loaded heavily on Factor 1. Their loadings on Factor 2 were negative in sign, indicating Factor 2 was far removed from Factor 1, which suggests that the Listening tasks were highly related to one factor only. In other words, the tasks shared one trait. A closer examination of the Listening tasks suggests that although listening ability had been operationalised in terms of the different tasks, a particular trait of language knowledge plus whatever process is unique to listening seemed to underlie these tasks and affected test taker performance. Factor 1 was named listening ability. The finding that all tasks were related to the factor of listening ability could support the construct validity of the Listening test.

With regard to the appropriateness of Task 4: *short-answers* as a listening task (Section 5.1.1.), the result indicates a loading of 0.768 to Factor 1, suggesting short-answers may be an appropriate and a valid listening task.

In Table 5-6b, two factors are extracted for the Grammar test. Part 2: *Verb forms* and Part 3(Gra3): *Sentence Transformation* are related to Factor 2 whereas Part 4(Gra4): *Filling in the missing words* and Part 5 (Gras 5a and 5b): *Giving the correct responses* are represented by Factor 1. This shows that the grammar tasks can be represented by two factors. Successful completion of Parts 2 and 3 required the use of explicit grammatical knowledge. However in order to score Parts 4 and 5, different language ability was required. Students may need not only linguistic knowledge but also rules and conventions for an appropriate response in the given context. In other words, their ability to interpret and create meaning in a given context was required. Therefore, Factor 1 seemed to be related to ability to use language appropriately, and Factor 2, explicit grammatical knowledge. This finding is in accord with the objectives listed in the Grammar test specifications (Section 3.4.) and provides evidence for the construct validity of the test. Regarding the appropriateness of Part V as a TG grammar task, the expert's comment on its inappropriateness could be disregarded.

Task 1: *multiple choice* was more strongly related to Factor 2 than Factor 1. However, the relationship was still a weak one. Two explanations are possible; first,

the weak relationship may be attributable to the fact that some of the MCQs were not well constructed; second, MCQs could be related to factors other than the two factors extracted. More investigation is needed. For the time being, the results seem to indicate that two traits were measured in the Grammar test: the test taker's grammatical knowledge and his/her ability for appropriate use, which is in accord with the test rationales specified in Section 3.4.

In terms of the weak relationship of MCQs with either one of the factors, it seems this task type could be omitted in future Grammar test development. The finding also seems to indirectly support the decision to include a variety of task types in the measurement of the constructs proposed (See Chapter 3).

To sum up, the findings of factor analysis seem to indicate

(1) the different listening tasks were all related to one factor which is the listening ability and
(2) grammar tasks were represented by two factors; one tapping grammatical knowledge and the other measuring ability for appropriate language use.

The results accord with the Listening and Grammar test rationales discussed in Chapter 3. The findings suggest that the two tests measured the traits described in the test specifications and therefore provide further evidence for the construct validity of the two tests.

Table 5-6a: Factor loadings on the Listening test

Structure Matrix

	Factor	
	1	2
LIST1	.406	-.273
LIST2	.657	-.296
LIST3A	.305	-.233
LIST3B	.631	-.583
LIST4	.768	-.439
LIST5A	.523	-.452
LIST5B	.692	-.477
LIST6A	.469	-.755
LIST6B	.804	-.859

Extraction Method: Maximum Likelihood.
Rotation Method: Oblimin with Kaiser Normalization.

Table 5-6b: Factor loadings on the Grammar test

Structure Matrix

	Factor	
	1	2
GRA1	.190	.373
GRA2	.408	.789
GRA3	.464	.605
GRA4	.684	.443
GRA5A	.891	.384
GRA5B	.590	.471

Extraction Method: Maximum Likelihood.
Rotation Method: Oblimin with Kaiser Normalization.

5.1.4. Consistency of response

The number of misfitting persons is another indication of the validity of a measurement instrument. A large number of person misfits indicates that the test may not be a valid instrument in the measurement of the proposed trait. The person fit statistics of the sub-tests are displayed in Table 5-7. The mean square values are expected to be one and the *t*-values are expected to be zero. Overall, the observed response pattern of the candidates fitted the expected response pattern, providing some evidence for construct validity. Specific person ability analyses (Section 4.2.3.) showed three test takers in the Listening and one in Grammar test had unmodelled response patterns. Their deviation from the model may have been caused by factors

other than the test itself. The small number of misfitting persons provides further support for the construct validity of the tests.

With regard to the Speaking test, the person separation index of 7.32 indicates the test can distinguish increments of the candidate ability about 7 times wider than the error in the ability scale, resulting in about 10 ability strata among the candidates. The ability stratum is calculated as:{(4x Separation +1)/3}. The mean square of 0.7 indicates that the overall person measures fitted the modelled expectation. The reliability estimate is about 0.98. The results seem to support the construct validity of the Speaking test.

Table 5-7: Person fit statistics of the TG test battery

	Listening	Grammar	Speaking
Infit Mean Square	1.00 (SD = 0.23)	1.01 (SD = 0.17)	0.7 (SD= 0.9)
Outfit Mean Square	0.96 (SD = 0.62)	0.99 (SD = 0.48)	0.7 (SD= 1.1)
Infit *t*	0.01 (SD = 1.12)	0.01 (SD = 1.01)	-1.5 (SD=2.5)
Outfit *t*	0.12 (SD = 0.83)	0.01 (SD = 0.81)	-1.6 (SD= 2.5)
Reliability Estimate	0.9	0.85	0.98
Separation index	---	---	7.32

5.1.5. Score relationship with external criteria

The final source of evidence about the TG's construct validity is its relationship with an established external criterion. Two sources are available: student TOEFL scores and their end-of-semester scores. Table 5-8 displays the correlation between the TOEFL scores and the TG sub-tests. Correlations between the end-of-semester scores and the TG scores are presented in Table 5-9. These results should be interpreted with caution for the following reasons (Also see Sections 1.2.9 & 1.3.5.):

(1) TOEFL and TG have different test purposes. TOEFL is an EAP test; its purpose is to establish if foreign students have sufficient language ability to cope academically in a U.S. college/university. The TG test is to measure if the test taker has sufficient language ability as a tour guide. The content coverage of the two tests is different.

(2) The language skills measured are different. The TOEFL consists of three objective sub-tests: Listening, Structure and Reading. Scores are reported

in terms of a numerical value. The Test of Written English (TWE) and Test of Spoken English (TSE) are optional though there is a tendency to include TWE in every TOEFL administration. TSE is administered separately. The TG test consists of three sub-tests of Listening, Grammar and Speaking tests, all of which are compulsory. Language ability is reported in terms of a band scale. In this present study, correlation was only possible with the TOEFL total score.

(3) The test method is different. The TOEFL uses MCQs whereas the TG test requires the test takers to complete different tasks.

(4) Regarding the use of end-of semester scores, there is no uniform consensus amongst schools on the levels of student performance. An A (90%+) from a given school does not necessarily equate to an A from another school. Further, end-of-term English scores are indicators of student achievement rather than the student's language proficiency.

Table 5-8: Correlation of TOEFL and TG sub-tests

	Listening	Speaking	Grammar	TOEFL
Listening	____	0.62	0.54	0.67
Speaking		____	0.55	0.79
Grammar			____	0.84
TOEFL				____

Spearman; N= TOEFL: 20; TG: 112 ; $p< 0.01$

Table 5-9: Correlation of end-of-semester score and TG sub-tests (Spearman)

	Listening	Speaking	Grammar	Grades
Listening	____	0.62	0.54	0.54
Speaking		____	0.55	0.05
Grammar			____	0.31
Grades				____

Spearman; N= School grade: 69; TG: 112; $p<0.01$

The results show TOEFL scores correlate moderately with the three sub-tests, indicating a marked relationship between the two tests. Given the different nature of the two tests, these results seem to suggest the two tests share some property but they are essentially different tests. However, given the number of TOEFL scores available (i.e., 20 students), the results should be interpreted with caution.

Correlation of the school grades with the TG sub-tests indicates a very low relationship (0.05) between the Speaking test and the grades. Listening correlates moderately with the school grades and Grammar correlates weakly with the final grade. The results seem to indicate the final grade does not seem to be a good predictor of the TG test performance. However, the results might be expected given present university classroom practice. Listening to the target language is provided during language lab sessions, and so listening ability is developed marginally; but the teaching of the oral skill is largely ignored in the classroom, and there is no explicit input in grammar. Instead, emphasis is placed on the development of reading and writing skills (Teng, 1994); these two skills are not measured in the TG test.

In sum, empirical results seem to indicate a moderate relationship between the TOEFL and the TG tests. The relationship of the TG tests with student grades seems to suggest a moderate relationship between the Listening test and the final grade but a low relationship with the Grammar test. The relationship with the Speaking test is so small that it can be ignored. These results seem to be attributable to current university classroom practice in that listening skill has been marginally emphasised; development of oral skill has been largely ignored. Grammar is not explicitly taught. However, for the reasons discussed, these results should be interpreted with caution.

5.2. Evidential bases of the TG test use

This section examines three issues concerning test use: (1) the appropriateness, (2) the fairness, and (3) the meaningfulness of the test.

5.2.1. Test appropriateness

Expert judgements and student evaluations on the appropriateness and fairness of the tests are reported. Five experts were asked to judge the appropriateness of the TG test as a measuring instrument for tour guides. All of them thought the TG test was appropriate and did a satisfactory job (Appendix 5-1 Qs15 & 16). But the experts had some comments on the appropriateness of specific test tasks such as short answers and providing appropriate responses (Section 5.1.1).

5.2.2. Test fairness

The fairness of the test is also examined. Test fairness in this project refers to the candidate's familiarity with the testing process and test content/tasks/item bias. Student feedback is examined and their comments are displayed in Table 5-10. Over half of the students (54%) thought the test was fair. 42% of the students indicated they did not know. The length of the test may have affected their reaction because most of the students were asked to fill in the questionnaire immediately after the test. But it is possible that these students may have had different criteria for fairness as suggested by Comments 3 and 6 in which students commented on the inclusion of specialist knowledge and skills other than language ability.

Table 5-10: Student comments on the fairness of the TG test

If you were to be a tour guide, do you think the test is fair?	Yes	No	Don't know
	27 (54%)	2 (4%)	21 (42%)

Comments:

1. "I think it may be useful for the government to test the English ability of tour guides."
2. "It's very hard to say. A tour guide must have very good command of English, but he doesn't have to know much grammar. The listening and speaking skills are more important."
3. "I don't know. I think that being a good guide, one must have a good idea about the place where he is guiding at. He must also have better communication skills, not just the usages of English speaking, writing and listening."
4. "Yes but after these test, I am not going to be a tour guide definitely."
5. "It's a fair test. Your English must be good otherwise you can't finish this test."
6. "I don't know. I think a tour guide should face and imagine all kinds of situation he or she would meet."

In general, student comments seem to suggest the following:

(1) The TG test was necessary and useful.

(2) Aural and oral skills were more important than a good display of grammatical knowledge.

(3) Specialist knowledge and communicative skills were also essential for a successful tour guide.

(4) A wider variety of contexts and situations should have been provided in the test.

Comment (3) suggests that some students may prefer a work sample approach to specific purpose testing (Section 1.2.8.). Provided there are sufficiently available

human resources and the cost is within acceptable range, Comment (3) is worth considering in the future.

Content bias

In regard to content and task bias, the tour guide informant and the raters were consulted before the pilot test and the main trial. None of them thought either the content or the tasks would show bias toward either gender or any other grouop. In addition, no specialist knowledge would be required to score the items.

Gender difference

To examine if there was a significant gender difference in test performance, an independent t-test was performed. The results, displayed in Table 5-11, indicated the difference between the mean scores was not statistically significant.

Table 5-11: Results of mean score difference between genders

	t value	df	*p*
Listening	1.29	109	ns
Grammar	-1.91	109	ns
Speaking*	z = -0.66	109	ns

*Mann-Whitney Test; $p<0.05$

Test method

Regarding the TG test method, the experts and raters preferred the assessment of separate skills and the inclusion of different tasks. Initially, the tour guide informant, Rater GG, did not like the test method; he preferred MCQs because they are economical and quick. In addition, newly qualified tour guides would undergo some training, during which time their language skills would be polished. Therefore, he did not think it necessary to change the present FLPT test. But during test development and test revision, he changed his mind. He preferred the new test for the following four reasons (personal communication and my own translation):

(1) The language use domain of tour guiding was reflected in the test.

(2) The inclusion of subject experts in the test development helps them better understand what is involved in language testing.

(3) The inclusion of different task types may seem to measure the test taker's language ability more thoroughly.

(4) The type of language use and content focused in the test gives the candidate some superficial taste of what tour guiding involves.

Rater GG has been involved in the administration of the Taiwanese Tour Guides Association. He is also the editor of the Tour Guide Journal published by the Association. With permission, Rater GG discussed the test battery in the staff meeting and published parts of the test battery to obtain more comments from other practising tour guides. It seems that some had reservations about the new test form because it had not been approved by the Tourism Bureau; others would like to try the new test as it seemed to have better face and content validity (Personal communication).

5.2.3. Test meaningfulness

The meaningfulness of the TG test is judged in terms of (1) how authentic the item context is, (2) how well matched the items/tasks to the test specifications, and (3) how well candidates and test users understand the nature of this type of test. (1) is interpreted in terms of degree of contextualisation judged by the examination of the test specifications and the test content. The purpose of the TG test is to enable the test user to draw inferences about the language use ability of potential tour guides. The concern was not candidate behaviour in one context but candidates' ability to use the knowledge skill consistently and appropriately in varied contexts. To achieve this, TG test specifications listed the item types and knowledge domains of interest to be assessed.

The tour guide informant was asked to review the test specifications and test instruments. He thought (1) the contexts provided were varied and reflected the language use domain of tour guiding and (2) the language skills had been sufficiently described. The kind of knowledge required to perform the grammar test successfully may be more than what a tour guide needs in real language use contexts. But the tour guide informant thought grammatical knowledge underlies good language performance. Therefore, the grammar tasks adequately measured the test taker's ability to use the language system.

In regard to (2) *how well matched test items/tasks to the descriptions of the test specifications*, the response of the five experts are listed in Table 5-12.

Table 5-12: Expert judgements on test items and test specifications

1. The Listening items fit the descriptions of the Listening test specifications.	Yes	No	Don't know
	5	0	0
2. The Speaking tasks fit the descriptions of the Speaking test specifications.	Yes	No	Don't know
	5	0	0
3. The Grammar items fit the descriptions of the Grammar test specifications.	Yes	No	Don't know
	4	0	1

Except for one judge who was not sure if the Grammar items fitted the Grammar test descriptions, the rest seemed to be positive that there was a match. In general, feedback from the five experts and the tour guide informant seemed to support the meaningfulness of the item context.

With respect to (3) *how well candidates and test users understand the nature of such test*, reactions from the experts, teachers and students to the test will be reported. In general, the 8 raters reacted to the tests positively. Some voiced concerns about the scoring procedure but agreed that such form of assessment may be more effective because students were forced to produce the target language. Initially Rater GG did not seem to understand the nature of the test. But once he became involved in test development, he seemed to understand the test better and learned to appreciate it.

Student opinions were divided (See Table 5-4). To minimise student anxiety and make them understand the nature of the test, a test syllabus with sample items had been given to the participating students a week before the test. In addition, their English teachers were given a copy of the test in case students had any queries. Interviews with some students afterwards suggested that while a large number of students knew the type of test they would do, some failed to read the test syllabus and anticipated a multiple-choice test. It seems those who read the test syllabus understood the purpose and the nature of the test.

To conclude this section, the ESP TG test seems an appropriate measuring instrument for the selection of tour-guides. Expert judgements suggest the test did not seem to be

biased toward any group of test takers. Statistical studies also suggest no significant gender difference in test performance for the main trial. Finally, regarding the meaningfulness of the TG tests, item contexts seemed authentic and meaningful and most students understood the nature and purpose of the test.

5.3. Consequential bases of the TG test interpretation

In a specific purpose test of language competence, issues such as replicability and generalisability have to be accounted for. The two refer to the interpretability of the test score and the consistency of score interpretation (Frederiksen et al., 1990; Linn et. al, 1991). Two issues will be investigated: (1) score stability and (2) score generalisability.

5.3.1. Score stability

Reliability is examined. Reliability involves quantifying the consistencies of test scores (Thompson & Vacha-Hasse, 2000). Since only one test administration was possible for the TG test, reliability of internal consistency was estimated and is reported below. α coefficients are reported for the Listening and Grammar Tests; inter-rater reliability and candidate performance are reported for the Speaking Test.

Table 5-13: Reliability coefficients of the TG sub-tests

Test	Listening	Grammar	Speaking
Method	α coefficient	α coefficient	Inter-rater reliability/ candidate performance
Reliability coefficient	0.91	0.87	0.77 – 0.78/--
IRT-based reliability coefficient	0.97	0.94	0.98/0.98 (Person separation index: 7.32)

From the point of view of replication, the α estimates of the Listening and Grammar tests suggest a strong relationship between the present test form and the hypothetical parallel form drawn randomly from a pool of similar tasks/items. But the marginal inter-rater reliability suggests a need for further rater training.

5.3.2. Score generalisability

To ensure score generalisability, test tasks were designed and developed according to the following guidelines suggested by educational researchers:

(1) to increase the number of performance tests for each student or to increase the number of tasks in each test (Linn, et. al., 1991:19)

(2) the use of task specifications so that at least transfer across topics within a domain is possible (Messick, 1994:15)

(3) a mix of structured items and open-ended tasks so different response forms are imposed onto the candidate (Messick, 1994:15) and

(4) the use of a matrix-sampling design with different performance tasks administered to different samples (Linn, et. al., 1991:19)

Except for (4), the rest were integrated in the development of the TG tests. Table 5-14 displays the number of tasks the types of items in the test battery.

Table 5-14: Number of TG tests and task types

No. of Tests (3)	Listening	Grammar	Speaking
No. of tasks	9	5	5
Task specifications	Yes	Yes	Yes
Item types	Structured and open-ended items	Structured and open-ended items	Structured and open-ended items

The TG test comprises three sub-tests; each requires performance in a particular language use dimension, namely, listening, speaking and the ability to use grammar accurately and appropriately. Second, each sub-test consists of at least five different tasks in different language use contexts to tap the construct proposed. Language functions to be measured and the purpose of the tasks have been pre-specified. Empirical evidence seems to support the claim (Section 5.1.3.).

To examine if the TG test is generalisable, expert judgements on the number of tasks in each sub-test and their appropriateness were collected (Appendix 5-1). Responses regarding the appropriateness of tasks are displayed in Table 5-15.

Table 5-15: Expert judgements on task appropriateness

7. The types of Listening tasks are	Appropriate 5	Not Appropriate 0	Don't know 0	for this test.
Comments: *(1) Not sure about the topic in Task 2. Perhaps an interesting adventure story or something about the culture of a country.* *(2) Multiple-choice questions are better.*				
8. The types of Speaking tasks are	Appropriate 5	Not Appropriate 0	Don't know 0	for this test.
Comments:				
9. The types of Grammar tasks are	Appropriate 3	Not Appropriate 1	Don't know 1	for this test.
Comments: *(1) Some parts are too hard, e.g., Part III* *(2) A little too hard.*				
10. Generally, the number of Listening tasks are listening ability of would-be tour guides.	Many 2	About right 3	Few 0	for measuring the
Comments:				
11. Generally, the number of Speaking tasks are speaking ability of would-be tour guides.	Many 1	About right 4	Few 0	for measuring the
Comments: *(1) Half of the questions are enough.*				
12. Generally, the number of Grammar tasks are grammatical knowledge of would-be tour guides.	Many 1	About right 4	Few 0	for measuring the
Comments:				

In terms of task appropriateness, all of the five experts considered the Listening and Speaking test types appropriate. Three of the experts thought the grammar tasks appropriate; one did not think them appropriate and one did not know; no reasons were given. In terms of the number of tasks, three thought the number of Listening tasks about right and four thought the number of Grammar and Speaking tasks about right.

To conclude, the following types of evidence seems to indicate the TG scores were generalisable and meaningful:

(1) The classical reliability coefficients and the Rasch estimate of reliability indicated high equivalence reliability for the TG Listening and Grammar tests.

(2) Guidelines to ensure score generalisability were used. Results of factor analysis suggested that the Listening and Grammar tests display internal

structure consistency. Rasch results showed a person separation index of 7.32 for the Speaking Test, indicating the test was capable of separating ability.

(3) Expert feedback seemed to suggest that task types were broadly appropriate but the number of tasks may be reduced.

5.4. Consequential bases of the TG test use

The final facet for the appraisal of the TG test validity is the examination of its consequences of use. Consequences of test interpretation and test use refer to the appropriateness and adequacy of the inferences and decisions made on the basis of the test score. Both long-term and short-term effects should be examined. However, since this test is only a trial, long-term effects of test use are not likely to appear. Therefore, I am only able to investigate short-term effects, which include the practicality of the test, test impact on individuals and value implications of the test use, i.e., students' and teachers' reaction to the test in relation to foreign language learning and teaching.

5.4.1. Test practicality

The cost and efficiency of the test will be examined. Efficiency refers to how easy the test administration is and cost refers to the total expenditure on the test. Table 5-16 lists the total cost and number of people involved in the test development.

Table 5-16: TG total expenditure

Total budget (in US$): ----				
Total cost (in US $): 21,982.00				
Human resources:				
	No. of people	Rate (US$)	Est. w/h	Est. cost (US$)
Test developers				
1. field researchers	2	10.00	150 hrs.	3,000.00
2. test writers	3	20.00	120 hrs.	7,200.00
Raters	8	20.00	58 hrs.	9,280.00
Monitors	2	8.00	12 hrs.	192.00
Technicians	2	8.00	12 hrs.	192.00
Expert informants	3	17.00	20 hrs.	1,020.00
EFL teachers to record test tapes	2	17.00	4 hrs.	136.00
Recording technician	1	8.00	4 hrs.	32.00
Clerks	2	8.00	20 hrs.	320.00
Material recourses:				
	Quantity	Rate (US$)		Cost (US$)
Language lab	4	10.00 p/h	3.5 hrs.	140 .00
Test materials	400 test booklets			210.00
	150 Blank tapes			165.00
Others				
1. Postage				25.00
2. Tea/biscuits etc.				70.00

Human resources made up most of the test expense, as is always the case when subjective scoring is involved. For this test administration, extra expenses had been incurred on field research and the consultation with expert informants, thus, increasing the total expense in about US$4,020.00. The average number of registrations has been about 350 people for the TG English test (Appendix 5-2). Each test taker would have had to pay at least about US$65.00, which is a fair test fee. Thus, in terms of cost, the test would be acceptable to test users.

In terms of test efficiency, the test was fairly easy to administer. Personnel involved in the test administration included two monitors and one lab technician per session. The monitors were there to observe any irregularity, help answer any questions the test taker may have had before the test; they also helped collect and distribute test booklets. The technician checked the machines before the Listening and the Speaking tests to make sure every test taker could hear the questions and record their answers.

The ease of test administration was partly due to the small number of applicants as well. If the test did not require complicated administration procedures and a large number of administrative staff, it should be acceptable for the testing institution.

Marking of the TG Listening and Grammar tests was less efficient in comparison with machine-scored tests. Raters took approximately 20 – 30 minutes to evaluate one Speaking answer tape. With regard to the marking of the two objective tests, it is hoped that by reducing the number of test tasks and items during test revision, marking would be made easier.

5.4.2. Test impact

Test impact on gender is examined. The main purpose of the new TG test was to assess the test taker's communicative language ability as a tour guide. Task types had been designed to highlight the language knowledge and communicative demands essential in conducting a tour. Results of the independent samples t-test indicated no significant gender difference in test performance. The results are listed in Table 5-17, reproduced from Table 5-11.

Table 5-17: Results of t-test on gender difference in test performance

	t value	df	*p*
Listening	1.29	109	ns
Grammar	-1.91	109	ns
Speaking*	z = -0.66	109	ns

*Mann-Whitney Test

5.4.3. Value implications of the TG Test

Value implications of the test refer to the students' and teachers' reactions to the test in relation to language learning and teaching. A thorough examination of the changes a test might bring about in teaching requires long-term empirical research (Khaniya, 1990; Wall & Alderson, 1993). Limited by time and the nature of the test being a trial, the following were investigated:

(1) the test takers' reflections on the test tasks and feedback on their language learning behaviour and

(2) the teachers' awareness of the TG test purpose and the types of test tasks in relation to classroom activities.

Hughes (1989) suggests a list of criteria that would help achieve beneficial washback effects. For the TG test, according to expert feedback, the following criteria have been met:

(a) Testing the abilities that the language testers want to encourage: the TG test attempts to assess the listening and oral skills, and therefore indirectly encourages the learning and teaching of the two skills in the classroom.

(b) Testing the skills the language testers are interested in: on the basis of field observations and interviews with the tour guide informants, listening and speaking are considered the two more essential skills required of a successful tour guide. Thus, the TG test is only interested in language use ability in the sub-dimensions of listening, speaking. The Grammar test is added to explicitly measure the test taker's ability for accurate and appropriate use of the grammar and to enhance the face validity.

(c) Making the test criterion-referenced: the language use domain required in tour guiding has been highlighted in the test specifications, and tasks are developed accordingly. In addition, language use ability in terms of a proficiency scale has been provided as a yardstick.

(d) Making sure the test is known and understood by students and teachers: sample test items are given to both the participating students and their teachers in advance. Further, teachers have been briefed on the purpose of the trial. Student feedback indicates most students read through the list and understood what is to be measured in the test.

A review session was arranged a week after the test. The purpose was to (1) go over the questions with the students and (2) to examine why students experience difficulty while their teachers and the experts did not think the test particularly difficult. Table 5-18 lists their feedback.

Table 5-18: Student feedback on their test performance

Listening:
1. Know the word but not the pronunciation.
2. I pronounce the word differently from the native speaker.
3. Know each of the words but don't understand the meaning of the phrase/sentence, e.g., *apart from, call at,* etc.
4. Don't know the words – only hear a string of foreign sounds.
5. Read too slowly, not able to answer the questions in the time given.

Grammar:
1. The Grammar test is not difficult but I don't know why I couldn't finish the test.
2. Grammar questions are more about everyday use. It is a bit different from the grammar I studied in school.
3. Part IV – know the story, and know which part of speech is missing but can't get the exact word.
4. I spent too much time on each question.

Speaking:
1. Generally questions are quite easy. I understand the questions in Part III but don't know how to reply.
2. I worry about grammar when answering the questions.

Their feedback on the Speaking test seems to explain the predictability of their performance (Section 4.2.). In general, student comments seem to point to the following three areas of inadequacy:

1. Insufficient vocabulary
2. Insufficient language knowledge in discourse and
3. Inadequate interaction skills

The TG sub-tests are also given to the six English teachers of the 112 students for review. Table 5-19 lists four of the teachers' oral comments.

Table 5-19: Response from 4 of college EFL teachers

- The items and tasks are realistic and related to every day use.
- It is appropriate to give students different task types so they know how well they are able to use the language.
- Some questions are difficult. But the test gives students some feel of how real language is used.
- I think we can teach the test and the tasks in the classroom.
- It seems a good test.

In general, the teachers' comments seem to indicate they understood that the TG test is measuring communicative language use ability and felt that test tasks accord with the test purpose. Comment (3) suggests that some teachers appreciate authentic language use. Comment (4) seems to indicate that this teacher thinks communicative competence is the focus of foreign language learning and teaching. However, for the

time being, it is not possible to know exactly what is taking place in the EFL classroom.

If language learning/teaching is for communication, these comments have the following implications in foreign language learning and teaching in the classroom:

(1) A direct approach to learning and teaching a foreign language, with communicative language teaching (CLT) as a general broad guideline rather than a teaching paradigm, should be implemented.

(2) A detailed description of what communicative competence entails should be provided so that the components relevant can be used as a content base in syllabus design. Classroom activities can then be developed for each of the selected language areas.

(3) Pedagogic tasks with a systematic focus on form have to be the organisational units in such a syllabus. The notion of focus on form should have a focus on both the linguistic codes and grammatical regularities and the higher level organizational principles/rules/patterns or conventions governing language use beyond the sentence level (Celce-Murcia, 1990). Further, focus on form should also include lexical formulaic phrases as some research on L1 has indicated that native-speakers store a large amount of chunks in their language repertoire which L2 learners lack (Nattinger & DeCarrico, 1992; Long & Robinson, 1998).

(4) Teachers can use *macrostrategies* as proposed by Kumaravadivelu (1994) as a framework to maximise learning opportunities, facilitate negotiated interaction, and foster language awareness.

These are only suggestions for classroom teaching on the basis of the comments collected. Empirical studies will have to be carried out to examine what is actually happening in the classroom.

5.5. Summary and conclusion

One purpose in validating the TG test is to ensure its usefulness as a measuring instrument and its usefulness to the test users so that appropriate inferences can be

made regarding a would-be tour guide's language ability. In this chapter, following Messick's progressive matrix (1989), different facets contributing to the validity of the new TG test have been examined. They include the evidential bases of the TG test interpretation (i.e., the construct validity) and test use (i.e., its relevance and utility), and the consequential bases of the TG test interpretation (i.e., the implications of the TG test) and test use (i.e., the impact of the TG test).

According to expert judgements, the TG test content and the tasks were relevant and appropriate for the purpose of the test. Statistical evidence collected from the Rasch item estimates indicates that items were working together cohesively, suggesting the homogeneity of the items. Results of factor analysis indicate one factor (i.e., listening ability) representing the listening tasks and two (i.e., grammatical knowledge and appropriate language use) for the grammar tasks. These results seem to accord with the rationales proposed in the Listening and Grammar test specifications, thus providing support for the construct validity of the test battery. In terms of concurrent validity of the TG test, correlational studies with the TOEFL suggest a moderate relationship between the two tests, which indicates that although the two tests share some of the traits being measured, they are still essentially two distinct kinds of tests. Results of correlations with the final grade indicate a moderate relationship with the Listening test, a weak relationship with the Grammar test and a very small relationship with the Speaking test. These results seem to reflect current university classroom teaching in that development of oral ability is largely ignored, listening ability is marginally taught, and grammar is not taught in the classroom. It is however also likely that the final university grade may not be an appropriate external criterion for the TG test, as it is indicative of student achievement whereas the TG test is a measuring instrument of language proficiency.

Test relevance and utility is examined as well. Expert and student feedback suggested that (1) they understood the purpose of this test, (2) despite the difficulty of the sub-tests, the test battery was an appropriate measuring instrument, and (3) test content and test tasks did not seem to bias against any group of candidates. These comments indicate the TG test seems an appropriate, meaningful and fair test. Results of a t-test further indicated no significant gender difference in the test performance, providing empirical evidence for the TG test's fairness regarding gender difference.

In terms of score consistency and generalisability, Cronbach's α values of 0.91 and 0.87 were obtained for the Listening and Grammar tests respectively, and suggested that the internal item structure seems to be stable and homogenous. As regards test replication, the correlation coefficients of 0.91 and 0.87 suggest a strong relationship between the present test forms with the hypothetical equivalent forms of the two tests. The inter-rater reliability of 0.77 - 0.78 for the Speaking test suggests significant variation in rating behaviour, and more rigorous rater training is needed.

The final facet of test validation examined the practicality of, and the teachers' and students' immediate reactions to, the test. Feedback from the administrative personnel suggests the test is fairly easy to administer. The overall expense of US$21,982.00 for test development and test administration seems to be acceptable (US$20,690.00 for test development and US$1,482.00 for test administration). Overall, the TG test seems to be practical. Regarding the impact of the TG test to the test takers, the t-test results indicate no significant gender difference in test performance.

Student comments suggest they thought the TG tests difficult and long. Some felt frustrated, but more than half of the students liked the test and thought it useful, which seemed to suggest the test was broadly acceptable to them, in spite of its difficulty. Teachers interviewed thought the test measured communicative language use. Student reflection on their test performance suggested they had to improve their vocabulary, pragmatics and interaction skills. The implications for teaching are that teachers should use a direct approach to language learning and teaching with the communicative language teaching as a broad guideline. In the classroom, teachers will have to maximise learning opportunities, facilitate negotiated interaction and foster language awareness.

To conclude, the evidence seems to indicate the TG test is empirically valid and broadly acceptable to the test users. It has not been possible to investigate the predictive validity and the washback effects the test may have on language learning and teaching. Therefore, I do not know how well this TG test predicts the language

use ability of those who pass the test and work as a tour guide. For now, on the basis of evidence available, this TG test seems a useful measuring instrument. The next step would be to develop parallel forms and implement the test on a national basis so that a full-scale validation would be possible.

Chapter 6: Conclusions and future directions

The two aims of this project were to (1) develop a specific purpose TG test of Listening, Speaking and Grammar sub-tests and (2) validate the test battery on a trial basis. I shall conclude the thesis by summarising the rationales behind the TG test design, stages of test development, and commenting briefly on test results presented in Chapter 4. I shall then suggest areas in which I think improvement is necessary. Finally, I shall briefly discuss future directions of specific purpose testing in the Taiwanese foreign language use contexts.

6.1. Design and development of the TG Test

The purpose of the specific purpose TG English test was to measure the test taker's English language use ability to work as a tour guide. The current existing TG Test, a general proficiency test, takes a structuralist/psychometric approach to item selection and is not adequate for the following reasons:

(1) The present TG test design does not relate the test taker's language knowledge to his/her ability to use the language.
(2) The general proficiency approach measures undefined language use ability. However, tour guiding constitutes a specific language use domain, which should be highlighted.
(3) The test items do not seem to provide meaning, that is items lack authenticity and
(4) The test does not seem to engage the test taker in the areas of language use necessary for successful performance as a tour guide.

A specific purpose approach, on the other hand, measures language ability in specific language use contexts, and will make the interpretation of candidates' test performance more meaningful in relation to performance in the target language use domain. In order to improve the TG test, the language construct has to be defined according to a model of language use so that language use required in the workplace is theoretically justified and the effect of interaction between language ability and the context explained. On the basis of the test rationales, the characteristics of domain

language use tasks can then be translated into test tasks, and the processes required for successful performance specified.

To best measure the candidate's language use ability as a tour guide, it was therefore decided to design an ESP TG test based on a framework of communicative language use (Bachman, 1990); tour guiding was the target language use domain to be assessed. Three language use dimensions were assessed: listening, speaking and the ability to use grammar accurately, meaningfully and appropriately.

This TG test was intended to engage the test takers authentically in test tasks that demonstrably related to areas of language use in tour guiding. The candidate's language knowledge, his/her world knowledge and strategic competence were part of the assessment. Language ability was to be inferred from the test taker's performance across different contexts and tasks. Strategic competence was not assessed but was considered the mediating factor between the candidate's language performance and the language use contexts. No specialist knowledge in tour guiding was required, but some topical/background knowledge relevant to the interpretation of the task and the context was required and would be called upon in the formulation of an appropriate response. Performance was then evaluated against a rating scale.

Based on a needs analysis, listening and speaking skills were considered to be the two abilities primarily required in tour guiding. The research review further indicated the complexity of the two skills; therefore, each skill seemed to deserve separate assessment (Chapter 3). From the communicative perspective, the valid assessment of grammar seems to include not only knowledge of the language system but knowledge of its meaning and values in relation to the context and the ability to use the system appropriately as well. Moreover, the inclusion of a grammar test would enhance the face validity. Therefore, a grammar test was included.

Content representativeness relates to task authenticity, which was a major concern in the TG test design. To ensure the content was sufficiently representative, a needs analysis was conducted and expert informants were consulted throughout the stages of test design and development.

6.2. Validation of the TG Test

Research question 1: Is the new TG Test empirically valid?

This question was examined in terms of content coverage and relevance, score reliability, model-fitting, relationship among the test tasks, and relationship with the student's TOEFL scores and with their end-of-year grades.

Content coverage and relevance

Expert judgements indicated that the content and the tasks were sufficiently relevant. Student feedback suggested, however, that specialist knowledge and a wider range of communicative skills should have been included.

Score reliability

Reliability coefficients of the Listening and Grammar tests (0.91 and 0.87 respectively) indicated high internal item consistency, which suggested the items were homogenous and were likely to correlate highly with a hypothetical parallel test. The inter-rater reliability of 0.77 – 0.78 indicated an unwelcome degree of variability among the raters, suggesting that additional rater training was needed.

Model-fitting

Item fit statistics of the Listening and Grammar tests showed that the variance of the listening and grammar items were within the acceptable range of 0.75 – 1.3, suggesting these items were working together cohesively. Person fit statistics indicated three persons in the Listening test and one person in the Grammar test fell outside the acceptable range of 0.75 – 1.3. The numbers, less than 2% of the sample population, indicated the variance of these persons' test performance might have been caused by factors other than the testing instruments such as fatigue or boredom. However, the majority of the response pattern seemed to form a coherent set. The person statistics of the two tests provided support for the construct validity of the two tests.

The person separation index of 7.32 suggested that the Speaking test was able to separate ability levels seven times wider than the measurement error, an indication of

the appropriatness of the test as a measuring instrument of oral ability. The rating scale was examined as well. Rasch results showed that each level was well separated by at least 5 logits, and suggested a clear separation of the different levels. The results implied that the raters seemed to have understood the descriptions of the levels and applied them with some clarity. The results were suggestive of the validity of the rating scale.

In terms of task relationships, a factor analysis was performed on the listening and grammar tasks. Results showed that the listening tasks were represented by one factor, interpreted as listening ability; and two factors, interpreted as grammatical knowledge and ability to use the knowledge appropriately, represented the grammar tasks. The results seemed to accord with the number of traits hypothesised in the test specifications (i.e., listening ability in the Listening test and grammatical knowledge and the ability to use it appropriately in the Grammar test), and gave further indication for the construct validity of the Listening and Grammar tests.

Correlational studies with two external criteria (i.e., student TOEFL scores and final grades) showed a moderate relationship with the TOEFL scores, indicating the two tests shared some trait property but they were essentially different tests. Comparison with student final grade showed a moderate relationship with the Listening Test, a weak relationship with the Grammar Test, and a very small relationship with the Speaking test. The results reflected the present university teaching practice but they may also suggest that the final grade may not be a good indicator of language use ability in tour guiding. However, these results should be interpreted with caution because of the different nature of the TOEFL and the end-of-year university grade (Section 5.1.5.). The results may also indicate that the two external criteria are not relevant indicators of the validity of the new TG test.

To conclude, classical and Rasch item stability estimates showed high internal consistency of the Listening and Grammar items. Statistical evidence on task relationship and person performance further suggested the two tests worked cohesively, indicative of the construct validity of the Listening and Grammar Tests. Rasch results of the Speaking test showed a wide person separation index and a high reliability estimate of 0.98, suggestive of the appropriateness of the Speaking test as a

measuring instrument of oral ability. Further, the clear separation of the rating scale levels suggested that the raters seemed to have understood the descriptions in each level and applied them well enough.

Finally, correlational analysis indicated a moderate relationship with the TOEFL but a moderate to small relationship with students' end-of-year grade. These results were considered satisfactory since the TG test is essentially different from the TOEFL in test design and test purpose and the final grade is indicative of student progress and achievement.

Research question 2a: Is the new TG test appropriate, fair and meaningful?

Appropriateness of the TG test as a measuring instrument

According to expert judgements, the test was appropriate and did a satisfactory job. However, some suggested that specific tasks such as short-answers (in the Listening test) and giving appropriate responses (in the Grammar test) may not be appropriate. That was not supported by the factor analysis which showed these two sets of tasks correlated well with the factors extracted, indicating they were able to measure the constructs proposed.

Student feedback suggested the three tests were in general difficult for this sample. They were not able to finish the Grammar test within the time allotment. Item analysis and expert judgements on individual items also suggested that the Listening and Grammar tests needed shortening and some of the items needed revision. Further research is necessary to establish the appropriate number of items and tasks in the new forms of the test.

Fairness of the TG test

Test fairness was examined in terms of the test takers' familiarity with the testing process. Gender bias was also examined; results of an independent samples t-test indicated no significant difference in the mean score between male and female students, suggesting this particular test does not bias against any sex. Student comments indicated over half of them considered the test fair. For those who were not sure of the fairness of the TG test, some suggested the assessment of specialist

knowledge and communicative skills and others indicated that the test should have included a wider variety of contexts and situations. These two comments indicate that test takers may have different criteria of test fairness from those of the language tester's.

Regarding the test takers' familiarity with the testing process, those who read the test syllabus understood the testing procedures. However, the procedures were explained again in Chinese before the test and time was given for questions.

With regard to the test method, expert judgements indicated that the inclusion of different task types forced students to produce language samples in different contexts and reduced the test method effect.

Meaningfulness of the TG test

The meaningfulness of the TG test was examined in terms of how realistic the item context was, how well matched the items were to the test specifications, and how well understood was the test purpose.

The tour guide informant was asked to comment on the item context. He thought the contexts were varied and reflected the target language use domain. Five experts (three university EFL lecturers, one language tester and one tour guide informant) examined the items against the test specifications. Except for one judge who was uncertain if Grammar items fitted the test specification, the rest thought there was a match between all the items with their respective test specifications. In general, the test seemed to match the descriptions outlined in the test specifications. Concerning the test takers' and the test users' understanding of the test purpose and the nature of the test, comments collected from the eight raters, the five experts and the students suggested the following:

(1) Most students were aware that the purpose of the test was to assess communicative language use ability in tour guiding; some students thought the test should have included the assessment of communicative skills, specialist knowledge and more contexts.

(2) EFL teachers considered that the test was communicative and the language authentic.

(3) The eight raters considered such a test would measure the test takers' language ability better because they were forced to produce language samples in different contexts.

Overall, students' and teachers' comments seemed to suggest they understood the purpose of the test.

Research question 2b: How practical is the TG test? What are the test's value implications of the use of the test (i.e., on foreign language learning and teaching)?

Practicality was evaluated in terms of cost and the efficiency of the TG test administration. The value implications of the test refer to the reactions of the students and teachers regarding the learning and teaching of English in the classroom.

The total expenditure (test development and test administration included) was US$21,982.00, of which US$20,500 went to test development and US$1,482.00 was spent on test administration. An average of 350 test takers would have to pay at least US$65.00 each, which would seem a fair test fee.

With regard to test efficiency, feedback from personnel involved suggested that the test was fairly easy to administer. Marking of the Listening and the Grammar tests was less easy than the current TG Listening and Use & Usage tests which are machine scored. However, once the number of items and tasks is reduced, marking of the new TG test may be less time-consuming.

The test takers' reaction to the test and their reflection on foreign language learning behaviour suggested the following areas of weakness:

(1) Insufficient vocabulary knowledge: Two areas of weaknesses were identified (a) no knowledge of a given word, e.g., students did not know the word they heard or read, and (b) incomplete or incorrect word knowledge, e.g., they misinterpreted or could not make sense of the context, or, they knew a given word but pronounced it differently from the native-speakers

(2) Insufficient pragmatic knowledge: The students were not able to relate a given structure to its appropriate use, e.g., Part IV of the Grammar test in which students were asked to reply appropriately according to the cue.

(3) Lack of interaction skills: Students heard an utterance and understood it but were not sure how to reply in an appropriate manner.

Teachers' comments suggested that overall they thought the test was communicative and the tasks authentic, suggesting they accepted that the test was measuring communicative language use ability. However, there is no way of knowing how their understanding of communicative foreign language learning is related to actual classroom teaching, which is a point worth investigating in the future.

Lacking in this project is a study of predictive validity and the washback effects, if any, on language learning and teaching. At the moment these are not researchable because this TG test was developed on a trial basis only. However, since the test is generally acceptable, implementation on a national basis will be suggested, and this will allow investigation of the washback effects and the test's predictive validity.

To conclude this section, classical test results of the trial TG test suggested the Listening and Grammar tests were difficult for this sample of students but the reliability estimates of the two tests indicated high internal consistency. Item fit statistics further showed the test items seemed to work cohesively; the small number of misfitting persons suggested that the overall candidate response patterns seemed coherent, implying that the tests would be workable measurement instruments for a large number of test takers. Finally, comments collected from the students and the experts showed that generally, the test was acceptable to them.

6.3. Areas for improvement

Although the TG test battery is broadly acceptable to the students and experts, and empirically valid, the following areas need improvement:

(1) The number and nature of test tasks: Further study is necessary to establish the appropriate number of tasks and the types of tasks most suitable for the

purpose of the test. In addition, alternative test methods such as computerised testing or life interview in the case of the Speaking test could be considered if resources are available.

(2) The rating scale: The rating scale was modified from the ESU Framework (Carroll & West, 1989) and was checked against recorded samples obtained from field observations. Rasch results suggest the validity of the rating scale. However, an empirically derived scale seems more appropriate. The next step will be to examine how well matched the descriptions in the rating scale are to criteria set in the workplace.

(3) Rater monitoring and regular rater training: Rater statistics showed some raters deviated from the guidelines suggested in the rating scale and caused variance. Although Rasch analysis can well separate rater error from candidate performance, constant rater monitoring and regular training is still necessary to ensure the quality and consistency of the rating.

(4) Revision of malfunctioning items: Some items, in particular MCQs, were poorly constructed and need further revision for future use. The number of poor items highlights the importance of quality control and team work in item construction.

(5) A committee consisting of language testers, EFL teachers and subject experts could be formed to supervise future test development, test revision and test validation.

In addition, the revised test should be piloted on a larger number of test takers again so that standardisation is possible before implementation on a national basis. Finally, parallel forms should be developed and piloted.

In terms of the expected outcomes discussed in Section 2.5., the following needs discussion:

(1) Collaboration between language testers and expert informants: Collaboration between the language tester, EFL teachers and the tour guide was central to the content representativeness and construct validity of the ESP TG test. However, language experts may have different opinions on how the test should be constructed and the appropriateness of certain tasks; their views may not be entirely valid. The test constructor

has to decide on the extent s/he would accept the comments on the basis of other considerations such as the theoretical justifications of the test design, statistical information obtained from the test, the relevance and acceptability of the comments, and economy/practicality.

(2) Constraints of this TG test: Three major difficulties were encountered in the course of the test development. One concerned convincing the experts of the possibility of a specific purpose TG test; the second concerned the test content and the assessment criteria; the third related to the attempt to validate the test in relation to a similar and well established external criterion.

The language experts reacted to this type of test more positively and quickly; however, it took the field expert, i.e., the tour guide some time before he thought positively of the test (Section 5.2.). The process was an educational experience for both the language tester and the expert informant.

Limited by the somewhat complex nature of interaction and performance in tour guiding, the tasks designed were restricted in range. More study is necessary to expand possible task types. The development of the assessment criteria was also a difficult task. For this administration, descriptors in the ESU Framework (Carroll & West, 1989) were adapted and modified. Empirical research is required on the generalisability of the descriptors (Section 3.5.).

The third difficulty was to establish the TG Test's relationship with external criteria. Validation of a performance-oriented test with an external criterion has not been found easy (Section 1.2.4.); nevertheless it is part of the validity inquiry. For this test administration, TOEFL scores and the students' end-of-year university scores were used as the reference point. However, the two did not seem appropriate for reasons discussed in Section 5.1.5.

(3) Students' awareness in their learning behaviour: Students experienced frustration, which was partly caused by their overall language ability and partly related to the test method. However, once they had the chance to read the script, the students realised that the test was not difficult (Section 5.2.). Their comments generally suggested a lack of communicative language ability (Section 5.4.3.) and pointed to a mismatch between the test purpose and the test taker's language learning experience, which may be related to the classroom teaching/learning practice in Taiwan, an issue worth examination in the future.

6.4. Future directions of LSP testing

LSP testing, according to arguments presented in 1.3., is about the measurement of the test takers' consistent ability to manipulate language in specific language use contexts. One assumption is that the candidates already possess a certain degree of the core language knowledge; the focus of assessment is therefore on the candidates' ability to demonstrate this knowledge in an appropriate manner in the language use contexts specified in the given language test.

In the early days of LSP testing, different language use/conventions/registers were considered to be tied to different subject areas (Section 1.3.1.). Needs analyses of the learner needs, linguistic features, and the target situations and incorporating these features in the test were considered essential in ensuring test authenticity. Direct testing and the inclusion of the specialist knowledge in the assessment have been two prominent features in test design. What had been overlooked was a theory of what language is and involves.

With the advances of our understanding of communicative language use ability, language ability is viewed as an interaction of an individual's language knowledge and the language use contexts. Therefore, in an LSP test design, instead of specifying linguistic features and routines in a behaviourist manner, the test construct is defined in terms of an interaction of the language knowledge, background/world knowledge and the strategies used in a test situation. In this approach to construct definition,

language ability is viewed as being influenced by both the test taker's language knowledge and the given language use contexts. Test takers, according to this approach, will have to vary their use of the language across contexts.

On the basis of my experience in the design of the ESP TG test, LSP tests seem a feasible alternative to the general-purpose TG test. By extension, LSP testing may be applied to measuring instruments in which a more clearly defined language use domain is present, for example, the assessment of officers' ability to survive in a training course in a foreign country or the assessment of diplomats' language ability.

The debate about the use of general proficiency tests and specific purpose tests is still vigorous. IELTS seems to be moving away to a less field specific approach (Douglas, 1995; Clapham, 1996, 2000; Davies, 2000b); McNamara (1997) has questioned the value of work sample ESP tests in the area of simulation of the target language use tasks and the assessment criteria. Specificity and authenticity underlie the debate; the assessment of specialist knowledge has its difficulties and problems (Section 1.3.7.), and research has called for more investigation into the issue of authenticity (Lewkowicz, 2000).

Davies (2001) challenges the assumption of "distinct varieties of a language" underlying some LSP test practice. My personal view is that the focus of an LSP test should be in the assessment of language ability through the language use context of interest and test design should address components of language use ability. A language test is only a sample of language behaviour and an indirect measure of an individual's language ability. In the choice between a general proficiency and a specific purpose test, the language tester has to take the test purpose, the stakes of the test, the type of performance required from the candidates, and practicality into consideration. Regardless of the choice of test type, the central issues in language testing are the representativeness of the language samples and the validity of the test claims. Perhaps in the end, as Davies (2001) suggests, there is no fundamental difference between a general proficiency test and a specific purpose test if the aim is to measure communicative characteristics through linguistic means. But from the communicative perspective, I think LSP testing is justified. In spite of some difficulties and practical problems in test development and test validation, there is a

place for specific purpose language testing; whether in Taiwan or elsewhere in the world.

Bibliography

Adams, R.J., P.E. Griffin and L. Martin (1987) A laten trait method for measuring a dimension in second language proficiency. *Language Testing* 3 (1): 9 - 27

Adams, R.J. & S.T. Khoo (1992) *Quest: The Interactive Test Analysis System.* Hawthorn:ACER

Alderson, J. C. (1981) Report on the discussion on communicative language testing. In J. C. Alderson & A. Hughes (eds.), pp 123 – 134

Alderson, J. C. (1988a) New procedures for validating proficiency tests of ESP? Theory and Practice. *Language Testing* 5(2): 220-232

Alderson, J. C. (1988b) Testing and its administration in ESP. In Chamberlain, D. & R. Baumgardner (eds.)1988 ELT Documents 128. London: Modern English Publications/ British Council, pp 87 – 97

Alderson, J. C. (1990) Testing reading comprehension skills (Part Two): Getting students to talk about taking a reading test (A pilot study). *Reading in a Foreign Language* 7(1): 465-502

Alderson, J. C. (1991) Bands and scores. In Alderson, J. C. & B. North (eds.), pp 71-86

Alderson, J. C. (1993) The relationship between grammar and reading in an English for academic purposes test battery. In Douglas, D. & C. Chapelle (eds.), pp 203-219

Alderson, J. C. (1997) Bands and scores. In Clapham, C & J.C. Alderson (eds.), pp 87-108

Alderson, J. C. (2000) Technology in testing: the present and the future. *System* 28(4): 593-603

Alderson, J.C. & A. Hughes (eds.) (1981) *Issues in language testing. ELT Document 111.* London: The British Council

Alderson, J.C. & A. H. Urquhart (1985) The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing* 2: 192 - 204

Alderson, J. C. & Y. Lukmani. (1989) Cognition and levels of comprehension as embodied in test questions. *Reading in a foreign language* 5(2): 253-270

Alderson, J.C. & North B. (eds.) (1991) *Language testing in the 1990s.* London: Macmillan Publishers Ltd.

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

Alderson, J.C. & D. Wall (1993) Deos washback exist? *Applied Linguistics* 14(2): 115-129

Alderson, J.C. & G. Buck. (1993) Standards in testing: A survey of the practice of UK Examination Boards in EFL testing. *Language Testing* 10(2): 1-26

Alderson, J.C., C. Clapham, & D. Wall (eds.) (1995) *Language test construction and evaluation*. Cambridge: Cambridge University Press

Allen, J. P. B. & H. G. Widdowson (1974) Teaching the communicative use of English in Swales (ed.) 1988, pp 69-89

Allen, J.P.B. & A. Davies (eds.) (1977) *The Edinburgh course in Applied Linguistics, vol.4: Testing and experimental methods*. Oxford: OUP

Allen, A. (1992) Development and validation of a scale to measure test-wiseness in EFL/ESL reading test takers. *Language Testing* 9:101-123

Anastasi, A. (1988) *Psychological testing*. London: Macmillan.

Anderson, A. & T. Lynch (1988) *Listening*. Oxford: Oxford University Press

Andrich, D. (1978) A rating formulation for ordered response categories. *Psychometrika*, 43(4): 561-573

Angoff, W. & A.J. Sharon. (1971) A comparison of scores earned on the Test of English as a Foreign Language by native American college students and foreign applicants. *TESOL Quarterly*, 5(2): 129-136

Bachman, L. F. (1986) The Test of English as a Foreign Language as a measure of communicative competence in C. W. Stansfield (ed.), pp 69-88

Bachman, L. F. (1990) *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. & A. S. Palmer (1981) A multitrait-multimethod investigation into the construct validity of six tests of listening and reading. In A.S. Palmer, P.J.M. Groot and G.A. Trosper (eds.), pp 149-165

Bachman, L. F. & Palmer, A. (1982) The construct validation of some components of communicative proficiency. *TESOL Quarterly* 16: 449-465

Bachman, L. F. & A. S. Palmer (1983) The construct validity of the FSI Oral Interview. In J. W. Oller (ed.), pp 156-169

Bachman, L. F. & S. J. Savignon (1986) The evaluation of communicative language proficiency: a critique of the ACTFL Oral Interview. *The Modern Language Journal* 70(4): 380-390

Bachman, L. F., B. K. Lynch & M. Mason (1995) Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing* 12(2): 238-257

Bachman, L. F. & Palmer, A. (1996) *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L. F., Davidson, F., Ryan, K. & Choi, I-C (1996) *An investigation into the comparability of two tests of English as a foreign language: The Cambrideg-TEOFL comparability study*. UCLES

Bachman, L. F., F. Davidson & M. Milanovic (1996) The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Test* 13(2): 125 – 150

Bachman, L. F., & A. D. Cohen (eds.) (1998) *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press

Bailey, K. (1983) Competitiveness and anxiety in adult second language learning: looking *at* and *through* the diary studies. In Seliger, H. & M. Long (eds) *Classroom oriented research in second language acquisition*. Rowley, Massachusetts: Newbury House Publishers, pp 67 –103

Bailey, K. (1996) Working for washback: a review of the washback concept in language testing. *Language Testing* 13(3): 257-279

Baker, R. (1987) An investigation of the Rasch model in its application to foreign language proficiency testing. Ph.D. Thesis: University of Edinburgh

Baker, R. (1997) *Classical test theory and item response theory in test analysis*. Special Report No 2: Language Test Update

Barber, C. L. (1962) Some measurable characteristics of modern scientific prose in Swales (ed.) 1988, pp 1 – 16

Benesch, S. (1996) Needs analysis and curriculum development in EAP: an example of a critical approach. *TESOL Quarterly* 30(4): 723-738

Bensoussan, M. (1982) Testing the test of advanced EFL reading comprehension: To what extent does the difficulty of a multiple-choice comprehension test reflect the difficulty of the test? *System* 10(3): 285 - 290

Bernhardt, E. B. (1991) *Reading development in a second language: Theoretical, empirical, & classroom perspectives*. Norwood, NJ: Ablex Publishing

Berwick, R.(1989) Needs assessment in language programming: from theory to practice in R.K. Johnson (ed.) *The second language curriculum*. Cambridge: Cambridge University Press. Pp 48 – 62

Bialystok, E. (1990) *Communication Strategies*. Cambridge, MA: Basil Blackwell

Blais, J. & M.D. Laurier (1995) The dimensionality of a placement test from several analytical perspectives. *Language Testing*, 12(1): 73 – 95

Bock, K. & W. J. Levelt (1993) Language production: Grammatical encoding. Gocnitive Science Technical Report UIUC-BI-CS-93-04. The Bechman Institute, University of Illinois: Urbana IL

Bolinger, D. (1977) *Meaning and form*. London: Longman

Brennan, R. L. (2000) An essay on the history and future of reliability from the perspective of replications. Paper presented at the Annual Meeting of the National Council on Measurement in Education. New Orleans.

Brindley, G. (1986) *The assessment of second language proficiency: Issues and approaches*. Adelaide: National Curriculum Resource Centre, Adult Migration Education Program.

Brindley, G. (1998) Describing language development? Rating scales and second language acquisition. In L. F. Bachman, & A. D. Cohen (eds.), pp 112-140

Brindley, G. (1998) Assessing listening abilities. *Annual Review of Applied Linguistics* 18: 171 - 191

Brown, A. (1993) The role of test-taker feedback in the test development process: test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing* 10(3): 277-303

Brown, A. (1994) LSP testing: the role of linguistic and real-world criteria. In R. Khoo (ed.) LSP; Problems and prospects. Singapore:RELC, pp 202 – 218

Brown, A. (1995) The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing* 12, 1 - 15

Brown, G. & G. Yule (1983a) *Discourse analysis*. Cambridge: Cambridge University Press

Brown, G. & G. Yule (1983b) *Teaching the spoken language*. Cambridge: Cambridge University Press

Brown, S. (1980) *What do they know: a review of criterion-referenced testing*. Edinburgh: HMSO.

Buck, G. (1991) The testing of listening comprehension: an introspective study. *Language Testing* 8(1): 67-91

Buck, G. (1992) Listening comprehension: Construct validity and trait characteristics. *Language Learning* 42(3): 313-357

Buck, G. (1994) The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing* 11:145 - 170

Bygate, M. (1987) *Speaking*. Oxford: Oxford University Press

Cambridge Certificates in Communicative Skills in English: Handbook (1995) UCLES: Cambridge

Campbell, D.T. & D.W. Fiske (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56: 81 – 104

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

Canale, M. (1983) On some dimensions of language proficiency. In Oller, J.W. (ed.), pp 333-342

Canale, M. (1988) The measurement of communicative competence. *Annual review of applied linguistics*, 8: 67-84

Candlin, C. N. (1986) Explaining communicative competence limits of testability? In C. W. Stansfield (ed.), pp 38 – 57

Carrell, P. L. (1983) Background knowledge in second language comprehension. *Language Learning and Communication* 2: 25 - 34

Carrell, P. L. (1983) Schema theory and ESL reading pedagogy. *TESOL Quarterly* 17: 553 - 573

Carroll, J.B. (1961) Fundamental considerations in testing for English language proficiency of foreign students. In J.B. Allen (ed.) *Teaching English as a second language*. Philippines: McGraw-Hill, pp 364 – 372

Carroll, J.B. (1980) *Testing communicative performance*. Pergamon: Oxford

Carroll, B.J. (1981) Specifications for an English language testing service in Alderson, J.C. and A. Hughes (eds.), pp 66-100

Carroll, B.J. (1983) Issues in the testing of language for specific purposes. In Hughes, A. and D. Porter (eds.), pp109-114

Carroll, B.J. (1985) Second language performance testing for university and professional contests in Hauptman, P.C. *et al.* (eds.), pp 73-88

Carroll, B. J. & R. West (1989) *ESU framework: Performance scales for English language examinations*. Essex: Longman.

Cazden, C. B. (1989) Contributions of the Bakhtin circle to 'communicative competence'. *Applied Linguistics* 10 (2): 116 - 127

Celce-Murcia, M. (1990) Discourse analysis and grammar instruction. *Annual Review of Applied Linguistics* 11:135-151

Celce-Murcia, M. (1997) Direct approaches in L2 instruction: A turning point in communicative language teaching? *TESOL Quarterly*: 31(1): 141-152

Certificates in English for International Business & Trade: Specifications and sample material for the Revised CEIBT (1998) UCLES: Cambridge

Chalhoub-Deville, M. (1995) A contextualized approach to describing oral language proficiency. *Language Learning* 45(2): 251-281

Chalhoub-Deville, M. (1997) Theoretical models, assessment frameworks and test construction. *Language Testing* 14: 1-22

Chapelle, C.A. (1998) Construct definition and validity inquiry in SLA research. In Bachman, L.F. & A.D. Cohen (eds.), pp 32-70

Chen, Z. & G. Henning (1985) Linguistic and cultural bias in language proficiency tests. *Language Testing* 1: 155-163

Chiang, C. S. & P. Dunkel (1992) The effect of speech modification, prior knowledge, and listening proficiency on EFL lecture learning. *TESOL Quarterly* 26:345 - 374

Choi, I.C. & L.F. Bachman (1992) An investigation into the adequacy of three IRT models for data from two EFL reading tests. *Language Testing* 9(1): 51-78

Chung, Jing-mei (1997) A comparison of two multiple-choice test formats for assessing English structure competence. *Foreign Language Annals*, 30(1): 111-122

Clapham, C. M. (1981) Reaction to the Carroll paper (1). In Alderson, J. C. and A. Hughes (eds.), pp 111 – 116

Clapham, C. M. (1993) Is ESP testing justified? In Douglas, D. and C. Chapelle (eds.) *A new decade of language testing research.* Virginia: TESOL, pp 257 – 271

Clapham, C. M. (1996) *The development of IELTS: A study of the effect of background knowledge on reading comprehension.* Cambridge: Cambridge University Press

Clapham, C. M. (2000) Assessment for academic purposes: where next? *System* 28(4): 511-521

Clapham, C. M. & J. C. Alderson (1997) *Constructing and trialling the IELTS Test.* IELTS Research Report 3. Cambridge: The British Council, University of Cambridge Local Examinations Syndicate and International Development Program of Australian Universities and Colleges.

Clark, J. L. D. (1972) *Foreign language testing: theory and practice.* Center for Curriculum Development, Philadelpha PA

Clark, J. L. D. & S. S. Swinton (1979) An exploration of the speaking proficiency measures in the TOEFL context (TEOFL Research Report 4). Princeton, NJ: ETS.

Clark, J. L. D. & R.T. Clifford (1988) The FSI/ILR/ACTFL proficiency scales and testing techniques: development, current status and needed research. *Studies in Second Language Acquisition* 10(2): 129-147

Connelly, M. (1997) Using C-tests in English with post-graduate students. *English for Specific Purposes*, 16(2): 139 - 150

Conrad, L. (1985) Semantic versus syntactic cues in listening comprehension. *Studies in Second Language Acquisition* 7: 59 - 72

Coppieters, R. (1987) Competence differences between native and near-native speakers. *Language* 63(3): 544 - 573

Criper, C. (1981) Reaction to the Carroll paper (2). In Alderson, J. C. and A. Hughes (eds.), pp 117-120

Criper, C. and A. Davies (1988) *ELTS validation project report.* The British Council/Cambridge: University of Cambridge Local Examinations Syndicate, London

Cronbach, L.J. (1990) *Essentials of psychological testing.* 5th edition. New York: Harper and Row

Cronbach, L. J., R. L. Linn & E. H. Haertel (1997) Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement* 57(3): 373 - 399

Crookes, G. and S. M. Gass (eds.) (1993) *Tasks and language learning: Integrating theory and practice.* Clevedon, Avon: Multilingual Matters

Cummins, J. (1980) The cross-lingual dimensions of language proficiency: implications for bi-lingual applications. *Applied Linguistics* 11(1): 27-43

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

Cziko, G. A. (1985) Some problems with empirically-based models of communicative competence. *Applied Linguistics* 5 (1): 23 - 37

Dandonoli, P. & G. Henning (1990) An investigation of the construct validity of the ACTFL Proficiency Guidelines and Oral Interview Procedure. *Foreign Language Annals*, 23 (1): 11 - 22

Davidson, F. (1998) Review of Clapham, C. 1996: The development of IELTS: a study of the effect of background knowledge on reading comprehension. *Language Testing* 15(2): 288-292

Davies, A. (1968) Introduction. In Davies A (ed.) *Language testing symposium: a psycholinguistic approach.* London: Oxford University Press, pp 1 – 18

Davies, A. (1977) The construction of language tests. In Allen, J.P.B. and A. Davies (eds.) *Testing and experimental methods. The Edinburgh Course in Applied Linguistics*, vol. 4 Oxford: Oxford University Press

Davies, A. (1978) Language Testing Parts 1 & 2. *Language Teaching & Linguistics Abstracts* 11(3): 145 - 159 & 11(4) 215 - 231

Davies A. (1981) Review of J. Munby, Communicative syllabus design. *TESOL Quarterly* 15(3): 50-69

Davies, A. (1981) Reaction to the Palmer and Bachman and the Vollmer Papers (2), in Alderson, J.C. and A. Hughes (eds.), pp 215 – 231

Davies, A. (1984) Validation three tests of English language proficiency. *Language Testing* 1: 50 – 69

Davies, A. (1984) ESL expectations in examining: the problem of English as a foreign language and English as a mother tongue. *Language Testing*, 1: 82-95

Davies, A. (1989) Communicative competence as language use. *Applied Linguistics* 10(2): 157-170

Davies, A. (1990) *Principles of language testing*. Oxford: Blackwell

Davies, A. (1990) Operationalising uncertainty in language testing: An argument in favour of content validity. In J. H. A. L. de Jong & D. Stevenson (eds.), pp 179 – 195

Davies, A. (2000a) Code of Ethics. ILTA

Davies, A. (2000b) The logic of LSP. Paper presented at the Second Asian Language Assessment Research Forum, The Hong Kong Polytechnic University

Davies, A. (2001) The logic of LSP testing. *Language Testing* 18(2): page numbers not known yet

Davies, A., A. Brown, C. Elder, K. Hill, T. Lumley & T. McNamara (1999) Studies in Language Testing 7: Dictionary of language testing. Cambridge: UCLES

de Jong, J. J. A. L. & D. K. Stevenson (eds.) (1990) *Individualizing the assessment of language abilities.* Clevedon: Multilingual Matters

Dirven, R. & J. Oakeshott-Taylor (1985) Listening comprehension (part II). *Language Teaching* 18: 2 - 20

Dixon, R. (1991) Listening comprehension: Textual, contextual, cognitive and affective considerations. ERIC Document Reproduction Service No. ED332513

Douglas, D. (1986) Communicative competence and tests of oral skills in C. W. Stansfield (ed.), pp 156 – 174

Douglas, D. (1988) Testing listening comprehension in the context of the ACTFL Proficiency Guidelines. *Studies in Second Language Acquisition*, 10: 245-262

Douglas, D. (1995) Developments in language testing. *ARAL* 15: 167 – 187

Douglas, D. (1998) Testing methods in context-based second language research. In Bachman, L.F. & A.D. Cohen (eds.), pp 141-155

Douglas, D. (2000) *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.

Douglas, D. & L. Selinker (1985) Principles for language tests within the discourse domains" theory of interlanguage: Research, test construction and interpretation. *Language Testing* 2: 205 - 226

Douglas, D. & L. Selinker (1992) Analyzing oral proficiency test performance in general and specific purpose tests. *System* 20 (3): 317 - 328

Douglas, D. & L. Selinker (1993) Performance on a general versus a field-specific test of speaking proficiency by international teaching assistants. In Douglas, D. & C. Chappelle (eds.), pp 235-256

Douglas, D. and C. Chappelle (eds.) (1993) *A new decade of language testing research*. Virginia: TESOL

Douglas, G. & B. Wright (1987) Response Patterns and their probabilites. *Rasch Measurement Transcations* 3(4):75

Dudley-Evans, T. & M. J. St. John (1998) *Developments in English for specific purposes: A multi-disciplinary approach*. Cambridge: Cambridge University Press

Dunkel, P. (1991) Listening in the native and second/foreign language: Toward an integration of research and practise. *TESOL Quarterly* 25: 431 - 457

Dunkel, P., G. Henning, G. Ghaudron (1993) The assessment of an L2 listening comprehension construct: a tentative model for test specification and development. *The Modern Language Journal*, 77: 180 - 191

Duranti A. & C. Goodwin (eds.) (1992) *Rethinking context: language as an interactive phenomenon*. Cambridge: Cambridge University Press

Elder C. (1993) Language proficiency as a predictor of performance in teacher education. *Melbourne Papers in Language Testing* 2(1): 68 - 89

Elder C. (1993) How do subject specialists construe classroom language proficiency? *Language Testing* 10(3): 235 - 254

Elder C. (1994) Performance testing as benchmark for LOTE teacher education. *Melbourne Papers in Language Testing* 3(1): 1- 25

Ellis, R. (1995) Interpretation tasks for grammar teaching. *TESOL Quarterly* 29(1): 87 - 106

English for the tourism industry: extended syllabus (1995) London: LCCI

Emmett A. (1985) The Associated Examining Board Test in English for Educational Purposes (TEEP). In Hauptman O.C., LeBlanc R. and Wesche M.B. (eds.), pp 131-152

Embretson, S. (1983) Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93: 179 - 197

Engel, D. M. & F. Myles (1996) Grammar teaching: The major concerns. In Engel, D.M. & F. Myles (eds.) *Teaching grammar: Perspectives in higher education.* London: CILT

Farhady, H. (1982) Measures of language proficiency from the learner perspective. *TESOL Quarterly* 16, 43-59

Faerch, C. & Kasper G. (1986) The role of comprehension in second-language learning. *Applied Linguistics* 7: 257-274

Fayer, J. M. & E. Drasinski (1987) Native and non-native judgements of intelligibility and irritation. *Language Learning* 37: 313 - 326

Ferguson, G. & E. White (1994) *IELTS research report: a predictive validity study of IELTS.* Edinburgh: University of Edinburgh

Feyten, G. M. (1991) The power of listening ability: an overlooked dimension in language acquisition. *The Modern Language Journal* 75(2):173-180

Fouly, K., L. Bachman, and G. Cziko (1990) The divisibility of language competence: a confirmatory approach. *Language Learning* 40(1): 1-21

Frederiksen J.R. and A. Collins (1989) A systems approach to educational testing. *Educational Researcher* 18(9): 27-32

Fruchter, B. (1954) *Introduction to Factor Analysis.* New York: D. Van Nostrand Company, Inc.

Fulcher, G. (1987) Tests of oral performance: the need for data-based criteria. *ELT Journal* 41(4); 287 - 291

Fulcher, G. (1996) Invalidating validity claims for the ACTFL oral rating scale. *System* 24(2): 163 - 172

Fulcher, G. (1999) Assessment in English for academic purposes: putting content validity in its place. *Applied Linguistics* 20(1): 221 – 236

Fulcher, G. (2000) The 'communicative' legacy in language testing. *System* 28(4): 483-497

Givon, T. (1979) *On understanding grammar.* New York: Academic Press

Goldstein, H. (1979) Consequences of using the Rasch model for educational assessment. *British Educational Research Journal*, 5(2): 211-220

Goldstein, M. & T. Lewis (eds.) (1996) *Assessment: problems, developments and statistical issues.* Chichester: Wiley

Graham J.G. (1987) English language proficiency and the prediction of academic success. *TESOL Quarterly* 21(3): 505-521

Grice, H. P. (1975) Logic and conversation. In P. Cole & J. L. Morgan (eds.) Syntax and semantics Volume 3: Speech act. New York: Academic Press pp 41-58

Griffin P.E., R.J. Adams, L. Martin & B. Tomlinson (1988) An algorithmic approach to prescriptive assessment in English as a Second Language. *Language Testing* 5(1): 1-8

Grotjahn, R. (1986) Test validation and cognitive psychology: some methodological considerations. *Language Testing*, 3(2): 159 - 185

Gruba, P. (1997) The role of video media in listening assessment. *System* 25(3): 335 - 345

Guilford, J.P. (1954) *Psychometric methods.* New York: McGraw-Hill

Guilford, J.P. & B. Fruchter (1978*Fundamental statistics in psychology and education.* New York:McGraw-Hill

Hale, G. A. (1988) Student major field and text content: interactive effects on reading comprehension in the Test of English as a Foreign Language. *Language Testing* 4: 49 - 61

Hamayan, E. V. (1995) Approaches to alternative assessment. *ARAL* 15: 212 – 226

Hamilton, J., M. Lopes, T.F. McNamara & E. Sheridan (1993) Rating scales and native speaker performance on a communicatively oriented EAP test. *Language Testing* 10(3): 337-353

Hambleton, R.K., H. Swaminathan & H.J. Rogers (1991) *Fundamentals of Item Response Theory.* Newbury Park, Calif.: Sage Publications.

Hare,V. C. & D. A. Devine. (1983) Topical knowledge and topical interest predictors of listening comprehension. *Journal of Educational Research* 76: 157 - 160

Harley, B., J. Cummins, M. Swain & P. Allen (eds.) (1990) *The development of second language proficiency.* New York: Cambridge University Press

Harrison, A. (1983) *A language testing handbook.* London: Macmillan

Harrison, A. (1983) Communicative testing: jam tomorrow. In Hughes, A. and D. Porter (eds.), 1983: 77-86

Hatch, E. & A. Lazaraton (1991) *The research manual: design and statistics for applied linguistics.* New York: Newbury House

Hauptman, P.C., R. Le Blanc & M.B. Wesche (eds.) (1985) *Second language performance testing.* Ottawa: University of Ottawa Press.

Heaton, J.B. (1988) *Writing English language tests.* London: Longman

Hellekant, J. (1994) Are multiple-choice tests unfair to girls? *System* 22(3): 349 - 352

Henning, G., Hudson T. & Turner, J. (1985) Item response theory and the assumption of unidimensionality for language tests. *Language Testing* 2(2): 141-145

Henning, G. (1987) *A guide to language testing: development, evaluation, research.* Cambridge MA: Newbury House.

Henning, G. (1987) Is the Bejar test of unidimensionality appropriate? A response to Spurling. *Language Testing* 4(1): 1-11

Henning, G. (1988) The influence of test and sample dimensionality on latent trait person ability and item difficulty calibrations. *Language Testing* 5(1): 83-99

Henning, G. (1992) Dimensionality and construct validity of language tests. *Language Testing* 9(1): 1-11

Henning, G. (1992) The ACTFL Oral Proficiency Interview: validity evidence. *System* 20(3): 365 - 372

Hock, T. S. (1990) The role of prior knowledge and language proficiency as predictors of reading comprehension among undergraduates. In J. H. A. L. de Jong & D. K. Stevenson (eds.), pp 214 – 224

Holliday, A. (1995) Assessing language needs within an institutional context: An ethnographic approach. *English for Specific Purposes* 14 (2): 115 - 126

Hornberger N. H. (1989) *Tramites* and *Transportes*: the acquisition of second language communicative competence for one speech event in Puno, Peru. *Applied Linguistics* 10(2): 214 - 230

Huang, X. & M. V. Naerssen (1987) Learning strategies for oral communication. *Applied Linguistics*, 8: 287-307

Hughes, A. (1981) Reaction to the Palmer and Bachman and the Vollmer Papers (1) in Alderson, J.C. and A. Hughes (eds.), 1981: 176-181

Hughes & Porter, D. (eds.) (1983) *Current developments in language testing*. London: Academic Press

Hughes, A, D. Porter & C. Weir (eds.) (1988*) ELTS Validation Project: Proceedings of a conference held to consider the ELTS Validation Project Report. English Language Testing Service Research Report 1 (ii).* British Council/ University of Cambridge Local Examinations Syndicate, London

Hughes, A. (ed.) (1988) *Testing English for university study, ELT Documents 127.* London: Modern English Publications, British Council

Hughes, A. (1989) *Testing for language teachers*. Cambridge: Cambridge University Press

Hughes, R. & M. McCarthy (1998) From sentence to discourse: Discourse grammar and English language teaching. *TESOL Quarterly* 32(2): 263-287

Hutchinson, T. & W. Waters (1987) *English for specific purposes*. Cambridge: Cambridge University Press

Hymes, D.H. (1967) Models of the interaction of language and social setting. *Journal of Social Issues* 23(2): 8-38

Hymes, D.H. (1972) On communicative competence. In Pride, J.B. and J. Holmes (eds.) *Sociolinguistics: selected readings*. Harmondsworth, Middlesex: Penguin pp269 – 293

Ingram, E. (1977) Basic concepts in testing. In Allen, J. P. B. and A. Davies (eds.), pp 11 – 37

Jacoby, S. & E. Ochs (1995) Co-construction: an introduction. *Research on Language and Social Interaction* 28(3): 171 - 183

Jakobovits L. (1969) A functional approach to the assessment of language skills. *Journal of English as a second language* 4:63-76

Jensen, A. R. (1980) *Bias in mental testing*. New York: The Free Press

John J. (1992) The Ontario Test of ESL Oral Interaction Test. *System* 20 (3): 305 - 316

Johnson, D.C. (1977) The TOEFL and domestic students: conclusively inappropriate. *TESOL Quarterly* 11(1): 79-86

Johnson, K. & D. Porter (1983) *Perspectives in communicative language teaching*. London: Academic Press.

Jones, R. L. & B. Spolsky (eds.) (1975) *Testing language proficiency*. Arlington, VA: Center for Applied Linguistics

Jones, R. L. (1985) Second language performance testing: an overview. In Hauptman, P. C., R. LeBlanc, & M. B. Wesche (eds.), pp 15 – 24

Kasper, G. (ed.) (1996) The development of pragmatic competence [Special issue]. *Studies in Second Language Acquisition*, 16(2).

Kelly, P. (1991) Lexical ignorance: the main obstacle to listening comprehension with advanced foreign language learners. *IRAL* 29: 135 - 149

Kelly, R. (1981) Aspects of communicative performance. *Applied Linguistics*, 11(2): 169-179

Khaniya, T.R. (1990) *Examinations as instruments for educational change: Investigating the washback effects of the Nepalese English exams*. Ph.D. Thesis: University of Edinburgh

Kim, J-O & C.W. Mueller (1978) Factor Analysis: Statistical Methods and Practical Issues. Beverly Hills, CA: Sage Publications

Kramsch, C. (1993) *Context and culture in language teaching*. Oxford: Oxford University Press

Kramsch, C. (1986) From language proficiency to interactional competence. *The Modern Language Journal,* 70: 366 - 371

Kumaravadivelu, B. (1994) The postmethod condition: (E)merging strategies for second/foreign language teaching. *TESOL Quarterly*, 28: 27-48

Kunnan, A.J. (ed.) (2000) Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquim, Orlando, Florida. UCLES: Cambridge University Press

Lackstorm, J., L. Selinker & L. P. Trimble (1972) Grammar and Technical English. In Swales (ed.), pp 60 – 66

Lado, R. (1961) *Language Testing*. London: Longmans

Language Testing (1997) Special Issue on ethics in language testing 14(3)

Lantolf, J.P. & W. Frawley (1988) Proficiency: understanding the construct. *Studies in second language acquisition* 10(2): 181-195

Lantolf, J. P. & W. Frawley (1992) Rejecting the OPI again: a response to Hagen. *ADFL Bulletin* 23: 34 - 37

Larsen-Freeman, D. (1986) A response to Sandra Savignon he meaning of communicative competence in relation to the TOEFL Program" in C. W. Stansfield (ed.), pp 31 – 37

Larsen-Freeman D. & M. H. Long (1991) *An introduction to second language acquisition research.* New York: Longman

Lazaraton, A. (1992) The structural organization of a language interview: a conversation analytic perspective. *System* 20(3): 373 - 386

Levelt, W. J. M. (1983) Monitoring and self-repair in speech. *Cognition* 14: 41- 104

Levelt, W. J. M. (1989) *Speaking: From intention to articulation.* Cambridge, Massachusetts: MIT Press

Lewkowicz, J. A. (2000) Aauthenticity in language testing: some outstanding questions. *Language Testing* 17(1): 43-64

Light, R. L., M. Xu & J. Mossop (1987) English proficiency and academic performance of international students. *TESOL Quarterly* 21: 251 - 261

Lightbown, P.M., N. Spada, & L. White, (eds.) (1993) The role of instruction in second language acquisition [Special issue]. *Studies in Second Language Acquisition*, 15(2).

Linacre, J.M. (1994) *Many-faceted Rasch measurement.* Chicago: MESA Press

Linacre, J.M. (1997) *A user's guide to FACETS: Rasch measurement computer program.* Chicago: MESA Press

Linacre, J.M. (1997) *A user's guide to Facform: Data formatting computer program for FACETS.* Chicago: MESA Press

Linacre, J.M.(1998) Rating, judges and fairness. *Rasch Measurement Transactions.* URL: www.rasch.org/rmt/rmt122f.htm

Linacre,J. & B. Wright (1992) Chi-square fit statistics. *Rasch Measurement Transactions* 8(2): 358

Linn, R.L. (1994) Performance assessment: Policy promises and technical measurement standards. *Educational Researcher* 23(9): 4-14

Linn, R.L, E.L. Baker and S.B. Dunbar (1991) Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher* 20(8): 12-21

Linn, R. L. & E. Burton (1994) Performance-based assessment: Implications for task specificity. *Educational Measurement: Issues and Practice* 13: 5 – 15

Long, D. R.(1989) Second language listening comprehension: A schema-theoretic perspective. *Modern Language Journal* 73: 32 – 40

Long, D. R. (1990) What you don't know can't help you: An exploratory study of background knowledge and second language listening comprehension. *Studies in Second Language Acquisition* 12: 65 – 80

Lowe, P. (1986) Proficiency: Panacea, framework, process? A reply to Kramsch, Schulz, and particular to Bachman and Savignon. *Modern Language Journal* 70: 391 - 397

Lowenberg, P. (1993) Issues of validity in test of English as a world language: whose standards? *World Englishes* 12(1): 95 - 106

Lumley, T. & T. F. McNamara (1995) Rater characteristics and rater bias: implications for training. *Language Testing* 12(1): 54-71

Lumley T. (1998) Perceptions of language-trained raters and occupational experts in a Test of Occupational English Language Proficiency. *English for Specific Purposes* 17 (4): 347 - 367

Lund, R. J. (1991) A comparison of second language listening and reading comprehension. *The Modern Language Journal* 75: 196 - 204

Lunz M. E. & J. Stahl (1990) Judge consistency and severity across grading periods. *Evaluation and the Health Professions* 13: 425 - 444

Lynch, B., F. Davidson & G. Henning. (1988) Person dimensionality in language test validation. *Language Testing* 5(2): 206-219

Maclean J. (1985) Descriptions of doctor-patient communication. *International Journal of the Sociology of Language* 51: 131 - 134

Madsen, H. (1982) Determining the debilitative impact of test anxiety. *Language Learning* 32: 133-143

Markham, P. (1988) Gender differences and the perceived expertness of the speaker as factors in ESL listening recall. *TESOL Quarterly* 22: 397-406

Markham, P & M. Latham. (1987) The influence of religion-specific background knowledge on the listening comprehension of adult second-language students. *Language Learning* 37: 157 - 170

Masters, G.N. (1982) A Rasch model for partial credit scoring. *Psychometrika* 47(2):149-174

Matthews, M. (1990) The measurement of productive skills: doubts concerning the assessment criteria of certain public examinations. *ELT Journal* 44(2): 117 – 121

Maxwell, J. A. (1992) Understanding and validity in qualitative research. *Harvard Educational Review* 62 (3):279 - 299

McKeown, A. (1997) The role of needs analysis in course design. Unpublished MSc project, University of Edinburgh

McNamara, T. F. (1989) ESP testing: general and particular. In Candlin C. N. and T. F. McNamara (eds.) *Language, learning and community*. Sydney: National Centre for English Language Teaching and Research, Macquarie University pp 125 - 142

McNamara, T. F. (1991) Test diminsionality: IRT analysis of an ESP listening test. *Language Testing* 8(2): 45 - 65

McNamara, T. F. (1996) *Measuring second language performance.* London: Longmans

McNamara, T.F. (1997) 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics* 18 (4): 446 - 466

McNamara, T. F. (1997) Problematizing content validity: the Occupational English Test (OET) as a measure of medical communication. *Melbourne Papers in Language Testing* 6(1): 19 – 39

Mead, R. (1982) Review of Munby, J. ommunicative syllabus design". *Applied Linguistics* 3: 70 - 77

Meisels, A., A. Dorfman & D. Steele (1995) Equity and excellence in group-administered and performance-based assessments in M. T. Nettles & A. L. Nettles (eds.) *Equity and Excellence in Educational Testing and Assessment*. London: Kluwer Academic Publishers, pp 243 - 261

Messick, S. (1989) Validity. In Linn R.L. (ed.) *Educational Measurement*. Third Edition. New York: Macmillan, pp 13 –104

Messick, S. (1994) The interplay of evidence and consequences in the validation of performance assessment. *Educational Researcher* 23(2):13 – 23

Milanovic, M. & N. Saville (eds.) (1996) Studies in language testing 3: Performance testing, cognition and assessment. UCLES:CUP

Mislevy, R. J. (1995) Test theory and language-learning assessment. *Language Testing* 12: 341-369

Morrow, K. (1979) Communicative language testing: revolution or evolution? In Brumfit C.J. and K. Johnson (eds.) *The communicative approach to language teaching*. Oxford: Oxford University Press, pp 143-157

Morrow, K. (1981) Communicative language testing: revolution or evolution? In *ELT Document 111:* 9-25

Morrow K. (1983) The Royal Society of Arts Examination in the Communicative Use of English as a Foreign Language. In Jordan R.R. (ed.) *Case studies in ELT.* London: Collins ELT, pp 102 – 107

Moss, P. (1992) Shifting conceptions of validity in educational measurement: implications for performance assessment. *Review of Educational Research* 62: 229-258

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

Muller, H. (1987) A Rasch model for continuous ratings. *Psychometrika*, 52(2):165-181

Munby, J. (1978) *Communicative syllabus design.* Cambridge: Cambridge University Press

Murphy, R. J. L. (1980) Sex differences in GCE examination entry statistics and success rates. *Educational Studies* 6(2): 169–178

Nevo, D. & E. Shohamy (1986) Evaluation standards for the assessment of alternative testing methods: an application. *Studies in Educational Evaluation* 12: 149-158

North, B. (1995) The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System* 23 (4): 445 - 465

North, B. & G. Schneider (1998) Scaling descriptors for language proficiency scales. *Language Testing* 15(2): 217 - 262

Nunan, D. (1988) *The learner-centred curriculum* Cambridge: Cambridge University Press

Oughlin, K. (1995) Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing* 12(2); 330-341

Omaggio A. (1983) Methodology in transition: the new focus on proficiency. *The Modern Language Journal* 67(4): 330-341

O'Malley, J. M., A. U. Chamot, & L. Kupper (1989) Listening comprehension strategies in second language acquisition. *Applied Linguistics* 10: 418 - 437

Oller, J.W. Jr. (1979) L*anguage tests at school.* London: Longmans

Oller, J.W. Jr. (ed.) (1983) *Issues in language testing research.* Rowley, Mass.: Newbury House

Oller, J. W. Jr (1986) Communication theory and testing: What and how in C. W. Stansfield (ed.), pp 104-155

Oller, J.W. Jr. & Perkins, K. (eds.) (1978) L*anguage in education: testing the tests.* Rowley, Mass.: Newbury House

Oller, J.W. Jr. & Perkins, K. (eds.) (1980) *Research in language testing.* Rowley, Mass.: Newbury House

Oxford International Business English Certificate. First level: Syllabus and specimen material. (1996) Oxford: UODLE

Oxford International Business English Certificate. Executive level: Syllabus and specimen material. (1996) Oxford: UODLE

Palmer, A. S., P. J. M. Groot & G. Trosper (eds.) (1981) *The construct validation of tests of communicative competence.* Washington, DC: TESOL

Papajohn, D. (1999) The effect of topic variation in performance testing: the case of the chemistry TEACH test for international teaching assistants. *Language Testing* 16(1): 52-81

Peirce, B.N. (1992) Demystifying the TOEFL reading test. *TESOL Quarterly* 25: 665-689

Perett, G. (1990) The language testing interview: a reappraisal. In de Jong J.H.A.L. and D.K. Stevenson (eds.), pp 225 – 238

Pienemannm, M. Johnston & G. Brindley (1988) Constructing an acquisition-based procedure for second language assessment. *Studies in Second Language Acquisition* 10(2): 217 – 243

Pollitt, A. & C. Hutchinson (1987) Calibrating graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing* 3: 72 – 72

Popham, W. J. (1993) Educational testing in America: What's right, what's wrong? *Educational Measurement: Issues and Practice* 21(1): 11 – 14

Portal, M. (ed.) (1986) *Innovations in language testing.* NFER-Nelson.

Porter D. (1991) ffective factors in language testing. In Alderson C.J. & B. North (eds.) *Language testing in the 1990s.* London: Modern English Publications and the British Council pp 32 – 40

Powers D. (1986) Academic demands related to listening skills. *Language Testing* 3(1): 1 – 38

Quinn T.J. (1993) The competency movement, applied linguistics and language testing: some reflections and suggestions for a possible research agenda. *Melbourne Papers in Language Testing* 2(2): 55 – 87

Raffaldini T. (1988) The use of situation tests as measures of communicative ability. *Studies in Second Language Acquisition* 10(2): 197 – 216

Rea Dickins, P. M. & E.G. Woods (1988) Some criteria for the development of communicative grammar tasks. *TESOL Quarterly* 22(4): 623 – 646

Read, J. (ed.) (1984) *Trends in language syllabus design* Anthology Series 13. Singapore: SEAMEO Regional English Language Centre: 55 – 67

Reed, D. J. (1992) The relationship between criterion-based levels of oral proficiency and norm-referenced scores of general proficiency in English as a second language. *System* 20 (3): 329 – 345

Reves, T. (1991) From testing research to educational policy: A comprehensive test of Oral Proficiency. In Alderson, J.C. & B. North (eds.), pp 178 – 188

Rice, F. (1959) The Foreign Service Institute tests language proficiency. *Linguistic Reporter* 1: 2 – 4

Richards, J. C. (1983) Listening comprehension: Approach, design, procedure. *TESOL Quarterly* 17(2): 219 - 239

Richards, J. C. (1985) Planning for proficiency. *Prospect* 1(2): 1-17

Rivera, C. (ed.) (1984) *Communicative competence approaches to language proficiency.* Clevedon: Multilingual Matters

Roach, J.O. (1945) *Some Problems of Oral Examinations in Modern Languages: An Experimental Approach Based on the Cambridge Examinations in English for Foreign Students, Being a Report Circulated to Oral Examiners and Local Examiners for Those Examinations.* Cambridge: University of Cambridge Local Examinations Syndicate.

Robinson, P. (1991) *ESP today.* Hemel Hempstead: Prentice Hall

Ross, S. (1992) Accommodative questions in oral proficiency interviews. *Language Testing* 9(2): 173 – 186

Ross, S. & R. Berwick (1992) The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition* 14: 159 – 176

Rost, M. (1990) *Listening Language Learning.* New York: Longman

Rubin, J. (1994) A review of second language listening comprehension research. *The Modern Language Journal* 78 (2): 199 – 221

Rumelhart, E. E. (1983) Understanding understand. In *Understanding Reading Comprehension* by J. Flood (ed.) pp 1 – 21. Neward, Delaware: International Reading Association.

Salaberry, M.R. & N. Lopez-Ortega (1998) Accurate L2 production across language tasks: focus on form, focus on meaning, and communicative control. *The Modern Language Journal* 82(4): 514 – 532

Salaberry, R. (2000) Revising the revised format of the ACTFL Oral Proficiency Interview. *Language Testing* 17(3): 289 – 310

Savignon S.J. (1972) *Communicative competence: an experiment in foreign language teaching*. Philadelphia PA: The Center for Curriculum Development

Savignon S.J. (1983) *Communicative competence: theory and classroom practice.* Reading MA: Addison-Wesley

Savignon S. J. (1985) Evaluation of communicative competence: the ACTFL Provisional Proficiency Guidelines. *The Modern Language Journal* 69(2): 129 – 134

Savignon, S. J. (1986) The meaning of communicative competence in relation to the TOEFL Program in C. W. Stansfield (ed.), pp 17 – 30

Scarcella, R. C., E. Andersen & S. d. Krashen (eds.) (1990) *Developing Communicative Competence in a Second Language* Boston, Heinle & Heinle Publishers

Schachter, J. (1990) Communicative competence revisited. In Harley et. al. pp 39 – 49

Schmidt, R.W. (1990) The role of consciousness in second language learning. *Applied Linguistics* 11(2): 129 – 158

Schmidt-Rinehart, B. C. (1994) The effects of topic familiarity on second language listening comprehension. *The Modern Language Journal* 78(2):179 –188

Scott, M. L., C. W. Stansfield and D. M. Kenyon (1996) Examining validity in a performance test: The Listening Summary Translation Exam (LSTE)-Spanish version. *Language Testing* 13(1): 83-110

Scott, W. & S. Muhlhaus (eds.) (1994) *Language for Specific Purposes*. Kingston: CiLT

Seedhouse, P. (1995) Needs analysis and the General English classroom. *ELT Journal*, 49(1): 59 – 65

Sharwood Smith, M. (1993) Input enhancement in instructed SLA. *Studies in Second Language Acquisition* 15(2): 165 – 179

Shavelson, R. J., N. M. Webb, & G. L. Rowley (1989) Generalizability theory. *American Psychologist* 44(6): 922 – 932

Shohamy, E. (1983) The stability of oral proficiency assessment on the oral interview testing procedures. *Language Learning* 33(4): 527 – 540

Shohamy, E. (1983) Rater reliability of the Oral Interview Speaking Test. *Foreign Language Annals* 16(3): 219 – 222

Shohamy, E. (1988) A proposed framework for testing the oral language of second/foreign language learners. *Studies in Second Language Acquisition* 10(2): 165 – 179

Shohamy, E. (1990) Language testing priorities: a different perspective. *Foreign Language Annals* 23(5): 385 – 394

Shohamy, E. (1994) The validity of direct versus semi-direct oral tests. *Language Testing* 11(2): 99 – 123

Shohamy, E. (1995) Performance assessment in language testing. *ARAL* 15: 188 – 211

Shohamy, E. & O. Inbar (1991) Validation of Listening comprehension tests: The effect of text and question type. *Language Testing* 8: 23 – 40

Shohamy, E., C. M. Gordon & R. Kraemer (1992) The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal* 76: 27 – 33

Skaggs, G. & R. Lissitz (1992) The consistency of detecting item bias across different test administrations: implications of another failure. *Journal of Educational Measurement* 29(3): 227 – 242

Skehan, P. (1984) Issues in the testing of English for specific purposes. *Language Testing* 1: 202 – 220

Skehan, P. (1989) Language testing Part I and Part II. *Language teaching* 22: 1 – 13

Smith, J. (1989) Topic and variation in ITA oral proficiency: SPEAK and field-specific oral tests. *English for Specific Purposes* 8: 155 – 168

Smith, R. M. (1987) Theory and practice of fit. *Rasch Measurement Transactions* 3(4): 78

Spoken language qualifications: extended syllabus (1998) London: LCCI

Spolsky, B. (1985) What does it mean to know how to use a language? *Language Testing* 2: 31 – 40

Spolsky, B. (1988) Test review: P.E. Griffin *et al.* 1986 Proficiency in English as a second language. (1) The development of an interview test for adult migrants. (2) The administration and creation of a test. (3) An interview test of English as a second language. *Language Testing* 5(1): 120 – 124

Spolsky, B. (1989) *Conditions for second language learning: introduction to a general theory.* Oxford: Oxford University Press

Spolsky, B. (1989) Competence, proficiency and beyond. *Applied Linguistics* 10(2):138-156

Spolsky B. (1995) *Measured words.* Oxford: Oxford University Press

Spurling S. (1987) Questioning the use of the Bejar method to determine unidimensionality. *Language Testing* 4(1): 93 – 95

SPSS Base 8.0: Users' Guide. (1998) Chicago: SPSS

Stansfield, C.W. (ed.) (1986) *Towards communicative competence testing: proceedings of the second TOEFL Invitational Conference.* TOEFL Research Report 21, Princeton, NJ: Educational Testing Service

Stansfield, C. W. (1993) Ethics, standards, and professionalism in language testing. *Issues in Applied Linguistics* 4(2): 189 – 206

Stansfield C.W. & D.M. Kenyon (1992) Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System* 20(3): 347 – 364

Stansfield C.W., W.M. Wu & M. van der Heide (2000) A job-relevant listening summary translation exam in Minnan. In A. J. Kunnan (ed.), pp 117 – 200

Stevenson, D.K. (1985) Authenticity, validity and a tea party. *Language Testing* 2(1): 41 – 47

Stiggins R.J. (1987) Design and development of performance assessments. *Education Measurement: Issues and Practice* 6(3): 33 – 42

Strevens, P. (1988) ESP after twenty years: a re-appraisal in Tickoo (1ed.), pp 1 – 13

Stufflebeam, D. L., C. H. McCormick, R. O. Brinkerhoff & C. O. Nelson (1985) *Conducting educational needs assessments.* Boston: Kluwer-Nijhoff Publishing

Swain, M. (1993) Second language testing and second language acquisition: Is there a conflict with traditional psychometrics? *Language Testing* 10(2): 193-207

Swales, J. (1988) *Episodes in ESP.* Cambridge: Prentice Hall International

Tarone, E. (1983) On the variability of interlanguage systems. *Applied Linguistics*, 4(2): 142-163

Tarone, E. (1988) Systematicity and attention in interlanguage. *Language Learning*, 32(1) 69-83

Taylor D.S. (1988) The meaning and use of the term ompetence' in linguistics and applied linguistics. *Applied Linguistics* 9(2): 148-168

Teasdale, A. (1996) Content validity in tests for well-defined LSP domains: an approach to defining what is to be tested. In M. Milanovic and N. Saville (eds.), pp 211 – 230

Teng, H. (1994) Effects of cultural schemata and visual cues on Chinese students' EFL listening comprehension. Paper, the Eleventh Conference on English Teaching and Learning in R.O.C.

Test of English for International Communication: Examinee handbook (1996) Princeton: ETS

The communicative use of English as a foreign language: Findings of review questionnaire and proposals for development (1989) UCLES: Cambridge

Thompson, B. & T. Vacha-Haase (2000) Psychometrics is datametrics: the test is not reliable. *Educational and Psychological Measurement* 60(2): 174 - 195

Thompson, G. (1996) Some misconceptions about communicative language teaching. *ELT Journal* 50(1): 9-15

Tickoo, M. L. (ed.) (1988) *ESP: State of the art* . Singapore: SEAMEO RLC

Turner, J. (1998) Assessing speaking. *Annual Review of Applied Linguistics* 18: 192 - 207

Tyndall B. & D.M. Kenyon (1995) Validation of a new holistic rating scale using Rasch multifaceted analysis. In Cumming A. and R. Berwick (eds.) *Validation in language testing*. Clevedon, Avon: Multilingual Matters, pp 39 – 57

Underhill, N. (1987) *Testing spoken language: A handbook of oral testing techniques.* Cambridge: Cambridge University Press

University of Cambridge Local Examinations Syndicate (1989) *The communicative use of English as a foreign language*

Upshur, J. A. & C. E. Turner (1995) Constructing rating scales for second language tests. *ELF Journal* 49(1): 3 - 12

Ur, P. (1988) *Grammar Practice Activities: a practical guide for teachers.* Cambridge: Cambridge University Press

Vollmer, H. (1983) The structure of foreign langauage competence. In A. Hughes & D. Porter (eds.), pp 3-29

Van Ek, J. (1975) Threshold level English: in a European unit/credit system for modern language learning by adults prepared for the Council of Europe. Oxford: Pergamon Press

van Lier, L. (1989) Reeling, writhing, drawling, stretching and fainting in coils: oral proficiency in interviews as conversation. *TESOL Quarterly* 23(3): 489-508

Van Patten, B. (1988) How juries get hung: problems with the evidence for a focus on form in teaching. *Language Learning* 38: 243-260

Wall, D. & J. C. Alderson. (1993) Examining washback. *Language Testing* 10(1): 41-69

Weir, C. J. (1981) Reaction to the Morrow paper (1). In *ELT documents 111 - Issues in language testing*. London: British Council.

Weir, C. J. (1988) Construct validity. In Hughes, A., D. Porter and C. Weir (eds.) *ELT Validation Project: Proceedings of a conference held to consider the ELTS Validation Project Report. English Language Testing Service Research Report 1(ii).* London: British Council/UCLES, pp15-25

Weir C.J. (1988) The specification, realization and validation of an English language proficiency test. In Hughes A. (ed.), pp 45 – 110

Weir, C.J. (1990) *Communicative language testing*. Wiltshire: Prentice Hall

Wesche, M. (1987) Second language performance testing: The Ontario Test of ESL as an example. *Language Testing* 4(2): 28 – 47

Wesche, M. (1992) Performance testing for work-related second language assessment in Shohamy, E. & R. Walton (eds.)

West, R. (1995) Needs analysis in language teaching. *Language Teaching*, 27(1): 1-19

Westaway, G., J. C. Alderson & C. M. Clapham (1990) Directions in testing for specific purposes. In J. H. A. L. de Jong & D. Stevenson (eds.), pp 239 – 256

White, R. (1988) *The ELT curriculum*. Oxford: Blackwell.

Widdowson, H.G. (1983) *Learning purpose and language use*. Oxford: Oxford University Press.

Widdowson, H. G. (1989) Knowledge of language and ability for use. *Applied Linguistics* 10(2): 128 –137

Wiggins, G. (1989) A true test: toward more authentic and equitable assessment. *Phi Delta Kappan* 70(9): 703 – 713

Wigglesworth, G. (1993) Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, 10(3): 305 – 335

Williams, J. (1995) Focus on form in communicative language teaching: Research findings and the classroom teacher. *TESOL Journal* 4:12 – 16

Wilson, K. (1992) *Relating TOEIC scores to Oral Proficiency Interview ratings. TOEIC Research Summaries No. 1.* Princeton: ETS

Wilson, M. (ed.) (1991) *Objective measurement: theory into practice.* Norwood NJ: Ablex

Wood, R. (1978) Sex differences in answers to English Language comprehension items. *Educational Studies* 4(2): 175 - 165

Woodford, P. E. (1982) *An introduction to TOEIC: the initial validity study. TOEIC Research Summaries* Princeton: ETS

Wright, B.D (1998) Interpreting reliabilities. Rasch Measurement Transactions 11(4): 602

Wright, B.D. & M.H. Stone (1979) Best Test Design. Chicago: MESA Press

Wright, B. D. & G. N. Masters (1982) *Rating scale analysis: Rasch Measurement.* Chicago: MESA Press

Young, R. & M. Milanovic (1992) Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition*: 403 – 424

Young, R. (1994) Conversational styles in language proficiency interviews. *Language Learning* 45(1): 3 – 24

Zeidner, M & M. Bensoussan (1988) College students' attitudes towards written versus oral tests of English as a foreign language. *Language Testing* 5(1): 100 – 114

Appendix 2-1: Candidate registration form

Registration number:
(Office Use Only)

Name: _____ *Sex: ____ M, _____ F

Date of birth: _____/_____/_____ I. D. Number: _____
month day year

*Education: _____ College, _____ Graduate School,

_____ other (Please specify.) _____

Present occupation: _____ How long in this job? _____

Previous occupation: _____ How long? _____

When did you start learning English?

Have you ever lived in an English-speaking country? _____ Yes, _____ No

If yes, which country did you live and how long? _____

English Language Proficiency tests taken: ____ TOEFL, Score _____

_____ IELTS, Score _____

Others: _____, Score _____

*Please check √.

Your address and phone number where you could be contacted

Address:

Phone number:

Appendix 2-2: English Language Listening and Speaking Skills Questionnaire

Name: _____ Sex: ___ M ___F

Registration No.: _____

English Language Listening and Speaking Skills Questionnaire

Instructions:

Thank you for taking part in the Tour Guide English Language (TG) pilot test. I would appreciate very much if you could take time to fill out the following questionnaire about your English listening and speaking abilities. This information will be very valuable and useful to me. All responses will be kept confidential.

In the questionnaire, you will be asked about your English language ability in listening and speaking skills. For each statement, decide if you can understand or speak in the situations/tasks described **(4) nearly always, (3) usually, (2) sometimes or (1) you cannot do** it in listening and **(4) easily, (3) with some difficulty, (2) with great difficulty** or you **(1) cannot do it** in speaking. Then circle the appropriate number.

Example:

Nearly always	**Usually**	**Sometimes**	**Cannot do it**
4	**3**	**2**	**1**

Listening

(4) 3 2 1 I can understand when told the time of the day

Easily	**With some difficulty**	**With great difficulty**	**Cannot do it**
4	**3**	**2**	**1**

Speaking

(4) 3 2 1 Give the time of the day

Some of the situations and tasks may be unfamiliar to you. If you have never been asked to do these tasks, please try to imagine if you would be able to do them, if it was necessary. Try not to leave any blanks.

Please think about these statements and fill out the questionnaire at home. Bring the completed questionnaire with you when you come to do the pilot test.

Please fill in your personal details at the top of this cover sheet. This information will be used to match your test scores. This information will be kept confidential. Again, thank you for taking time to do the pilot test and fill out this questionnaire.

I. Listening ability

	Nearly always	Usually	Sometimes	Cannot do it
	4	3	2	1

4 3 2 1 I can understand when told the time of the day and/or the days of the week.

4 3 2 1 I can understand simple questions (e.g. What is this? Where is the station? etc.).

4 3 2 1 I can understand someone's simple requests (e.g. Asking for the time, borrowing pens etc.).

4 3 2 1 I can understand questions about my name, work, address, phone number etc.

4 3 2 1 I know if an utterance is a statement, a question or a request.

4 3 2 1 When surrounded by people speaking different languages, I know who is speaking English and can catch a few phrases.

4 3 2 1 I can understand when a speaker is referring to the past, present or future in a face-to-face conversation.

4 3 2 1 I can understand and remember when someone gives his/her biographical information (e.g. date of birth, occupation, experience, and/or family members etc.).

4 3 2 1 I can understand radio advertisement (e.g. know what's being advertised, where, and at what prices if stated etc.).

4 3 2 1 I can understand details such as the place, time and persons to meet for an appointment.

4 3 2 1 On the phone, when the caller only speaks English, I can understand what the person wants.

4 3 2 1 Generally I can understand people when speaking on the phone with them.

4 3 2 1 When listening, I can understand conversation conducted at normal speed by native speakers.

4 3 2 1 In a face-to-face conversation, I can understand people I am speaking with so I do not have to ask them to speak slowly.

4 3 2 1 In a face-to-face conversation, I can understand questions asked and do not have to ask the speaker to repeat or rephrase the question(s).

4 3 2 1 I can understand rapid conversations of native speakers.

4 3 2 1 In casual conversation, I can understand everything said to me.

	Nearly always	Usually	Sometimes	Cannot do it
	4	3	2	1

4	3	2	1	I can understand when people explain regulations, or significance of historical places, people or events or his/her religious/political beliefs.
4	3	2	1	I can understand instructions directed at me in a face-to-face conversation.
4	3	2	1	In emergency, I can understand people's requests for help.
4	3	2	1	I can understand voice instructions on the phone.
4	3	2	1	I can understand a customs officer's questions.
4	3	2	1	I can understand and am able to pick out the essential points in people's conversation.
4	3	2	1	I can understand when told by a secretary that the person I am meeting with is not in the office and understand when told the expected time for his/her return.
4	3	2	1	I can understand questions asked by a tax official when filing the income tax.
4	3	2	1	I can understand questions asked by a directory operator when making a business enquiry.
4	3	2	1	When the sound is good and other noises around me do not distract me greatly, I can understand public announcements in places like the department stores, train stations or airports etc.
4	3	2	1	I can understand public announcements with poor speakers and echoes.
4	3	2	1	I can understand the factual content of news broadcasts like traffic reports, weather forecasts and major events around the world.
4	3	2	1	I can understand radio news broadcasts, commentaries, interviews (e.g. know what is happening, when, where and knows the points of view of the host and the person(s) being interviewed etc.).
4	3	2	1	I can understand play-by-play sports commentary.
4	3	2	1	I can understand what is said in a meeting concerning my own area of specialisation or interest.
4	3	2	1	I can understand main points of a speech concerning my own area of interest or specialisation.
4	3	2	1	I can understand supporting details of a speech concerning my own area of interest or specialisation.

	Nearly always	Usually	Sometimes	Cannot do it
	4	**3**	**2**	**1**

4	3	2	1	I know the relationship of characters and major developments in a film or TV programme when there is no sub-titles.
4	3	2	1	I can understand the nuance, cultural references or jokes in a film or a TV programme when there is no sub-titles.
4	3	2	1	I can understand a political speech.
4	3	2	1	I can understand most of the jokes told in a social setting.
4	3	2	1	I can follow and comprehend a debate among several speakers.
4	3	2	1	I can understand in context numbers 0 - 1,000.
4	3	2	1	I can understand in context numbers 1,000 - 100,000.
4	3	2	1	I can understand in context numbers 100,000 - 1,000,000.
4	3	2	1	Understand in context common household words (e.g. door, telephone, bathroom etc.)
4	3	2	1	Understand in context technical language concerning my own area of specialisation or interest

II. Speaking ability

	Easily		With some difficulty		With great difficulty		Cannot do it	
	4		**3**		**2**		**1**	
4	3	2	1	Give the time and date				
4	3	2	1	Say what the weather is like				
4	3	2	1	Ask how much something costs				
4	3	2	1	Tell a taxi driver where I want to go				
4	3	2	1	Ask for directions on the street				
4	3	2	1	Properly introduce myself and people in a social situation				
4	3	2	1	Describe a person or a place I am familiar with				
4	3	2	1	Describe a typical day				
4	3	2	1	Describe a person to someone so that s/he can recognise this person				
4	3	2	1	Give simple route instructions				
4	3	2	1	Give instructions on how to cook a simple Chinese dish				
4	3	2	1	Give instructions to people working with me				
4	3	2	1	Ask and answer questions about daily life occurrences				
4	3	2	1	Thank someone for their help				
4	3	2	1	Apologise for being late				
4	3	2	1	Use appropriate greetings or leave-taking expressions				
4	3	2	1	Reply appropriately when receive an apology				
4	3	2	1	Express sympathy properly when told about an unfortunate incident				
4	3	2	1	Order a meal in a restaurant and know what I will get				
4	3	2	1	Tell about my favourite author, political or sports personality etc.				
4	3	2	1	Buy clothes in a shop and discuss sizes, colours and prices etc.				
4	3	2	1	Describe symptoms to a doctor, if sick				
4	3	2	1	Book tickets for a plane, or a train through telephone				

	Easily	With some difficulty	With great difficulty	Cannot do it	
	4	3	2	1	
4	3	2	1		At an accident site, call emergency services and request for medical help
4	3	2	1		Explain how to use the public phone in my country
4	3	2	1		Describe an object (e.g., a beautiful clock, a painting, an antique vase etc.)
4	3	2	1		Tell an appropriate joke
4	3	2	1		Talk in length about a hobby or special interest I have
4	3	2	1		Tell the factual content of something that has just been reported on TV news or in the newspaper
4	3	2	1		Tell in detail about some everyday event that happened recently
4	3	2	1		Give detailed information about my future plan
4	3	2	1		If late for an appointment, can explain what happened and get another one
4	3	2	1		Make a public announcement (e.g., change of venue for a meeting or change of plans etc.)
4	3	2	1		Make a complaint to a restaurant manager about food service
4	3	2	1		Explain to a police officer about an accident just happened
4	3	2	1		Discuss at length and in detail a subject area familiar to me
4	3	2	1		Discuss current economic or political issues
4	3	2	1		Explain with examples how my country has changed over the past ten years
4	3	2	1		Talk through a complicated situation, e.g., a missed flight, a stolen wallet etc.
4	3	2	1		Make a presentation on a subject I am fairly familiar with
4	3	2	1		Explain how my country's law came about
4	3	2	1		Compare present and previous jobs, schools, etc.
4	3	2	1		Compare two relatively similar objects or ideas
4	3	2	1		Explain in detail a concept in my culture (e.g., Fengsui, religious practice, significance of a Chinese festival etc.)

	Easily		**With some difficulty**		**With great difficulty**		**Cannot do it**
	4		**3**		**2**		**1**
4	3	2	1	Explain current social, economic and/or political issues in my country			
4	3	2	1	Explain and defend my position on my country's social or political issues			
4	3	2	1	Conduct a meeting at work or school			
4	3	2	1	Take part in a panel discussion with native speakers of English			
4	3	2	1	Interpret between a doctor and his/her patient			
4	3	2	1	Interpret in a conference on a subject familiar to me			
4	3	2	1	Interpret in a law court			

Appendix 2-3: Rater profile

Name: _____ Sex: _____ Age: _____

Nationality: _____ Education: _____

Native language: _____ Present occupation: _____

1. Do you speak other languages? What are they?

2. Do you have second/foreign language teaching experience?

3. If *yes* to Question 2, where did you teach and how long did you teach there?

4. If your native language is not English, have you ever lived in an English-speaking country?

5. If *yes* to Question 4, how long have you lived there?

Appendix 2-4: Questionnaire (Test takers)

April 12, 1999

Dear test takers,

Thank you for taking part in the Tour Guide English Language Pilot Test. Before you leave the test room, you will be given a questionnaire about the pilot test. I would appreciate very much if you could answer the questions in the questionnaire and hand it to your English teacher. Your comments will be used for future revision of the Tour Guide English Language Test. Should you have any questions or doubts, please feel free to contact me at alicykp@srv0.arts.ed.ac.uk.

Again, thank you for your participation in the test.

Best wishes,

chung-yao Kao

Questionnaire

The Tour Guide English Language Test

1. The time allotment of the entire test is: (Please check √.)

_____ long _____ about right _____ short

2. Difficulty levels of the test: (Please check √)

- Generally, I think the Listening Test is:

_____ difficult _____ about right _____ easy

- Generally, I think the Use and Usage Test is:

_____ difficult _____ about right _____ easy

- Generally, I think the Speaking Test is:

_____ difficult _____ about right _____ easy

3. Question types: (*If you are not sure of the question types, please look at the examples given in the last page of the questionnaire.*)

3.1. The number of questions:

• Listening Test: (Please check √.)

I think the questions in **Identification and Labelling** are:

_____ too many _____ about right _____ few

I think the questions in **Matching** are:

_____ too many _____ about right _____ few

I think the questions in **Information Transfer** are:

_____ too many _____ about right _____ few

I think the questions in **True or False** are:

_____ too many _____ about right _____ few

I think the questions in **Sentence Completion** are:

_____ too many _____ about right _____ few

I think the questions in **Short Answers** are:

_____ too many _____ about right _____ few

• English Use and Usage Test (Please check √ .)

I think the **Multiple-choice questions** are:

_____ too many _____ about right _____ few

I think the questions in **Verb Forms** are:

_____ too many _____ about right _____ few

I think the questions in **Sentence Transformation** are:

_____ too many _____ about right _____ few

I think the questions in **Fill in the blanks** are:

_____ too many _____ about right _____ few

I think the questions in **Complete the Conversation** are:

_____ too many _____ about right _____ few

3.2. Do you like the question types? (Please check √ __Yes or __No.)

- Listening Test

I like the questions in **Identification and Labelling**

_____ Yes _____ No

I like the questions in **Matching**

_____ Yes _____ No

I like the questions in **Information Transfer**

_____ Yes _____ No

I like the questions in **True or False**

_____ Yes _____ No

I like the questions in **Sentence Completion**

_____ Yes _____ No

I like the questions in **Short Answers**

_____ Yes _____ No

- English Use and Usage Test

I like the **Multiple-choice questions**

_____ Yes _____ No

I like the questions in **Verb Forms**

_____ Yes _____ No

I like the questions on **Sentence Transformation**

_____ Yes _____ No

I like the questions in **Fill in the blanks**

_____ Yes _____ No

I like the questions in **Complete the Conversation**

_____ Yes _____ No

• Speaking Test

I like the questions in **Part II: Interpretation and Translation**

_____ Yes _____ No

I like the questions in **Part III: Answering Short Questions**

_____ Yes _____ No

I like the questions in **Part IV: Making an Announcement**

_____ Yes _____ No

I like the questions in **Part V: Solving a Problem**

_____ Yes _____ No

4. On the whole, do you like the test? (Please check √.)

_____ Yes _____ No _____ Don't know

Comments:

5. If you were to be a tour guide, do you think the test is fair? (Please check √.)

_____ Yes _____ No _____ Don't know

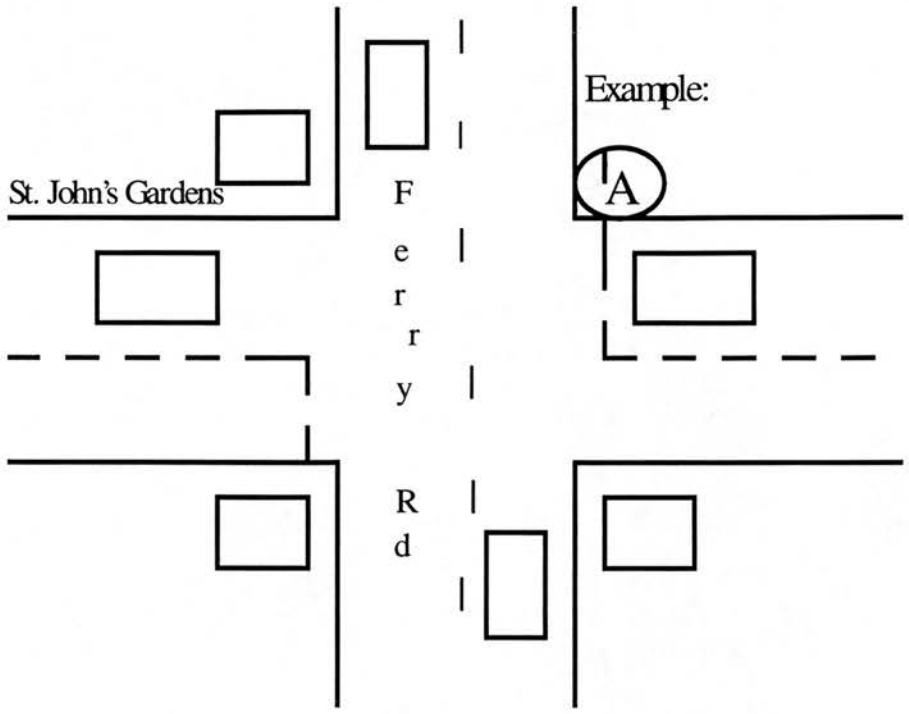
Comments:

6. If you choose "No", please tell me the type of test you like the best.

- The End -

Appendix 3--Marking scheme

Listening



4. 1891

5. 1914

6. 1930

7. 1976

8. 1926

9. Murder on the Orient Express

10. 1940

11. F

12. T

13. F

14. T

15. J

16. D

Marking criteria:

1. Each question is worth ONE point. But, for Questions 66 - 70, each correct response is worth 0.5 point. The total mark is 70 points.
2. For questions requiring constructed answers (i.e., Q26 - Q45, Q61 - Q70), suggested answers have been provided. They are by no means exhaustive. In case of ambiguities or other possible answers, please consult the chief assessor.
3. Contractions (e.g. is not = isn't) are acceptable.
4. Spelling mistakes are not acceptable.
5. Consult the chief assessor if you have any doubts on the answers provided or queries concerning the marking of the paper.

54. Well, do you want me to go to the post office for you?
I could help you post those letters (if you'd like).

55. OK, see you soon/then.
That's fine. Bye for now.

Vb: some suggested answers

56. Don't forget (to bring) your umbrella.
Remember your umbrella.

57. Would you like some (mineral) water?
Can I offer you some (mineral) water?
How about some (mineral) water?

58. If you just come this way, (you'll see the famous -----).
(The Jade Cabbage is this way) if you'd like to follow me.
This way please.
Follow me. We're going to see the Jade Cabbage.

(Answers like *Attention please. I would like to introduce something special to you.* or *Attention please.* are not acceptable.)

59. Please be quiet and respectful. It's a very solemn place for us.
Please show your respect by keeping your voices low.
Please remember to be quiet to show our respect.
Here, we must show our respect and be quiet.
It is very much appreciated if visitors are quiet and respectful here.
Shhhh. Please show respect. This is a solemn place.

(Answers like *May we lower down our voices?* or *May we be quiet please?* Are not acceptable. But an answer like *Could you please lower down your voice in this place?* is acceptable.)

60. Would you mind if I borrowed your newspaper for a moment?
Could you lend me your newspaper for a moment?
Please could I have a quick look at your newspaper?

35.Joan sends her apologies (for not attending the meeting).

36. At no time is this door to be opened/ can this door be opened.

37.The suitcase was too heavy for Tommy to lift.

IV: Fill in the blanks

38. as

39. from

40. and

41. on

42. at

43. a/another

44. the

45. how/that (meaning is different)

46. are

47. expelled/conquered/defeated/beat(?)

48. In

49. state

50. were

V: Complete the conversation

Va: (Some suggested answers) As Sandra is just an acquaintance, the candidates are expected to use appropriate registers when they talk with such a person.

51.Why don't we go to the Taipei Zoo?
Let's go to the Taipei Zoo.
How about the Taipei Zoo?

52.How about the Taipei Zoo?
I said how about the Taipei Zoo.

53.It's in Mucha, across from the Mucha MRT Station.

Marking scheme – Grammar

I.

1. C	2. D	3. A	4. B	5. C	6. C
7. C	8. D	9. D	10. D	11. D	12. C

II.

13. will/can see	14. Works/is working	15. Built
16. was called	17. Was used	18. Renamed
19. facing	20. Retains/has retained	21. is situated
22. has		

III. Sentence transformation

23. We seem to have run out of film.

24. What Emma says is different from/ says is not/ says is not the same as what she does.

25. It won't be long before Mary returns from her business trip to Dublin.

26. The fact that many people have seen the show shows how popular it is.

27. Further information can be found at the Tourism Bureau Information Center.
can be obtained from
can be had from
is available from
is provided by

28. The company car park is restricted to senior staff members only.
to be used by
for
reserved for

29. The is the statement (that) I cannot agree with.

30. John invited me for lunch.

31. I've yet to meet a more annoying person than my cousin.

32. I didn't expect the new boss to be so approachable.

33. Under no circumstances will I ever lend money to Robert.

34. John hasn't been seen for two weeks.

Part III: Answering short questions

Candidates will answer 10 short questions. They are expected to speak in complete sentences. Again raters should look for clarity, coherence and directness in candidates' responses. Circumlocution in opinions questions (e.g. Questions 5, 7, 9 and 10) could be expected from some S3 candidates.

Part IV: Making an announcement

In this part, raters should look for coherent presentation of the temporal sequence in the itinerary. Raters should also pay attention to the manner of presentation, i.e., the *overall effectiveness*. Raters are encouraged to refer to the NS speech sample for *overall effectiveness*.

Part V: Solving a Problem

In this part, raters should look for

(1) coherence and clarity of utterance:

if candidates show sympathy towards the dissatisfied client but remain firm and yet persuasive.

Marking scheme – Speaking

Part I: Warming up
Candidates will talk about themselves in 30 seconds. This part is not to be assessed.

Part II: Interpretation and Translation

In this part, the candidates are required to interpret 5 short signs (Q1 - Q5) and 5 sentences in Chinese (Q6 - Q10). Please refer to the Tour Guide Oral Proficiency levels for assessment. Suggested responses are as follows:

1. "Smoking is not allowed in this building. Please refrain from smoking."
2. "Please keep your voice down."
3. "No flash photograph or video-taping is permitted during the program/the show."
4. "Clearance sale; 70% off"
5. "Keep right."
6. We'll spend about three hours in the National Palace Museum.
7. The Change of Guards at the Martyrs' Shrine is held every hour. The next one will take place at 10.
8. The Guan-ying Statue in Chung-shang Park in Keelung City is the highest (in Taiwan). It's about 25 metres high.
9. Thank you for joining the Tainan City tour. We hope to see you (and your friends) again soon.
10. Don't forget your belongings when you leave the bus.
    Please remember to take everything with you.

In this part, raters should look for

(1) clarity, coherence and directness in candidates' responses and

(2) complete sentences for Questions 6 - 10.

Candidates should be marked down in cases of circumlocution.

17. G

18. C

19. E

20. F

21. Roof and framework

22. Horse Back

23. His social status

24. To protect the house/ ward off evil spirits

25. Early morning

26. 14

27. 11:30

28. 11:25

29. shuttle

30. Thomas Cook/ the concourse

31. in the front

32. F

33. G

34. D

35. E

36. every day/ all year round

37. free

38. noisiest

39. walk

40. July

Marking criteria:

1. Each question is worth ONE point.
2. Spelling mistakes are not to be penalised.
3. Except for proper nouns, the use of both lower and upper cases are permitted.
4. Answers in Chinese are acceptable.

Appendix 4-1: Listening item threshold

TG Listening

Item Estimates (Thresholds) In input Order 12/28/99
11:30:10
all on all (N = 112 L = 45)

ITEM	NAME	SCORE	MAXSCR	THRSH 1	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
1	item 1	15	112	1.39 .31	1.06	.87	.4	-.1
2	item 2	15	112	1.39 .31	1.09	.94	.5	0.0
3	item 3	11	112	1.81 .35	1.17	1.17	.8	.5
4	item 4	65	112	-1.46 .22	1.20	1.39	2.0	2.0
5	item 5	22	112	.82 .27	1.28	1.34	1.7	1.0
6	item 6	7	112	2.37 .42	1.06	1.87	.3	1.1
7	item 7	76	112	-2.02 .24	1.47	1.63	3.7	2.3
8	item 8	16	112	1.30 .30	.93	1.10	-.3	.4
9	item 9	20	112	.97 .28	1.05	1.33	.3	.9
10	item 10	1	112	4.49 1.02	.80	.07	.1	.1
11	item 11	20	112	.97 .28	1.14	1.24	.9	.7
12	item 12	95	112	-3.27 .30	1.42	2.09	2.1	1.9
13	item 13	88	112	-2.74 .26	.98	1.16	-.1	.6
14	item 14	59	112	-1.17 .22	1.23	1.41	2.4	2.2
15	item 15	74	112	-1.91 .23	1.47	2.40	3.8	4.7
16	item 16	64	112	-1.41 .22	1.00	1.00	0.0	.1
17	item 17	55	112	-.98 .22	1.10	1.07	1.1	.5
18	item 18	40	112	-.25 .23	.94	.86	-.6	-.6
19	item 19	50	112	-.74 .22	1.06	1.05	.7	.3
20	item 20	35	112	.02 .24	1.19	1.05	1.6	.3
21	item 21	17	112	1.21 .30	1.00	.95	.1	0.0

22	item 22	0	0	Item has zero score			
23	item 23	17	112	1.21 .30	.81	.50	-1.1 -1.2
24	item 24	10	112	1.93 .36	.72	.29	-1.2 -1.3
25	item 25	7	112	2.37 .42	.82	.33	-.5 -.8
26	item 26	10	112	1.93 .36	.89	.40	-.4 -1.0
27	item 27	65	112	-1.46 .22	1.05	1.06	.6 .4
28	item 28	81	112	-2.30 .24	.72	.73	-2.5 -1.0
29	item 29	77	112	-2.07 .24	.81	.68	-1.7 -1.4
30	item 30	23	112	.75 .27	.84	.63	-1.1 -1.1
31	item 31	37	112	-.09 .23	.95	1.01	-.4 .1
32	item 32	6	112	2.56 .45	.81	.88	-.5 .2
33	item 33	63	112	-1.36 .22	.89	.86	-1.1 -.8
34	item 34	72	112	-1.81 .23	.94	.88	-.6 -.5
35	item 35	72	112	-1.81 .23	.88	.81	-1.1 -.9
36	item 36	57	112	-1.07 .22	1.01	1.02	.2 .2
37	item 37	74	112	-1.91 .23	.91	.81	-.8 -.8
38	item 38	52	112	-.83 .22	.80	.69	-2.3 -2.0
39	item 39	52	112	-.83 .22	.88	.97	-1.3 -.1
40	item 40	46	112	-.55 .22	.88	.75	-1.3 -1.4
41	item 41	16	112	1.30 .30	.88	.52	-.6 -1.0
42	item 42	23	112	.75 .27	.93	.77	-.4 -.6
43	item 43	13	112	1.59 .33	.81	.44	-.9 -1.1
44	item 44	46	112	-.55 .22	.81	.69	-2.0 -1.9
45	item 45	14	112	1.48 .32	.79	.41	-1.0 -1.3
Mean				0.00	.99	.96	0.0 0.0
SD				1.72	.19	.46	1.4 1.3

Appendix 4-2: TG Listening person ability

Case Estimates In input Order 12/28/99
11:30:37
all on all (N = 112 L = 45)

NAME	SCORE	MAXSCR	ESTIMATE	ERROR	INFIT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
1 1001	34	44	1.85	.43	1.18	1.15	.92	.44
2 1002	31	44	1.33	.40	1.41	2.12	2.00	1.83
3 1003	12	44	-1.54	.40	1.40	1.36	1.96	.69
4 1005	38	44	2.69	.50	1.26	.72	.90	.05
5 1006	22	44	-.03	.38	.97	.95	-.06	-.01
6 1007	14	44	-1.22	.39	.75	.51	-1.39	-.82
7 1008	15	44	-1.07	.39	1.21	1.07	1.10	.31
8 1009	28	44	.86	.39	1.57	1.75	2.55	1.59
9 1010	5	44	-2.94	.52	.82	.45	-.51	0.00
10 1011	17	44	-.77	.39	1.03	.82	.21	-.25
11 1013	22	44	-.03	.38	1.06	1.20	.36	.64
12 1014	16	44	-.92	.39	1.23	1.51	1.18	1.05
13 3002	13	44	-1.38	.40	1.22	1.00	1.17	.21
14 3003	12	44	-1.54	.40	1.07	.74	.44	-.18
15 3005	10	44	-1.88	.42	1.10	1.14	.56	.44
16 3007	2	44	-4.07	.76	.82	.29	-.13	.45
17 3009	13	44	-1.38	.40	.77	.54	-1.27	-.67
18 3010	17	44	-.77	.39	.90	.78	-.45	-.35
19 3011	27	44	.71	.39	1.19	1.71	.94	1.61
20 3013	8	44	-2.25	.45	1.09	1.12	.47	.45
21 3014	20	44	-.32	.38	.74	.59	-1.31	-1.08
22 3015	14	44	-1.22	.39	.78	.64	-1.24	-.50
23 3016	9	44	-2.06	.43	.72	.41	-1.39	-.54
24 3021	21	44	-.18	.38	.95	.83	-.18	-.35
25 3022	7	44	-2.46	.47	.94	.50	-.14	-.18
26 3024	6	44	-2.68	.49	.95	1.31	-.10	.64
27 3025	7	44	-2.46	.47	.88	.45	-.40	-.26
28 3026	3	44	-3.60	.64	.83	.32	-.25	.20
29 3027	5	44	-2.94	.52	1.13	.58	.51	.13
30 3028	4	44	-3.24	.57	1.10	.50	.37	.19
31 3030	13	44	-1.38	.40	.73	.48	-1.57	-.81
32 3031	6	44	-2.68	.49	.96	4.28	-.04	1.90
33 3032	14	44	-1.22	.39	1.00	1.02	.05	.23
34 3033	14	44	-1.22	.39	1.02	1.18	.17	.48
35 3034	11	44	-1.70	.41	1.06	.71	.37	-.18
36 3035	17	44	-.77	.39	.83	.64	-.83	-.72
37 3038	18	44	-.62	.38	.57	.41	-2.54	-1.59
38 3039	8	44	-2.25	.45	.92	.51	-.31	-.26
39 3040	2	44	-4.07	.76	.82	.29	-.13	.45
40 3041	2	44	-4.07	.76	.82	.29	-.13	.45
41 3042	4	44	-3.24	.57	.94	.81	-.04	.43
42 3043	5	44	-2.94	.52	1.25	.76	.85	.30
43 3044	9	44	-2.06	.43	.88	.59	-.53	-.23
44 3045	11	44	-1.70	.41	1.75	3.43	3.26	2.26
45 3046	19	44	-.47	.38	.81	.64	-.91	-.84
46 3047	21	44	-.18	.38	1.32	1.21	1.47	.65
47 3051	20	44	-.32	.38	1.04	1.19	.25	.58
48 3052	13	44	-1.38	.40	1.01	.73	.14	-.26
49 3054	15	44	-1.07	.39	.79	.57	-1.12	-.76
50 3055	15	44	-1.07	.39	.79	.57	-1.12	-.76
51 3059	10	44	-1.88	.42	1.05	.95	.30	.22
52 3063	6	44	-2.68	.49	.94	.52	-.13	-.05
53 3064	10	44	-1.88	.42	.83	.74	-.84	-.07
54 3068	2	44	-4.07	.76	.81	.26	-.15	.43
55 3069	6	44	-2.68	.49	.79	.40	-.72	-.20
56 4001	13	44	-1.38	.40	.63	.40	-2.29	-1.02
57 4002	27	44	.71	.39	1.17	1.14	.87	.46
58 4003	19	44	-.47	.38	1.29	1.20	1.37	.58
59 4004	13	44	-1.38	.40	.93	.85	-.30	-.03
60 4005	18	44	-.62	.38	.95	.88	-.17	-.12
61 4006	22	44	-.03	.38	.83	.73	-.77	-.69
62 4007	16	44	-.92	.39	.71	.54	-1.65	-.92
63 4008	13	44	-1.38	.40	.91	.66	-.45	-.39
64 4009	10	44	-1.88	.42	.99	1.46	.03	.76
65 4010	14	44	-1.22	.39	.74	.49	-1.47	-.89

66	4011	5	44	-2.94	.52	.83	.43	-.47	-.03
67	4012	21	44	-.18	.38	.85	.75	-.70	-.60
68	4013	13	44	-1.38	.40	.71	.48	-1.65	-.81
69	4014	17	44	-.77	.39	1.38	1.30	1.80	.75
70	4015	20	44	-.32	.38	.87	.76	-.61	-.53
71	4016	6	44	-2.68	.49	1.29	1.60	1.05	.82
72	4017	22	44	-.03	.38	1.20	1.27	.96	.79
73	4018	13	44	-1.38	.40	1.03	1.26	.21	.59
74	4019	10	44	-1.88	.42	1.22	1.83	1.09	1.09
75	4020	20	44	-.32	.38	1.22	1.26	1.07	.75
76	6001	17	44	-.77	.39	.65	.48	-1.96	-1.22
77	6002	16	44	-.92	.39	.73	.51	-1.52	-1.02
78	6003	19	44	-.47	.38	1.01	.95	.13	.02
79	6004	15	44	-1.07	.39	1.03	.96	.24	.11
80	6005	26	44	.56	.39	.89	.78	-.50	-.50
81	6006	13	44	-1.38	.40	1.29	1.47	1.49	.85
82	6007	25	44	.41	.38	.45	.34	-3.31	-2.32
83	6009	17	44	-.77	.39	.83	.80	-.88	-.29
84	6010	29	44	1.01	.39	1.05	.78	.32	-.39
85	6011	11	44	-1.70	.41	.90	.79	-.50	-.05
86	6012	25	44	.41	.38	.99	1.09	.04	.36
87	6013	14	44	-1.22	.39	.97	.82	-.12	-.12
88	6014	24	44	.26	.38	1.15	1.16	.76	.54
89	6015	30	44	1.17	.40	1.03	1.20	.20	.54
90	6016	21	44	-.18	.38	.98	.92	-.01	-.08
91	6017	23	44	.12	.38	.91	.86	-.35	-.29
92	6018	11	44	-1.70	.41	.83	.73	-.88	-.14
93	6019	31	44	1.33	.40	1.00	1.18	.09	.50
94	6020	13	44	-1.38	.40	.94	.71	-.26	-.29
95	6021	4	44	-3.24	.57	.77	.34	-.56	.02
96	6022	21	44	-.18	.38	.97	1.44	-.09	1.14
97	6023	18	44	-.62	.38	.95	.87	-.20	-.17
98	6024	20	44	-.32	.38	.76	.62	-1.20	-.98
99	6025	14	44	-1.22	.39	1.18	1.17	1.00	.47
100	6026	17	44	-.77	.39	.60	.47	-2.33	-1.26
101	7001	27	44	.71	.39	1.51	1.78	2.29	1.74
102	7002	32	44	1.50	.41	1.04	1.51	.27	.98
103	7003	30	44	1.17	.40	1.35	3.32	1.72	3.24
104	7004	30	44	1.17	.40	1.27	1.11	1.38	.39
105	7005	20	44	-.32	.38	1.00	.94	.08	-.02
106	7006	5	44	-2.94	.52	.83	.43	-.47	-.03
107	7007	23	44	.12	.38	1.09	1.02	.49	.17
108	7008	34	44	1.85	.43	1.09	.87	.48	.03
109	7009	8	44	-2.25	.45	.99	.60	.01	-.13
110	7010	18	44	-.62	.38	1.08	.89	.44	-.10
111	7014	28	44	.86	.39	1.77	2.35	3.26	2.49
112	7020	14	44	-1.22	.39	1.18	1.38	.97	.78
Mean				-1.04		1.00	.96	.01	.12
SD				1.39		.23	.62	1.12	.83

Appendix 4-3: Listening item analysis

```
------------------------------------------------------------------------------
Item Analysis Results for Observed Responses                        12/28/99
11:31:29
all on all (N = 112 L = 45)
------------------------------------------------------------------------------
..............................................................................

Item    1: item 1                                 Infit MNSQ = 1.06

Categories          0         1      missing

Count              97        15         0
Percent (%)      86.6      13.4
Pt-Biserial      -.38       .38
p-value          .000      .000
Mean Ability    -1.24       .26        NA

Step Labels                  1

Thresholds                1.39
Error                      .31
..............................................................................

Item    2: item 2                                 Infit MNSQ = 1.09

Categories          0         1      missing

Count              97        15         0
Percent (%)      86.6      13.4
Pt-Biserial      -.36       .36
p-value          .000      .000
Mean Ability    -1.23       .20        NA

Step Labels                  1

Thresholds                1.39
Error                      .31
..............................................................................

Item    3: item 3                                 Infit MNSQ = 1.17

Categories          0         1      missing

Count             101        11         0
Percent (%)      90.2       9.8
Pt-Biserial      -.26       .26
p-value          .002      .002
Mean Ability    -1.15       .04        NA

Step Labels                  1

Thresholds                1.81
Error                      .35
..............................................................................

Item    4: item 4                                 Infit MNSQ = 1.20

Categories          0         1      missing

Count              47        65         0
Percent (%)      42.0      58.0
Pt-Biserial      -.35       .35
p-value          .000      .000
Mean Ability    -1.63      -.61        NA

Step Labels                  1

Thresholds               -1.46
Error                      .22
..............................................................................

Item    5: item 5                                 Infit MNSQ = 1.28
```

Categories	0	1	missing
Count	90	22	0
Percent (%)	80.4	19.6	
Pt-Biserial	-.28	.28	
p-value	.001	.001	
Mean Ability	-1.23	-.25	NA
Step Labels		1	
Thresholds		.82	
Error		.27	

Item 6: item 6 Infit MNSQ = 1.06

Categories	0	1	missing
Count	105	7	0
Percent (%)	93.8	6.3	
Pt-Biserial	-.29	.29	
p-value	.001	.001	
Mean Ability	-1.13	.38	NA
Step Labels		1	
Thresholds		2.37	
Error		.42	

Item 7: item 7 Infit MNSQ = 1.47

Categories	0	1	missing
Count	36	76	0
Percent (%)	32.1	67.9	
Pt-Biserial	-.19	.19	
p-value	.020	.020	
Mean Ability	-1.44	-.85	NA
Step Labels		1	
Thresholds		-2.02	
Error		.24	

Item 8: item 8 Infit MNSQ = .93

Categories	0	1	missing
Count	96	16	0
Percent (%)	85.7	14.3	
Pt-Biserial	-.41	.41	
p-value	.000	.000	
Mean Ability	-1.26	.32	NA
Step Labels		1	
Thresholds		1.30	
Error		.30	

Item 9: item 9 Infit MNSQ = 1.05

Categories	0	1	missing
Count	92	20	0
Percent (%)	82.1	17.9	
Pt-Biserial	-.37	.37	
p-value	.000	.000	
Mean Ability	-1.27	.04	NA
Step Labels		1	
Thresholds		.97	
Error		.28	

```
Item   10: item 10                              Infit MNSQ =  .80

Categories          0         1      missing

Count             111         1            0
Percent (%)      99.1        .9
Pt-Biserial      -.26       .26
p-value          .003      .003
Mean Ability    -1.07      2.69           NA

Step Labels                  1

Thresholds                4.49
Error                     1.02

.....................................................................

Item   11: item 11                              Infit MNSQ = 1.14

Categories          0         1      missing

Count              92        20            0
Percent (%)      82.1      17.9
Pt-Biserial      -.34       .34
p-value          .000      .000
Mean Ability    -1.25      -.05           NA

Step Labels                  1

Thresholds                 .97
Error                      .28
.....................................................................

Item   12: item 12                              Infit MNSQ = 1.42

Categories          0         1      missing

Count              17        95            0
Percent (%)      15.2      84.8
Pt-Biserial      -.17       .17
p-value          .040      .040
Mean Ability    -1.53      -.95           NA

Step Labels                  1

Thresholds               -3.27
Error                     .30
.....................................................................

Item   13: item 13                              Infit MNSQ =  .98

Categories          0         1      missing

Count              24        88            0
Percent (%)      21.4      78.6
Pt-Biserial      -.40       .40
p-value          .000      .000
Mean Ability    -2.20      -.72           NA

Step Labels                  1

Thresholds               -2.74
Error                     .26
.....................................................................

Item   14: item 14                              Infit MNSQ = 1.23

Categories          0         1      missing

Count              53        59            0
Percent (%)      47.3      52.7
Pt-Biserial      -.35       .35
p-value          .000      .000
Mean Ability    -1.56      -.57           NA

Step Labels                  1
```

```
Thresholds                -1.17
Error                       .22
.....................................................................

Item   15: item 15                          Infit MNSQ = 1.47

Categories          0         1      missing

Count              38        74         0
Percent (%)      33.9      66.1
Pt-Biserial      -.14       .14
p-value          .076      .076
Mean Ability    -1.30      -.90        NA

Step Labels                 1

Thresholds                -1.91
Error                       .23
.....................................................................

Item   16: item 16                          Infit MNSQ = 1.00

Categories          0         1      missing

Count              48        64         0
Percent (%)      42.9      57.1
Pt-Biserial      -.49       .49
p-value          .000      .000
Mean Ability    -1.85      -.43        NA

Step Labels                 1

Thresholds                -1.41
Error                       .22
.....................................................................

Item   17: item 17                          Infit MNSQ = 1.10

Categories          0         1      missing

Count              57        55         0
Percent (%)      50.9      49.1
Pt-Biserial      -.44       .44
p-value          .000      .000
Mean Ability    -1.65      -.40        NA

Step Labels                 1

Thresholds                -.98
Error                       .22
.....................................................................

Item   18: item 18                          Infit MNSQ =  .94

Categories          0         1      missing

Count              72        40         0
Percent (%)      64.3      35.7
Pt-Biserial      -.55       .55
p-value          .000      .000
Mean Ability    -1.59      -.04        NA

Step Labels                 1

Thresholds                -.25
Error                       .23
.....................................................................

Item   19: item 19                          Infit MNSQ = 1.06

Categories          0         1      missing

Count              62        50         0
Percent (%)      55.4      44.6
Pt-Biserial      -.47       .47
p-value          .000      .000
```

```
Mean Ability      -1.62       -.31         NA

Step Labels                    1

Thresholds                   -.74
Error                         .22
.....................................................................

Item   20: item 20                              Infit MNSQ = 1.19

Categories           0          1       missing

Count                77         35          0
Percent (%)        68.8       31.3
Pt-Biserial        -.38        .38
p-value            .000       .000
Mean Ability      -1.40       -.23         NA

Step Labels                    1

Thresholds                    .02
Error                         .24
.....................................................................

Item   21: item 21                              Infit MNSQ = 1.00

Categories           0          1       missing

Count                95         17          0
Percent (%)        84.8       15.2
Pt-Biserial        -.41        .41
p-value            .000       .000
Mean Ability      -1.27        .27         NA

Step Labels                    1

Thresholds                   1.21
Error                         .30
.....................................................................

Item   22: item 22                              Infit MNSQ =  .00

Categories           0          1       missing

Count               112          0          0
Percent (%)       100.0         .0
Pt-Biserial         .00         NA
p-value            .500        NA
Mean Ability      -1.04         NA         NA

Step Labels

Thresholds
Error
.....................................................................

Item   23: item 23                              Infit MNSQ =  .81

Categories           0          1       missing

Count                95         17          0
Percent (%)        84.8       15.2
Pt-Biserial        -.56        .56
p-value            .000       .000
Mean Ability      -1.35        .71         NA

Step Labels                    1

Thresholds                   1.21
Error                         .30
.....................................................................

Item   24: item 24                              Infit MNSQ =  .72

Categories           0          1       missing

Count              102         10          0
```

Percent (%)	91.1	8.9	
Pt-Biserial	-.55	.55	
p-value	.000	.000	
Mean Ability	-1.26	1.26	NA
Step Labels		1	
Thresholds		1.93	
Error		.36	

Item 25: item 25 Infit MNSQ = .82

Categories	0	1	missing
Count	105	7	0
Percent (%)	93.8	6.3	
Pt-Biserial	-.44	.44	
p-value	.000	.000	
Mean Ability	-1.19	1.23	NA
Step Labels		1	
Thresholds		2.37	
Error		.42	

Item 26: item 26 Infit MNSQ = .89

Categories	0	1	missing
Count	102	10	0
Percent (%)	91.1	8.9	
Pt-Biserial	-.47	.47	
p-value	.000	.000	
Mean Ability	-1.23	.92	NA
Step Labels		1	
Thresholds		1.93	
Error		.36	

Item 27: item 27 Infit MNSQ = 1.05

Categories	0	1	missing
Count	47	65	0
Percent (%)	42.0	58.0	
Pt-Biserial	-.45	.45	
p-value	.000	.000	
Mean Ability	-1.81	-.48	NA
Step Labels		1	
Thresholds		-1.46	
Error		.22	

Item 28: item 28 Infit MNSQ = .72

Categories	0	1	missing
Count	31	81	0
Percent (%)	27.7	72.3	
Pt-Biserial	-.59	.59	
p-value	.000	.000	
Mean Ability	-2.44	-.50	NA
Step Labels		1	
Thresholds		-2.30	
Error		.24	

Item 29: item 29 Infit MNSQ = .81

```
Categories          0         1      missing

Count              35        77         0
Percent (%)      31.3      68.8
Pt-Biserial      -.58       .58
p-value          .000      .000
Mean Ability    -2.28      -.47        NA

Step Labels                  1

Thresholds                -2.07
Error                       .24
```

```
Item   30: item 30                          Infit MNSQ =  .84

Categories          0         1      missing

Count              89        23         0
Percent (%)      79.5      20.5
Pt-Biserial      -.58       .58
p-value          .000      .000
Mean Ability    -1.42       .46        NA

Step Labels                  1

Thresholds                 .75
Error                      .27
```

```
Item   31: item 31                          Infit MNSQ =  .95

Categories          0         1      missing

Count              75        37         0
Percent (%)      67.0      33.0
Pt-Biserial      -.52       .52
p-value          .000      .000
Mean Ability    -1.53      -.03        NA

Step Labels                  1

Thresholds                -.09
Error                      .23
```

```
Item   32: item 32                          Infit MNSQ =  .81

Categories          0         1      missing

Count             106         6         0
Percent (%)      94.6       5.4
Pt-Biserial      -.38       .38
p-value          .000      .000
Mean Ability    -1.16      1.07        NA

Step Labels                  1

Thresholds                2.56
Error                      .45
```

```
Item   33: item 33                          Infit MNSQ =  .89

Categories          0         1      missing

Count              49        63         0
Percent (%)      43.8      56.3
Pt-Biserial      -.57       .57
p-value          .000      .000
Mean Ability    -1.93      -.34        NA

Step Labels                  1

Thresholds                -1.36
Error                      .22
```

Item 34: item 34 Infit MNSQ = .94

Categories	0	1	missing
Count	40	72	0
Percent (%)	35.7	64.3	
Pt-Biserial	-.52	.52	
p-value	.000	.000	
Mean Ability	-2.04	-.48	NA
Step Labels		1	
Thresholds		-1.81	
Error		.23	

Item 35: item 35 Infit MNSQ = .88

Categories	0	1	missing
Count	40	72	0
Percent (%)	35.7	64.3	
Pt-Biserial	-.56	.56	
p-value	.000	.000	
Mean Ability	-2.10	-.44	NA
Step Labels		1	
Thresholds		-1.81	
Error		.23	

Item 36: item 36 Infit MNSQ = 1.01

Categories	0	1	missing
Count	55	57	0
Percent (%)	49.1	50.9	
Pt-Biserial	-.49	.49	
p-value	.000	.000	
Mean Ability	-1.74	-.36	NA
Step Labels		1	
Thresholds		-1.07	
Error		.22	

Item 37: item 37 Infit MNSQ = .91

Categories	0	1	missing
Count	38	74	0
Percent (%)	33.9	66.1	
Pt-Biserial	-.54	.54	
p-value	.000	.000	
Mean Ability	-2.11	-.49	NA
Step Labels		1	
Thresholds		-1.91	
Error		.23	

Item 38: item 38 Infit MNSQ = .80

Categories	0	1	missing
Count	60	52	0
Percent (%)	53.6	46.4	
Pt-Biserial	-.64	.64	
p-value	.000	.000	
Mean Ability	-1.85	-.09	NA
Step Labels		1	

```
Thresholds                -.83
Error                      .22
.....................................................................

Item   39: item 39                          Infit MNSQ =  .88

Categories          0         1      missing

Count              60        52         0
Percent (%)      53.6      46.4
Pt-Biserial      -.56       .56
p-value          .000      .000
Mean Ability    -1.76      -.20        NA

Step Labels                1

Thresholds                -.83
Error                      .22
.....................................................................

Item   40: item 40                          Infit MNSQ =  .88

Categories          0         1      missing

Count              66        46         0
Percent (%)      58.9      41.1
Pt-Biserial      -.59       .59
p-value          .000      .000
Mean Ability    -1.71      -.06        NA

Step Labels                1

Thresholds                -.55
Error                      .22
.....................................................................

Item   41: item 41                          Infit MNSQ =  .88

Categories          0         1      missing

Count              96        16         0
Percent (%)      85.7      14.3
Pt-Biserial      -.52       .52
p-value          .000      .000
Mean Ability    -1.32       .64        NA

Step Labels                1

Thresholds                1.30
Error                      .30
.....................................................................

Item   42: item 42                          Infit MNSQ =  .93

Categories          0         1      missing

Count              89        23         0
Percent (%)      79.5      20.5
Pt-Biserial      -.51       .51
p-value          .000      .000
Mean Ability    -1.38       .30        NA

Step Labels                1

Thresholds                 .75
Error                      .27
.....................................................................

Item   43: item 43                          Infit MNSQ =  .81

Categories          0         1      missing

Count              99        13         0
Percent (%)      88.4      11.6
Pt-Biserial      -.53       .53
p-value          .000      .000
Mean Ability    -1.29       .89        NA
```

```
Step Labels                  1

Thresholds                1.59
Error                      .33
------------------------------------------------------------------------------

Item   44: item 44                          Infit MNSQ =  .81

Categories          0         1      missing

Count                66        46         0
Percent (%)        58.9      41.1
Pt-Biserial        -.64       .64
p-value            .000      .000
Mean Ability      -1.76      0.00        NA

Step Labels                  1

Thresholds                -.55
Error                      .22
------------------------------------------------------------------------------

Item   45: item 45                          Infit MNSQ =  .79

Categories          0         1      missing

Count                98        14         0
Percent (%)        87.5      12.5
Pt-Biserial        -.56       .56
p-value            .000      .000
Mean Ability      -1.31       .89        NA

Step Labels                  1

Thresholds                1.48
Error                      .32
------------------------------------------------------------------------------
........

Mean test score       15.87
Standard deviation     8.17
Internal Consistency    .91

The individual item statistics are calculated
using all available data.

The overall mean, standard deviation and internal
consistency indices assume that missing responses
are incorrect.  They should only be considered useful when
there is only a limited amount of missing data.
==============================================================================
```

Appendix 4-4: Grammar item threshold

Item Estimates (Thresholds) In input Order
04/05/00
all on all (N = 112 L = 65)

	ITEM NAME	SCORE	MAXSCR	THRSH 1	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
1	item 1	30	112	-.07 .23	1.24	1.39	2.1	2.2
2	item 2	18	111	.59 .27	1.20	1.52	1.2	1.9
3	item 3	40	111	-.58 .21	1.08	1.14	1.0	1.2
4	item 4	62	112	-1.46 .21	1.09	1.10	1.3	1.0
5	item 5	34	112	-.27 .22	1.16	1.34	1.7	2.2
6	item 6	95	112	-3.17 .28	.94	2.02	-.3	2.9
7	item 7	47	112	-.84 .21	1.42	1.57	5.3	5.0
8	item 8	36	112	-.36 .22	1.16	1.22	1.7	1.6
9	item 9	34	112	-.27 .22	1.00	.98	0.0	-.1
10	item 10	10	112	1.36 .34	1.19	2.97	.8	3.5
11	item 11	53	112	-1.09 .20	.99	1.06	-.1	.7
12	item 12	31	112	-.12 .22	1.09	1.10	.9	.7
13	item 13	62	112	-1.46 .21	1.06	1.25	.9	2.4
14	item 14	58	112	-1.29 .20	.89	.85	-1.6	-1.7
15	item 15	36	112	-.36 .22	1.14	1.14	1.5	1.0
16	item 16	48	112	-.88 .21	.95	.91	-.7	-.9
17	item 17	60	112	-1.37 .21	.88	.83	-1.8	-1.9

18	item 18	93	112	-3.02 .27	1.12	1.26	.8	1.0
19	item 19	43	112	-.67 .21	1.00	.99	0.0	0.0
20	item 20	45	112	-.75 .21	.92	.88	-1.0	-1.1
21	item 21	11	112	1.25 .33	.97	.99	0.0	.1
22	item 22	68	112	-1.71 .21	1.05	1.17	.7	1.6
23	item 23	1	112	3.80 1.01	.85	.09	.2	-.7
24	item 24	38	112	-.45 .21	1.00	.90	0.0	-.8
25	item 25	1	112	3.80 1.01	.85	.09	.2	-.7
26	item 26	42	112	-.63 .21	1.11	1.20	1.4	1.8
27	item 27	10	112	1.36 .34	.92	.76	-.2	-.5
28	item 28	11	112	1.25 .33	.98	.68	0.0	-.9
29	item 29	39	112	-.50 .21	.98	.95	-.2	-.3
30	item 30	59	112	-1.33 .20	1.13	1.15	1.9	1.6
31	item 31	2	112	3.10 .72	.88	.42	0.0	-.4
32	item 32	0	0	Item has zero score				
33	item 33	3	112	2.68 .59	.87	.35	-.1	-.8
34	item 34	16	112	.79 .28	1.13	1.05	.7	.3
35	item 35	0	0	Item has zero score				
36	item 36	4	112	2.37 .52	.94	.88	0.0	0.0
37	item 37	19	112	.57 .26	1.01	1.00	.1	.1

38	item 38	26	112	.14 .24	.91	.84	-.7	-.9
39	item 39	70	112	-1.80 .21	.96	1.00	-.5	.1
40	item 40	45	112	-.75 .21	1.07	1.15	1.0	1.5
41	item 41	38	112	-.45 .21	1.15	1.26	1.8	1.9
42	item 42	18	112	.64 .27	.94	.77	-.3	-.9
43	item 43	27	112	.09 .24	.96	.88	-.3	-.6
44	item 44	42	112	-.63 .21	1.03	1.15	.4	1.3
45	item 45	38	112	-.45 .21	.98	.97	-.2	-.2
46	item 46	42	112	-.63 .21	.99	1.00	-.2	0.0
47	item 47	21	112	.44 .25	1.03	.99	.2	0.0
48	item 48	55	112	-1.17 .20	1.04	1.05	.6	.6
49	item 49	3	112	2.68 .59	.86	1.05	-.1	.3
50	item 50	41	112	-.58 .21	.93	.89	-.9	-1.0
51	item 51	68	112	-1.71 .21	.85	.79	-2.0	-2.1
52	item 52	45	112	-.75 .21	.98	.99	-.3	-.1
53	item 53	29	112	-.02 .23	.95	.91	-.4	-.5
54	item 54	22	112	.38 .25	.94	1.12	-.4	.6
55	item 55	62	112	-1.46 .21	.89	.89	-1.6	-1.1
56	item 56	42	112	-.63 .21	.79	.75	-2.9	-2.5
57	item 57	34	112	-.27 .22	.82	.74	-2.0	-2.0

58	item 58	27	112	.09 .24	.83	.69	-1.5	-1.8
59	item 59	16	112	.79 .28	.89	.64	-.5	-1.4
60	item 60	21	112	.44 .25	.88	.82	-.8	-.8
61	item 61	12	112	1.15 .32	.91	1.33	-.3	1.0
62	item 62	18	112	.64 .27	.79	.55	-1.3	-2.0
63	item 63	9	112	1.48 .36	.82	.46	-.6	-1.5
64	item 64	17	112	.72 .28	.83	.69	-1.0	-1.2
65	item 65	10	111	1.32 .34	.87	.55	-.4	-1.3
Mean				0.00	.98	.99	0.0	.1
SD				1.41	.12	.41	1.2	1.5

Appendix 4-5: Grammar person measures

Case Estimates In input Order 04/05/00
all on all (N = 112 L = 65)

NAME	SCORE	MAXSCR	ESTIMATE	ERROR	INFIT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
1 1001	40	62	.70	.31	1.49	1.83	2.92	2.56
2 1002	29	63	-.29	.29	1.01	1.16	.11	.64
3 1003	0	63	Case has zero score					
4 1005	47	61	1.59	.36	2.10	4.49	4.01	4.87
5 1006	34	63	.13	.29	1.02	.94	.19	-.17
6 1007	25	63	-.62	.29	1.08	1.16	.74	.59
7 1008	24	63	-.71	.29	1.20	1.22	1.69	.74
8 1009	33	63	.05	.29	.93	.88	-.58	-.41
9 1010	18	63	-1.25	.31	1.04	1.37	.31	.91
10 1011	19	63	-1.16	.31	.93	.76	-.47	-.50
11 1013	24	63	-.71	.29	.81	.75	-1.78	-.76
12 1014	29	63	-.29	.29	.92	1.00	-.67	.09
13 4001	16	63	-1.45	.32	1.00	1.04	.06	.25
14 4002	20	63	-1.07	.30	.98	.86	-.11	-.25
15 4003	22	63	-.88	.30	.82	.66	-1.58	-1.00
16 4004	26	63	-.54	.29	.88	.91	-1.03	-.21
17 4005	18	63	-1.25	.31	1.00	.86	.04	-.19
18 4006	16	63	-1.45	.32	1.06	1.12	.42	.40
19 4007	15	63	-1.56	.33	.90	.86	-.56	-.12
20 4008	9	63	-2.32	.39	.86	.61	-.51	-.36
21 4009	17	63	-1.35	.32	1.08	1.15	.58	.47
22 4010	8	63	-2.48	.41	.90	.66	-.31	-.21
23 4011	9	63	-2.32	.39	1.12	1.40	.54	.72
24 4012	22	63	-.88	.30	1.10	.95	.88	-.03
25 4013	8	63	-2.48	.41	.96	1.00	-.05	.27
26 4014	17	63	-1.35	.32	.89	.78	-.70	-.39
27 4015	16	63	-1.45	.32	1.00	.97	.04	.09
28 4016	29	63	-.29	.29	.85	.77	-1.35	-.83
29 4017	26	63	-.54	.29	1.12	1.02	1.09	.16
30 4018	8	63	-2.48	.41	.96	.66	-.04	-.21
31 4019	11	63	-2.03	.36	.95	.91	-.15	.08
32 4020	17	63	-1.35	.32	1.03	.98	.28	.11
33 6001	14	63	-1.67	.34	.96	.79	-.16	-.26
34 6002	16	63	-1.45	.32	.98	.80	-.07	-.31
35 6003	27	63	-.46	.29	.98	.87	-.18	-.40
36 6004	13	63	-1.78	.34	.88	.66	-.62	-.50
37 6005	17	63	-1.35	.32	.93	.72	-.44	-.54
38 6006	17	63	-1.35	.32	.87	.69	-.86	-.62
39 6007	27	63	-.46	.29	1.00	.88	.06	-.33
40 6009	24	63	-.71	.29	1.08	1.09	.75	.36
41 6010	23	63	-.80	.30	1.08	1.02	.67	.18
42 6011	20	63	-1.07	.30	.99	.92	-.03	-.08
43 6012	19	63	-1.16	.31	.97	.87	-.20	-.20
44 6013	28	63	-.37	.29	.82	.70	-1.65	-1.13
45 6014	20	63	-1.07	.30	.90	.85	-.72	-.28
46 6015	21	63	-.97	.30	.82	.66	-1.52	-.93
47 6016	29	63	-.29	.29	.82	.69	-1.73	-1.23
48 6017	11	63	-2.03	.36	1.04	.84	.25	-.04
49 6018	24	63	-.71	.29	.76	.61	-2.28	-1.30
50 6019	16	63	-1.45	.32	.74	.54	-1.79	-1.01
51 6020	22	63	-.88	.30	.98	.86	-.13	-.30
52 6021	11	63	-2.03	.36	1.01	.79	.12	-.14
53 6022	14	63	-1.67	.34	.80	.56	-1.18	-.82
54 6023	26	63	-.54	.29	.92	.80	-.73	-.63
55 6024	27	63	-.46	.29	.84	.71	-1.49	-1.03
56 6025	14	63	-1.67	.34	.75	.52	-1.54	-.92
57 6026	11	63	-2.03	.36	.74	.49	-1.28	-.78
58 7001	33	63	.05	.29	.94	.84	-.51	-.58
59 7002	38	63	.48	.30	1.05	.95	.44	-.12
60 7003	27	63	-.46	.29	1.15	1.10	1.32	.45
61 7004	26	63	-.54	.29	1.10	1.06	.92	.30
62 7005	23	63	-.80	.30	1.34	1.32	2.68	.97
63 7006	5	63	-3.08	.50	1.07	1.20	.30	.54
64 7007	34	63	.13	.29	.83	1.19	-1.43	.80
65 7008	37	63	.39	.30	1.15	1.05	1.13	.29

66	7009	25	63	-.62	.29	1.00	.85	.04	-.40
67	7010	23	63	-.80	.30	1.14	1.10	1.19	.39
68	7014	28	63	-.37	.29	1.14	1.09	1.23	.42
69	7020	12	63	-1.90	.35	1.00	2.32	.09	1.81
70	3002	10	63	-2.17	.38	.95	.71	-.13	-.24
71	3003	10	63	-2.17	.38	.96	.72	-.09	-.22
72	3005	3	63	-3.69	.62	1.01	3.01	.20	1.38
73	3007	11	63	-2.03	.36	1.08	.88	.42	.02
74	3009	30	63	-.20	.29	1.06	.99	.55	.06
75	3010	13	63	-1.78	.34	1.20	1.45	1.07	.89
76	3011	31	63	-.12	.29	1.00	.90	.08	-.32
77	3013	26	63	-.54	.29	1.14	1.06	1.22	.31
78	3014	23	63	-.80	.30	.93	.90	-.58	-.21
79	3015	15	63	-1.56	.33	.92	.78	-.43	-.32
80	3016	8	63	-2.48	.41	1.03	.73	.21	-.09
81	3021	26	63	-.54	.29	1.18	1.42	1.52	1.36
82	3022	23	63	-.80	.30	.94	.80	-.53	-.53
83	3024	15	63	-1.56	.33	.78	.57	-1.41	-.87
84	3025	16	63	-1.45	.32	.83	.63	-1.12	-.76
85	3026	14	63	-1.67	.34	1.12	.93	.74	.04
86	3027	11	63	-2.03	.36	1.24	1.14	1.11	.43
87	3028	16	63	-1.45	.32	.88	.67	-.75	-.64
88	3030	20	63	-1.07	.30	1.17	1.42	1.29	1.10
89	3031	11	63	-2.03	.36	1.20	1.50	.94	.89
90	3032	13	63	-1.78	.34	1.16	1.20	.90	.52
91	3033	23	63	-.80	.30	.97	.96	-.21	-.01
92	3034	15	63	-1.56	.33	1.09	1.15	.60	.45
93	3035	24	63	-.71	.29	1.00	.98	.07	.05
94	3038	18	63	-1.25	.31	.96	.73	-.24	-.55
95	3039	8	63	-2.48	.41	1.17	.83	.67	.05
96	3040	12	63	-1.90	.35	.78	.57	-1.11	-.64
97	3041	11	63	-2.03	.36	.94	.75	-.24	-.20
98	3042	28	63	-.37	.29	1.00	.91	.08	-.24
99	3043	22	63	-.88	.30	1.16	1.05	1.34	.26
100	3044	21	63	-.97	.30	.91	.77	-.68	-.56
101	3045	20	63	-1.07	.30	1.07	.91	.54	-.12
102	3046	19	63	-1.16	.31	1.07	.94	.55	-.03
103	3047	26	63	-.54	.29	1.03	1.55	.27	1.70
104	3051	17	63	-1.35	.32	1.05	1.01	.39	.17
105	3052	19	63	-1.16	.31	.76	.62	-1.89	-.97
106	3054	8	63	-2.48	.41	1.15	1.47	.62	.78
107	3055	7	63	-2.65	.43	1.13	1.59	.51	.86
108	3059	8	63	-2.48	.41	1.03	.70	.19	-.14
109	3063	13	63	-1.78	.34	1.14	.89	.77	-.01
110	3064	19	63	-1.16	.31	.89	.76	-.82	-.50
111	3068	11	63	-2.03	.36	.85	.58	-.65	-.55
112	3069	9	63	-2.32	.39	1.08	.71	.38	-.18
Mean				-1.20		1.01	.99	.01	.01
SD				.85		.17	.48	1.01	.81

Appendix 4-6: Grammar item analysis

```
-------------------------------------------------------------------------------
Item Analysis Results for Observed Responses                    04/05/00
all on all (N = 112 L = 65)
-------------------------------------------------------------------------------
Item    1: item 1                                  Infit MNSQ = 1.24

Categories          0         1      missing

Count              82        30            0
Percent (%)      73.2      26.8
Pt-Biserial      -.04       .04
p-value          .333      .333
Mean Ability    -1.22     -1.14           NA

Step Labels                   1

Thresholds                 -.07
Error                       .23
.....................................................................
Item    2: item 2                                  Infit MNSQ = 1.20

Categories          0         1      missing

Count              93        18            1
Percent (%)      83.8      16.2
Pt-Biserial       .01      -.01
p-value          .448      .448
Mean Ability    -1.22     -1.24         1.59

Step Labels                   1

Thresholds                  .59
Error                       .27
.....................................................................
Item    3: item 3                                  Infit MNSQ = 1.08

Categories          0         1      missing

Count              71        40            1
Percent (%)      64.0      36.0
Pt-Biserial      -.25       .25
p-value          .004      .004
Mean Ability    -1.37     -.97          1.59

Step Labels                   1

Thresholds                 -.58
Error                       .21
.....................................................................
Item    4: item 4                                  Infit MNSQ = 1.09

Categories          0         1      missing

Count              50        62            0
Percent (%)      44.6      55.4
Pt-Biserial      -.27       .27
p-value          .002      .002
Mean Ability    -1.45     -1.00           NA

Step Labels                   1

Thresholds                -1.46
Error                       .21
.....................................................................
Item    5: item 5                                  Infit MNSQ = 1.16

Categories          0         1      missing

Count              78        34            0
Percent (%)      69.6      30.4
Pt-Biserial      -.15       .15
p-value          .063      .063
```

```
Mean Ability      -1.27      -1.04         NA

Step Labels                    1

Thresholds                  -.27
Error                        .22
.....................................................................
Item    6: item 6                              Infit MNSQ =  .94

Categories           0         1      missing

Count               17        95         0
Percent (%)       15.2      84.8
Pt-Biserial       -.23       .23
p-value           .007      .007
Mean Ability      -1.70     -1.11         NA

Step Labels                    1

Thresholds                 -3.17
Error                        .28
.....................................................................
Item    7: item 7                              Infit MNSQ = 1.42

Categories           0         1      missing

Count               65        47         0
Percent (%)       58.0      42.0
Pt-Biserial        .12       -.12
p-value           .107      .107
Mean Ability      -1.11     -1.32         NA

Step Labels                    1

Thresholds                  -.84
Error                        .21
.....................................................................
Item    8: item 8                              Infit MNSQ = 1.16

Categories           0         1      missing

Count               76        36         0
Percent (%)       67.9      32.1
Pt-Biserial       -.16       .16
p-value           .044      .044
Mean Ability      -1.29     -1.01         NA

Step Labels                    1

Thresholds                  -.36
Error                        .22
.....................................................................
Item    9: item 9                              Infit MNSQ = 1.00

Categories           0         1      missing

Count               78        34         0
Percent (%)       69.6      30.4
Pt-Biserial       -.35       .35
p-value           .000      .000
Mean Ability      -1.39     -.76          NA

Step Labels                    1

Thresholds                  -.27
Error                        .22
.....................................................................
Item   10: item 10                             Infit MNSQ = 1.19

Categories           0         1      missing

Count              102        10         0
Percent (%)       91.1       8.9
Pt-Biserial        .07       -.07
p-value           .224      .224
Mean Ability      -1.17     -1.48         NA
```

```
Step Labels                 1

Thresholds                1.36
Error                      .34
.....................................................................
Item   11: item 11                          Infit MNSQ =  .99

Categories          0         1      missing

Count              59        53         0
Percent (%)      52.7      47.3
Pt-Biserial      -.37       .37
p-value          .000      .000
Mean Ability    -1.49      -.88        NA

Step Labels                 1

Thresholds               -1.09
Error                      .20
.....................................................................

Item   12: item 12                          Infit MNSQ = 1.09

Categories          0         1      missing

Count              81        31         0
Percent (%)      72.3      27.7
Pt-Biserial      -.24       .24
p-value          .006      .006
Mean Ability    -1.32      -.88        NA

Step Labels                 1

Thresholds               -.12
Error                      .22
.....................................................................
Item   13: item 13                          Infit MNSQ = 1.06

Categories          0         1      missing

Count              50        62         0
Percent (%)      44.6      55.4
Pt-Biserial      -.27       .27
p-value          .002      .002
Mean Ability    -1.45     -1.00        NA

Step Labels                 1

Thresholds               -1.46
Error                      .21
.....................................................................
Item   14: item 14                          Infit MNSQ =  .89

Categories          0         1      missing

Count              54        58         0
Percent (%)      48.2      51.8
Pt-Biserial      -.49       .49
p-value          .000      .000
Mean Ability    -1.63      -.81        NA

Step Labels                 1

Thresholds               -1.29
Error                      .20
.....................................................................
Item   15: item 15                          Infit MNSQ = 1.14

Categories          0         1      missing

Count              76        36         0
Percent (%)      67.9      32.1
Pt-Biserial      -.19       .19
p-value          .020      .020
Mean Ability    -1.32      -.95        NA

Step Labels                 1
```

```
Thresholds                -.36
Error                      .22
.....................................................................
Item   16: item 16                          Infit MNSQ =  .95

Categories          0         1      missing

Count              64        48         0
Percent (%)      57.1      42.9
Pt-Biserial      -.43       .43
p-value          .000      .000
Mean Ability    -1.51      -.79        NA

Step Labels                1

Thresholds                -.88
Error                      .21
.....................................................................
Item   17: item 17                          Infit MNSQ =  .88

Categories          0         1      missing

Count              52        60         0
Percent (%)      46.4      53.6
Pt-Biserial      -.50       .50
p-value          .000      .000
Mean Ability    -1.66      -.81        NA

Step Labels                1

Thresholds               -1.37
Error                      .21
.....................................................................
Item   18: item 18                          Infit MNSQ = 1.12

Categories          0         1      missing

Count              19        93         0
Percent (%)      17.0      83.0
Pt-Biserial      -.18       .18
p-value          .027      .027
Mean Ability    -1.46     -1.15        NA

Step Labels                1

Thresholds               -3.02
Error                      .27
.....................................................................
Item   19: item 19                          Infit MNSQ = 1.00

Categories          0         1      missing

Count              69        43         0
Percent (%)      61.6      38.4
Pt-Biserial      -.37       .37
p-value          .000      .000
Mean Ability    -1.44      -.82        NA

Step Labels                1

Thresholds                -.67
Error                      .21
.....................................................................
Item   20: item 20                          Infit MNSQ =  .92

Categories          0         1      missing

Count              67        45         0
Percent (%)      59.8      40.2
Pt-Biserial      -.45       .45
p-value          .000      .000
Mean Ability    -1.51      -.74        NA

Step Labels                1

Thresholds                -.75
```

```
Error                        .21
.....................................................................
Item   21: item 21                          Infit MNSQ =  .97

Categories          0         1      missing

Count             101        11         0
Percent (%)      90.2       9.8
Pt-Biserial      -.24       .24
p-value          .006      .006
Mean Ability    -1.27      -.58        NA

Step Labels                  1

Thresholds                1.25
Error                      .33
.....................................................................
Item   22: item 22                          Infit MNSQ = 1.05

Categories          0         1      missing

Count              44        68         0
Percent (%)      39.3      60.7
Pt-Biserial      -.25       .25
p-value          .004      .004
Mean Ability    -1.49     -1.02        NA

Step Labels                  1

Thresholds               -1.71
Error                      .21
.....................................................................
Item   23: item 23                          Infit MNSQ =  .85

Categories          0         1      missing

Count             111         1         0
Percent (%)      99.1        .9
Pt-Biserial      -.31       .31
p-value          .000      .000
Mean Ability    -1.22      1.59        NA

Step Labels                  1

Thresholds                3.80
Error                     1.01
.....................................................................
Item   24: item 24                          Infit MNSQ = 1.00

Categories          0         1      missing

Count              74        38         0
Percent (%)      66.1      33.9
Pt-Biserial      -.36       .36
p-value          .000      .000
Mean Ability    -1.43      -.76        NA

Step Labels                  1

Thresholds                -.45
Error                      .21
.....................................................................
Item   25: item 25                          Infit MNSQ =  .85

Categories          0         1      missing

Count             111         1         0
Percent (%)      99.1        .9
Pt-Biserial      -.31       .31
p-value          .000      .000
Mean Ability    -1.22      1.59        NA

Step Labels                  1

Thresholds                3.80
Error                     1.01
.....................................................................
```

```
Item   26: item 26                                  Infit MNSQ = 1.11

Categories          0         1      missing

Count              70        42         0
Percent (%)      62.5      37.5
Pt-Biserial      -.22       .22
p-value          .009      .009
Mean Ability    -1.34      -.97        NA

Step Labels                  1

Thresholds                -.63
Error                      .21
.....................................................................
Item   27: item 27                                  Infit MNSQ =  .92

Categories          0         1      missing

Count             102        10         0
Percent (%)      91.1       8.9
Pt-Biserial      -.34       .34
p-value          .000      .000
Mean Ability    -1.29      -.31        NA

Step Labels                  1

Thresholds                1.36
Error                      .34
.....................................................................
Item   28: item 28                                  Infit MNSQ =  .98

Categories          0         1      missing

Count             101        11         0
Percent (%)      90.2       9.8
Pt-Biserial      -.34       .34
p-value          .000      .000
Mean Ability    -1.29      -.37        NA

Step Labels                  1

Thresholds                1.25
Error                      .33
.....................................................................
Item   29: item 29                                  Infit MNSQ =  .98

Categories          0         1      missing

Count              73        39         0
Percent (%)      65.2      34.8
Pt-Biserial      -.38       .38
p-value          .000      .000
Mean Ability    -1.43      -.77        NA

Step Labels                  1

Thresholds                -.50
Error                      .21
.....................................................................
Item   30: item 30                                  Infit MNSQ = 1.13

Categories          0         1      missing

Count              53        59         0
Percent (%)      47.3      52.7
Pt-Biserial      -.22       .22
p-value          .010      .010
Mean Ability    -1.39     -1.03        NA

Step Labels                  1

Thresholds               -1.33
Error                      .20
.....................................................................
Item   31: item 31                                  Infit MNSQ =  .88
```

```
Categories          0         1      missing

Count             110         2         0
Percent (%)      98.2       1.8
Pt-Biserial      -.28       .28
p-value          .002      .002
Mean Ability    -1.23       .52        NA

Step Labels                  1

Thresholds                3.10
Error                      .72
------------------------------------------------------------------------------
Item   32: item 32                              Infit MNSQ =  .00

Categories          0         1      missing

Count             112         0         0
Percent (%)     100.0        .0
Pt-Biserial       .00        NA
p-value          .500        NA
Mean Ability    -1.20        NA        NA

Step Labels

Thresholds
Error
------------------------------------------------------------------------------
Item   33: item 33                              Infit MNSQ =  .87

Categories          0         1      missing

Count             109         3         0
Percent (%)      97.3       2.7
Pt-Biserial      -.34       .34
p-value          .000      .000
Mean Ability    -1.25       .48        NA

Step Labels                  1

Thresholds                2.68
Error                      .59
------------------------------------------------------------------------------
Item   34: item 34                              Infit MNSQ = 1.13

Categories          0         1      missing

Count              96        16         0
Percent (%)      85.7      14.3
Pt-Biserial      -.16       .16
p-value          .050      .050
Mean Ability    -1.26      -.86        NA

Step Labels                  1

Thresholds                 .79
Error                      .28
------------------------------------------------------------------------------
Item   35: item 35                              Infit MNSQ =  .00

Categories          0         1      missing

Count             112         0         0
Percent (%)     100.0        .0
Pt-Biserial       .00        NA
p-value          .500        NA
Mean Ability    -1.20        NA        NA

Step Labels

Thresholds
Error
------------------------------------------------------------------------------
Item   36: item 36                              Infit MNSQ =  .94

Categories          0         1      missing
```

Count	108	4	0
Percent (%)	96.4	3.6	
Pt-Biserial	-.19	.19	
p-value	.021	.021	
Mean Ability	-1.23	-.32	NA
Step Labels		1	
Thresholds		2.37	
Error		.52	

Item 37: item 37 — Infit MNSQ = 1.01

Categories	0	1	missing
Count	93	19	0
Percent (%)	83.0	17.0	
Pt-Biserial	-.29	.29	
p-value	.001	.001	
Mean Ability	-1.31	-.67	NA
Step Labels		1	
Thresholds		.57	
Error		.26	

Item 38: item 38 — Infit MNSQ = .91

Categories	0	1	missing
Count	86	26	0
Percent (%)	76.8	23.2	
Pt-Biserial	-.44	.44	
p-value	.000	.000	
Mean Ability	-1.40	-.54	NA
Step Labels		1	
Thresholds		.14	
Error		.24	

Item 39: item 39 — Infit MNSQ = .96

Categories	0	1	missing
Count	42	70	0
Percent (%)	37.5	62.5	
Pt-Biserial	-.39	.39	
p-value	.000	.000	
Mean Ability	-1.63	-.95	NA
Step Labels		1	
Thresholds		-1.80	
Error		.21	

Item 40: item 40 — Infit MNSQ = 1.07

Categories	0	1	missing
Count	67	45	0
Percent (%)	59.8	40.2	
Pt-Biserial	-.28	.28	
p-value	.001	.001	
Mean Ability	-1.38	-.93	NA
Step Labels		1	
Thresholds		-.75	
Error		.21	

Item 41: item 41 — Infit MNSQ = 1.15

Categories	0	1	missing
Count	74	38	0
Percent (%)	66.1	33.9	

```
Pt-Biserial       -.17       .17
p-value           .033       .033
Mean Ability     -1.30      -1.01        NA

Step Labels                   1

Thresholds                  -.45
Error                        .21
------------------------------------------------------------------------------
Item   42: item 42                              Infit MNSQ =  .94

Categories          0         1      missing

Count              94        18         0
Percent (%)      83.9      16.1
Pt-Biserial      -.39       .39
p-value          .000       .000
Mean Ability     -1.34      -.47        NA

Step Labels                   1

Thresholds                  .64
Error                        .27
------------------------------------------------------------------------------
Item   43: item 43                              Infit MNSQ =  .96

Categories          0         1      missing

Count              85        27         0
Percent (%)      75.9      24.1
Pt-Biserial      -.38       .38
p-value          .000       .000
Mean Ability     -1.38      -.64        NA

Step Labels                   1

Thresholds                  .09
Error                        .24
------------------------------------------------------------------------------
Item   44: item 44                              Infit MNSQ = 1.03

Categories          0         1      missing

Count              70        42         0
Percent (%)      62.5      37.5
Pt-Biserial      -.33       .33
p-value          .000       .000
Mean Ability     -1.40      -.86        NA

Step Labels                   1

Thresholds                 -.63
Error                        .21
------------------------------------------------------------------------------
Item   45: item 45                              Infit MNSQ =  .98

Categories          0         1      missing

Count              74        38         0
Percent (%)      66.1      33.9
Pt-Biserial      -.38       .38
p-value          .000       .000
Mean Ability     -1.42      -.77        NA

Step Labels                   1

Thresholds                 -.45
Error                        .21
------------------------------------------------------------------------------
Item   46: item 46                              Infit MNSQ =  .99

Categories          0         1      missing

Count              70        42         0
Percent (%)      62.5      37.5
Pt-Biserial      -.38       .38
p-value          .000       .000
```

```
Mean Ability      -1.44       -.80        NA

Step Labels                    1

Thresholds                   -.63
Error                         .21
.....................................................................
Item   47: item 47                              Infit MNSQ = 1.03

Categories          0          1      missing

Count              91         21          0
Percent (%)      81.3       18.8
Pt-Biserial      -.27        .27
p-value          .002       .002
Mean Ability     -1.31       -.72        NA

Step Labels                    1

Thresholds                    .44
Error                         .25
.....................................................................
Item   48: item 48                              Infit MNSQ = 1.04

Categories          0          1      missing

Count              57         55          0
Percent (%)      50.9       49.1
Pt-Biserial      -.32        .32
p-value          .000       .000
Mean Ability     -1.46       -.93        NA

Step Labels                    1

Thresholds                  -1.17
Error                         .20
.....................................................................
Item   49: item 49                              Infit MNSQ =  .86

Categories          0          1      missing

Count             109          3          0
Percent (%)      97.3        2.7
Pt-Biserial      -.27        .27
p-value          .002       .002
Mean Ability     -1.24        .13        NA

Step Labels                    1

Thresholds                   2.68
Error                         .59
.....................................................................
Item   50: item 50                              Infit MNSQ =  .93

Categories          0          1      missing

Count              71         41          0
Percent (%)      63.4       36.6
Pt-Biserial      -.44        .44
p-value          .000       .000
Mean Ability     -1.48       -.72        NA

Step Labels                    1

Thresholds                   -.58
Error                         .21
.....................................................................
Item   51: item 51                              Infit MNSQ =  .85

Categories          0          1      missing

Count              44         68          0
Percent (%)      39.3       60.7
Pt-Biserial      -.54        .54
p-value          .000       .000
Mean Ability     -1.76       -.84        NA
```

```
Step Labels                  1

Thresholds                -1.71
Error                       .21
.....................................................................
Item   52: item 52                          Infit MNSQ =  .98

Categories          0         1      missing

Count              67        45         0
Percent (%)      59.8      40.2
Pt-Biserial      -.38       .38
p-value          .000      .000
Mean Ability    -1.46      -.81        NA

Step Labels                  1

Thresholds                 -.75
Error                       .21
.....................................................................
Item   53: item 53                          Infit MNSQ =  .95

Categories          0         1      missing

Count              83        29         0
Percent (%)      74.1      25.9
Pt-Biserial      -.40       .40
p-value          .000      .000
Mean Ability    -1.39      -.65        NA

Step Labels                  1

Thresholds                 -.02
Error                       .23
.....................................................................
Item   54: item 54                          Infit MNSQ =  .94

Categories          0         1      missing

Count              90        22         0
Percent (%)      80.4      19.6
Pt-Biserial      -.35       .35
p-value          .000      .000
Mean Ability    -1.34      -.64        NA

Step Labels                  1

Thresholds                  .38
Error                       .25
.....................................................................
Item   55: item 55                          Infit MNSQ =  .89

Categories          0         1      missing

Count              50        62         0
Percent (%)      44.6      55.4
Pt-Biserial      -.48       .48
p-value          .000      .000
Mean Ability    -1.65      -.84        NA

Step Labels                  1

Thresholds                -1.46
Error                       .21
.....................................................................
Item   56: item 56                          Infit MNSQ =  .79

Categories          0         1      missing

Count              70        42         0
Percent (%)      62.5      37.5
Pt-Biserial      -.60       .60
p-value          .000      .000
Mean Ability    -1.58      -.57        NA

Step Labels                  1
```

```
Thresholds                -.63
Error                      .21
.....................................................................
Item   57: item 57                          Infit MNSQ =  .82

Categories          0         1      missing

Count              78        34         0
Percent (%)      69.6      30.4
Pt-Biserial      -.57       .57
p-value          .000      .000
Mean Ability    -1.51      -.50        NA

Step Labels                1

Thresholds                -.27
Error                      .22
.....................................................................
Item   58: item 58                          Infit MNSQ =  .83

Categories          0         1      missing

Count              85        27         0
Percent (%)      75.9      24.1
Pt-Biserial      -.55       .55
p-value          .000      .000
Mean Ability    -1.45      -.41        NA

Step Labels                1

Thresholds                 .09
Error                      .24
.....................................................................
Item   59: item 59                          Infit MNSQ =  .89

Categories          0         1      missing

Count              96        16         0
Percent (%)      85.7      14.3
Pt-Biserial      -.45       .45
p-value          .000      .000
Mean Ability    -1.35      -.30        NA

Step Labels                1

Thresholds                 .79
Error                      .28
.....................................................................
Item   60: item 60                          Infit MNSQ =  .88

Categories          0         1      missing

Count              91        21         0
Percent (%)      81.3      18.8
Pt-Biserial      -.45       .45
p-value          .000      .000
Mean Ability    -1.37      -.45        NA

Step Labels                1

Thresholds                 .44
Error                      .25
.....................................................................
Item   61: item 61                          Infit MNSQ =  .91

Categories          0         1      missing

Count             100        12         0
Percent (%)      89.3      10.7
Pt-Biserial      -.32       .32
p-value          .000      .000
Mean Ability    -1.28      -.50        NA

Step Labels                1

Thresholds                1.15
Error                      .32
```

```
.....................................................................................
Item   62: item 62                              Infit MNSQ =  .79

Categories          0         1      missing

Count              94        18            0
Percent (%)      83.9      16.1
Pt-Biserial      -.58       .58
p-value          .000      .000
Mean Ability    -1.40      -.14           NA

Step Labels                  1

Thresholds                 .64
Error                      .27
.....................................................................................
Item   63: item 63                              Infit MNSQ =  .82

Categories          0         1      missing

Count             103         9            0
Percent (%)      92.0       8.0
Pt-Biserial      -.49       .49
p-value          .000      .000
Mean Ability    -1.32       .12           NA

Step Labels                  1

Thresholds                1.48
Error                      .36
.....................................................................................
Item   64: item 64                              Infit MNSQ =  .83

Categories          0         1      missing

Count              95        17            0
Percent (%)      84.8      15.2
Pt-Biserial      -.51       .51
p-value          .000      .000
Mean Ability    -1.37      -.25           NA

Step Labels                  1

Thresholds                 .72
Error                      .28
.....................................................................................
 Item   65: item 65                               Infit MNSQ =  .87

Categories          0         1      missing

Count             101        10            1
Percent (%)      91.0       9.0
Pt-Biserial      -.44       .44
p-value          .000      .000
Mean Ability    -1.33      -.12          .70

Step Labels                  1

Thresholds                1.32
Error                      .34
.....................................................................................

Mean test score       19.26
Standard deviation     8.42
Internal Consistency    .86

The individual item statistics are calculated
using all available data.

The overall mean, standard deviation and internal
consistency indices assume that missing responses
are incorrect.  They should only be considered useful when
there is only a limited amount of missing data.
=====================================================================================
```

Appendix 4-7: TG Listening response patterns

sum	CODE	L22	L10	L32	L25	L06	L24	L26	L03	L43	L45	L01	L02	L08	L41	L21	L23	L09	L11	L05	L30
38	1005	0	1	1	1	0	1	0	0	1	1	0	0	1	1	1	1	1	1	1	0
34	1001	0	0	1	0	0	1	1	1	0	1	1	1	0	1	0	1	1	0	1	0
34	7008	0	0	1	1	0	1	0	0	1	1	0	0	1	1	1	0	1	0	0	1
32	7002	0	0	0	0	0	1	1	0	1	1	0	0	1	0	0	1	1	1	0	1
31	1002	0	0	0	1	1	0	1	1	0	0	0	0	1	0	1	1	1	1	0	1
31	6019	0	0	0	0	1	0	0	0	1	0	1	1	1	0	1	0	1	1	0	1
30	6015	0	0	0	0	0	0	0	0	1	0	1	1	1	0	1	1	1	1	0	1
30	7003	0	0	0	0	1	0	1	0	0	1	1	0	1	1	1	1	0	0	1	1
30	7004	0	0	0	1	0	1	1	0	1	1	0	1	0	0	1	1	0	0	1	0
29	6010	0	0	0	0	0	1	0	1	1	0	1	1	0	1	0	0	0	0	0	1
28	1009	0	0	1	0	0	1	1	0	0	1	1	1	1	1	0	0	0	1	0	1
28	7014	0	0	1	1	0	0	1	0	0	0	1	0	1	1	1	1	0	1	0	1
27	4002	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1
27	7001	0	0	0	0	1	1	0	0	1	1	0	0	1	1	0	1	1	0	0	0
27	3011	0	0	0	0	1	0	1	0	1	0	0	1	0	1	0	1	0	0	0	0
26	6005	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	1
25	6007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
25	6012	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	1	1
24	6014	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	1
23	6017	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1
23	7007	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	1
22	1006	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0
22	1013	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0
22	4006	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0
22	4017	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1
21	4012	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0
21	6016	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1
21	6022	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
21	3021	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1
21	3047	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	1	0
20	4015	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
20	4020	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0
20	6024	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
20	7005	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
20	3014	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
20	3051	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
19	4003	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1
19	6003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
19	3046	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
18	4005	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
18	6023	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
18	7010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
18	3038	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	1011	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
17	4014	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	1	0	0
17	6001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	6009	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
17	6026	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
17	3010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
17	3035	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	1014	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
16	4007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
16	6002	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	1008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0
15	6004	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0

15	3054	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	3055	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	1007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	4010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	6013	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
14	6025	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
14	7020	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0
14	3015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
14	3032	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
14	3033	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
13	4001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	4004	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
13	4008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	4013	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	4018	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
13	6006	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
13	6020	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	3002	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
13	3009	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	3030	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	3052	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	1003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
12	3003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	6011	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
11	6018	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
11	3034	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	3045	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	4009	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
10	4019	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0
10	3005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
10	3059	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
10	3064	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	3016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	3044	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	7009	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	3013	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
8	3039	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	3022	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	3025	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	4016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
6	3024	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
6	3031	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	3063	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	3069	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	4011	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	7006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	3027	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	3043	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	6021	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	3028	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	3042	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	3026	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	3007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	3040	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	3041	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

2	3068	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

L42	L20	L31	L18	L44	L40	L19	L38	L39	L17	L36	L14	L33	L16	L27	L04	L35	L34	L15	L37	L07	L29	L28
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1
1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	0	1	1	0	1	1	1	1
0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
0	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1
0	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1
0	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1
1	0	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	0	0	1	1	1	1	0	1	0	1	1	1	1	0	1	1	1	1	1	0	1	1
1	1	1	0	1	1	1	1	0	0	1	1	0	1	1	1	0	0	0	1	1	1	1
1	0	0	1	1	1	0	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1
1	1	0	1	0	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	0
0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
1	0	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1
1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1
1	1	1	0	1	1	1	1	1	0	1	0	1	1	1	0	1	1	1	1	0	1	1
0	0	0	1	1	0	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1
1	0	0	1	0	1	1	1	1	1	1	0	1	1	0	1	1	1	0	1	1	1	1
0	1	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	0	1	0	1	1
1	0	0	1	1	1	1	1	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1
0	0	0	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1
0	1	1	1	0	1	0	1	1	0	1	0	1	0	1	1	1	1	0	1	1	1	1
0	0	1	0	1	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
0	0	1	0	1	1	1	1	1	1	1	0	0	1	0	1	1	0	1	1	1	1	1
1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1
0	1	1	0	0	0	0	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1
1	1	1	0	0	0	1	1	1	1	0	1	1	1	0	0	1	1	1	1	0	0	1
1	0	0	1	1	1	0	1	1	1	1	1	1	1	0	0	1	1	1	1	0	1	1
0	0	0	1	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	0
0	0	1	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1
0	0	1	1	1	1	1	0	1	1	0	1	1	1	1	0	1	1	1	0	0	1	1
0	1	1	0	1	0	1	1	1	1	0	0	1	0	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	0	0	1	1	1	0	1	0	0	1	1	1	1	1	1	1	1
0	0	1	1	1	1	0	0	0	0	1	1	1	0	1	0	1	1	1	0	0	1	1
0	1	0	1	0	0	1	1	1	1	1	1	0	1	1	1	0	0	1	1	1	1	1
1	0	1	0	0	0	1	1	1	1	0	1	1	1	0	0	1	1	1	1	1	1	1
0	0	1	0	1	1	1	0	0	0	0	1	1	0	1	1	1	0	1	1	1	1	1
0	0	1	0	0	0	1	1	1	0	1	0	1	1	1	1	1	1	0	1	0	1	1
0	1	1	1	0	1	0	1	0	0	1	0	1	1	0	0	1	1	1	1	0	1	1
0	0	1	0	0	1	0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
1	0	0	1	0	1	1	1	1	0	1	0	1	0	1	0	1	1	1	0	0	1	1
0	1	0	1	1	0	0	0	0	1	0	1	0	0	1	1	0	0	1	0	1	1	1
0	0	0	1	0	0	1	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1
0	0	0	0	0	1	0	0	0	1	1	1	1	0	0	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1
0	0	0	0	1	1	0	1	0	0	1	0	1	1	1	0	1	1	1	1	1	0	1
0	0	1	0	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	0	1	0
0	0	0	0	1	0	1	0	1	1	0	1	0	0	0	1	1	1	1	1	0	1	1
0	0	0	0	0	0	1	0	1	1	1	0	1	1	0	0	1	1	1	1	1	1	1
0	1	0	0	1	0	1	1	0	0	1	0	1	0	1	1	1	1	1	0	1	1	1
0	1	0	0	0	1	1	1	0	1	0	1	0	1	0	1	0	0	0	1	1	0	1
0	0	0	0	0	1	0	1	1	1	0	0	0	1	1	1	1	0	0	0	1	1	1

0	1	0	0	0	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1	0	0
0	1	0	0	0	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1	0	0
0	1	0	0	0	0	0	1	1	0	1	1	0	0	0	1	0	1	1	1	1	1
0	0	0	0	0	1	0	1	1	0	1	1	1	0	0	0	1	1	1	1	0	1	1
1	0	0	0	0	0	1	0	0	1	0	1	0	1	1	1	0	0	1	1	0	1	1
0	0	0	0	1	0	1	0	0	0	1	0	1	1	1	1	1	1	0	0	1	1	1
0	0	0	1	0	0	0	0	1	0	0	0	0	1	1	0	0	1	1	1	1	1	1
0	0	0	0	0	0	0	1	0	1	1	0	0	0	1	1	1	1	1	1	1	0	1
0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	0	1	1	0	0	1	0	1
0	0	0	0	1	1	0	1	1	0	0	0	1	1	0	0	1	1	0	1	1	1	1
0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	1	1	1	1	1
0	0	0	0	1	0	0	0	1	1	1	0	0	0	1	1	0	0	0	1	1	1	1
0	0	0	0	1	0	1	0	0	1	1	0	1	1	1	0	1	1	0	1	1	0	1
0	0	1	0	0	0	0	0	1	0	1	0	0	1	1	0	1	0	1	1	1	1	1
0	1	0	0	0	0	0	0	0	0	0	1	0	1	1	1	0	0	1	0	1	1	1
0	0	1	0	0	1	1	0	0	1	0	0	0	0	1	1	1	0	1	0	0	1	1
0	1	0	0	0	1	0	1	0	0	0	1	0	1	1	1	1	0	1	1	0	1	1
0	1	0	1	0	0	1	0	0	1	1	1	0	1	1	0	0	0	0	1	0	1	1
0	1	0	1	0	0	0	0	0	1	0	0	0	0	1	1	1	1	1	0	1	1	1
0	0	0	1	0	0	0	0	0	1	0	1	1	0	1	1	0	1	1	0	1	1	1
0	0	0	0	0	1	1	1	1	1	0	0	1	1	0	1	1	1	0	0	1	0	1
1	1	0	1	1	0	0	0	0	1	1	0	0	1	1	0	0	0	0	1	0	0	0
0	1	0	1	0	0	1	0	0	1	1	0	1	1	0	0	1	0	0	1	0	1	1
0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	1	1	1	1
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	1	1
0	1	0	0	1	0	0	1	1	1	0	0	0	1	0	0	0	1	0	1	1	0	0
0	0	1	0	0	0	1	0	0	1	1	0	1	1	0	1	0	0	1	1	1	0	0
0	0	0	0	1	0	0	0	0	0	0	1	0	1	1	0	0	1	1	0	0	1	0
0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	1
0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	1	1	1	1	0	0	1	0
0	0	0	0	0	1	0	1	0	0	0	0	1	0	1	0	0	1	1	0	0	1	1
1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	1	0	1	1	1
0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1	1	0	1	0	1
0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	1	1	1
0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1
0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	1	1	0	0
0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	1	1	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	1	1	0	1	0	0
0	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	1	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0
0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	1	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	1	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0

L13	L12
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	0
1	1
1	1
1	1
1	1
1	1
1	1
0	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
0	1
1	1
1	1
1	1
1	1
0	0
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	0
1	1
0	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	1
1	0
1	1
1	1
1	1
1	1

1	1
1	1
1	1
1	1
1	1
0	0
0	1
1	1
1	1
1	0
1	1
1	1
1	0
1	1
1	1
1	0
1	0
0	1
1	1
1	1
1	0
1	1
0	1
1	1
1	1
1	1
1	0
1	1
1	1
0	1
0	1
1	1
1	1
1	0
1	0
0	1
1	1
0	1
0	1
0	1
1	1
1	1
1	0
1	1
1	1
1	1
1	1
0	0
0	0
1	1
0	0
0	1
0	1
0	1
0	1
0	1

0 1

Appendix 4-8: TG Grammar response patterns

RAWDATA

SUM	CODE	G35	G32	G25	G23	G31	G33	G49	G36	G63	G27	G10	G65	G21	G28	G61	G59	G34
50	1005	1	0	0	0	0	1	0	1	1	1	0	1	1	1	1	1	1
43	1001	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	1	0
38	7002	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0
37	7008	0	0	0	0	0	0	0	0	0	1	0	1	0	1	1	1	0
34	1006	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1
34	7007	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
33	1009	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
33	7001	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
31	3011	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
30	3009	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0
29	1002	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0
29	1014	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
29	4016	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0
29	6016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	6013	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
28	7014	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0
28	3042	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
27	6003	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
27	6007	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
27	6024	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	7003	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0
26	4004	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
26	4017	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
26	6023	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
26	7004	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0
26	3013	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
26	3021	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
26	3047	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
25	1007	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1
25	7009	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
24	1008	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	1
24	1013	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
24	6009	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	1
24	6018	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	3035	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0
23	6010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	7005	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0
23	7010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
23	3014	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
23	3022	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	3033	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
22	4003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	4012	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	6020	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
22	3043	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
21	6015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	3044	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	4002	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
20	6011	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
20	6014	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
20	3030	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
19	1011	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	6012	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
19	3046	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	3052	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

19	3064	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	1010	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
18	4005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	3038	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	4009	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
17	4014	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
17	4020	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
17	6005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	6006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	3051	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
16	4001	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
16	4006	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
16	4015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	6002	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	6019	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	3025	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	3028	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	4007	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
15	3015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
15	3024	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	3034	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
14	6001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	6022	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	6025	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	3026	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	6004	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	3010	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
13	3032	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
13	3063	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	7020	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
12	3040	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	4019	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
11	6017	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	6021	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	6026	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	3007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	3027	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	3031	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
11	3041	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	3068	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	3002	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	3003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	4008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	4011	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
9	3054	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	3069	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	4010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	4013	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	4018	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	3016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	3039	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	3055	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	3059	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	7006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	3005	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	1003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

0	3045	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0

G64	G62	G42	G02	G37	G47	G60	G54	G38	G43	G58	G53	G01	G12	G57	G05	G09	G15	G08	G41
1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1
1	1	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1
1	1	1	0	0	0	0	1	1	0	1	1	0	0	1	1	1	0	1	1
1	1	0	0	0	0	1	1	1	1	1	1	0	1	1	0	1	0	0	0
1	1	0	0	0	0	1	0	1	1	0	1	0	0	0	1	1	1	0	0
0	0	0	0	0	1	0	1	0	1	1	0	0	0	1	0	1	1	1	0
0	1	1	0	0	0	1	0	1	0	0	1	1	0	1	1	1	0	1	0
1	0	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	1
1	1	0	0	1	0	1	1	0	0	1	0	0	0	1	0	0	0	1	0
1	1	1	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0
0	0	0	0	1	1	0	1	0	0	0	0	0	0	1	1	1	0	0	1
0	0	0	0	1	1	0	0	0	1	0	1	0	1	0	0	0	0	1	1
0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	1	0	0	1	0
0	1	0	0	1	0	1	1	0	0	0	0	0	0	1	1	0	1	0	1
0	0	1	0	0	0	0	0	0	0	1	1	0	0	1	1	1	0	1	1
1	1	0	0	0	0	1	0	1	0	0	0	0	1	1	1	0	0	0	1
1	0	1	1	0	0	0	0	0	0	1	1	0	0	1	1	1	0	0	0
0	0	0	0	0	0	1	0	1	0	1	1	1	1	1	0	0	1	0	0
0	0	0	0	0	0	1	1	1	1	1	0	0	1	1	1	0	1	1	0
0	0	0	0	0	1	0	1	0	0	1	0	0	1	1	1	0	0	1	0
0	1	0	1	0	0	0	0	0	1	1	1	1	0	1	0	0	0	0	1
0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	0	0	0	0	1
1	1	0	0	0	0	1	0	1	0	1	1	0	0	0	0	0	1	1	0
0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	1	1	1	1	0
0	1	1	0	0	1	1	0	0	0	0	0	1	0	0	0	1	0	0	1
1	1	0	0	0	0	0	0	1	0	1	0	1	1	1	1	1	0	0	0
1	1	1	0	1	1	0	0	0	0	1	0	0	0	0	1	1	1	1	0
0	0	0	0	0	1	0	0	1	1	0	1	0	1	1	0	0	1	0	0
1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0
0	0	0	1	0	0	0	1	0	1	1	0	1	1	1	0	0	1	0	1
0	1	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	1	0	0
0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0
0	0	0	1	0	1	0	1	1	1	0	0	0	1	0	0	1	0	0	1
1	1	0	0	0	1	1	0	1	0	1	0	1	1	1	0	0	0	0	0
0	0	1	0	0	0	0	1	0	0	1	1	1	0	0	0	0	1	1	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
0	1	0	0	1	0	1	0	0	0	1	0	0	1	0	0	0	0	1	0
0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	1	0	1	1	1
0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1
0	0	1	0	1	1	0	0	1	1	0	1	1	0	0	0	0	1	0	0
0	0	1	0	1	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0
0	0	0	1	0	0	0	0	1	0	1	0	0	1	1	1	0	0	0	1
0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0
0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	1	0
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
0	0	0	0	1	0	0	1	0	1	0	1	1	1	0	0	0	0	1	1
0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	1	0	1	0	1
0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0
0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1
0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	1	1	0
0	0	1	0	0	1	0	0	0	1	0	1	0	1	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	1	0	1	0	0
0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	0	1
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	1
0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
0	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0
0	0	1	1	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1
0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	1	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	1
0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0
1	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	1	0
0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	1
0	0	0	0	0	0	1	0	1	0	0	1	0	0	1	1	0	0	1	0
0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

RAWDATA

0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 1

G24	G45	G29	G03	G50	G56	G26	G46	G44	G19	G52	G40	G20	G07	G16	G11	G48	G14	G30
1	1	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
1	1	0	1	1	0	1	0	1	1	1	0	0	0	1	1	1	1	1
0	1	0	1	1	0	1	1	1	1	0	1	1	0	1	1	1	1	1
0	1	0	0	1	1	0	1	1	0	1	1	1	0	1	1	1	1	1
1	1	1	0	0	1	1	1	0	1	1	0	1	1	1	1	0	1	1
0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
1	1	1	1	1	1	1	0	0	1	1	1	0	1	1	1	1	0	1
0	1	1	0	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1
1	0	1	0	0	1	1	1	0	1	0	0	1	1	1	1	1	1	1
1	1	1	1	1	1	1	0	0	0	1	0	1	1	1	0	1	1	0
1	1	1	0	0	1	1	1	0	1	1	1	0	0	1	1	0	1	0
1	1	1	0	1	1	0	1	1	1	0	1	0	0	1	1	0	1	1
0	0	1	1	1	1	1	1	1	0	1	1	1	0	1	0	1	0	1
0	1	1	1	1	1	0	0	0	1	0	1	1	0	0	1	1	1	1
0	0	0	1	1	1	0	0	0	1	1	0	0	1	1	1	1	1	0
1	1	0	0	1	1	0	1	1	0	0	0	0	0	1	1	1	1	0
0	1	1	1	0	1	0	0	1	1	1	1	0	0	1	1	1	0	1
0	0	0	1	0	1	0	1	1	0	1	0	0	0	1	1	1	1	0
0	1	0	0	0	1	1	0	0	1	1	0	1	0	0	1	0	0	1
0	1	0	1	1	1	0	1	1	1	0	0	1	0	1	0	1	1	1
0	0	0	0	1	1	1	1	1	0	0	0	0	1	1	1	0	1	1
0	1	1	0	1	1	1	0	1	0	1	1	1	0	0	1	1	1	0
1	0	1	0	0	1	0	0	0	0	0	1	0	0	1	0	1	1	0
0	0	0	1	0	0	1	1	1	1	1	0	1	0	1	1	1	1	1
1	0	0	0	0	0	0	0	1	1	0	1	1	0	0	1	1	1	1
1	1	0	0	0	1	0	0	0	1	0	1	0	1	1	0	0	1	0
1	0	0	1	0	1	0	1	0	0	1	0	0	1	0	1	0	1	1
1	0	0	1	1	1	0	1	0	1	1	1	1	0	0	0	1	1	0
1	1	1	0	0	1	0	0	0	0	0	1	1	0	0	0	1	1	1
0	0	0	0	1	1	0	1	1	0	0	1	0	0	0	1	1	1	0
1	0	1	1	0	1	1	0	1	0	1	0	0	0	0	1	1	0	1
1	1	1	1	1	1	1	0	0	0	0	0	1	0	1	1	0	1	1
1	0	0	1	1	0	0	0	0	1	0	1	1	1	1	1	0	1	1
0	0	0	1	1	0	1	1	1	1	1	0	1	0	1	1	0	1	1
1	1	0	0	0	1	0	0	0	0	1	0	0	1	1	0	0	1	0
1	0	0	0	0	0	0	1	1	1	1	0	1	0	1	1	1	1	0
0	1	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0
1	0	0	0	1	1	0	0	1	1	1	1	1	0	0	1	1	1	1
1	1	0	1	1	0	1	0	1	1	1	1	1	0	0	0	1	0	1
1	0	1	0	1	0	1	1	0	0	1	0	1	0	0	0	0	1	1
0	0	0	0	0	0	1	1	0	1	1	1	0	0	0	1	1	0	1
1	1	1	0	1	0	1	1	0	0	1	1	1	0	0	0	1	1	1
1	0	0	0	0	0	1	1	1	1	1	0	1	0	0	0	1	0	1
1	1	1	0	1	0	1	0	0	0	0	0	1	1	0	0	0	1	1
0	1	1	1	0	1	1	1	0	1	0	0	0	1	0	1	1	0	1
0	0	0	1	1	0	0	1	0	0	0	1	1	1	1	0	1	1	0
1	1	1	1	1	0	0	0	0	1	0	0	0	1	1	0	1	0	1
0	1	0	0	0	1	1	0	1	0	1	0	1	0	0	0	1	0	0
0	1	1	1	0	0	0	0	1	1	0	1	0	1	1	1	0	1	1
0	0	1	1	0	0	1	0	1	0	0	1	1	0	1	0	0	1	1
1	0	1	0	0	0	0	0	0	0	1	1	0	1	0	0	1	0	0
1	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	0	0	0
0	0	1	0	1	0	0	0	1	0	1	1	0	0	1	1	1	0	0
0	1	0	1	1	0	0	1	1	1	1	1	0	1	0	0	1	0	0
1	0	0	0	0	0	0	1	1	0	1	0	1	1	0	1	1	0	1

0	0	1	1	0	0	1	0	0	0	0	1	0	1	0	0	0	0	1
0	0	0	0	1	0	1	0	1	0	0	1	1	1	0	0	0	0	0
1	0	0	0	1	0	0	0	0	0	1	0	1	0	0	1	0	1	0
0	0	0	1	0	1	1	1	0	1	1	0	0	1	1	0	0	0	0
0	1	1	0	0	0	1	0	0	0	0	0	0	1	1	0	0	1	0
1	0	1	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1	0
0	1	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	1	0
0	0	0	0	0	1	0	1	1	0	1	0	0	0	1	1	0	0	0
1	0	0	1	1	0	0	0	0	0	1	0	0	1	1	1	1	1	1
0	0	1	0	0	0	0	0	1	1	0	1	0	0	0	1	0	0	1
1	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	1	0	0	0	0	1	0	0	1	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1
0	0	0	0	0	0	1	0	0	0	1	0	0	1	1	1	0	1	1
0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	1	1
0	1	0	0	0	0	0	1	1	0	0	1	1	0	0	0	1	1	0
1	0	0	0	0	1	0	0	0	1	1	0	1	0	1	1	0	1	1
1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	0	1	1	0
0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	1	1	1	0
0	0	1	0	0	0	1	0	0	0	0	0	1	1	1	1	0	1	1
0	1	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	1	0	0	0	1	1	1	0	1	1
0	0	0	0	0	0	0	0	1	1	0	1	1	0	0	1	1	1	1
1	0	1	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	1
0	0	0	1	0	1	1	1	1	0	0	0	0	1	0	1	0	0	1
1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1
0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	1	0	0	1
0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0
0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1
0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	1	0	1
0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0
0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0
0	0	0	1	0	0	0	0	0	1	1	0	0	1	0	1	0	0	0
0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	0	1
0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0
0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0
0	0	1	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1
0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0
0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	1	0	0
0	0	1	0	0	0	0	0	0	1	0	1	0	1	1	1	1	1	0
0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0
0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	1
0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0
0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1
0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	1
0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0
0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1
0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	1	0	1
0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0
0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0
0	0	0	1	0	0	0	0	0	1	1	0	0	1	0	1	0	0	0
0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	0	1
0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0
0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0
0	0	1	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1
0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0
0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	1	0	0
0	0	1	0	0	0	0	0	0	1	0	1	0	1	1	1	1	1	0
0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0
0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	1
0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0
0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1
0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

RAWDATA

1 1 1 1 0 0 0 1 1 1 1 0 0 1 0 1 1 0 1

G17	G55	G13	G04	G51	G22	G39	G18	G06
1	1	1	0	1	1	1	1	1
1	1	0	1	1	1	1	1	1
1	1	1	0	1	1	1	1	1
1	1	1	1	1	0	1	1	1
1	1	1	1	1	0	1	1	1
1	1	1	1	1	1	1	0	1
1	1	1	1	1	0	0	1	1
1	1	0	0	1	1	1	1	1
1	1	1	0	1	1	1	1	1
1	1	1	1	1	0	1	1	1
1	1	1	1	1	0	1	1	1
1	1	1	1	1	1	1	0	1
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1
1	1	0	0	1	1	1	1	1
1	1	0	1	1	0	1	1	1
0	1	1	1	1	1	1	1	1
1	1	1	0	1	1	1	1	1
1	1	1	1	1	1	0	1	1
1	1	1	0	0	1	0	1	1
1	1	1	1	1	1	1	1	0
1	0	1	1	1	1	1	1	1
1	0	1	1	1	0	0	1	1
0	1	1	0	1	1	1	1	1
1	1	0	1	1	0	1	1	1
0	1	0	1	1	1	1	1	0
1	0	1	1	1	1	1	0	1
1	1	1	1	1	1	1	1	0
0	1	1	0	1	1	1	1	1
0	1	0	0	1	1	0	1	1
1	1	1	1	1	1	1	1	1
1	0	1	1	0	1	0	1	1
1	0	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1
0	1	0	0	1	1	1	1	0
0	1	1	0	1	0	1	1	1
0	0	0	0	0	1	1	1	0
0	1	1	1	1	0	1	1	1
1	1	1	1	1	0	1	1	1
1	1	1	0	1	1	1	0	1
0	1	0	1	1	1	1	1	1
0	1	0	0	1	1	0	1	1
0	1	0	1	1	1	1	1	1
1	0	0	0	0	0	0	1	1
1	1	1	1	0	1	1	1	1
1	1	1	0	0	1	1	1	1
0	1	1	0	1	1	1	1	1
1	0	1	0	0	0	1	1	1
1	0	1	1	0	1	1	1	1
0	1	0	1	0	0	1	1	1
1	0	0	1	1	1	1	1	1
1	1	1	0	1	1	1	0	1
0	1	0	1	1	0	1	0	1
1	1	1	0	1	1	1	1	1

0	1	1	1	1	1	1	1	1
1	0	0	1	0	1	1	1	1
0	0	1	1	1	0	1	1	1
1	1	0	1	1	0	0	1	1
1	0	0	1	0	1	0	1	1
1	1	1	0	1	1	0	1	1
1	1	1	0	0	0	1	1	1
0	1	0	0	1	1	1	1	1
0	0	1	1	0	1	0	1	1
1	0	0	1	0	1	0	1	1
1	0	0	1	1	1	1	1	1
0	1	1	0	1	1	1	0	1
1	0	0	0	0	1	1	1	1
1	0	0	0	1	0	0	1	1
1	1	1	1	0	1	1	1	1
0	1	1	1	1	1	0	1	1
1	1	1	0	1	1	0	0	1
0	0	1	1	0	1	1	1	1
1	0	0	1	0	0	1	1	1
0	0	1	0	1	1	1	1	1
1	0	0	0	0	1	1	1	1
1	0	1	0	0	0	0	1	1
0	0	1	0	0	1	1	1	1
0	0	1	1	1	1	1	1	1
1	0	0	1	0	0	0	0	1
1	0	1	1	0	1	0	1	1
0	1	0	0	1	0	1	0	1
0	1	0	0	1	0	0	0	1
0	0	1	0	0	1	0	0	1
0	1	0	0	0	1	0	1	1
1	0	1	1	0	1	1	1	1
1	1	0	0	1	0	0	1	1
0	0	0	0	1	0	1	0	1
0	1	1	1	1	0	0	0	1
0	0	1	1	1	1	1	1	1
0	0	0	0	0	0	0	1	1
0	0	0	1	0	0	0	1	0
0	0	0	0	0	0	1	1	0
0	1	0	1	1	0	0	1	1
1	1	1	1	0	0	0	1	1
0	0	0	0	0	1	0	1	1
0	0	0	0	0	1	0	1	1
0	0	1	1	0	1	0	1	1
0	0	1	1	0	1	0	1	0
0	1	0	0	0	0	1	1	0
0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	1	1	1
0	0	0	0	0	0	1	1	1
1	1	0	1	1	0	0	1	0
0	1	0	1	0	0	0	0	1
0	0	0	0	1	1	0	0	0
0	1	0	0	0	0	1	1	0
0	0	1	1	0	0	0	0	1
0	0	1	0	0	0	0	1	0
0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0

0 0 0 1 1 1 0 0 1

Appendix 4-9: Speaking test candidate measures

TG Oral Ability 06-03-2000 18:35:45
Table 7.1.1 candidate Measurement Report (arranged by N).

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Num	candidate
72	24	3.0	3.24	6.86	.48	0.3	-2	0.2	-2	1001	1001
92	24	3.8	3.97	11.30	.65	0.2	-2	0.1	-2	1002	1002
44	24	1.8	2.06	.97	.52	0.1	-3	0.1	-3	1003	1003
92	24	3.8	3.88	9.93	.57	1.6	1	7.5	4	1005	1005
44	24	1.8	2.06	.97	.52	0.1	-3	0.1	-3	1006	1006
44	24	1.8	2.08	1.10	.52	2.8	2	2.6	2	1007	1007
48	24	2.0	2.23	1.99	.50	0.3	-2	0.2	-2	1008	1008
72	24	3.0	3.24	6.86	.48	0.9	0	0.8	0	1009	1009
44	24	1.8	2.06	.97	.52	0.1	-3	0.1	-3	1010	1010
44	24	1.8	2.06	.97	.52	0.1	-3	0.1	-3	1011	1011
44	24	1.8	2.06	.97	.52	0.1	-3	0.1	-3	1013	1013
0	24			(-12.22	1.80)	Minimum				1014	1014
0	24			(-14.10	1.84)	Minimum				3002	3002
0	24			(-14.19	1.83)	Minimum				3003	3003
28	24	1.2	1.18	-3.45	.65	2.5	1	2.2	1	3005	3005
0	24			(-14.10	1.84)	Minimum				3007	3007
28	24	1.2	1.18	-3.45	.65	0.1	-2	0.1	-2	3009	3009
28	24	1.2	1.18	-3.45	.65	0.1	-2	0.1	-2	3010	3010
72	24	3.0	3.03	5.73	.46	3.3	4	3.0	3	3011	3011
0	24			(-14.19	1.83)	Minimum				3013	3013
52	24	2.2	2.20	1.84	.53	1.8	1	1.7	1	3014	3014
0	24			(-14.19	1.83)	Minimum				3015	3015
0	24			(-14.19	1.83)	Minimum				3016	3016
72	24	3.0	3.03	5.73	.46	0.4	-2	0.3	-2	3021	3021
24	24	1.0	1.02	-5.13	.66	0.2	-2	0.1	-2	3022	3022
28	24	1.2	1.18	-3.45	.65	0.1	-2	0.1	-2	3024	3024
24	24	1.0	1.02	-5.13	.66	0.2	-2	0.1	-2	3025	3025
4	24	0.2	0.41	-9.40	.80	3.8	2	2.1	0	3026	3026
4	24	0.2	0.41	-9.40	.80	0.1	-2	0.1	-1	3027	3027
48	24	2.0	2.29	2.29	.47	1.1	0	1.4	1	3028	3028
40	24	1.7	2.04	.84	.41	1.9	5	2.0	5	3030	3030
0	24			(-13.30	1.84)	Minimum				3031	3031
4	24	0.2	0.41	-9.40	.80	3.8	2	2.1	0	3032	3032
52	24	2.2	2.54	3.25	.53	0.0	-4	0.0	-4	3033	3033
20	24	0.8	1.03	-4.97	.49	1.7	2	1.6	1	3034	3034
28	24	1.2	1.45	-2.19	.69	2.9	2	1.9	0	3035	3035
28	24	1.2	1.45	-2.19	.69	0.1	-2	0.0	-2	3038	3038
8	24	0.3	0.81	-7.63	.54	0.5	-1	0.4	-1	3039	3039
0	24			(-13.30	1.84)	Minimum				3040	3040
0	24			(-13.30	1.84)	Minimum				3041	3041
0	24			(-13.30	1.84)	Minimum				3042	3042
0	24			(-13.30	1.84)	Minimum				3043	3043
68	24	2.8	3.16	6.48	.41	2.2	4	2.1	4	3044	3044
40	24	1.7	2.04	.84	.41	1.9	5	2.0	5	3045	3045
44	24	1.8	2.14	1.52	.42	1.1	0	1.3	1	3046	3046
24	24	1.0	1.28	-2.95	.71	0.1	-2	0.1	-2	3047	3047
0	24			(-11.42	1.83)	Minimum				3051	3051
0	24			(-11.42	1.83)	Minimum				3052	3052
24	24	1.0	1.16	-3.65	.68	2.7	2	2.0	1	3054	3054
0	24			(-11.42	1.83)	Minimum				3055	3055
28	24	1.2	1.43	-2.28	.52	0.7	-1	0.5	-1	3059	3059
0	24			(-11.42	1.83)	Minimum				3063	3063
0	24			(-11.42	1.83)	Minimum				3064	3064
0	24			(-11.42	1.83)	Minimum				3068	3068
0	24			(-11.42	1.83)	Minimum				3069	3069
44	24	1.8	2.01	.62	.47	0.3	-3	0.3	-3	4001	4001
72	24	3.0	3.08	6.04	.47	0.3	-3	0.2	-3	4002	4002
44	24	1.8	2.01	.62	.47	0.3	-3	0.3	-3	4003	4003
48	24	2.0	2.14	1.54	.50	1.0	0	0.7	0	4004	4004
44	24	1.8	2.01	.62	.47	0.3	-3	0.3	-3	4005	4005
24	24	1.0	1.28	-2.95	.71	0.1	-2	0.1	-2	4006	4006
44	24	1.8	2.24	2.04	.46	0.3	-3	0.3	-3	4007	4007
24	24	1.0	1.28	-2.95	.71	0.1	-2	0.1	-2	4008	4008
40	24	1.7	2.10	1.26	.43	1.2	1	1.7	2	4009	4009
24	24	1.0	1.28	-2.95	.71	0.1	-2	0.1	-2	4010	4010
24	24	1.0	1.05	-4.70	.70	0.1	-2	0.1	-2	4011	4011
48	24	2.0	2.07	1.06	.50	0.2	-3	0.2	-3	4012	4012
44	24	1.8	1.93	.08	.50	0.2	-3	0.1	-3	4013	4013
44	24	1.8	1.93	.08	.50	0.2	-3	0.1	-3	4014	4014
52	24	2.2	2.16	1.62	.48	1.5	1	1.6	1	4015	4015
48	24	2.0	2.07	1.06	.50	0.2	-3	0.2	-3	4016	4016
68	24	2.8	2.92	5.00	.48	0.1	-3	0.1	-3	4017	4017
44	24	1.8	1.93	.08	.50	0.2	-3	0.1	-3	4018	4018
28	24	1.2	1.21	-3.29	.53	1.5	1	3.1	3	4019	4019
48	24	2.0	2.07	1.06	.50	1.0	0	0.8	0	4020	4020
44	24	1.8	2.00	.54	.54	0.0	-4	0.0	-4	6001	6001
44	24	1.8	2.00	.54	.54	0.0	-4	0.0	-4	6002	6002

44	24	1.8	2.00	.54	.54	0.0	-4	0.0	-4	6003	6003
24	24	1.0	1.02	-5.13	.71	0.1	-2	0.1	-2	6004	6004
44	24	1.8	2.00	.54	.54	0.0	-4	0.0	-4	6005	6005
44	24	1.8	2.00	.54	.54	0.0	-4	0.0	-4	6006	6006
68	24	2.8	2.99	5.46	.51	0.0	-4	0.0	-4	6007	6007
44	24	1.8	2.00	.54	.54	0.0	-4	0.0	-4	6008	6008
44	24	1.8	2.00	.54	.54	0.0	-4	0.0	-4	6010	6010
44	24	1.8	2.00	.54	.54	0.0	-4	0.0	-4	6011	6011
44	24	1.8	2.00	.54	.54	0.0	-4	0.0	-4	6012	6012
44	24	1.8	2.00	.54	.54	0.0	-4	0.0	-4	6013	6013
48	24	2.0	2.15	1.60	.50	1.0	0	1.1	0	6014	6014
44	24	1.8	2.00	.55	.54	0.1	-4	0.0	-4	6015	6015
48	24	2.0	2.15	1.60	.50	1.0	0	1.1	0	6016	6016
48	24	2.0	2.26	2.11	.50	1.0	0	0.9	0	6017	6017
48	24	2.0	2.15	1.60	.50	1.0	0	1.1	0	6018	6018
28	24	1.2	1.29	-2.89	.50	0.9	0	0.9	0	6019	6019
60	24	2.5	2.68	3.81	.42	1.2	0	1.4	1	6020	6020
28	24	1.2	1.43	-2.28	.52	0.7	-1	0.5	-1	6021	6021
48	24	2.0	2.01	.67	.50	0.2	-3	0.1	-3	6022	6022
48	24	2.0	2.07	1.06	.50	1.0	0	0.8	0	6023	6023
48	24	2.0	2.07	1.06	.50	1.0	0	0.8	0	6024	6024
52	24	2.2	2.22	1.96	.46	0.4	-2	0.4	-2	6025	6025
44	24	1.8	1.93	.08	.50	0.2	-3	0.1	-3	6026	6026
48	24	2.0	2.01	.67	.50	0.2	-3	0.1	-3	7001	7001
52	24	2.2	2.22	1.96	.46	0.4	-2	0.4	-2	7002	7002
52	24	2.2	2.22	1.96	.46	0.4	-2	0.4	-2	7003	7003
88	24	3.7	3.77	9.13	.48	1.5	1	1.5	1	7004	7004
52	24	2.2	2.22	1.96	.46	0.4	-2	0.4	-2	7005	7005
44	24	1.8	1.93	.08	.50	0.2	-3	0.1	-3	7006	7006
88	24	3.7	3.77	9.13	.48	1.5	1	1.5	1	7007	7007
76	24	3.2	3.15	6.43	.47	0.2	-3	0.2	-3	7008	7008
48	24	2.0	2.07	1.06	.50	1.0	0	0.8	0	7009	7009
68	24	2.8	2.92	5.01	.48	2.4	2	2.2	2	7010	7010
92	24	3.8	3.92	10.31	.61	0.3	-2	0.2	-2	7014	7014
68	24	2.8	2.92	5.01	.48	0.1	-3	0.1	-3	7020	7020

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Num	candidate
37.6	24.0	1.6	1.71	.74	.54	0.7	-1.5	0.7	-1.6	Mean (Count: 112)	
24.0	0.0	1.0	1.02	4.04	.09	0.9	2.5	1.1	2.5	S.D.	

RMSE (Model) .55 Adj S.D. 4.00 Separation 7.32 Reliability .98
Fixed (all same) chi-square: 4458.0 d.f.: 92 significance: .00
Random (normal) chi-square: 91.5 d.f.: 91 significance: .47

Appendix 4-10: Revised TG Listening and Grammar Tests

Form Code: TGE L001

Name: _____ Registration Number: _____

The Tour Guide English Language Listening Test

Test Booklet

Time allowed: About 40 minutes

Number of questions: 35

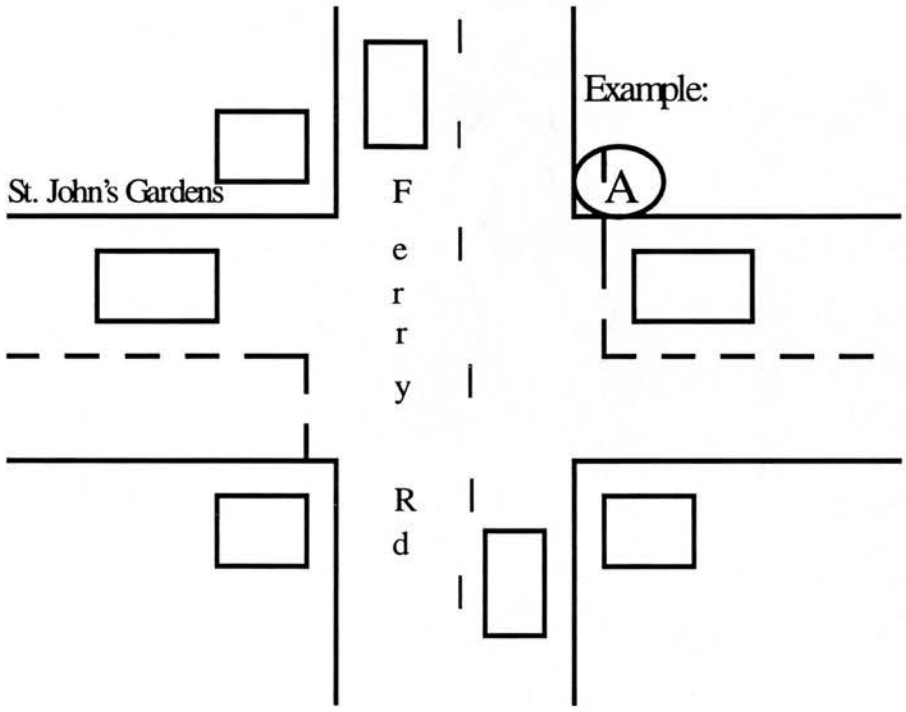
General instructions:

There are 6 listening passages with 35 questions altogether in 8 tasks. The tasks are printed in this test booklet. In each of the tasks there are several questions. Listen to each passage and then do the task. You will be given time to read through the questions before you listen to each passage. After you listen, you will also be given time to write your answers. You may take notes while you are listening to the passages. Write your notes on the paper provided.

(The test continues on the next page.)

Task 1 (Listening Passage 1)

For Questions 1 - 3, listen to the conversation between a policewoman and the driver and indicate the positions of the Volkswagen, the Honda Civic and the witness before the accident in the appropriate boxes in the picture below. The "SLOW" sign has been indicated for you as an example.



Example:

A: "SLOW" sign

(1) B: Volkswagen

(2) C: Honda Civic

(3) D: the witness

(The test continues on the next page.)

Task 2 (Listening Passage 2)

For Questions 4 - 7, listen to the story of Agatha Christie and fill in the missing information.

Section A: Personal data

Date of birth:	**Example: 1891**
Place of birth:	Torquay, England
Education:	At home, by her mother School in Paris
Dates of marriages:	(4) _ _ _ _ _ _ _ _ _ _ 1930
Names of her husbands:	Archibald Christie Max Mallowan
Date of death:	1976

Section B: Some important dates in her life

Date	What happened/ What she did
1914 - 1918	worked in a hospital
(5) _ _ _ _ _ _ _ _ _ _	disappeared from her home and was found in Yorkshire suffering from amnesia.

(The test continues on the next page.)

Section C: Four of her books mentioned in the talk

Titles	Dates published
The Mysterious Affair at Styles	(6) _____
The Murder of Roger Achroyd	1926
Murder on the Orient Express	(7) _____
And Then There Were None	1940

(The test continues on the next page.)

Task 3 (Listening Passage 3)

For Questions 8 - 12, match the colours used in Chinese opera in Box A with their associated meanings in Box B. Not all meanings will be used. The first one has been done for you as an example.

Box A

Example: _____ Red		
_____ (8) Green	_____ (9) Black	_____ (10) White
_____ (11) Purple	_____ (12) Yellow	

Box B

A: Righteousness	B: Pride	C: Reliability	D: Honesty
E: Cunning	F: Danger	G: Cheerfulness	H: Insincerity
I: Faithfulness	J: Greed	K: Anger	

(The test continues on the next page.)

Task 4 (Listening Passage 4)

For Questions 13 - 17, listen to the talk on traditional architecture in Taiwan and while you are listening, fill in the missing words to complete each of the sentences. The first one is an example.

Example: Traditional Taiwanese architecture originates in <u>southern China</u>.

(13) The elements of popular belief and thinking in *I Ching* ( ) are usually realise in the _____ and the _____ of a building.

(14) In traditional Taiwanese architecture, the type of roof representing one of the five elements in the Chinese philosophy is _____.

(15) The owner's official status in the government is represented in the _____ roof.

(16) The pottery figurines set on the roof are used to _____.

(17) The ceremony to give thanks to Lu Ban, Patron of Construction, is usually held _____ on a favourable day.

(The test continues on the next page.)

Tasks 5a and 5b (Listening Passage 5)

Task 5a

Listen to those announcements at Edinburgh Railway Station and fill in the details of the train on the grid. The first one has been done for you as an example.

	Destination	**Departing from**	**Platform**	**Time of departure**
Ex.	Glasgow (Queen Street)	Edinburgh	_ _ _ _ _ _	11:30
(18)	Edinburgh	Glasgow (Queen Street)	_ _ _ _ _ _	11:20
	London (King's Cross)	Edinburgh	19	11:30
(19)	Plymouth	Edinburgh	10	_ _ _ _ _ _

Task 5b

Listen again to the announcements and complete the sentences in Questions 20 - 23.

(20) The 11:30 shuttle service to Glasgow Queen Street offers _ _ _ _ _ _ _ _ _ _ _ _ and light refreshments.

(21) Mr. Jonathan Brown should go to _ _ _ _ _ _ _ _ _ _ _ _ _ _ to meet his mother.

(22) _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ on the train to London King's Cross can be found at the front.

(23) People can buy _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ on the train to Plymouth.

(The test continues on the next page.)

Tasks 6a and 6b (Listening Passage 6)

Task 6a

Listen to the tour guide talk about famous places on the Royal Mile in Edinburgh, Scotland's capital. Put the correct letter next to the places mentioned in the box. The first one has been done for you as an example. The tour starts from the castle and is heading east to Holyroodhouse Palace.

(map of Royal Mile)

This stretch of road is called the Royal Mile.

Example: _ _ _ _ _ Writers' Museum

(24) _ _ _ _ _ St. Giles Cathedral

(25) _ _ _ _ _ Museum of Childhood

(26) _ _ _ _ _ John Knox House

(27) _ _ _ _ _ People's Story Museum

(The test continues on the next page.)

Task 6b

Listen to the talk again and fill in the missing information. The first one has been done for you as an example.

Example:

Length of the Royal Mile: *a mile and 200 yards*

(28) Edinburgh Castle: Opens _____.

(29) Queen Margaret's Chapel in the Castle: _____.

(30) The Writers' Museum: Admission is _____.

(31) John Knox House is now a _____.

(32) The Museum of Childhood: Opens _____.

(33) The word GATE in Canongate means _____.

(34) The People's Museum used to be _____.

(35) Holyroodhouse Palace is the official residence of the Queen who usually comes in _____.

-The End -

The TG Listening Test

(Tape script - listening)

The Tour Guide English Language Listening Test

General instructions:

There are 6 listening passages with 35 questions altogether in 8 tasks. The tasks are printed in your test booklet. In each task there are several questions. Listen to each passage and then do the task. You will be given time to read through the questions before you listen to each passage. After you listen, you will also be given time to write your answers. You may take notes while you are listening to the passages. Write your notes on the paper provided. Do you have any questions? If you do, please raise your hand and an assistant will come and help you. ...(pause 5 seconds)... If you don't have any more questions, let us begin the test ...(pause 2 seconds). Now turn to Task number 1 in your test booklet and listen. (5 seconds)

Listening Passage 1

Listen to the conversation between a policewoman and the driver of a car which has been involved in a road accident. The policewoman is asking the driver for details of the accident. Listen and mark the positions of the Volkswagen, the Honda Civic and the witness before the accident in the appropriate boxes in the picture in Task number 1. The location of the SLOW sign has been indicated for you as an example. Now look at the picture for 10 seconds...(Pause 10 seconds.)..... *Now listen carefully to the conversation.*

Policewoman: And what's your name, sir?

Drummond: Drummond. John Drummond.

Policewoman: Do you live around here, Mr. Drummond?

Drummond: Yes, I do. I live on this road, actually. Ferry Road. Number 18.

Policewoman: So your address is 18, Ferry Road. And are you the driver of this Volkswagen?

Drummond: Yes, I am. And just look at this....

Policewoman: The Honda Civic is in pretty bad shape too.

Drummond: The idiot drove too fast!

Policewoman: Can I see your driver's licence, sir?

Drummond: Yes. Here you are.

Policewoman: How fast were you driving?

Drummond: I was under the speed limit. I know this road very well. I go up and down it every day.

Policewoman: I see. Well, exactly what happened?

Drummond: I was driving up Ferry Road. I slowed down a little as I was passing the junction with St. John's Gardens.

Policewoman: Did you see the other car?

Drummond: Yes. I saw it and I thought it was going to slow down. There's a "SLOW" sign on St. John's Gardens before the junction with Ferry Road so I had the right of way. But the Honda came right across the road and ran into my right side. He didn't slow down at all.

Policewoman: Did you try to brake?

Drummond: Yes, I tried but I didn't have a chance.

Policewoman: Do you have any witnesses?

Drummond: Yes, the lady who lives in the corner house on this side of the road was just coming out. She saw everything.

(3 seconds)

Now indicate the positions of the two cars and the witness just before the accident by writing down B, C and D in the appropriate boxes in the picture. You will have 20 seconds to complete this task.

(20 seconds)

Now, let's move on to Listening Passage 2.

(text modified from *Developing Strategies*, 9)

Listening Passage 2

Now listen to a talk on Agatha Christie, who is known throughout the world as the Queen of Crime. Her 77 novels and books of stories have been translated into every major language in the world and her sales are calculated in tens of millions. Listen carefully to the talk. While you are listening, complete the table marked Task Number 2 in your test book. The first one has been done for you as an example. Before we begin, read through Sections A, B and C for 30 seconds... (Pause 30 seconds.) *Now listen.*

Agatha Christie was born in 1891 in Torquay, England. She was educated by her mother at home and she also went to school in Paris. In 1914 she married Archibald Christie. During the First World War she worked in a hospital and began to write detective stories. In 1920 her first book, the *Mysterious Affair at Styles* was published. In 1926 her most famous book, *The Murder of Roger Achroyd* was published and in the same year she disappeared from her home. She was finally found in Yorkshire, suffering from amnesia. She wrote many books, some of which were made into plays, for example *And Then There Were None*, published in 1940 and some of which were made into films, for example, *Murder on the Orient Express*, published in 1934. In 1928 she divorced her husband and two years later she married Max Mallowan, an archaeologist. She accompanied him on some of his expeditions abroad and one or two of her books were based on this experience. She died at the age of 85 in 1976.

(3 seconds)
Now, you will be given 30 more seconds to complete this task.

(30 seconds)
Now, let's move on to Listening Passage 3.

(text modified from *Focus Listening*, Unit 14)

Listening Passage 3

Listen to the talk on facial symbolism in Chinese opera and then answer the questions in Task 3. You may take notes during the talk. Before we begin, read through the questions in Task 3 for 20 seconds...(Pause 20 seconds.).....
Now let's begin.

Facial Symbolism in Chinese Opera

I'd like to talk about facial symbolism which I think is one of the most fascinating elements in Chinese opera. The actors use a variety of colours to tell the audience about the characters they are playing. A little understanding of facial symbolism will certainly help us more deeply appreciate Chinese opera.

Making up faces in Chinese opera is a very specialised skill. It is also an art. The actors and actresses use colourful paints to paint a variety of facial symbols and lines on their faces. Different facial symbols stand for different characteristics, social positions, and ages.

Basically there are four types of symbols: "Zheng lian" *whole face*, "San kwai wa" *three parts*, "Hua san kwai wa" *complicated three parts* and "Sui lian" *most complicated.* Each of the types requires different designs on the forehead, the sides of the nose, the cheeks, eyebrows and mouth. They also depict four types of characters in Chinese Opera: furious and angry people, villains, ghosts and fairies, and historical figures. In general, the less colourful and complicated the faces are, the higher the characters' positions. The more colourful and complicated the faces, the lower in social position and the more cheerful the characters are in the play.

The colours used to draw facial symbols stand for the personality of the characters. For example, red means loyalty and righteousness. *Kuan Yu* in Chinese opera is depicted with a red face. A green face means an angry man. Fierce, unpleasant and dangerous people have blue faces. A black face signifies openness and honesty. A yellow face means the character is cunning but a purple face indicates the character is a reliable person. Fairies usually have gold and silver faces. A white face like the character *Tsau Tsau* means insincerity and deception, but a white face with a square

or a reverse U-shaped black spot on the nose means the character is a clown in the play.

I have just given you a very general description of facial symbolism in Chinese opera. There are many other varieties and exceptions that I haven't mentioned. In order to understand more about the facial symbolism, you simply have to see a Chinese opera or two and observe the characters in the play. You'll find out how interesting it is.
(3 seconds)
Now let's do Questions 8 - 12 in Task number 3. Each of the facial colours in Chinese opera has a particular significance. Match the meaning with the colour it stands for. You will not use all the choices in Box B . The first one has been done for you as an example. You have 45 seconds to complete this task.
(45 seconds) *Now, let's move on to Listening Passage 4.*

(modified from promotional brochure by National Fu Hsin Dramatic Arts Academy)

Listening Passage 4

Listen to the mini lecture on Taiwan's traditional architecture. While you're listening, complete the sentences in Questions 13 - 17 in Task number 4. Before we begin, read through Questions 13 – 17 for 30 seconds. (Pause 30 seconds.)
Now, listen carefully.

The beauty of Taiwan's traditional architecture: Part 1, the roof

The origin of Taiwan's traditional architecture comes from southern China. Apart from the ideology of architecture brought by the first mainland settlers, Taiwanese architecture has added elements in a style that reflects the thinking in I-Ching and some popular beliefs prevalent in our everyday life. These elements may be best realised in the work of the roof and the framework of a building.

The roof is the most exquisite and most important part of traditional Taiwanese architecture. The four most distinguishing features of the roof are: the Horse Back, the Swallow Tail, the Talisman and the Tong-wai tiles.

Horse Back refers to the two ends of the ridges and is usually in the shape to represent one of the five elements in *I-Ching*. The five elements refer to Metal, Wood, Water, Fire and Earth.

Swallow Tail means that both ends of the ridges raise up like the tail of a swallow. It symbolises the house owner's official status in the government.

Talisman refers to the pottery figures set on the roof. They are there to ward off evil spirits.

Tong-wai Tiles refer to the glazed semi-cylinder shaped tiles on the slope of the roof. They are typical tiles used on a Chinese roof.

The roof is always the last part to be completed. Upon its completion, a solemn ceremony is held to give thanks to Master Lu Ban, patron of Construction. The ceremony is usually in the early morning on an auspicious day.
(Pause 3 seconds)

Now, let's move on to Listening Passage 5.

(text modified from a pamphlet on Taiwanese architectural features)

Listening Passage 5

Listen to the railway announcements in Edinburgh Railway Station and fill in the details of the train on the grid. Then listen again and complete the sentences in Questions 18 - 23. Before we begin, read through the questions for 20 seconds... (Pause 20 seconds.)... *Now, listen and fill in the missing information in Task number 5a.*

Platform 14 for the 11: 30 Scot Railways shuttle service to Glasgow Queen Street calling at Haymarket, Linlithgow, Falkirk High and Glasgow Queen Street. Service of drinks and light refreshments is available. Platform 14 for the 11: 30 Scot Railways shuttle service to Glasgow Queen Street.

This is a railway customer call for the person by the name of Mr. Jonathan Brown. Mr. Jonathan Brown. Meeting his mum please go to Thomas Cook which is opposite the main concourse. Mr. Jonathan Brown meeting his mother please go to Thomas Cook which is situated at the main concourse. Mr. Jonathan Brown.

The 11:20 Scot Railway terminating service from Glasgow Queen Street is approaching Platform 14.

Platform 19 for the 11:30 Great North-Eastern Railway service to London King's Cross calling at Newcastle, York, Peterborough, and London King's Cross. Hot food facilities are available. First class accommodation is situated at the front. Platform 19 for the 11:30 Great North-Eastern Railway service to London King's Cross.

Platform 10 for the 11:25 Virgin Train service to Plymouth calling at Birmingham New Street, Cheltenham Spa, Gloucester, Bristol Parkway, Bristol Temple Meads, Taunton, Torquay, Totnes and Plymouth. Hot food facilities are available. First class accommodation is situated at the front. Platform 10 for the 11:25 Virgin Trains service to Plymouth.

(Pause 3 seconds.)

Now listen again. While you're listening, complete the sentences in Questions 30 to 32.

(Play Announcements again.)

(Pause 5 seconds)
Now, let's move on to Listening Passage 6.

(text modified from 27/2/99 recording in Edinburgh Railway Station)

Listening Passage 6 ***(Edinburgh Tour 28/2/99 2:30pm (the script has been modified.))*

Listen to the tour guide talk about famous places on the Royal Mile in Edinburgh, Scotland's capital. While you are listening, put the correct letter next to the places mentioned in the box. Then listen again and fill in the missing information about each place. The tour starts from the castle and is heading east towards Holyroodhouse Palace. Now, before you listen, read through the questions for 30 seconds.....(Pause 30 seconds)..... *Now listen.*

Now, we're just beginning the ..... Royal Mile now. It's called the Royal Mile because it has the Castle at the one end and the Palace of Holyroodhouse at the other. It's actually an old Scots mile..... a mile and 200 yards.

We're leaving the castle esplanade now. Edinburgh Castle is a working castle. It opens all year round. The castle is on top of an extinct volcano. The oldest part of the castle dates back to the eleventh century. That's where Queen Margaret's Chapel is. The chapel is also the oldest building in the castle.

Now just alongside the left hand side of the bus is a little lane. Lady Stair's Close it's called and that's the way through to the Writers' Museum.... and it's free, no charge, open from Monday to Saturday. The museum holds the artefacts of three of the most famous writers: Sir Walter Scott, Robert Louis Stevenson and Robert Burns.

Now, ahead of you on your right hand side is the St. Giles Cathedral. It's a very large church. St. Giles is the high kirk of the Church of Scotland. Parts of this building date back to the twelfth century. And the cathedral goes hand in hand with the history of Edinburgh, the history of Scotland in fact.

Directly coming ahead of us on the right hand side is my favourite museum, the Museum of Childhood. The museum's been described as "the noisiest place in the world". And it is one of the most frequently visited attractions in Scotland. All the toys you, your parents, your grandparents played with could be found here. Just

passing on your right hand side, the Museum of Childhood. Open Monday to Saturday and the entrance is free. It's worth a visit.

Almost directly across the street is John Knox House, which is a museum now. John Knox was a famous Protestant reformer and had much to do with the reformation of the Church in Scotland in the sixteenth century. The house has hand-painted ceilings and it's entered by forestairs which was once a common architectural feature in the Royal Mile. But there are only a few surviving examples now.

The place where we are now was at one time the country side. The area is called the Canongate.

"Gate" is the Scots word for "walk". This is where the canons, or monks used to walk up and down from Holyrood Abbey to St. Giles Cathedral. Outside Edinburgh it was a separate burgh or town and had its own town council. And indeed it had its own jail. Prisoners were housed in the Tolbooth. The Tolbooth nowadays is the People's Story Museum. It's open Monday to Saturday and it shows how ordinary people, not kings and queens but how ordinary people would have lived from the eighteenth century onwards up to about the nineteen fifties.

Now those staircases on your left hand side are the entrance to the People's Story Museum. Open Monday to Saturday. The entrance is .... free.

Now the next stop is the Holyroodhouse Palace. The Queen usually visits during the month of July, during which time she stays at the palace. When the queen is not here, it's open to the public. A very interesting place to visit. Currently there's an exhibition of water colours from the private collection of Prince Albert and Queen Victoria. This is where we stop for anyone wishing to visit Holyroodhouse Palace.

(Pause 5 seconds.)

Now listen again and while you're listening, complete the description of the places in task number 6b.

(Play the talk again.)

(Pause 5 seconds)

This is the end of the listening test. Thank you for your participation. Now, please remain seated until the test assistant tells you to leave. Test assistant, please collect the test booklets. Thank you.
(Pause 3 seconds.)

This is the end of the recording.

Form Code: TGE G001

Name: _____

Registration Number: _____

The Tour Guide English Language Grammar Test

Time allowed: 40 minutes

Number of questions: 45

General instructions:

There are five parts to this test. Instructions for each part of the test will appear before the test questions along with an example. A recommended time to do each part of the test will also be given.

Write your answers on the answer sheet provided. Do not write on the test paper!

Part I: Multiple choice questions

Recommended time: 3 minutes

Instructions:

For Questions 1 - 5, choose A, B, C or D to complete the sentence in each question.

Write your answers on the answer sheet provided.

Example:

What time _____ it now?

A: be

B: is

C: are

D: were

The best answer to the question above is B: is. Therefore you should choose B.

(The test continues on the next page.)

1. _____ usually have great difficulty in getting a job. They need more help from the government.

A: Homeless

B: Homelessness

C: The homeless

D: The homelessness

2. I wish the room _____ a bit bigger.

A: is

B: will

C: were

D: would be

3. Jane _____ before crossing the road but she didn't.

A: must look

B: should look

C: might have looked

D: should have looked

4. In fact, we heard Susie ____ the whole piece from beginning to end on the piano.

A: to play

B: play

C: was playing

D: played

5. While playing tennis, Alice says, "I _____ the game. I think it's going to suit me."

A: enjoy

B: enjoyed

C: am enjoying

D: have enjoyed

(The test continues on the next page.)

Part II: Verb forms

Recommended time: 3 minutes

For questions 6 - 15, read the two passages carefully and supply the correct verb form according to the meaning of the passages. You may need to write one to three words for each question. The first one has been done for you as an example. Write your answers on the answer sheet provided.

Passage 1: *A tour guide is giving a brief history of the Presidential Office Building to his clients. Read the passage and supply the correct verb forms.*

If you _ _ _ _ _ _ _ ahead of you, you *(6. see)* the Presidential Office Building, where President Lee *(7. work)*. *(8. Build)* in 1919, originally, the building *(9. call)* the Supreme Office and it *(10. use)* by the Japanese Governor during the Japanese occupation. We renamed the building after the defeat of Japan in 1945.... Now, here is a good place for you to take pictures.

Passage 2: *The following is a description of a cottage in a travel brochure. Read the passage and supply the correct forms of the verbs.*

Surrounded by pine trees, with direct access to a sandy beach and a heated pool *(11. face)* the sea, this charming 19th century manor house *(12. retain)* all of its original character. It *(13. situate)* just 4 km from Royan in the small resort of Vaux sur Mer. Beautifully *(14. decorate)* and with fine furnishings, this elegant hotel *(15. have)* a bar, lounge and tennis court, and there is a golf course 4 km away.

(The test continues on the next page.)

Part III: Sentence transformation

Recommended time: 15 minutes

For questions 16 - 25, finish the second sentence so it has the same meaning as the sentence before. Write your answers on the answer sheet provided.

Example 1:

Be careful or you'll hurt yourself.

If *you are not careful* you'll hurt yourself.

Example 2:

I wasn't comfortable. Then I had the painful tooth extracted.

Until I *had the tooth extracted, I was in pain*.

(The test continues on the next page.)

16. Emma says something different. It is not the same as she does.

What Emma _____ she does.

17. Many people have seen the show. It shows how popular it is.

The fact that _____ shows how popular it is.

18. Only senior staff members are allowed to use the company car park.

The company car park is _____ senior staff members only.

19. It won't be long before Mary returns from her business trip to Dublin.

Mary will _____.

20. I cannot agree with that statement.

That is the statement _____.

21. "Would you like to have lunch with me?" John said.

John invited _____.

22. I've yet to see a more annoying person than my cousin.

My cousin is _____.

23. I was surprised at how approachable the new boss is.

I didn't expect _____.

24. It has been two weeks since anyone saw John.

John _____.

25. The suitcase was so heavy that Tommy could not lift it.

The suitcase was too _____.

(The test continues on the next page.)

Part IV: Fill in the blanks

Recommended time: 8 minutes

In the newspaper article below (Questions 26 - 35), a word is missing in each of the blanks. Read the article and fill in the missing words. Each blank requires **one word** only. The first blank has been done for you as an example. Write your answers on the answer sheet provided.

Tour of Historical Sites of Taiwan:

Ruins of An-ping Fort

The ruins of Taiwan city are located in An-ping, Tainan City. Built by _ _ _ _ _ Dutch in 1624, An-ping Fort was the first to be built in Taiwan. The fort was first known (26) Fort Orange. Later, the name was changed to Fort Zeelandia. The fort was built with red bricks brought (27) Indonesia. The bricks were mortared with a mixture of sugar syrup, glutinous rice (28) crushed oyster shells. They made a very strong foundation. The square shaped fort is built (29) top of a two-storey platform with lookout towers on the four corners giving the place a grand appearance. At the north-west corner of the fort, there used to be (30) city surrounded by a ten-meter high wall. From (31) remains of the wall, we can still see (32) the interior was constructed with wooden beams, and there (33) still traces of metal studs. After Cheng Cheng-kung defeated the Dutch in 1661, the fort was used as Cheng's residence and it was renamed *Wang Cheng*, the City of the Prince. (34) the late 19th century the fort was in a dilapidated state and during the Japanese occupation (1895 - 1945), the houses (35) levelled to form steps. A platform and a lighthouse were built on top. Today, only the outer walls survive from the original construction by the Dutch.

(The test continues on the next page.)

Part V: Complete the conversation

Recommended time: 6 minutes

There are two parts to *Complete the conversation.*

Part Va:

For questions 36 - 40, read the conversation between you and Sandy, an acquaintance of yours. Some of your part has been deliberately left out. Read the conversation carefully and write the sentences numbered 36 - 40 according to the suggested cues. The first one has been done for you as an example. Write your answers on the answer sheet provided.

Sandy: Hello! Isn't it a lovely day!

You: Yes, beautiful!

Sandy: Can you find somewhere to sit? I'm sorry this room is so untidy.

You: **Disagree** No, it isn't. It's fine.

Sandy: Well, I've got this afternoon off. Where shall we go?

You: **(36) Suggest a visit to Taipei Zoo** _____

Sandy: Sorry, I didn't quite catch that.

You: **(37) Repeat what you said** _____

Sandy: Where is that exactly?

You: **(38) Explain** _____

Sandy: All right. That sounds fine. But I must go to the bank first and I've got all these letters to post too.

You: **(39) Offer to help** _____

Sandy: Oh, that's kind of you. Thanks. Well, shall we meet in half an hour then?

You: **(40) Agree and say goodbye** _____

(The test continues on the next page.)

Part Vb:

For questions 41 - 45, a situation is given to you. Read the situation and write one appropriate response if you were in that situation. Write your answer on the answer sheet provided.

Example:

You are buying some "Thank You" cards in a stationer's for the office. You want a receipt. You ask the shop assistant for it. What would you say? Write the answers on the answer sheet provided.

(a): ***Could you give me a receipt, please?***

(The test continues on the next page.)

41. You are taking some foreign visitors to Keelung tomorrow. The city is well known for its rainfall. You want to remind your guests to bring an umbrella. What would you say?

(41): _____.

42. You are taking a group of visitors to hike the Tsauling Historic Trail. It's a very warm day. You offer an elderly lady some mineral water. What would you say?

(42): _____.

43. You and your foreign clients are visiting the National Palace Museum. You want to show them the well-known jade exhibit, the Jade Cabbage. You are trying to get your clients to follow you. What would you say?

(43): _____.

44. You and your foreign visitors are visiting the Martyrs' Shrine. It is a solemn place. You are reminding your visitors to be quiet and respectful. What would you say?

(44): _____.

45. You want to borrow a client's newspaper so you can check the cinema times. What would you say?

(45): _____.

- The End -

Appendix 5-1: Questionnaire for the experts

Please read the following questions and indicate if you agree/disagree by encircling the appropriate option. If you have any comments, please write them in the space provided.

1. The Listening items fit the descriptions of the Listening test specifications. Yes No Don't know
5 0 0

Comments:

2. The Speaking tasks fit the descriptions of the Speaking test specifications. Yes No Don't know
5 0 0

Comments:

3. The Grammar items fit the descriptions of the Grammar test specifications. Yes No Don't know
4 0 1

Comments:

4. I think the time given for the Listening test is: Long Short About right Don't know
0 0 4 1

Comments:

5. I think the time given for the Speaking test is : Long Short About right Don't know
0 2 3 0

Comments:

6. I think the time given for the Grammar test is: Long Short About right Don't know
0 3 2 0

Comments:

7. The types of Listening tasks are Appropriate Not Appropriate Don't know for this test.
5 0 0

Comments: *(1) Not sure about the topic in Task 2. Perhaps an interesting adventure story or something about the culture of a country.*
(2) Multiple-choice questions are better.

8. The types of Speaking tasks are Appropriate Not Appropriate Don't know for this test.
5 0 0

Comments:

9. The types of Grammar tasks are Appropriate Not Appropriate Don't know for this test.
3 1 1

Comments: *(1) Some parts are too hard, e.g., Part III*
(2) A little too hard.

10. Generally, the number of Listening tasks are Many About right Few for measuring the listening ability of would-be tour guides. 2 3 0

Comments:

11. Generally, the number of Speaking tasks are Many About right Few for measuring the speaking ability of would-be tour guides. 1 4 0

Comments: *(1) Half of the questions are enough.*

12. Generally, the number of Grammar tasks are Many About right Few for measuring the grammatical knowledge of would-be tour guides. 1 4 0

Comments:

13. Please indicate the appropriateness of individual items by checking (√) the right box.

Listening Test:

Item No.	Appropriate	Not appropriate	Don't know		Item No.	Appropriate	Not appropriate	Don't know
1	5				24	2	3	
2	5				25	2	3	
3	5				26	2	3	
4	5				27	5		
5	4	1			28	5		
6	4	1			29	5		
7	4	1			30	4		1
8	5	1			31	5		
9	5				32	5		
10	5				33	4		1
11	4	1			34	3	1	1
12	2	1	2		35	4		1
13	3	2			36	4		1
14	3	1	1		37	5		
15	2	3			38	5		
16	4	1			39	4	1	
17	4	1			40	5		
18	4	1			41	5		
19	4	1			42	5		
20	4	1			43	5		
21	4	1			44	3	2	
22	2	3			45	4		1
23	2	2	1					

Comments: *Questions 12 – 20: I feel I don't really need to listen to the whole thing to answer these questions.*

Questions 21 – 26: Maybe more appropriate for a reading test?

Grammar Test:

Item No.	Appropriate	Not appropriate	Don't know		Item No.	Appropriate	Not appropriate	Don't know
1	3	2			34	4	1	
2	4		1		35	5		
3	5				36	5		
4	5				37	5		
5	5				38	5		
6	5				39	5		
7	4	1			40	4		
8	5				41	3	1	1
9	5				42	4	1	
10	3	1	1		43	5		
11	5				44	4	1	

12	5				45	5		
13	5				46	5		
14	5				47	3	2	
15	5				48	3	2	
16	5				49	3	2	
17	5				50	4	1	
18	4		1		51	2	2	1
19	4		1		52	4		1
20	5				53	4		1
21	5				54	4		1
22	5				55	4		1
23	2	3			56a	4		1
24	3	2			56b	4		1
25	4		1		57a	4		1
26	5				57b	4		1
27	5				58a	4		1
28	5				58b	4		1
29	3		2		59a	4		1
30	3	2			59b	4		1
31	2	3			60a	4		1
32	2		3		60b	4		1
33	4	1						

Comments: *For Questions 51 – 55: testing grammar in spoken language?*
For Questions 56a – 60b: more like a speaking test!

Speaking Test:

Task No.	Appropriate	Not appropriate	Don't know
Part I	5		
Part II	5		
Part III	4		1
Part IV	4		1
Part V	4		1

Comments:
For Parts IV and V, some people may just read from the test book without producing much of their own language.

14. Overall, the test battery is Difficulty About right Easy to measure language use ability of would-be tour guides. 1 4 0

Comments: 1. *Should try avoiding modal verbs in Grammar.* 2. *The Grammar test looks practical.* 3. *Marking the Grammar test can be time-consuming.*

15. On the whole, the test is appropriate not appropriate don't know.
5 0 0

Comments:

16. On the whole, the test does a satisfactory job. Yes No Don't know
5 0 0

Comments:

17. Other comments: *There are too many questions. I don't think students will have the patience to sit through the test.*

Appendix 5-2: Number of candidates taking the TG English test

Year	No. of Applicants	No. of candidates certified
2000	161	37
1999	179	28
1998	567	44
1997	323	31
1996	250	24
1995	379	44
1994	380	50
1993	419	42
1992	401	26