VALIDATION OF THE TEST OF ENGLISH CONVERSATION PROFICIENCY

by

Timothy Paul Moritoshi

A dissertation submitted to the
School of Humanities
of the University of Birmingham
in part fulfilment of the requirements
for the degree of

Master of Arts
in
Teaching English as a Foreign or Second Language (TEFL/TESL)

This dissertation consists of approximately 12,000 words
Supervisor: Corony Edwards

Centre for English Language Studies
Department of English
University of Birmingham
Edgbaston, Birmingham B15 2TT
United Kingdom

March 2002

Paul Moritoshi, The University of Birmingham, 2002.

ABSTRACT

The increasingly common use of more communicative language teaching styles has brought with it a concomitant requirement for more communicative, interactive tests that can accurately measure students' foreign language conversation proficiency. This study validates one such test, the Test of English Conversation Proficiency (TECP), designed in-house for use at a Japanese university. After reviewing the literature relating to language test development and a range of techniques for testing oral proficiency, the paper examines various aspects of the TECP's validity, reliability and 'usefulness' (Bachman and Palmer, 1996: 18), using students' end-of-semester tests as the raw data source (n = 56). The evaluation found that the TECP has good-high validity, quite high inter-rater reliability ($\rho = 0.84$, at $p<0.01$) and moderate test-retest reliability ($\rho = 0.55$, at $p<0.01$). It also found the TECP to be practical within the constraints imposed by the setting in which it is used and to be of quite a high usefulness overall. Given these findings, the TECP's continued use is sanctioned, though recommendations for its improvement are made. The study's limitations are acknowledged and several areas for future research are also identified.

DEDICATION

For my children Emma Hinako and Stephen Kotaro, by way of encouragement to ask ‘how?’, ‘why?’ and ‘what if.....?’

ACKNOWLEDGEMENTS

There are several people to whom I am indebted for their assistance in completing this project. I would like to thank the students who took part in this study, since it was their hard work that provided the raw data on which it is based. I am also grateful to Jennifer Scott, both for her assistance during data collection and for her suggestions on how to improve the coding sheet and spreadsheet contained within. I would like to express my thanks to Cindy Cunningham, for her extensive and conscientious efforts in second marking, without which the inter-rater reliability estimates could not have been obtained. I would also especially like to thank my supervisor, Corony Edwards. This paper is undoubtedly the better for her prompt, concise and constructive advice. Finally, particular thanks are due to my wife, both for her work in translation and for her personal support.

Paul Moritoshi, The University of Birmingham, 2002.

CONTENTS

LIST OF TABLES AND FIGURES

Tables

Figures

LIST OF ABBREVIATIONS

ACTFL American Council on the Teaching of Foreign Languages.

EFL English as a foreign language.

ELT English language teaching.

FVQ Face Validity Questionnaire: the 13-item questionnaire administered to subjects after their final test sitting to assess their reactions to various aspects of the Test of English Conversation Proficiency (TECP).

L2 Second language: in the context of this study, it usually refers to English.

NNS Non-native speaker (of English).

NS Native speaker (of English).

OPI Oral Proficiency Interview: a technique for measuring 'functional speaking ability' (ACTFL, 2001).

SGU Sanyo Gakuen University: the institution at which the TECP was designed and is administered.

TECP The Test of English Conversation Proficiency: a test designed in-house at Sanyo Gakuen University for end-of-semester assessment purposes.

TLU Target Language Use: the 'TLU domain' is the real-life situation in which language will be used.

CHAPTER 1
INTRODUCTION AND BACKGROUND

1.1 Introduction

Since the advent of the Communicative Language Teaching approach in Britain in the late 1960's (Richards and Rodgers, 1986: 64), there has been an increasing requirement for practical test instruments that can validly and reliably measure examinees' oral 'communicative competence' (Spolsky, 1995: 103), however one defines the term. However, there still seems to be no validated, affordable, accessible and practical test available for measuring non-native English-speakers' (NNS) ability to hold a casual conversation in English. Some instruments, such as the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI) have been validated but require would-be examiners to attend expensive and time-consuming training courses, often in distant locations, to attain accreditation. There are also concerns as to whether interviews as a discourse style can adequately simulate a casual conversation at all (Shohamy et al., 1986: 213; Bachman, 1988; van Lier, 1989). Those seeking to measure NNSs' proficiency in English conversation therefore require a more accessible and appropriate, but still valid and reliable alternative test instrument.

The purpose of this study is to validate one such test, designed in-house at Sanyo Gakuen University (SGU), Okayama city, Japan, as an end-of-semester assessment tool: the Test of English Conversation Proficiency (TECP). Without such a validation study, an informed decision as to its continued use cannot be made. Beyond this justification, it is also intended that the study should serve to highlight the measure's strengths and flaws so that, if continued use is feasible, improvements can be made.

As with any research conducted within a restricted time-span and budget, this study has some limitations that should be acknowledged here. Firstly, due to the context in which the TECP is used, the sample sizes are modest, though sufficient for statistical validity (test n = 56, retest n = 52). Secondly, that these samples come from a single academic institution will also reduce the generalisability of the study's findings to

Paul Moritoshi, The University of Birmingham, 2002.

other settings. Also, that while this study examines the TECP's validity, reliability, practicality and other aspects of its overall usefulness, it has not attempted to evaluate the test's fairness.

1.2 Background

The test originated from the requirement to assess and grade the English conversation proficiency of students taking SGU's single semester English conversation courses. In line with the view that direct measurement is the most desirable (Canale, 1983: 18-19; Hughes, 1989: 16), the test required examinees to participate individually with the teacher in a 4-minute conversation in English, which was video-recorded for later analysis. However, after two test administrations it became clear that the test was seriously flawed. With the exception of the time limit, the test procedure was entirely unstandardised, so could not guarantee provision of the necessary opportunities for candidates to demonstrate their level of proficiency in essential discourse skills, such as opening and closing a conversation, topic-shifting and turn-taking. Also, the rating scheme which accompanied the original form was ambiguous, highly inferential, inconsistent, incomplete and often penalised examinees for the same error more than once, (a problem also noted by Lynch and Davidson, 1994: 737). The potential for students to be awarded unrepresentative grades was therefore unacceptably high and an improved form was sought.

After an extensive literature search relating to language test development, conversational discourse analysis and oral proficiency testing, a second version of the TECP was created and administered for a third semester-end assessment. It is the students' scores from that administration that have provided the data for this validation study.

The research null hypotheses are that the TECP does not have the following properties as a measure of NNSs' English conversation proficiency:

1. high construct validity,
2. high content validity,

Paul Moritoshi, The University of Birmingham, 2002.

3. high face validity,
4. high inter-rater reliability,
5. high test-retest reliability,
6. is practical to administer and score,
7. high overall usefulness.

Following Bachman (1990: 257, citing Cronbach, 1980: 103) and Nunan (1992: 13, citing Popper, 1972 and Chalmers, 1982: 38-39), these hypotheses are falsifiable. Before attempting to prove them, I will review the literature in relevant fields of study in chapter 2, then go on to outline the TECP in chapter 3.

Paul Moritoshi, The University of Birmingham, 2002.

CHAPTER 2
ISSUES IN CONVERSATION TEST DEVELOPMENT

To better understand the rationale behind the TECP's development it is helpful to examine the literature in two relevant areas: language test development and conversation as a discourse genre within the context of oral proficiency testing.

2.1 Language test development

Though several writers have highlighted the importance of certain key aspects of the test development process, it is Bachman and Palmer's (1996) work that appears to offer by far the most comprehensive and integrated framework. It operationalises Bachman's (1990) earlier, though largely theoretical text and covers the entire process from conceptualisation and description, through construction and administration, to scoring and test validation. A brief review of this framework will give some understanding of the complexities involved in language test development. To this end I will examine five keys stages in the process: construct definition; test tasks and topical contents selection; test administration; scoring and validation (sections 2.1.1- 2.1.5 respectively). Beyond that, a concise overview is offered by Bachman and Palmer (1996: 87).

2.1.1 Construct definition

The first stage is to identify and conceptually define that which the test seeks to measure: the construct. Without this specification a test has no focus, thus Bachman (1990: 255); Nunan (1992: 15) and Bachman and Palmer see it as an 'essential activity' (1996: 115). Unfortunately, it is also one of the most problematic. Hughes' definition of 'construct' hints at why this is so:

> The word 'construct' refers to any underlying ability (or trait) which is hypothesised in a theory of language ability.
>
> (Hughes, 1989: 26).

In other words, constructs are abstract conceptualisations of human behaviours, faculties or characteristics, (for example conversation, language learning and motivation respectively), definitions of which are based upon an understanding of

Paul Moritoshi, The University of Birmingham, 2002.

related language theory. They are predominantly psychological entities (Nunan, 1992: 15), though they have physical manifestations by which they can be observed (Cronbach and Meehl, 1955: 283, cited by Oller, 1979: 110). The difficulty arises in being able to say with confidence that what is being observed is due to the construct under examination and not some other, perhaps totally unrelated one.

2.1.2 Test tasks and topical contents selection

The next key stage in test development is tasks and topical contents selection. Most, if not all of the texts on language testing reviewed here make some mention of this stage (for example: Oller, 1979: 50-51; Underhill, 1987: 106; Hughes, 1989: 49; Bachman, 1990: 244-247; Weir, 1993: 30-32; Bachman and Palmer, 1996: 43-59; Genesee and Upshur, 1996: 65-66; McNamara, 2000: 50-51). That this is so highlights the importance of well-considered tasks and topical contents selection to the test development process. Messick goes straight to the heart of the matter by identifying the two issues that must be addressed at this stage:

> the specification of the behavioural domain in question, and the attendant specification of the task or test domain.
>
> (Messick, 1980: 1017, quoted by Bachman, 1990: 244).

That is to say that one must first identify the real-life behaviours (i.e. the physical manifestations) associated with the theoretical construct under examination, then detail the tasks and topical contents that will be used to elicit them. If the latter adequately reflect the former, tasks and contents can be considered authentic (Bachman, 1990: 301; Bachman and Palmer, 1996: 23-24; McNamara, 2000: 131) and will facilitate more accurate measurement of the construct. What makes this part of test development so problematic and contentious is that there appears to be no consensus as to the right (or wrong) way(s) to select a test's tasks or topical contents (Owen, 1997: 31), except that they should where possible be relevant (Bachman, 1990: 244); authentic (Underhill, 1987: 8; Hughes, 1989: 15); of interest to candidates and account for their topical knowledge base (Bachman and Palmer, 1996: 65).

Paul Moritoshi, The University of Birmingham, 2002.

Choice of task type strongly influences a test's characteristics. Those requiring examinees to perform the actual, real-life task (albeit under test conditions) are called 'direct' measures (Hughes, 1989: 15), while those attempting to access the abilities which underpin the construct being tested, but without performing the real-life task itself are termed 'indirect' (ibid). A further distinction exists between discrete (or divisible) tasks and those that are integrative (or unitary). The former are based upon the assumption that the ability being tested can be split into distinct components for testing purposes (McNamara, 2000: 133), without compromising the construct's integrity. The latter presupposes that the construct can only be measured accurately when it is used as an integrated whole, and often as part of a larger set of language systems and knowledge (ibid).

2.1.3 Test administration

The literature also highlights certain administrative practices that can help examinees to give an optimum test performance. They include:

1. Adequate pre-test examinee preparation (Underhill, 1987: 40; Hughes, 1989: 46).
2. Provision of adequate, clear, on-test instructions (Underhill, 1987: 40-41; Bachman, 1990: 123; Bachman and Palmer, 1996: 181-191; Genesee and Upshur, 1996: 201-203).
3. The use of a clear layout or format (Genesee and Upshur, 1996: 203-206).

2.1.4 Scoring foreign language oral production

The substantial body of literature on rating scales that has built up since the 1970's (Upshur and Turner, 1995: 4), clearly supports Sheal's (1989: 97) view that they are 'the most popular type of form' for scoring examinees' foreign language oral production. (A 'rating' is a judgement about an examinee's performance, made by a human 'rater', McNamara, 2000: 35). This section therefore focuses on this method of scoring and touches briefly upon some of the main issues and concerns commonly expressed in the literature regarding their design and application.

Paul Moritoshi, The University of Birmingham, 2002.

One design issue is how to decide what to rate and how many separate 'analytic' (Hughes, 1989: 91) scales to use. Bachman and Palmer offer the following advice:

> In designing rating scales, we start with componential construct definitions and create *analytic* scales, which require the rater to provide separate ratings for different components of language ability in the construct definition. In developing analytic rating scales we will have the same number of separate scales as there are distinct components in the construct definition.
>
> (Bachman and Palmer, 1996: 211).

In other words, that these issues are determined by one's conceptual definition of the construct under examination. However, Fulcher (1987) and North and Schneider (1998: 217) have noted that scale designs are usually based upon intuition, rather than empirical evidence and Matthews has written that 'the selected criteria are at times arbitrary and inconsistent' (1990: 118). Despite advances in our understanding of language learning and testing, these criticisms are probably fair even today. There is then a wealth of experience in designing and using analytic scales, but apparently little solid empirical data on which to base well-reasoned designs. This represents a wide area of study for future research.

The literature commonly notes that a scale's levels should describe varying degrees of proficiency in a selected performance criterion, usually ranging from a complete absence (minimum score), to mastery of that component (maximum score) (Bachman and Palmer, 1996: 195). Also, that the level descriptions should be 'low-inference', i.e. unambiguous (ibid: 209-210), since the less raters are required to infer from the descriptions, the more consistently the scale will be applied (Hughes, 1989: 19; Nunan, 1992: 60). The converse is of course also true for 'high-inference descriptors'. Further, that consideration should be given to each scale's weighting so that it best represents that scale's relative importance within the rating scheme as a whole (Underhill, 1987: 97-98; Hughes, 1989: 94; Bachman and Palmer, 1996: 210).

Beyond the design issues, two other concerns relate to rating scales' accuracy and practicality. The former has been criticised on the following grounds:

Paul Moritoshi, The University of Birmingham, 2002.

1. Reliance upon human subjectivity (Oller, 1979: 392; Sheal, 1989: 97; McNamara, 2000: 37).
2. Susceptibility to bias (Sheal, 1989: 97).
3. Low inter-rater reliability (Bachman, 1990: 180), in part due to inconsistent rater severity (Underhill, 1987: 88; Bachman and Palmer, 1996: 221).
4. Low intra-rater reliability (Underhill, 1987: 88; Bachman, 1990: 178-179).

Though these arguments are to varying degrees valid, the literature also suggests ways in which their impact on rater accuracy can be reduced. These include the use of:

1. Low-inference descriptors (Hughes, 1989: 110).
2. Rater training and preparation (Underhill, 1987: 90-92; Hughes, 1989: 113-114; Bachman and Palmer, 1996: 221-222; McNamara, 2000: 44).
3. Multiple raters (Underhill, 1987: 89-90; Hughes, 1989: 114).
4. Sufficiently large, representative (i.e. 'ratable') samples of examinees' proficiency level (Hughes, 1989: 16; Weir, 1993: 29; Spolsky, 1995: 104; Bachman and Palmer, 1996: 218-219).
5. Test recordings that can be viewed as often as necessary (Underhill, 1987: 92; Richards and Lockhart, 1996: 11).
6. Marking keys or protocols, in an attempt to standardise scale application (Underhill, 1987: 94-95; Brown, 1994: 254).
7. Several analytic scales (exemplified in Appendix I), rather than one 'global impression' or 'holistic' scale (Weir, 1993: 42 and Genesee and Upshur, 1996: 206 respectively) (exemplified in Appendix II) (Weir, 1993: 45; Bachman and Palmer, 1996: 211). This increases the accuracy of the average score (Hughes, 1989: 94; Bachman and Palmer, 1996: 220).

Regarding practicality, the literature notes that analytic rating scales make greater demands upon resources than other, more objective, discrete point marking systems

Paul Moritoshi, The University of Birmingham, 2002.

(Hughes, 1989: 94; Upshur and Turner, 1995: 5). This is probably true in most cases, but the additional burden has been justified by Bachman and Palmer who point out that:

> this demand on human resources must be recognized as an unavoidable cost of obtaining the kinds of information that ratings can provide.
>
> (Bachman and Palmer, 1996: 220).

That is that the reduced practicality associated with the use of analytic scales is a necessary price to pay for the capacity to assess examinees' language proficiency in a range of performance criteria (i.e. being criterion-referenced), rather than simply comparing examinees against each other, or against native speakers (i.e. being norm-referenced) (ibid: 212). To put it another way, being able to show what each individual can and cannot do, rather than how much better or worse one examinee is compared to another. This additional information is often presented as a performance profile that can be used to provide highly specific feedback to individual test-takers (Genesee and Upshur, 1996: 206; Bachman and Palmer, 1996: 223).

Constructing a valid and reliable rating scheme is clearly a complex undertaking. However, that is in itself insufficient justification for using other, cheaper, more convenient, less time-consuming scoring methods (Hughes, 1989: 19). After all, a more practical marking system does not necessarily equate to a more valid or reliable one. Despite certain design and application problems, rating scales appear to be the most appropriate system currently available for scoring oral production (Oller, 1979: 392; Bachman, 1990: 27; Bachman and Palmer, 1996: 208).

2.1.5 Test validation

Validation examines certain qualities inherent within a test's design, principally to establish whether or not it is valid and reliable. Most texts on language testing appear to cover this process to some degree (for example: Oller, 1979: 50-69; Underhill, 1987: 104-108; Hughes, 1989: 22-42; Bachman, 1990: 236-291; Weir, 1993: 169-70; Bachman and Palmer, 1996: 17-40; McNamara, 2000: 47-54). The principal test qualities of interest here are:

Paul Moritoshi, The University of Birmingham, 2002.

1. *Construct validity* - the extent to which a test measures the construct that it claims to (Brown, 1994: 256; Bachman and Palmer, 1996: 21).
2. *Content validity* - the degree to which a test's tasks and topical contents are relevant to, and proportionately representative of the real-life domain to which the test corresponds (Hughes, 1989: 22; Bachman, 1990: 306).
3. *Face validity* - the extent to which test-users, particularly examinees, perceive the test to measure what it claims to (Underhill, 1987: 105; Hughes, 1989: 27).
4. *Inter-rater reliability* - the degree to which the scores from two or more markers agree (Nunan, 1992: 14-15; Weir and Roberts, 1994: 172).
5. *Test-retest reliability* - the degree of consistency with which a test measures individuals' performances across administrations (Underhill, 1987: 9; McNamara, 2000: 136).

Bachman (1990: 258) introduces the correlational method for evaluating construct validity. This technique examines the strength of the relationship (i.e. the correlation) between the scores obtained for each individual test item (or in the TECP's case, each analytic scale) and the percentage scores. A high correlation indicates that that scale is related to the ability being tested, while a low correlation indicates that it is not. The technique yields a set of correlation coefficients that might be called a correlation profile. The higher the coefficients are, and the more numerous the high correlations in the profile, the stronger the case for high construct validity becomes. If the TECP's construct validity is high, the correlation profile should yield numerous high, positive correlations between individual scales' ratings (each measuring a key performance criteria associated with conversation proficiency) and the percentage scores.

The correlational technique also shows the extent to which each scale is contributing to the construct's measurement as a whole: the larger the coefficient, the greater the contribution. Hughes (1989: 160) states that a correlation coefficient of 0.3 is an acceptable minimum in this regard. Though he does not explain how this critical value was derived, the results presented in tables 4.1-4.3 suggest that it simply

Paul Moritoshi, The University of Birmingham, 2002.

represents the approximate point at which coefficients become statistically significant.

It is conceded though that one weakness inherent in the correlational approach to construct validation is that it only evaluates the relevance of those performance criteria that are already included. It cannot identify others that are relevant to the construct but which have been omitted. It is further conceded that construct validity can also be assessed via the experimental or analytical methods (Bachman, 1990, 266-271), but these techniques require resources beyond those available for this study, and so are not detailed here.

Content validity can be evaluated in part by showing the relevance of tasks and topical contents to the construct being tested and/or to the 'target language use' (TLU) domain (Bachman and Palmer, 1996: 44-45), i.e. the real-life situation in which the language will be used. However, demonstrating the representativeness of tasks and topical contents coverage is much more problematic (Bachman, 1990: 244-247), as it would almost certainly require substantial empirical research into the frequency with which tasks and topical contents actually arise in the TLU domain. Brown suggests that conversation proficiency tests have a high content validity whenever they require candidates to converse within 'some sort of authentic context' (1994: 255), which perhaps refers to direct testing. This is true up to a point but as Oller notes, there is a little more to it than that:

> [while] the conversation should be natural, it [should also emphasize] the point that 'it is not simply a friendly conversation on whatever topics come to mind.....It is rather, a specialized procedure which efficiently uses the relatively brief testing period to explore many different aspects of the student's language competence'.
>
> (Oller, 1979: 321, quoting the Educational Testing Service, 1970: 11).

In other words a test must have a standardised procedure requiring all examinees to perform the same tasks. Further, that topical contents should be, if not the same for everyone, then at least as even-handed as possible. The literature has also identified a strong link between a test's content validity and the degree of confidence examinees

Paul Moritoshi, The University of Birmingham, 2002.

have in the test itself, i.e. its face validity (Jafarpur, 1987: 204-205; Brown, 1994: 256; Bachman and Palmer, 1996: 24).

Face validity can be evidenced via administration of a suitable questionnaire. Given that examinees are not usually privy to the rationale underlying a test's development, this type of validity neither confirms nor negates a test's actual (i.e. construct and content) validity (Stevenson, 1985: 111; Bachman, 1990: 285; Bachman and Palmer, 1996: 42). However, low face validity might adversely affect successful test implementation (Hughes, 1989: 27; Brown, 1994: 256).

Concurrent validity is not estimated here for two reasons. Firstly, there appears to be no validated equivalent measure that could act as a concurrent test in the present setting. Secondly, it is not enough to show concurrent validity between two tests claiming to measure the same thing. It is also necessary to show that the new instrument *does not* have concurrent validity with measures of other, entirely unrelated abilities (Bachman, 1990: 250). Collection of such evidence is beyond the scope of this paper.

Inter-rater and test-retest reliability can be calculated via correlation procedures performed upon test score data (Bachman, 1990: 180-182), though for subjectively scored tests such as the TECP, Upshur and Turner note that:

> Due to the nature of the assessment setting, lower reliability, with an attendant limit on validity, is to be expected.
>
> (Upshur and Turner, 1995: 5).

That is to say that a test's 'validity' (or what might perhaps be more appropriately labelled 'overall test accuracy', since validity is clearly a separate concept), is limited by its reliability. Further, that in subjectively scored tests the reliability (and therefore the overall accuracy), will be lower than that for objectively marked tests. However, obtaining estimates for both types of reliability does at least offer insight into how a test performs within and across administrations. Hughes (1989: 32, citing Lado, 1961) suggests that a minimum acceptable correlation coefficient for test-retest reliability for measures of oral production is between 0.7-0.79.

Paul Moritoshi, The University of Birmingham, 2002.

The literature also increasingly makes mention of the importance of test practicality, (particularly with respect to financial, material and human resources), and fairness (for example: Underhill, 1987: 15-18; Weir, 1993: 21-22; Bachman and Palmer, 1996: 35-37). This is perhaps due to a heightened awareness in recent years of both the decreasing availability of resources and the increasing requirement to produce ethically and professionally sound tests.

One final consideration in validation is that of Bachman and Palmer's 'test usefulness':

> Usefulness = Reliability + Construct validity +
> Authenticity + Interactiveness + Impact + Practicality
> (Bachman and Palmer, 1996: 18).

In other words, this composite provides a good overall impression of a test's utility. Authenticity has been outlined in section 2.1.2, while reliability, construct validity and practicality have been described above. 'Interactiveness' is the degree to which and ways in which a test engages examinees' second language (L2) ability, their affective schemata of the construct being tested and their topical knowledge (ibid: 25). 'Impact' relates to the effects that a test has on individuals, classes, institutions and society at large (ibid: 29-30). Of particular interest with respect to impact is washback (or backwash), which Hughes defines as 'the effect of testing on teaching and learning' (1989: 1), and which can be either beneficial or detrimental, depending upon the circumstances (ibid).

To summarise, this section has detailed five areas of test development that have been particularly influential in the rationale underlying the TECP's design:

1. Construct definition, whereby the ability under examination is identified and conceptually defined, and without which a test has little or no focus.
2. Test tasks and topical contents selection, which identifies the real-life behaviours and subject matter associated with the theoretical construct and TLU domain and then details the tasks and topical contents that will be used to elicit them.

Paul Moritoshi, The University of Birmingham, 2002.

3. Test administration, whereby decisions are made pertaining to pre-test examinee preparation and procedural matters such as test instructions and format.
4. Scoring, particularly with respect to the design of analytic rating scales.
5. Validation, whereby the test's validity, reliability, practicality, fairness and usefulness are evaluated.

2.2 Testing conversation proficiency

When is an oral test a measure of conversation proficiency? A fair answer might be: when it conforms to the norms and rules of conversational discourse, and that anything else is testing something else. Genesee and Upshur offer a technique by which those that do conform might be distinguished from those that do not:

> Like reliability, the validity of assessment procedures can often be judged by identifying the possible factors that can invalidate them.
>
> (Genesee and Upshur, 1996: 67).

That is, where an oral proficiency test can be shown to contravene a norm or rule of conversational discourse, it can also be shown to be conceptually and/or procedurally invalid as a measure of conversation proficiency. By way of exemplification, the following oral proficiency testing techniques are taken from Underhill (1987) and Weir (1993).

Reading aloud, oral reports, speeches, presentations, story re-telling and other monologues require only one speaker. Conversation may be **limited to a small number of people** (Cook, 1989: 51), but it always requires **at least two interlocutors** (Richards, 1980: 414). Further, where there is only one speaker there can be no **turn-taking** which, though one of the most complex phenomenon to occur during conversation, is also one of its most fundamental features (Scollon and Scollon, 1983: 159; Wardhaugh, 1998: 295). These techniques are therefore not valid measures of conversation proficiency, though of course they are unlikely to be mistaken for such.

Techniques such as sentence repetition and mini-tape situations, which require examinees to be totally subordinate to the examiner or a recording, run counter to the

Paul Moritoshi, The University of Birmingham, 2002.

conversational norm that any pre-existing **disparity in power between the interlocutors is 'partially suspended'** (Cook, 1989: 51; Fairclough, 1992: 19). Further, **opening and closing rituals** have been shown to be an integral part of the conversational process (Robb, 1980: 10; Badovi-Harlig et al., 1991: 6-7; Dornyei and Thurrell, 1994: 42-43). Without them, interlocutors may appear impolite or conversationally inept (Richards and Schmidt, 1983: 135). Sentence repetition and (particularly) mini-tape situations do not require candidates to participate in these rituals in a meaningful way, if at all.

Also, where the part of one of the interlocutors is played by a tape (audio or video), there can be no scope for the use of **back-channelling** or **non-verbal communication** behaviours between participants. 'Back-channelling' is the provision of feedback by which listeners indicate their reaction to what the current speaker is saying (Montgomery, 1995: 114) and can be conveyed verbally (Nunan, 1993: 96) and/or non-verbally (Hatch, 1992: 14). Non-verbal communication such as gestures also enhances correct interpretation of utterances (Sinclair and Brazil, 1982: 20; Cook, 1989: 9). While both behaviours have some commonality across cultures, each also has its own culture-specific idiosyncrasies (LoCastro, 1987: 112; Holmes, 1992: 304). Both, though particularly the former, are recognised as important mechanisms by which conversational discourse is maintained (vom Saal, 1983: 494; Soudek and Soudek, 1985; Montgomery, 1995: 114) and where a test lacks the potential for these behaviours to occur, there is probably little meaningful interaction occurring.

Procedures such as information gap-filling, which require examinees to use the target language to obtain information in order to complete a task, do display certain conversational features. Conversation is **purposeful** (Canale, 1983: 3; Brazil, 1992: 4), **interactive** (Brazil, ibid; McCarthy, 1991: 70) and is **performed in real-time** (Brazil, 1992: 3-4; Weir, 1993: 33). Where such techniques are administered without examinees' prior knowledge of the subject matter, they may also produce **unplanned, spontaneous and unpredictable** discourse, in line with conversational norms (collectively Goffman, 1981: 14-15, cited by van Lier, 1989: 495; Canale,

Paul Moritoshi, The University of Birmingham, 2002.

1983: 3; Valdman, 1989: 19-20; Cook, 1989: 116; Sacks, 1968, cited by Fairclough, 1992: 155; Brazil, 1992: 3-4). However, the topical contents in such techniques are predetermined by the examiner, (who may not even be participating in the dyad), and thus they contravene the conversational norm that **topics are 'relatively unconstrained'** (Coulthard, 1985: 49) or 'up for grabs' (Nunan, 1987: 137).

The above oral proficiency testing techniques can be invalidated relatively easily. One intuitively understands that they do not simulate conversation and it is not difficult to find at least one conversational norm that they contravene. However, two other techniques do have at least a *prima facie* appeal and might appear more difficult to invalidate: interviews and role-plays.

The communicative purpose of a language test interview is the summative assessment of one participant's language ability by another. (That this is also true for conversation tests, whereas the communicative purpose of normal conversation is rarely if ever summative assessment, is the contradiction which designers of such tests strive to overcome, since it is perhaps the biggest single hurdle to high validity.) To fulfil this purpose the interviewer and interviewee have clearly distinct and mutually agreed and understood roles inherent in their role titles. The interviewer takes the 'controlling role' (Brazil, 1995: 103) while the interviewee is subordinate. This contravenes the norm of 'power suspension' mentioned earlier. Indeed such a controlling role is unnecessary and even undesirable during conversation since it is essentially '**self-regulating**' (van Lier, 1989: 499; Seedhouse, 1996: 17).

Furthermore, conversation has a '**potentially equal distribution of rights and duties in talk**' (van Lier, 1989: 495, citing Goffman, 1981: 14-15; Fairclough, 1992: 155), with respect to topic nomination, questions, interruptions, length of turn etc. However, it is the interviewer that nominates the topics, even where their relevance is enhanced through 'personalized questions' to the interviewee, as in the ACTFL OPI's procedure (ACTFL, 2001). Also, it is the interviewer who asks the questions. Further, it would take a very confident interviewee indeed to interrupt their examiner. Finally, rather than share the responsibility of speech, an interviewer must

Paul Moritoshi, The University of Birmingham, 2002.

of necessity seek to elicit quite extended turns from the interviewee in order to obtain a ratable sample, whilst simultaneously minimising their own talk. Not only does this result in an unconversational imbalance in the quantity of speech each participant contributes to the dyad, but interviewee turns also often have the potential to be overly long, counter to the conversational norm that '**turns are quite short**' (Coulthard, 1985: 61; Cook, 1989: 51). To sum up, it is the interviewer's unwillingness to allow the interviewee to guide the discourse or to take the initiative in these ways that makes the resulting discourse unconversational (Underhill, 1987: 45; Stern, 1992: 318).

Where role-plays are performed spontaneously (i.e. unscripted and unrehearsed) and where the topic is not predetermined or restrained, they may indeed offer a more conceptually sound alternative to interviews. In common with interviews and role-plays, conversation requires participants to be **co-operative** (Richards and Schmidt, 1983: 119-124; Cook, 1989: 29; Brazil, 1992: 4-5), particularly in the negotiation of meaning. The principles that underlie this co-operation are called the Gricean maxims and are explained and exemplified by Richards (1980: 415-417). Further, Kormos (1999) has demonstrated that role-playing produces a more symmetrical type of discourse than interviews and affords candidates greater opportunities to show their abilities in key conversational behaviours such as opening, holding the floor, interrupting and closing.

Though there are concerns regarding fairness (Underhill, 1987: 52-54; Hughes, 1989: 107-108) and practical application (Weir, 1993: 62), what invalidates role-playing as a technique for measuring conversation proficiency, is that the discourse lacks authenticity (McCarthy, 1991: 128). Swan (1985: 84) points out that the roles examinees play (and consequently the discourse they produce) are ‘fictional’ and offers ‘You are George - ask Mary what she does at Radio Rhubarb’, by way of an example (ibid). Unsurprisingly, candidates question how meaningful such roles are (Al-Arishi, 1994: 337, citing Taylor, 1982 and Piper, 1984), to the point where it might be better if ‘[they] are simply asked to talk about themselves’ (Swan, 1985:

Paul Moritoshi, The University of Birmingham, 2002.

84). This might be interpreted as a recommendation to participate in conversational discourse.

This section has examined various techniques for testing NNSs' English oral proficiency. By identifying the conversational norms and rules that these techniques contravene, they have been shown to be conceptually and/or procedurally invalid as measures of conversation proficiency. In doing so, not only has a gap been identified in the English language teaching (ELT) profession's capacity to measure this ability in a conceptually valid way, but also a set of features (highlighted in bold) has been generated, by which conversation as a discourse genre might be conceptually defined. There remains then a very real requirement for a test that can measure NNSs' English conversation proficiency whilst simultaneously conforming to these features of conversational discourse. The TECP is a tentative attempt to meet this demand and it is described in the following chapter.

Paul Moritoshi, The University of Birmingham, 2002.

CHAPTER 3
THE TEST OF ENGLISH CONVERSATION PROFICIENCY

The TECP is a direct, criterion-referenced, integrative test of English conversation proficiency. The following sections detail how examinees and examiners prepare for it and how is it administered and scored.

3.1 Candidate preparation

Following the requirement to fully prepare and inform examinees, each course group was given a comprehensive presentation a week prior to their first test sitting. The presentation detailed the necessary candidate preparation (which included consideration of the topics they would use during the test), the test's procedure, things to bring, the scoring criteria and hints on what constitutes a good conversation. To standardise the presentation, all students were given a handout (in English) (Appendix III), which was then explained carefully in Japanese to maximise comprehension. Students were also advised to review their course notes and handouts, which covered many of the discoursal features highlighted in section 2.2. They were then given ample thinking time and opportunities to ask questions and to clarify anything they did not understand.

3.2 Test room preparation

Following Hughes' (1989: 106) call for a quiet test room, candidates waited in a separate room while tests were in progress. Before starting each testing session, the test room was cleared (as much as possible) of any materials that may have assisted or distracted candidates during the test. Spare dictionaries, a pencil and clean paper were placed on the table for either the candidate or examiner to use if required. These items were deemed acceptable because they would likely be available in a real-life situation where these students were expecting to converse with a native-speaker. The video recording apparatus was then set up in the configuration shown in figure 3.1 below.

Paul Moritoshi, The University of Birmingham, 2002.

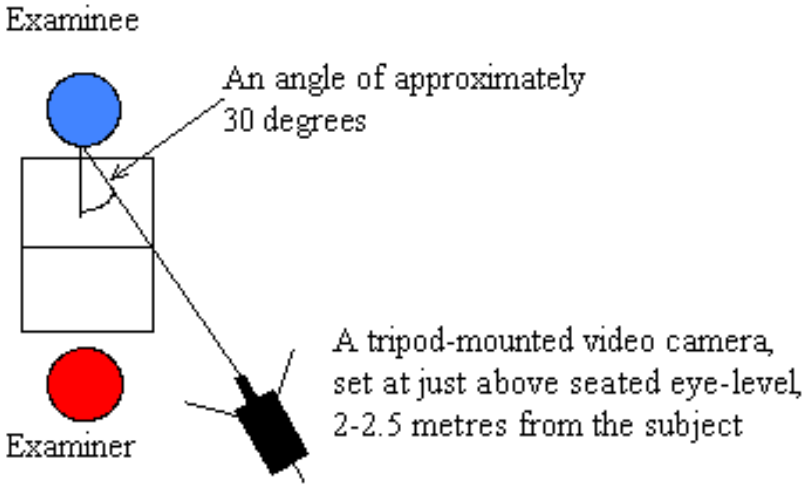


Figure 3.1 - Plan view of the TECP video recording configuration

This camera position was deemed to offer the best compromise between obtaining a sufficiently high quality visual and audio recording of the candidate's performance for rating on the one hand and not intruding upon that performance on the other.

Despite concerns that the presence of a video camera may disrupt or in some way affect the proceedings (Richards and Lockhart, 1996: 11), several justifications exist for the use of such equipment in this type of situation. Firstly, the recording can be reviewed as often as necessary (ibid). Secondly, it detaches the rater from the event, so may help to decrease potential bias and increase marker objectivity (Nunan, 1989: 6). Further, any number of markers can view the recording, which can increase the accuracy of the final scores. Also, replacing a second examiner with a video camera makes the test more practical. Finally, even a trained rater could not possibly rate 19 separate scales in real-time.

3.3 Test procedure

The TECP procedure takes approximately $6^{1}/_{2}$ minutes to administer. This duration offers the best practicable compromise between Hughes' (1989: 105) recommendation of 15-30 minutes and Spolsky's (1995: 104) suggestion of 5 minutes for oral tests, whilst also remaining within the practical limitations imposed by SGU. Candidates take the test individually, pairing with the examiner. The decision to pair in this way complies with the examinees' own wishes, which support

Paul Moritoshi, The University of Birmingham, 2002.

Foot's (1999) findings regarding students' pairing preferences, whilst simultaneously countering those of Egyud and Glover (2001), who suggest that students prefer to pair with each other. Upon entering the test room, candidates are invited to sit down and the test starts immediately, without preamble. The procedure shown in figure 3.2 (reproduced from Appendix III for ease of reference) is then administered.

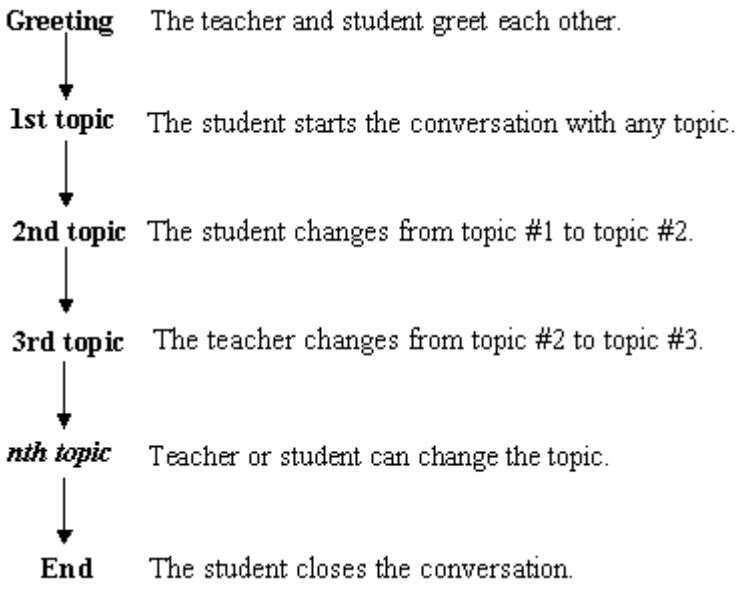


Figure 3.2 - The TECP procedure

Either party can initiate the greeting, after which the examinee must start the conversation on any topic of her choice. There is no prescribed duration for each topic, the participants simply move from one to the next as each topic is exhausted. Topic can shift gradually, or more abruptly via a discourse marker (e.g. 'By the way...'), whichever occurs naturally within the conversation. Where the three topics run their course before the time is up, either party may initiate new topic(s). If the examinee is deemed not to have shifted or changed topic after 5 minutes, it is assumed that she is unlikely to do so and the examiner changes topic to see if the candidate can follow the shift.

If candidates wish to check a translation equivalent, they may use a dictionary. Alternatively, they may ask the examiner, on the condition that the question is presented in English as exemplified in the following (fabricated) exchanges:

Paul Moritoshi, The University of Birmingham, 2002.

Candidate: What's *takai* in English?
Examiner: Expensive.

or

Candidate: What's difficult in Japanese?
Examiner: *Muzukashi*.

This is considered equivalent to, but faster than using a dictionary and optimises the available test time. It is the only occasion on which anything other than English is accepted. If examinees use any other language at any other time, the examiner should react as if he/she does not understand and candidates must negotiate meaning through whatever alternative strategies they have at their disposal.

The examiner should initially use native-speaking speed, but modify as necessary where required. He/she must also use English deemed to be consistent with the candidate's level. It is felt that these concessions conform to usual native-speaker (NS) behaviour during NS-NNS interaction.

A stopwatch, set to give an audible alarm after six minutes is used to time the tests. When the alarm sounds, the examiner closes the current topic at the earliest polite opportunity then gives a 'T' hand gesture to indicate that the time is up. This cues the examinee to close the conversation. (A full transcript from an illustrative TECP test is provided in Appendix IV).

3.4 Scoring and grading systems

Following the literature, an analytic rating scheme was developed to rate examinees' English conversation performances. The scheme (Appendix V) comprises 19 scales for assessing candidates' proficiency in many of those performance criteria identified in section 2.2 as relevant to conversation, as well as rating English ability in terms of vocabulary, grammar and pronunciation. Following Bachman and Palmer (1996: 214-216), each scale is accompanied by a brief description of the performance criterion the scale rates for and how it is operationalised, as well as any necessary

Paul Moritoshi, The University of Birmingham, 2002.

notes for clarifying how the scales are to be applied. An additional rating is calculated automatically using the following Microsoft Excel (Version 7.0) spreadsheet formula:

=SUM(IF(J7>0,SUM(J7-1),0)+IF(M7>0,SUM(M7-1),0)+IF(P7>0,SUM(P7-1),0)+O7+R7)*0.75

This formula takes the ratings for those performance criteria deemed to evidence a subject's willingness to contribute pro-actively to the conversation (i.e. full participation in the greeting ritual, conversation initiation, topic shifting/changing and asking questions) and multiplies them by a factor of 0.75. When added to the original ratings, this gives these performance criteria a weighting of 1.75 and is an attempt to account for and reward this pro-active attribute in an objective way.

The 20 ratings are summed to give a compensatory composite score (Bachman and Palmer, 1996: 224) and then converted into a percentage of the total marks available. Examinees' alphabetical grades are derived directly from this percentage score.

It might be helpful at this point to explain SGU's grading system. The university requires staff to report students' grades expressed as letters, not percentage scores. Table 3.1 shows the percentage bandwidths for each grade, as specified by SGU.

Table 3.1 - The Sanyo Gakuen University grading system

Grade	**Percentage**	**Level**
A	80 - 100 %	Distinction
B	70 - 79 %	Merit
C	60 - 69 %	Pass
D	Less than 60 %	Fail

However, since all students take the same conversation test regardless of their course level, it would be more difficult for lower level students to obtain any given grade than it would be for higher level students. To account for this, the pass mark and

Paul Moritoshi, The University of Birmingham, 2002.

grade bands for lower level groups are shifted further down the percentage scale (albeit arbitrarily) to make it easier for lower level students to obtain any given grade. This in turn requires the above percentage bandwidths to be recalculated for each group, while remaining faithful to the original bandwidth proportions specified by SGU. The recalculated grade bands are presented in table 3.2 below.

Table 3.2 - The recalculated grade bands for each course level

Course level	Grade	Percentage	Level
Level 2 (Beginner)	A	70 - 100%	Distinction
	B	55 - 69.99%	Merit
	C	40 - 54.99%	Pass
	D	< 40%	Fail
Level 3 (Intermediate)	A	77 - 100%	Distinction
	B	64 - 76.99%	Merit
	C	50 - 63.99%	Pass
	D	< 50%	Fail
Level 4 (Upper intermediate)	A	80 - 100%	Distinction
	B	70 - 79.99%	Merit
	C	60 - 69.99%	Pass
	D	< 60%	Fail

For example, level 4 students scoring 67% would get a grade C, while level 2 students with the same percentage score would get a grade B.

This chapter has described the steps by which candidates and examiners prepare for the TECP and by which it is administered, scored and graded. The following chapter goes on to describe the analyses by which the TECP was validated and the results of those analyses.

Paul Moritoshi, The University of Birmingham, 2002.

CHAPTER 4
INVESTIGATING THE TEST OF ENGLISH CONVERSATION PROFICIENCY

4.1 Subjects

The subjects used for this study were those SGU students enrolled on first semester English conversation courses (levels 2-4) for the 2001-2002 academic year. Only those students who had qualified for end-of-semester assessment (by virtue of attending 67% or more of their course lessons) participated in the study (*n* = 56 after the exclusions described in section 4.3.1). Of these, 48 were Japanese, 7 were Chinese and 1 was Korean. All subjects were female.

At the time of testing, the subjects' were 18-30 years of age (mean age = 19 years, standard deviation = $2^{1}/_{3}$ years) and their average length of formal study in English as a foreign language (EFL) was 7 years (standard deviation = $1^{1}/_{4}$ years).

4.2 Situation

The institutional context within which this research is set has already been outlined in section 1.2.

4.3 Data collection

After describing in section 4.3.1 how the dataset for this study was obtained, I will outline how various aspects of the TECP's validity, reliability and practicality were analysed and present the results from those analyses in sections 4.3.2 to 4.3.7.

4.3.1 Candidates' English conversation proficiency

Data on subjects' English conversation proficiency were collected through the test procedure outlined in section 3.3. Some tests ran substantially over-time, (by as much as 5 minutes in some cases). These have been excluded from the dataset since they do not conform to the test protocol and would offer invalid comparisons between test performances.

Paul Moritoshi, The University of Birmingham, 2002.

The tests' video recordings were then transferred on to videotape, viewed and rated. Three administrative instruments were used in the rating process. The first (the rating scheme) has already been detailed in section 3.4 (Appendix V). The second was a coding sheet (Appendix VI), designed for use in conjunction with the rating scheme. It helps raters to code for an examinee's conversational behaviours as the test progresses. Without such a device it would be very difficult (if not impossible) for raters to keep an accurate, relatively objective tally of a candidate's performance in such a wide range of performance criteria, many of which occur very rapidly in sequence or even simultaneously. A fresh coding sheet was used for each subject. The final rating document was a spreadsheet onto which the examiner's ratings were input (Appendix VII). These three administrative instruments were designed with a view to enhancing the TECP's practicality by reducing rating time, whilst also maximising rater accuracy.

In an attempt to reduce data input error, the raw data then underwent a validation check to ensure that they fell within the numerical parameters for each scale.

4.3.2 Construct validity

The correlational method introduced in section 2.1.5 was used to assess the test's construct validity. This technique was chosen in preference to others because it was the most practical method available, given the constraints imposed upon this study. The ratings for each of the TECP's scales were correlated in turn with the percentage scores to examine the strength of the relationships between them. For these analyses the first and second markers' ratings for the first test sitting were combined and averaged. This averaged data should theoretically be more rounded and so more accurately reflect any correlates that may exist, but which might otherwise be masked if only one rater's potentially extreme ratings were used. The results are presented in table 4.1 as a correlation profile.

Paul Moritoshi, The University of Birmingham, 2002.

Table 4.1 - Spearman correlation coefficients for construct validity

Scale		ρ	Significant	
			at p<0.01	at p<0.05
1a	Quantity of greeting	0.26	-	-
1b	Quality of greeting	0.13	-	-
2a	Start (hesitation)	0.18	-	-
2b	Starting technique	0.12	-	-
3a	Range of topic changing techniques used	0.71	✓	-
3b	Number of student-initiated topic changes	0.73	✓	-
4a	Question's type	0.75	✓	-
4b	Question's purpose	0.82	✓	-
4c	Number of questions	0.85	✓	-
5a	Non-verbal communication	-0.04	-	-
5b	Back-channel feedback	0.29	-	✓
6	Conversation's cohesion	0.75	✓	-
7	Turn length	0.52	✓	-
8	Turn-taking ability	0.69	✓	-
9	Vocabulary	0.78	✓	-
10a	Grammatical accuracy	0.67	✓	-
10b	Grammatical complexity	0.51	✓	-
11	Pronunciation	0.19	-	-
12	Closing	0.12	-	-
13	Proactive participation	0.87	✓	-

4.3.3 Content validity

Content validity was assessed by the evaluative technique outlined in section 2.1.5. Since this method does not yield numerical data, no results are presented here. Instead, the test's content validity is discussed in section 5.2.

Paul Moritoshi, The University of Birmingham, 2002.

4.3.4 Face validity

A 13-item face validity questionnaire (FVQ) was designed and administered to ascertain subjects' views on how conversation should be tested generally and their reactions to certain aspects of the TECP in particular. Though only item 11 pertains specifically to subjects' perceptions of the test's face validity, the supplementary items were intended to provide additional quantitative and qualitative information which might offer insight into why subjects responded as they did for item 11. This additional information might also prove useful in subsequent test development and helping to further enhance the test's face validity.

Though the FVQ was originally written in English (Appendix VIII), a Japanese version (Appendix IX) was administered to maximise comprehension and depth of subjects' responses. Translation of both the FVQ into Japanese and subjects' comments back into English were performed in close consultation with an English-speaking Japanese person.

Where subjects took the test twice, the FVQ was administered after the second test, but where subjects took the test only once, it was administered after that one sitting. This gave subjects maximum exposure to the test before completing the FVQ, helping them to give more informed responses. In all cases, upon conclusion of each examinee's final test, they were given a copy of the Japanese FVQ before leaving the test room and asked to complete it in the room where the other candidates were waiting. They were asked to take their time, to answer the questions as honestly as possible and, to ensure anonymity, not to write their name or student number on the questionnaire. Completed forms where collected at the end of each testing session.

Of the 56 subjects to take the test, 55 answered item 11 relating directly to their perception of the TECP's face validity. The results for that item are presented in figure 4.1.

Paul Moritoshi, The University of Birmingham, 2002.

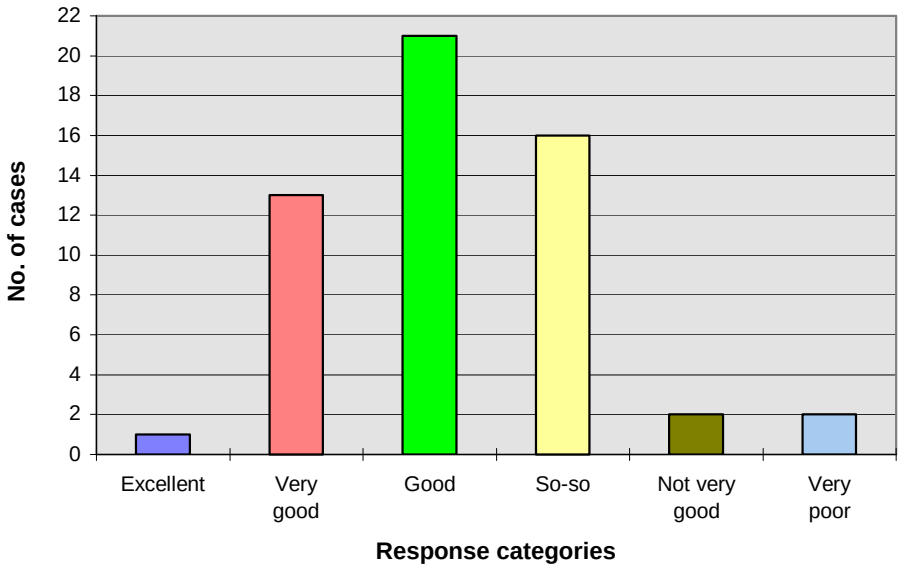


Figure 4.1 - Subjects' perceptions of the TECP's face validity (FVQ item 11)

Since these data are at ordinal level they cannot be averaged to find a mean value. However, 35 cases (64% of the sample) think that the TECP has a good-excellent, i.e. satisfactory face validity, while 20 cases (36%) think that its face validity is so-so or lower, i.e. not satisfactory. To assist in the interpretation of this result further analyses were performed on other FVQ items, the results of which are shown in figures 4.2 to 4.5.

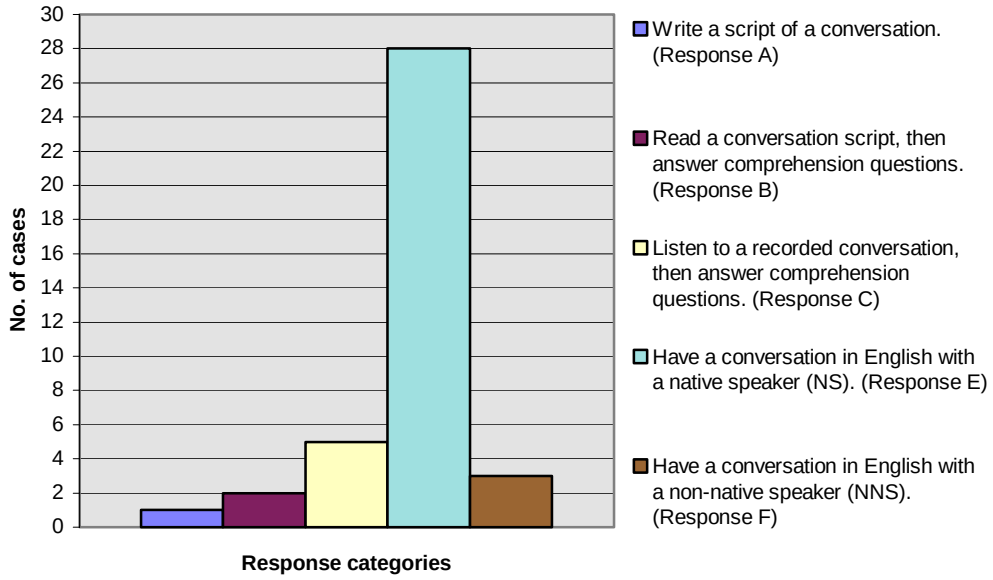


Figure 4.2 - Subjects' opinions on the most accurate way to check English conversation ability (FVQ item 1)

Paul Moritoshi, The University of Birmingham, 2002.

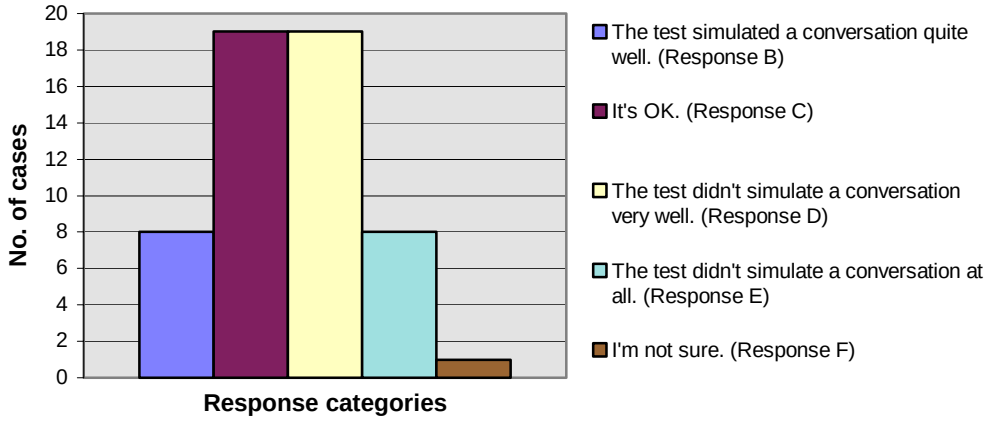


Figure 4.3 - Subjects' perceptions of how well the TECP simulates a conversation (FVQ item 2)

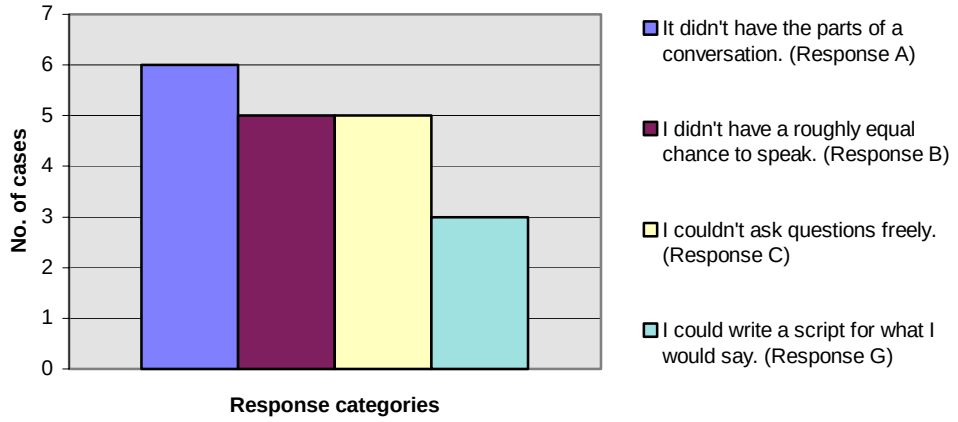


Figure 4.4 - Reasons given as to why the TECP does not simulate a conversation (FVQ item 4)

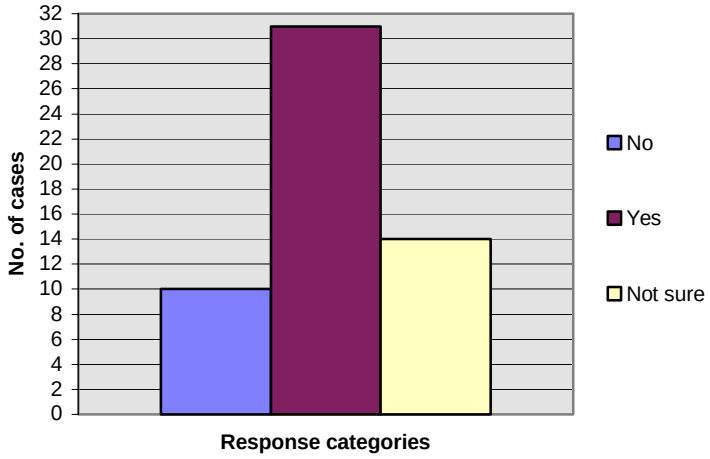


Figure 4.5 - Subjects' perceptions of whether they had enough chances to speak during the TECP (FVQ item 5)

Paul Moritoshi, The University of Birmingham, 2002.

4.3.5 Inter-rater reliability

To measure inter-rater reliability, a native English-speaking second marker independently rated all the tests from the first sitting, using the same rating procedures and instruments described in sections 3.4 and 4.3.1.

Since the second marker lived too far away to make the rater training recommended by Bachman and Palmer (1996: 222) feasible, she was given a set of 'training notes' (Appendix X). These notes detailed how the rating scales, coding sheet and spread-sheet should be used and served as an attempt to standardise rating practices between the two markers. Also, regular contact with the second rater by telephone and e-mail, served to answer questions and to clarify points regarding use and interpretation. Care was taken throughout the entire process not to lead her as to the ratings she should provide or to disclose the ratings given by myself as the first marker.

Two analyses were conducted on the two sets of scores. Firstly, the two markers' ratings were correlated, scale-for-scale to provide an estimate of inter-rater reliability for each separate scale, as well as for the percentage scores and the alphabetical grades. Though it is the correlations for percentage scores and the alphabetical grades that are of principal interest here, it was hoped that performing the additional scale-for-scale correlations might provide some insights into how each scale performed when used by different raters. These insights might prove useful in subsequent test development.

The second analysis expressed as a percentage of the whole sample the number of subjects who were given the same grade by both raters. Though this duplicates to some extent the work already done by the grade-for-grade correlation, it presents the same result in a non-technical way, which those unfamiliar with statistics may find easier to interpret.

The correlation coefficients are presented in table 4.2. Of the 56 subjects who took the test, 42 (75% of the sample) were awarded the same alphabetical grade by both raters. Of the 14 remaining subjects (25%), all obtained grades that had one level of difference between them (e.g. C-B or C-D).

Paul Moritoshi, The University of Birmingham, 2002.

Table 4.2 - Scale-for-scale correlation coefficients for inter-rater reliability

Scale		ρ or r	Significant	
			at p<0.01	at p<0.05
1a	Quantity of greeting	ρ = 0.85	✓	-
1b	Quality of greeting	ρ = 0.24	-	-
2a	Start (hesitation)	ρ = 0.35	✓	-
2b	Starting technique	ρ = 0.57	✓	-
3a	Range of topic changing techniques used	ρ = 0.64	✓	-
3b	Number of student-initiated topic changes	ρ = 0.65	✓	-
4a	Question's type	ρ = 0.62	✓	-
4b	Question's purpose	ρ = 0.57	✓	-
4c	Number of questions	ρ = 0.84	✓	-
5a	Non-verbal communication	ρ = -0.02	-	-
5b	Back-channel feedback	ρ = 0.14	-	-
6	Conversation's cohesion	ρ = 0.70	✓	-
7	Turn length	ρ = 0.62	✓	-
8	Turn-taking ability	ρ = 0.52	✓	-
9	Vocabulary	ρ = 0.63	✓	-
10a	Grammatical accuracy	ρ = 0.53	✓	-
10b	Grammatical complexity	ρ = 0.78	✓	-
11	Pronunciation	ρ = 0.26	-	-
12	Closing	ρ = 0.61	✓	-
13	Proactive participation	ρ = 0.87	✓	-
Percentage score		r = 0.88	✓	-
Alphabetic grade		ρ = 0.84	✓	-

Paul Moritoshi, The University of Birmingham, 2002.

4.3.6 Test-retest reliability

Following the generally accepted procedure, the TECP's test-retest reliability was estimated by correlating the scores that the same subjects received from two test administrations (Hughes, 1989: 32; Bachman, 1990: 181-182; Weir and Roberts, 1994: 172; Bachman and Palmer, 1996: 222).

Of the 56 subjects to take the initial test, 52 took the retest. The period between test sittings ranged from 7 to 10 days, during which time no lessons were given. In order to maximise the accuracy of the results, it was necessary to motivate subjects to perform at their best during both tests. To this end, they were told during the preparatory presentation that their end-of-semester English conversation course grade would be derived from the better of their two test performances and that they should use both chances to get the best grade they possibly could.

To avoid informative feedback that might improve retest performance, only the most cursory of comments were made to subjects as they finished their first test (e.g. "OK. Thanks. See you on Monday."). The second round of tests duplicated the first in every respect but one: to avoid a simple repetition of the first tests' topical contents, subjects were asked to use different conversational topics during the retest.

As with inter-rater reliability, though only the correlations for percentage scores and alphabetical grades are essential to this study, performing additional scale-for-scale correlations might offer some insights into how each scale behaves across test administrations. Also as with inter-rater reliability, in addition to the correlation estimates, the number of subjects receiving the same grade in both test sittings was calculated as a percentage of the whole sample.

The correlation coefficients for test-retest reliability are presented in table 4.3. Of the 52 subjects who were retested, 26 (50% of the sample) received the same alphabetical grade for both tests. Of the remaining 26 subjects, 24 (46%) obtained grades that had one level of difference between them (e.g. C-B, or A-B) and 2 subjects (4%) received grades that had two levels of difference (i.e. C-A and A-C).

Paul Moritoshi, The University of Birmingham, 2002.

Table 4.3 - Scale-for-scale correlation coefficients for test-retest reliability

Scales		ρ or r	Significant	
			at p<0.01	at p<0.05
1a	Quantity of greeting	ρ = 0.27	-	-
1b	Quality of greeting	ρ = -0.05	-	-
2a	Start (hesitation)	ρ = 0.29	-	✓
2b	Starting technique	ρ = -0.02	-	-
3a	Range of topic changing techniques used	ρ = 0.25	-	-
3b	Number of student-initiated topic changes	ρ = 0.26	-	-
4a	Question's type	ρ = 0.26	-	-
4b	Question's purpose	ρ = 0.37	✓	-
4c	Number of questions	ρ = 0.50	✓	-
5a	Non-verbal communication	ρ = -0.02	-	-
5b	Back-channel feedback	ρ = 0.20	-	-
6	Conversation's cohesion	ρ = 0.65	✓	-
7	Turn length	ρ = 0.49	✓	-
8	Turn-taking ability	ρ = 0.39	✓	-
9	Vocabulary	ρ = 0.59	✓	-
10a	Grammatical accuracy	ρ = 0.41	✓	-
10b	Grammatical complexity	ρ = 0.66	✓	-
11	Pronunciation	ρ = 0.24	-	-
12	Closing	ρ = 0.24	-	-
13	Proactive participation	ρ = 0.35	-	✓
Percentage score		r = 0.67	✓	-
Alphabetic grade		ρ = 0.55	✓	-

Finally, it is useful to note that 2 subjects (4% of the sample) failed one test but passed the other.

Paul Moritoshi, The University of Birmingham, 2002.

4.3.7 Test practicality

To objectively measure the test's practicality I timed how long it took to rate each test. The average rating time could then be added to the approximate time required to administer the test itself, to give an approximate average total time required per test. The results from this analysis are presented in table 4.4.

Table 4.4 - Approximate average total time required per test

Approximate individual test time	$6^{1}/_{2}$ minutes
Mean rating time	16 minutes (SD = 2, min. = 11, max. = 21, range = 10)
Approximate average total time per test	$22^{1}/_{2}$ minutes

These various results and their implications for the TECP's continued use and development are discussed in the following chapter.

Paul Moritoshi, The University of Birmingham, 2002.

CHAPTER 5
DISCUSSION OF THE VALIDATION ANALYSES

Much of the following discussion relies upon the ability to interpret results from a particular type of statistical procedure called correlations. It might therefore be useful to the reader if a brief explanation was given as to how correlation results are interpreted, before going on to discuss the results presented in chapter 4.

Where correlations are used, two values are usually cited. The first is the correlation coefficient (ρ or r), expressed as a value between 0 and 1 (or -1), which indicates the extent to which two (or more) variables are related. Where the coefficient is positive, the variables are directly related (i.e. one increases as the other increases, or one decreases as the other decreases). Where a coefficient is negative, the variables are inversely related (i.e. one increases as the other decreases, or *vice versa*). The closer the coefficient is to 1 (or -1), the stronger the relationship appears to be and the more potentially persuasive the result becomes until, at a value of 1 (or -1), the relationship is considered directly (or inversely) perfect.

The second value cited is the level of significance (p), which indicates the amount of confidence with which one can say that the coefficient is important or meaningful, i.e. significant. For example, $\rho = 0.9$, $p<0.01$ indicates a very strong, direct relationship between the variables tested and that this coefficient has a 99% chance of being significant. If $p<0.05$, the chance that it is significant is reduced to 95%. It is then for the individual to decide whether or not these two values together constitute persuasive evidence.

5.1 Construct validity

I will discuss the TECP's construct validity by addressing two issues:

1. How well do the scale ratings correlate with the percentage scores, i.e. how relevant is each scale to the measurement of conversation as a construct?
2. What is the TECP's overall construct validity?

Paul Moritoshi, The University of Birmingham, 2002.

The correlation profile presented in table 4.1 gives quite a mixture of results. Since it is beyond the scope of this paper to examine each correlation in detail, it must suffice only to highlight and discuss certain trends that the profile exhibits.

Perhaps the most striking finding is that the ratings for greeting and closing behaviours (scales 1a, 1b and 12 respectively) are not significantly correlated with examinees' percentage scores. Are these behaviours therefore irrelevant to conversation proficiency? Not necessarily. The lack of correlations here is likely due not to these performance criteria being irrelevant to the construct, so much as to them being so easy to master that most of the examinees got the maximum ratings on these scales. When Spearman's Rank Order Correlation is run on such data, it gives the impression of irrelevance. What these paradoxical findings do indicate however is that rating for such easy or well-practised behaviours will not necessarily help to discriminate between the conversationally proficient and non-proficient.

The non-significant and, more revealingly, negative correlation between non-verbal communication (scale 5a) and percentage scores might indeed indicate that it is irrelevant to the construct. It seems unlikely after all that the worse someone is at using non-verbal communication, the better they are at conversation. An alternative explanation however might be that it is relevant, but that the scale's descriptors do not sufficiently discriminate between those who use non-verbal communication effectively during conversation and those who do not. With the data currently available it is not possible to show which explanation is correct, though further research might resolve the issue.

The remaining non-significant correlations relating to how candidates start the conversation (scales 2a and 2b) and pronunciation (scale 11) are in all likelihood irrelevant to the construct. These scales might therefore safely be eliminated from the rating scheme, especially since the hesitation and questioning elements contained within scales 2a and 2b are also covered by other, apparently relevant scales, namely questioning behaviours (scales 4a-4c) and conversational cohesion (scale 6).

Paul Moritoshi, The University of Birmingham, 2002.

The remaining 13 scales relating to topic shifting/changing (scales 3a and 3b); questioning behaviours (4a-4c); back-channel feedback (5b); conversational cohesion (6); turn length (7); turn-taking (8); lexical and grammatical competence (9 and 10a-10b respectively) and the computed score for proactive participation (13) are all significantly correlated to varying degrees with the percentage scores, usually at a very high level of significance ($p<0.01$). These results indicate that these performance criteria are relevant to the construct of conversation.

Given that the irrelevant scales constitute only 14 out of the total of 85 points available on the test (i.e. 16.5%), their impact on percentage scores might be considered minor. However, their inclusion does somewhat reduce the TECP's construct validity in its current form. Nonetheless, given also that 13 of the 20 scales have been shown to be relevant to the construct, and that a further 3 scales for greeting and closing are likely to be relevant, the TECP's construct validity might fairly be described as good-high. The null hypothesis is therefore tentatively rejected. It is important to remember though that the correlational technique used to evaluate the TECP's construct validity has its limitations, most notably that it cannot identify those performance criteria which are relevant to the construct, but which have not been incorporated into the test's design.

5.2 Content validity

I will discuss the TECP's content validity with respect to two criteria: test tasks and topical contents. Since the TECP's content is not based upon empirical research evidencing the relative frequencies with which conversational tasks and topics occur in the TLU domain, it is conceded that no claim can be made for content *representativeness* beyond the fact that greeting and closing occur only once in the test, just as they normally do in real conversations. A strong intuitive case though can be presented for the tasks' and topical contents' *relevance*.

The procedure depicted in figure 3.2 shows that candidates are required to greet, initiate and close the conversation and to change or shift the topic at least once during the test. These are mandatory tasks. Additional behaviours such as asking

Paul Moritoshi, The University of Birmingham, 2002.

questions, supplying back-channel feedback and turn-taking, are not tasks in the strictest sense of the word, though of course without them the dyad would cease to be a conversation. It appears therefore that the tasks and supporting behaviours that the TECP requires candidates to perform and on which they are rated, mirror those behaviours identified in section 2.2 as necessary or useful during natural conversation. The TECP's content validity with respect to task relevance thus appears to be quite high.

Appendix XI provides a sample of the topics talked about during the TECP tests used for this study. (I have attempted to make this sample as representative of the whole as possible by indicating relative frequency.) Two points are worthy of note regarding this sample. Firstly, that while the topics used range from the commonplace (e.g. weekend plans), to the obscure (e.g. differences between Korea and Japan), they seem to be what one might expect during conversation with (on average) 19 year old female university students, of various Asian nationalities. Topics that might be expected to fall outside their topical knowledge base or to be of little or no relevance or interest to them, such as scientific, technical or political subject matter, simply did not arise. This can be largely explained by the second point: that two of the three topics used during the TECP procedure are candidate-initiated. In other words, the topical contents are relevant and/or of interest to candidates and are generally within their topical knowledge base because it is they who nominate most of the topics used. Even those topics that only I initiated (i.e. the end-of-semester tests and students' future careers) are relevant, since they affect students directly, either in the short- or long-term.

In short, the tasks the TECP requires candidates to perform are highly relevant to the construct of conversation and its topical contents generally appear to be self-selectively relevant and/or of interest to examinees and within their topical knowledge base. The TECP's content therefore appears to meet the criteria set by the literature for high authenticity (Underhill, 1987: 8; Hughes, 1989: 15; Bachman and Palmer, 1996: 23-25) and content relevance (Bachman, 1990: 244-245). The null hypothesis regarding the TECP's content validity is therefore rejected, though with

Paul Moritoshi, The University of Birmingham, 2002.

the concession that the evidence presented here is intuitive and does not account for content validity with respect to representativeness.

As noted in section 2.1.5, there is often a strong link between content validity and examinees' perceptions of the test itself. If, as seems to be the case, the TECP's content validity is high, at least with respect to relevance, does this also show itself in the form of high face validity?

5.3 Face validity

Some 64% of the sample perceived the TECP to have satisfactory face validity, while 36% did not. In so far as this finding and the results presented in figure 4.1 can be 'averaged' to provide an assessment of this quality, it might most fairly be described as good. The null hypothesis that it does not have high face validity is therefore accepted, though with the concession that it does possess this quality to an adequate degree.

Why does such a direct measure have only moderate face validity, particularly when the results presented in figure 4.2 show that 31 subjects (72% of the sample) think that the best way to measure English conversation ability is to have a conversation with either a NS or NNS? The results shown in figure 4.3 provide one likely explanation: that half the sample felt that the TECP did not simulate a conversation very well, if at all. One reason given for this view was that they did not have a roughly *equal* chance to speak (see figure 4.4), though this is somewhat contradicted by the finding presented in figure 4.5 that only 10 subjects (18%) felt they did not have *enough* chances to speak.

Figure 4.4 also shows another reason why some subjects perceived the TECP not to simulate a conversation: that they could not ask questions freely. The reason(s) for this are unclear, but possible causes include the following. Firstly, some examinees did not recognise or utilise potential avenues of inquiry as they arose during the conversation. This was apparent in several cases. Secondly, it is possible that as an examiner I did not consistently create sufficient lines of potential inquiry for the

Paul Moritoshi, The University of Birmingham, 2002.

examinees to utilise. Thirdly, that some subjects did not attempt to create such lines of inquiry for themselves, as they would be required to do in normal conversation. Finally, that perhaps some of the subjects could not overcome what is, for Asians, a very real and substantial psychological and cultural hurdle: that of talking to a teacher conversationally, on a more-or-less equal basis in a test setting. Instead, perhaps they continued to apply the rules that they would normally use when talking to a teacher: the rules of 'traditional' classroom discourse, whereby the teacher asks the questions and the students answer them, not *vice versa*. Whatever the cause, if the TECP is to simulate a normal conversation effectively, it is important that candidates understand and use their right to ask questions during the test. Apparently some work is required to improve students' understanding on this point.

Two other reasons as to why the TECP does not simulate a conversation, i.e. that it does not have the component parts of a conversation and that candidates could script what they would say beforehand (see figure 4.4), are less easily explained and warrant further investigation.

As noted in section 2.1.5, where examinees lack confidence in a test, this might adversely affect test implementation. The main concern is that such candidates will not perform at their peak and that their scores will not therefore accurately reflect their true proficiency (Hughes, 1989: 27; Brown, 1994: 256). Since the results presented in figure 4.1 show that this may apply to a substantial minority of subjects (36% of the sample), clearly some work will have to be done to enhance the TECP's face validity. It is hoped that the enhancements in the test's procedure that result from this validation study, in tandem with the additional information gained from the FVQ will help in this task. There is some cause for optimism. Firstly, it is important to remember that this somewhat disappointing finding does not in itself negate those above for construct and content validity (Bachman and Palmer, 1996: 42). Secondly, that 64% of the subjects think that the test is good or better is a promising starting point from which to improve the perceptions of the remaining 36%.

Paul Moritoshi, The University of Birmingham, 2002.

5.4 Inter-rater reliability

Table 4.2 shows the inter-rater reliability estimates for percentage scores and grades. At $r = 0.88$ and $\rho = 0.84$ respectively (equivalent to 75% commonality in grades awarded), these are quite high and the associated null hypothesis is rejected.

That the two markers' scale-for-scale ratings usually correlated quite well is a strong indication that the scales are generally very clear, unambiguous and well supported by keys and notes to assist in their consistent application. It might also suggest that each scale has a suitable number of levels for describing and rating a candidate's performance: not too few that they lack discriminatory power, but not so many that raters cannot easily select the most appropriate rating. Further, since many of the suggestions made in section 2.1.4 for enhancing the objectivity and reliability of rating scales were taken up in the design and application of the TECP rating system, these findings offer circumstantial support for the efficacy of those suggestions.

Feedback from the second rater raised an interesting point relating to the scales for non-verbal communication (5a) and back-channel feedback (5b). In a personal communication she pointed out that she initially rated for these behaviours in line with Western communication style, i.e. she focused on 'eye contact, strong facial expressions, hand-movement and body language'. It was later agreed to also include such things as nods, grunts and smiles, which are common to the Japanese communication style (Tannen, 1994: 69, citing Hayashi, 1988), even when Japanese are speaking in English to native speakers (Takahashi, 1989: 250; Hattori, 1987, cited by Takahashi, ibid). This reinforces the literature's observation that, while non-verbal communication and back-channelling behaviours may have some commonality across cultures, they also each have their own idiosyncrasies (LoCastro, 1987: 112; Holmes, 1992: 304). One implication this has for the measurement of conversation proficiency is that scales intended to measure these (and perhaps other) performance criteria may have to be tailored to the cultural group (the nationality?) being tested. Other than that however, the second rater noted that the rating scheme 'was basically quite simple to use and follow'.

Paul Moritoshi, The University of Birmingham, 2002.

The TECP's inter-rater reliability could still be improved. For example, table 4.2 shows that reliability for scales 1b, 5a, 5b and 11 is very low. Others (6, 8, 9, 10a and 11) that currently use high-inference descriptions could be redesigned to use lower levels of inference. However, it appears that in its present form, the TECP generally yields quite reliable scores and grades between raters. How then does it perform across administrations? What is its test-retest reliability?

5.5 Test-retest reliability

The results presented in table 4.3 show the TECP to have statistically significant test-retest reliability with respect both to percentage scores and their equivalent grades ($r = 0.67$ and $\rho = 0.55$ respectively, both at $p<0.01$). However, in line with Upshur and Turner (1995: 5) in section 2.1.5, these correlations are not particularly strong. While the coefficient for the percentage scores approaches the acceptable minimum of 0.7 suggested by Hughes (1989: 32, citing Lado, 1961), that for the grades does not. What this means is perhaps clearer if one considers that only 50% of the subjects were given the same grade for both test sittings and that two subjects' grades differed by as much as two levels (i.e. grades A-C and C-A). Test-retest reliability might therefore at best be considered only moderate and the associated null hypothesis is accepted.

This disappointing result is probably the TECP's greatest weakness, but comes as little surprise. Weir (1993: 57) predicted that 'candidate performances are likely to vary from one [test] to the next' in procedures which permit 'constructive interplay with unpredictable stimuli' (Weir, ibid, quoting Carroll, 1980), in other words spontaneity of target language use. Also, one of the main reasons why there are so few (if any?) validated tests of conversation proficiency widely available is that they are, among other things, notoriously difficult to make reliable. It is for this reason that Bachman and Palmer offer the following advice concerning reliability in general:

> Reliability is clearly an essential quality of test scores, for unless test scores are relatively consistent, they cannot provide us with any information at all about the ability we want to measure. At the same time, we need to recognize that it is not possible to eliminate inconsistencies entirely. What we can do, however, is

Paul Moritoshi, The University of Birmingham, 2002.

> try to minimize the effects of those potential sources of inconsistency that are under our control, through test design.
>
> (Bachman and Palmer, 1996: 20).

That is to say that firstly, one should be realistic about how reliable tests (particularly those like the TECP which measure productive abilities) can ever be and secondly, that it should be possible to maximally enhance reliability by identifying and overcoming all of those causes of unreliability over which designers have some influence. What then are the likely causes of the TECP's disappointing test-retest reliability?

The results listed in table 4.3 offer one probable cause. The test-retest reliability is relatively (if not very) weak for all the scale-for-scale estimates, but most revealing are those for the scales for topic changing/shifting (3a and 3b) and questioning behaviours (4a-4c). These scales contribute a substantial (and in retrospect perhaps excessive) amount to examinees' percentage scores because they have a weighting of 1.75. If, for example, a candidate did not shift topic or ask any questions during the first test, then shifted topic once and asked only one question during the second test, their score would go up by 11.8%. Table 3.2 shows that this alone would probably be enough to bridge the gap between one grade-band and the next and in extreme cases, where candidates asked several more questions in one test than in another, it could even bridge the gap between two grade-bands.

Reducing the influence these scales have upon percentage scores and grades by reducing their weighting to 1 might help to enhance test-rest reliability. A further analysis along these lines revealed that while the test-retest coefficient for the percentage scores remained unchanged ($r = 0.67$, significant at $p<0.01$), that for the grades increased from $\rho = 0.55$ (significant at $p<0.01$) to $\rho = 0.63$. In other words, the percentage of subjects getting the same grade in both tests increased from 50% to 60%. Given this promising result, the estimate for inter-rater reliability was also recalculated in the same way. While the coefficient for percentage scores remained essentially unchanged (from $r = 0.88$ to 0.87), that for inter-rater grade reliability went down considerably from $\rho = 0.84$ to 0.70, (i.e. the percentage of subjects

Paul Moritoshi, The University of Birmingham, 2002.

getting the same grade from both raters decreased from 75% to 66%). This suggests that the 'tension' that exists between validity and reliability (Hughes, 1989: 42; Bachman, 1990: 289; Bachman and Palmer, 1996: 18; Owen, 1997: 21-22) might also be present between inter-rater and test-retest reliabilities, i.e. between how a test performs within and across administrations. Why this is so is unclear, but warrants further investigation. Consequently, careful consideration will be required in deciding whether or not to eliminate scale 13's complicated formula (see section 3.4) from future TECP tests to bring the weighting for questioning and topic shifting down to 1. Alternatively, further trials with weightings between 1-1.75 might offer an acceptable compromise between gains in test-retest reliability and reductions in inter-rater reliability.

Another probable source of unreliability is the inclusion of what section 5.1 has shown to be irrelevant scales. Future research will determine whether or not eliminating these from the scoring system further increases test-retest reliability.

These are two sources of test-retest unreliability in the TECP's present design and there are almost certainly others. In short, the TECP does not currently provide candidates with sufficiently reliable results. The implications for assigning unrepresentative grades are of course very serious. Academic transcripts that do not accurately reflect students' abilities will give a false impression to prospective employers. Also, extreme cases of unreliability can cause a few examinees to be dubiously failed, as happened to 2 subjects in this study (4% of the sample). Such students lose course credits and would have to retake the module, adversely affecting their capacity to take other subjects later in their degree course.

5.6 Test practicality

Sections 2.1.4 and 2.1.5 highlighted the requirement that a test be practical, i.e. that the financial, material and human resources it requires do not exceed those that are available. The TECP appears to meet this requirement. It does not call for additional funds beyond those already allocated for the examiner's regular wages and the material resources required (a video camera, tripod, two rooms, some tables and

Paul Moritoshi, The University of Birmingham, 2002.

chairs, a video player and a television) are already on SGU's inventory. Further, the approximate average of $22^{1}/_{2}$ minutes required to administer and score each test (see table 4.4), falls well within that allocated by SGU for testing (30 minutes per student). The TECP's practicality therefore appears to be quite high and the associated null hypothesis is cautiously rejected. It should be added however that this finding's generalisability might not be very high. If the TECP were used in other settings where, for example, there were much larger classes, it might exceed the (particularly human) resources available.

Though the TECP's practicality is quite high, as with other aspects of its design, this area could be further improved. The greatest gains here may come from reducing the time required for scoring, but such reductions should not be achieved at the expense of the test's validity or reliability.

5.7 Test usefulness

Bachman and Palmer's (1996: 18) composite 'usefulness' (section 2.1.5) comprises: reliability; construct validity; authenticity; interactiveness; impact and practicality. Reliability, construct validity and practicality have already been addressed in previous sections. How does the TECP fare with respect to the remaining components?

Section 2.1.4 has already described a test's authenticity as being dependant upon the extent to which its tasks mirror the corresponding TLU domain behaviours (Bachman and Palmer, 1996: 23-24). Since the TECP's tasks have been shown in section 5.2 to be very similar (if not identical) to those behaviours that would occur during real-life conversation, it might be reasonable to say that the TECP has a high authenticity by virtue of the high relevance of its tasks. In support, Underhill (1987: 8) writes that 'To engage in free conversation is an authentic task' and higher authenticity is likely to be one of the advantages inherent in direct tests such as the TECP. However, no test can ever be totally authentic since, as Jones (1985: 81) writes, all are 'to some degree contrived' (Hughes, 1989: 15; Baker, 1989: 11, cited by Owen, 1997: 15).

Paul Moritoshi, The University of Birmingham, 2002.

Evaluating interactiveness is more problematic, but it might be reasonable to suppose that the TECP engages candidates' L2 abilities, affective schemata of 'conversation' and topical knowledge to quite a high degree because examinees must:

1. process both incoming and outgoing messages almost exclusively in English,
2. do so in real-time,
3. negotiate meaning,
4. participate in an unpredictable dyad,
5. conform to the Gricean conversational principles and
6. nominate two of the topics used during the procedure.

The TECP's interactiveness is therefore tentatively considered to be quite high.

Moritoshi (2001) demonstrates how the TECP's development has dramatically improved the content relevance and overall coherence of SGU's conversation courses' syllabi. Initially, these courses used 'conversation' textbooks which focused on grammar, vocabulary and fixed phrases, i.e. linguistic knowledge, but which apparently did little (if anything) to develop students' conversation proficiency. As a direct result of the TECP's introduction, materials and activities have been produced to improve students' proficiency in greeting; back-channelling; non-verbal communication; questioning; topic shifting/changing; and closing behaviours in English. These materials and activities are also used to highlight features of sociolinguistic, discourse and strategic competence, in attempts to raise students' awareness of these areas of conversational 'communicative competence' as defined by Canale (1983: 7-14). Though no empirical evidence is available to show that these developments in syllabus and material design have led to greater gains in conversation proficiency than the syllabi based predominantly on conversation textbooks, it seems likely that they have the potential to do so. If this is the case, beneficial washback could be considered to be very high and it would follow that impact on individuals (both students and teacher) and classes would also be

Paul Moritoshi, The University of Birmingham, 2002.

considerable. Impact might therefore be cautiously estimated as high, though this does not account for the TECP's impact at the institutional or social levels.

Substituting these various findings back into Bachman and Palmer's (1996: 18) formula for test usefulness gives the following profile:

- Test-retest reliability (moderate).
- Inter-rater reliability (quite high).
- Construct validity (good-high).
- Authenticity (high).
- Interactiveness (quite high).
- Impact at the individual and class level (high).
- Practicality (quite high).

Assuming that each factor has an equal weighting, overall test usefulness might cautiously be described as quite high and the associated null hypothesis is cautiously rejected.

5.8 Future enhancements to the TECP

This validation study has highlighted various ways in which the TECP can be further enhanced and some recommendations have already been made in the relevant sections. Additional improvements include:

- Producing a short video explaining the test's format, rules, necessary preparation etc., both to further enhance the pre-test presentation's standardisation and to ensure that no necessary information is omitted.
- Asking candidates to provide a list of topics that interest them (e.g. hobbies and free-time activities) to enhance the topical contents' fairness and relevance.

Paul Moritoshi, The University of Birmingham, 2002.

- Drawing more overt attention to the links that exist between the TECP's design on the one hand and the conversation courses' contents and conversation as a unitary concept on the other, in an attempt to increase face validity.
- Including a warm-up period at the start of each testing session, which would be particularly useful for the first few candidates who would otherwise have to go in 'cold'. This might further enhance test fairness.
- Using the improved test configuration shown in figure 5.1.

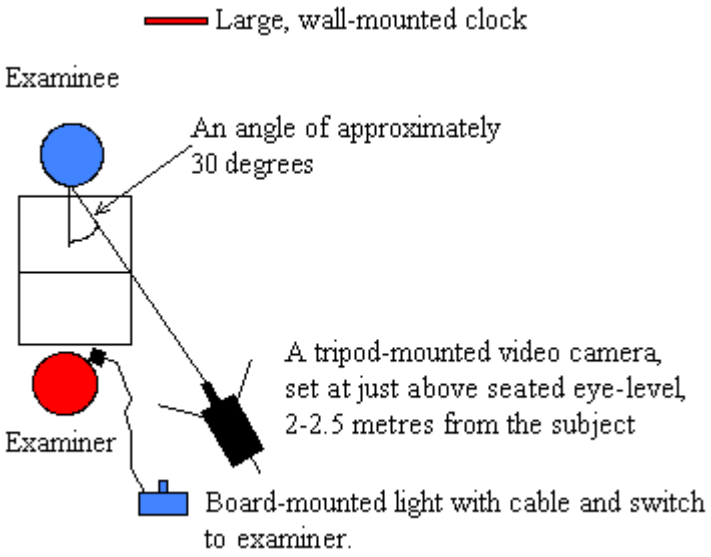


Figure 5.1 - Suggested improvements in the test configuration

The board-mounted light operated by the examiner offers a more natural (or at least less unnatural) method of indicating test termination than the one currently in use (described at the end of section 3.3 and exemplified at the end of Appendix IV). The lamp goes on after 6 minutes, regardless of how the discourse stands with regard to the present topic. Candidates must then create the opportunity to terminate the current topic at the earliest suitable moment and then close the conversation. This innovation transfers responsibility for this former task from the examiner to the candidate. In addition, the wall-mounted clocks will be less distracting for both participants than the hand-held stopwatch currently used.

Paul Moritoshi, The University of Birmingham, 2002.

- Eliminating irrelevant scales (2a, 2b and 11) from the scoring scheme and re-evaluating the number of levels and descriptions used by the remaining scales, particularly with regard to the level of inference required. Any subsequent reduction in rating time could then either be used to increase practicality or to extend the test time for collecting a larger ratable sample.

5.9 This study's limitations

It is conceded that though the sample size used here is sufficient for statistical inference and represents 65% of the population from which it is taken, it is still relatively small. It is also accepted that the generalisability of these findings will be low, though in its defence, the TECP is not currently intended for use beyond the SGU setting.

This study has examined the TECP's validity, reliability, practicality and overall usefulness as thoroughly as possible within the prevailing constraints. However, no attempt was made to evaluate the TECP's fairness. This is a particularly important omission for validation of a topic-based test such as the TECP (Jennings et al., 1999). Also, the correlational technique used to evaluate construct validity is itself limited. Further, the evaluation of the test's usefulness, particularly with respect to its authenticity, interactiveness and impact, is based largely upon intuition, rather than being empirically derived.

Another omission is that little attempt was made to evaluate the representativeness of the TECP's tasks or topical contents. That it was not possible to do so perhaps indicates a gap in the ELT literature relating to the frequency with which certain conversational tasks and topics arise. Though research in this area would make a valuable contribution to the field, it is quite possible that the findings of such research would not be generalisable beyond the speech community from which the data were collected. The implication is that in order to evidence the TECP's content validity with respect to representativeness, such research would probably have to be conducted in-house. Subsequent validations should where possible attempt to overcome these various limitations.

Paul Moritoshi, The University of Birmingham, 2002.

CHAPTER 6
CONCLUSION

Working within the prevailing limited resources, this study has applied a series of validation tests to the TECP with a view to facilitating an informed decision as to the advisability of its continued use. The analyses indicate that the TECP has the following qualities:

1. Good-high construct validity.
2. High content validity with respect to test tasks' and topical contents' relevance.
3. Good (though not high) face validity.
4. Quite high inter-rater reliability.
5. Moderate test-retest reliability.
6. Quite high practicality and overall usefulness.

These findings indicate that the TECP is based upon a relatively sound conceptual definition of conversation; that the tasks and topical contents it uses are relevant to that construct and that the rating scheme employed is relatively clear and unambiguous. There are however concerns pertaining to candidates' perceptions of the test and to its ability to award consistent grades across administrations. The implications these weaknesses have for examinees' test performance in the case of the former, and for students' employment and for their prospective employees in the case of the latter, make continued use of the TECP in its current form ill-advised. However, with the implementation of the recommendations made in section 5.8, it should be possible to enhance the TECP's performance in these areas. It seems acceptable therefore to sanction the TECP's continued use, on the condition that these recommendations are implemented and that their effect(s), if any, are ascertained through ongoing research.

The TECP in its current form is the culmination of over 18 months of work, but it still has its weak-points. This is in no small part due to the sheer complexities

Paul Moritoshi, The University of Birmingham, 2002.

involved, not only in understanding what conversation is and how it works, but also in designing a useful test that can measure conversation proficiency validly, reliably and practically. It does however raise three questions for professional ELT testing practices. Firstly, are the tests we administer (or take?) on a regular basis as accurate as we think they are, or as accurate as they should be? Secondly, can we, as professionals, justify the use of tests that have not undergone at least small-scale validation in order to evidence the extent of their validity, reliability and practicality, or to highlight their weaknesses? Finally, and perhaps most importantly, how confident can we be that the grades we give our students are an accurate reflection of their true abilities?

Paul Moritoshi, The University of Birmingham, 2002.

APPENDIX I
EXAMPLES OF ANALYTIC RATING SCALES

Grammar

Points	Description
1	Grammar almost entirely inaccurate phrases.
2	Constant errors showing control of very few major patterns and frequently preventing communication.
3	Frequent errors showing some major patterns uncontrolled and causing occasional irritation and misunderstanding.
4	Occasional errors showing imperfect control of some patterns but no weakness that causes misunderstanding.
5	Few errors, with no patterns of failure.
6	No more than two errors during the interview.

Vocabulary

Points	Description
1	Vocabulary inadequate for even the simplest conversation.
2	Vocabulary limited to basic personal and survival areas (time, food, transportation, family, etc.).
3	Choice of words sometimes inaccurate, limitations of vocabulary prevent discussion of some common professional and social topics.
4	Professional vocabulary adequate to discuss special interests; general vocabulary permits discussion of any non-technical subject with some circumlocutions.
5	Professional vocabulary broad and precise; general vocabulary adequate to cope with complex practical problems and varied social situations.
6	Vocabulary apparently as accurate and extensive as that of an educated native speaker.

(Hughes, 1989: 111).

Paul Moritoshi, The University of Birmingham, 2002.

APPENDIX II
AN EXAMPLE OF A GLOBAL IMPRESSION MARKING SCHEME

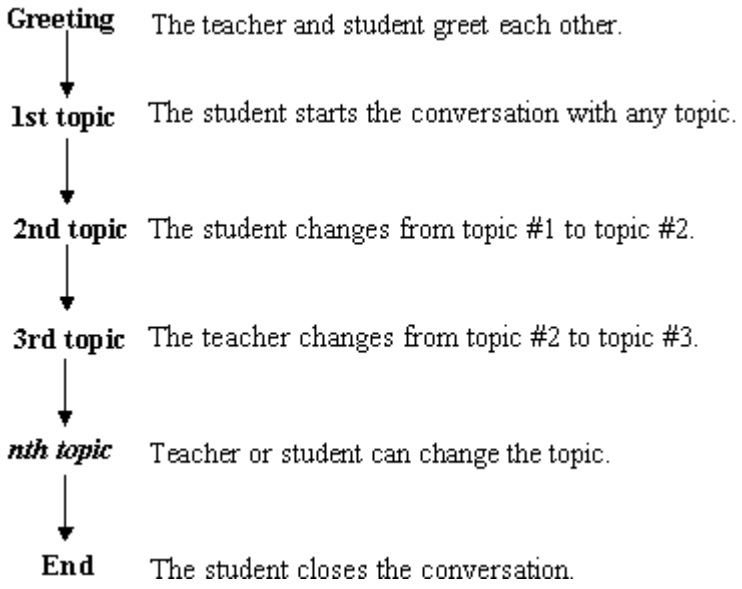
Band	Description
9	Expert speaker. Speaks with authority on a variety of topics. Can initiate, expand, and develop a theme.
8	Very good non-native speaker. Maintains effectively his own part of a discussion. Initiates, maintains and elaborates as necessary. Reveals humour where needed and responds to attitudinal tones.
7	Good speaker. Presents case clearly and logically and can develop the dialogue coherently and constructively. Rather less flexible and fluent than Band 8 performer but can respond to main changes of tone or topic. Some hesitation and repetition due to a measure of language restriction but interacts effectively.
6	Competent speaker. Is able to maintain theme of dialogue, to follow topic switches and to use and appreciate main attitude markers. Stumbles and hesitates at times but is reasonably fluent otherwise. Some errors and inappropriate language but these will not impede exchange of views. Shows some independence in discussion with ability to initiate.
5	Modest speaker. Although gist of dialogue is relevant and can be basically understood, there are noticeable deficiencies in mastery of language patterns and style. Needs to ask for repetition or clarification and similarly to be asked for them. Lacks flexibility and initiative. The interviewer has to speak rather deliberately. Copes but not with great style or interest.
4	Marginal speaker. Can maintain dialogue but only in a rather passive manner, rarely taking initiative or guiding the discussion. Has difficulty in following English at normal speed; lacks fluency and probably accuracy in speaking. The dialogue is therefore neither easy nor flowing. Nevertheless, gives the impression that he is in touch with the gist of the dialogue even if not wholly master of it. Marked L1 accent.
3	Extremely limited speaker. Dialogue is a drawn-out affair punctuated with hesitations and misunderstandings. Only catches part of normal speech and unable to produce continuous and accurate discourse. Basic merit is just hanging on to discussion gist, without making major contribution to it.
2	Intermittent speaker. No working facility; occasional, sporadic communication.
1/0	Non-speaker. Not able to understand and/or speak.

(Carroll, 1980, cited by Weir, 1993: 44).

Paul Moritoshi, The University of Birmingham, 2002.

APPENDIX III
INSTRUCTIONS TO TECP CANDIDATES

1 This is the test format:



2 For the test, please prepare:

- Two topics and the vocabulary and grammar necessary to talk about them.
- A reason to finish the conversation.

3 Please bring the following to the test:

- Japanese-English and English-Japanese dictionaries.
- A pen / pencil.
- A clean sheet of paper.
- You ***can't*** bring anything else into the test room (e.g. photos or magazines).

4 Points to remember:

- This is a conversation, not an interview, so you can change the topic, ask questions and interrupt. Please take an active part in the conversation and relax.
- This is a conversation, not a speech. You can't write a script and you must give the other person a chance to speak sometimes.
- You can't use Japanese, but you can ask the meaning of a Japanese word in English.

Paul Moritoshi, The University of Birmingham, 2002.

5 The checking system.

Your conversation will be checked on these points:

Point 1 Greeting.

Point 2 Starting the conversation.

Point 3 Changing the topic.

Point 4 Asking questions.

Point 5 Using non-verbal communication and giving reactions.

Point 6 The conversation's pace.

Point 7 Length of speech.

Point 8 Taking turns to speak.

Point 9 Vocabulary.

Point 10 Grammar.

Point 11 Pronunciation.

Point 12 Finishing the conversation politely.

Point 13 Your overall contribution to the conversation.

Paul Moritoshi, The University of Birmingham, 2002.

APPENDIX IV
A TRANSCRIPT OF AN ILLUSTRATIVE TECP TEST

Italics indicate words spoken with additional emphasis. Text in square brackets '[ ]' provides additional information to aid the reader's interpretation.

Examiner: How are you then Yo?

Candidate: I'm fine. How are you?

E: Oh, I'm, all right

C: Now is very hot.

E: Yeah, too hot.

C: Teacher Paul, when did you go to Okayama?

E: When did I *come* here?

C: Hmm.

E: Oh, five years ago.

C: Ahh.

E: Five years ago.

C: When you come Okayama, did you feel lonely?

E: Lonely?

C: Hmm.

E: Umm, no, not really. Not at first. For the first three years I had many foreign friends Americans, Canadians, Australians, but now they went home back to their country.

C: Ahhhh.

E: So now, sometimes, it's a bit lonely.

C: Now I'm very lonely. My mother and father is not in here.

E: Ah, that's sad.

C: I'm very missing of them.

E: You miss home.

C: Eh?

E: Do you miss home?

C: Hmm.

E: So, have they come to Okayama?

Paul Moritoshi, The University of Birmingham, 2002.

C: I don't know.

E: Did they come to Okayama to see you?

C: I think two years...umm...later. They came...come here to meet me.

E: Well that'll be good. That'll be fun. But...er...you have...um, there's Shin, and Keichin...do you know Keichin?

C: Ugh?

E: Keichin?

C: Keichin? Hmm. [Yes].

E: So there're some Chinese students here, so you can...you can meet them.

C: Now I have many friend...er friends. I have a question. Can I ask teacher Paul?

E: Yeah, if you want to.

C: When did you meet your wife?

E: My wife?

C: Mmm.

E: When?

C: Mmm.

E: Oh, nine years ago? Yeah, nine years ago.

C: She's very beautiful.

E: Have you seen a picture?

C: Hmm. [Yes]

E: Yeah, she's OK...She's OK. She's kinda cute, I suppose.

C: Happiness. [Big smile].

E: Er...sometimes..... |

C: | Sometimes.

E: Only sometimes [Smiles and laughs].

C: Er...Paul teacher, teacher Paul...have two, have two children?

E: That's right, yeah.

C: Erm...all girl?

E: No.

C: One girl |

E: | One girl. My daughter is two and a half.

C: Ah.

E: and my son is three months.

Paul Moritoshi, The University of Birmingham, 2002.

C: Is very little.

E: Yeah, he's very small.

C: Pretty.

E: Yeah, I think so.

C: I very like.

E: Yeah, they're cute.

C: Er, by the way.....um, last weekend, what did you do?

E: Last weekend....um....I went to the home centre.

C: Home centre?

E: I bought some flowers.....and some things for my garden.

C: Hmmm.

E: So, I'm making an English garden, with flowers.

C: Do you like?

E: Yeah, I like....I like gardening. I like flowers.

C: Very good. I'm very pity. Last weekend....hmmm....I every day, every day is do my part-time job. And I'm hurt [The candidate puts her left hand on the table and shows the examiner her lacerated thumb].

E: [Looks at the thumb] Ooooh, Yo |

C: | Oooh.

E: Yo, Yo, oww. Was that with a knife?

C: Hmm. Knife. Cutting food.

E: Oh. Cutting food. Oh man! That looks really painful.

C: But now is OK.

E: Be careful, when you're chopping, when you're cutting with a knife. Don't go crazy! [Both laugh].

C: Don't go crazy.

E: Slowly.....slowly.

C: But it's very busy.

E: Busy?

C: Mmm.

E: Where do you work? In a restaurant?

C: Mmmm.....ramen.....ramen store.....ramen shop [Ramen is the Japanese word for Chinese noodles served in a watery soup with vegetables and meat].

Paul Moritoshi, The University of Birmingham, 2002.

E: So it's very busy?

C: Mm....Umm, Paul, teacher Paul. When you children....when you is children, what is .....er.....oooh.....er.....dif.....what do, what did you want to do?

E: My job?

C: Mmm.

E: [Smiles] I wanted to be a pilot.

C: Be pilot?

E: Yeah, I wanted to fly.

C: But you're a teacher.

E: Yeah, now I'm a teacher, but being a teacher's good fun too. But I still enjoy flying. I still like flying.

C: I want to be a teacher too [At this point the stopwatch alarm sounded, indicating that 6 minutes had elapsed]. Eh?

E: It's OK.

C: But now in China, very good teacher is very.....very.....very,very.....

E: Few?

C: Hmmm. No. Good teacher is very.....[indicates that she'd like to use her dictionary]

E: Yeah, that's OK.

C: I'm sorry. [Checks the word in her dictionary] Ahh, little. Is very little.

E: Mmm, is very few.

C: Very few. Ah, very few. Ah, sorry.

E: So you can.....well maybe you could be a teacher then. You can be a teacher.

C: Thank you. I hope. Teacher is very..very...a lot of...but a good teacher is very few.

E: I think you can do it.

C: Thank you.

E: [The examiner gives the 'T' time-up gesture with his hands].

C: Er...finish?

E: Yeah.

C: Hmm. Er, umm, I must..er..go to our class. I'm very glad to speak to you. Thank you.

E: OK, I'll speak to you later.

Paul Moritoshi, The University of Birmingham, 2002.

APPENDIX V
THE TECP ANALYTIC RATING SCHEME

Scale 1 - Ability to exchange greetings

Theoretical construct definition: the ability to exchange greetings.

Operational construct definition:

a). The level of participation in the greeting ritual through use of adjacency pairs and

b). The appropriateness of the reply to the greeting offered by the examiner.

These are rated on separate sub-scales.

Notes:

#1 - Greetings usually consist of two adjacency pairs (e.g. “How are you?”, “I'm OK. You?”, “Yeah, I'm fine thanks.”).

#2 - '*Appropriateness*' here means that the reply constitutes what is considered to be an acceptable response to the greeting, e.g. “How are you doing?” – “I'm fine” is appropriate, whereas “How are you doing?” – “It's very hot today” is not, as this evades the question.

Sub-scale 1a - Participation in the greeting adjacency pair ritual

Level	Description
0	**Non-participation** - Fails to offer a greeting and does not supply a reply of any kind.
1	**Partial participation** - Offers a greeting but does not supply a reply of any kind. *OR* - Does not greet but supplies a reply of some kind.
2	**Faltering participation** - Offers a greeting and supplies a reply, but with some false starts and / or need for repetition.
3	**Full participation** - Offers a greeting and supplies a reply, with no false starts and / or need for repetition.

Sub-scale 1b – Appropriateness of the reply

0	Fails to offer a reply so cannot be rated.
1	Replies with an inappropriate response.
2	Replies with an appropriate response.

Paul Moritoshi, The University of Birmingham, 2002.

Scale 2 - Ability to initiate a conversation

Theoretical construct definition: the ability to initiate the first topic of a conversation.

Operational construct definition: the level of ability with which the examinee can independently initiate a conversation without hesitation and through use of specified techniques.

Sub-scale 2a - Degree of hesitation

0	Fails to open the conversation as required.
1	Opens the conversation with substantial hesitation, i.e. more than 10 seconds.
2	Opens the conversation with some hesitation, but less than 10 seconds.
3	Opens the conversation with no hesitation.

Sub-scale 2b - Initiation technique used

1	Initiates the conversation through an explanation or opinion of something.
2	Initiates the conversation by using a question with no related preamble.
3	Initiates the conversation by providing an explanation or opinion of some sort, then follows up with a *related* question.

Scale 3 - Ability to change topics during a conversation

Theoretical construct definition: the ability to change topics during a conversation.

Operational construct definition: the level of ability with which the examinee can change from one topic to another during a conversation.

Notes:

#1 - Topic means 'what is talked about'.

#2 - A change of topic can be indicated through questions, non-verbal communication or discourse markers (e.g. “By the way....”, “Anyway....”, “That reminds me....”, “So....” etc.), or may be 'unmarked'.

Sub-scale 3a - Range of topic-changing techniques used

0	Does not demonstrate the ability to change topic.
1	Demonstrates the ability to change topic through the use of only 1 technique or discourse marker.
2	Demonstrates the ability to change topic through the use of 2 different techniques and / or discourse markers.
3	Demonstrates the ability to change topic through the use of 3 different techniques and / or discourse markers.
4	Demonstrates the ability to change topic through the use of 4 or more different techniques and / or discourse markers.

Paul Moritoshi, The University of Birmingham, 2002.

Sub-scale 3b - Number of examinee-initiated topic changes

0	The examinee does not change the topic.
1	The examinee changes the topic once.
2	The examinee changes the topic twice.
3	The examinee changes the topic three times or more.

Scale 4 - Use of questions

Theoretical construct definition: the ability to use questions during a conversation.

Operational construct definition: evidence that the examinee can use closed- and open-ended questions for purposes that enhance a conversation.

Notes:

#1 - The grammatical accuracy and complexity of the questions used are rated independently through scales 10a and 10b. The sub-scales here pertain only to question type, purpose and quantity.

#2 - Questions used during the greeting phase (e.g. “How are you?”) do not contribute to this rating.

#3 - 'Translation questions' (e.g. “What is ~ in English / Japanese?” or “What does ~ mean in English / Japanese?”) are those used only to request a translation equivalent and are considered equivalent to the use of a dictionary.

Sub-scale 4a - question type

0	The examinee asks no questions.
1	The examinee asks only translation question(s).
2	Excluding translation question(s), the examinee asks only close-ended (e.g. yes / no) question(s).
3	Excluding translation question(s), the examinee asks only open-ended (e.g. 'wh') question(s).
4	The examinee asks both close- (e.g. yes / no) and open-ended (e.g. 'wh') questions.

Sub-scale 4b - question's purpose

0	The examinee asks no questions.
1	The examinee asks only translation question(s).
2	The examinee asks question(s) (other than translation questions) to initiate the conversation and / or to change the topic.
3	The examinee asks question(s) (other than translation questions) to maintain the present topic.
4	The examinee asks questions both to change the topic and to maintain the current topic.

Paul Moritoshi, The University of Birmingham, 2002.

Sub-scale 4c - number of questions

0	The examinee asks no questions.
1	Excluding translation question(s), the examinee asks only 1 question.
2	Excluding translation question(s), the examinee asks 2 questions.
3	Excluding translation question(s), the examinee asks 3 questions.
4	Excluding translation question(s), the examinee asks 4 questions.
5	Excluding translation question(s), the examinee asks 5 questions.
6	Excluding translation question(s), the examinee asks 6 or more questions.

Scale 5 - Use of non-verbal communication and 'back-channelling'

Theoretical construct definition: use of non-verbal communication and back-channelling behaviour.

Operational construct definition: evidence that the examinee can use non-verbal communication (NVC) strategies to help convey or enhance meaning and can also provide feedback via the 'back-channel'.

Notes:

#1 - 'Non-verbal communication' means any method, excluding speech, employed to convey or enhance meaning. It includes, but is not necessarily limited to: facial expressions, gestures, eye contact, head or body movements and use of pen and paper.

#2 - 'Back-channelling' behaviour is that whereby the listener gives feedback, either verbal or non-verbal (e.g. grunts, smiles, laughter, head movement, facial expression), as to how what the speaker is saying is being received.

Sub-scale 5a - Use of NVC to convey or enhance meaning

0	The examinee makes no use of NVC strategies in any form to convey or enhance meaning.
1	The examinee uses NVC in any form to convey or enhance meaning 3 times or less during the test.
2	The examinee uses NVC in any form to convey or enhance meaning between 4-6 times during the test.
3	The examinee uses NVC in any form to convey or enhance meaning between 7-9 times during the test.
4	The examinee uses NVC in any form to convey or enhance meaning 10 times or more during the test.

Paul Moritoshi, The University of Birmingham, 2002.

Sub-scale 5b - Use of back-channel feedback

0	The examinee does not provide back-channel feedback in any form during the test.
1	The examinee provides back-channel feedback 3 times or less during the test.
2	The examinee provides back-channel feedback between 4-6 times during the test.
3	The examinee provides back-channel feedback between 7-9 times during the test.
4	The examinee provides back-channel feedback 10 times or more during the test.

Scale 6 - Cohesion

Theoretical construct definition: the ability to maintain a conversation's cohesion.

Operational construct definition: the level to which the examinee demonstrates the ability to keep a conversation going. This is viewed in terms of the conversation test's pace (a function mainly of dictionary use and thinking time), degree of reparability, examinee hesitation, their need for repetition and their level of comprehension.

1	**No - low cohesion** - The conversation proceeds slowly, is generally of poor quality and may or may not be reparable. There is substantial hesitation and / or need for repetition and / or lack of comprehension.
2	**Moderate - satisfactory cohesion** - The conversation proceeds at a moderate - satisfactory pace and is generally reparable. There may be some hesitation and / or need for repetition and / or lack of comprehension.
3	**High - excellent cohesion** - The conversation proceeds at a good - normal (i.e. native speaker) or near normal pace with little or no need for repair. There is little or no hesitation, need for repetition or lack of comprehension.

Scale 7 - Turn length

Theoretical construct definition: turn length.

Operational construct definition: the level to which the examinee can produce extended turns in order to fulfil his/her speaking obligations.

Notes:

#1 - 'Turn' is defined as the opportunity an interlocutor has to speak before 'the floor' is taken by another interlocutor.

#2 - Verbal back-channel feedback (e.g. "Hmm", "Aah") does not constitute a turn.

Paul Moritoshi, The University of Birmingham, 2002.

0	The examinee provides no responses during offered turns or uses only single-word responses.
1	The examinee uses single-sentence turns or lower, with no extended (i.e. multiple-sentence) turns.
2	The examinee uses extended (i.e. multiple-sentence) turns less than $^1/_3$ of the time.
3	The examinee uses extended (i.e. multiple-sentence) turns more than $^1/_3$ of the time.

Scale 8 - Turn-taking ability

Theoretical construct definition: turn-taking ability.

Operational construct definition: the level to which the examinee demonstrates the ability to take and relinquish 'the floor' smoothly.

1	**None - little** - Shows no or little ability to take and / or relinquish the floor smoothly, characterised by confusion in speaking rights on most or many of the turn changes during the test.
2	**Moderate - satisfactory** - Shows moderate - satisfactory ability to take and / or relinquish the floor smoothly, characterised by confusion in speaking rights on some of the turn changes during the test.
3	**Good - high** - Shows good ability to take and / or relinquish the floor smoothly, characterised by confusion in speaking rights only rarely, if ever during the conversation.

Scale 9 - Vocabulary[1]

Theoretical construct definition: knowledge of vocabulary.

Operational construct definition: the range and appropriateness of the vocabulary used by the examinee.

Note: Lexical errors that subjects self-correct should be ignored.

1	**Extremely limited vocabulary** - Cannot converse on any topic due to an extremely limited vocabulary. Can only use a few basic words and formulaic phrases (e.g. "How are you?", "I'm fine". "Yes", "No").
2	**Small vocabulary** - Often has difficulty conversing due to a lack of necessary vocabulary. Frequently misses or searches for words and / or uses unsuitable or inappropriate words.
3	**Good vocabulary** - Sometimes has difficulty conversing due to a lack of necessary vocabulary. Occasionally misses or searches for words and / or uses unsuitable or inappropriate words.
4	**Large vocabulary** - Rarely, if ever has difficulty conversing due to a lack of necessary vocabulary. Rarely, if ever misses or searches for words or uses unsuitable or inappropriate words.

[1] Adapted from Bachman (1990: 327).

Paul Moritoshi, The University of Birmingham, 2002.

Scale 10 - Grammar

Theoretical construct definition: knowledge of grammar.

Operational construct definition: the level of grammatical accuracy and complexity demonstrated by the examinee.

Notes:

#1 - Grammatical errors which examinees self-correct should be ignored.

#2 - Grammatical accuracy and complexity are rated separately on the sub-scales below.

Sub-scale 10a - grammatical accuracy

0	The examinee fails to supply sufficient sentential level production to allow assessment.
1	The examinee makes errors in most of the grammars used.
2	The examinee makes errors in some of the grammars used.
3	The examinee makes errors in few of the grammars used.
4	The examinee rarely, if ever, makes errors in the grammars used.

Sub-scale 10b - grammatical complexity

Notes:

#1 - This sub-scale should only be used to rate utterances produced during the examinee's turn and not back-channel feedback utterances.

#2 - A multiple-clause sentence uses ***mid-sentence*** conjunctions, e.g. 'and', 'so', 'but'.

0	The examinee fails to supply sufficient production to allow assessment.
1	The examinee's utterances generally consist of only a single word.
2	The examinee's utterances generally consist of more than one word, but are not structured through grammar.
3	The examinee's utterances are generally single-clause sentences with no multiple-clause sentences,
4	The examinee uses 1 or 2 multiple-clause sentences, indicated by ***mid-sentence*** conjunctions such as "and", "so", "but".
5	The examinee uses 3 or 4 multiple-clause sentences, indicated by ***mid-sentence*** conjunctions such as "and", "so", "but".
6	The examinee uses 5 or more multiple-clause sentences, indicated by ***mid-sentence*** conjunctions such as "and", "so", "but".

Paul Moritoshi, The University of Birmingham, 2002.

Scale 11 - Pronunciation

Theoretical construct definition: pronunciation.

Operational construct definition: the level to which the rater perceives the examinee's accent and / or pronunciation has inhibited intelligibility or communication during the test.

Note: This scale is not intended to compare the examinee's accent or pronunciation with that of a native English speaker, nor should the rater try to second-guess what problems native speakers generally might have understanding the examinee's pronunciation.

0	The examinee fails to supply sufficient production to allow assessment. *OR* The examinee's accent or pronunciation appeared to inhibit intelligibility or communication all of the time.
1	The examinee's accent or pronunciation appeared to inhibit intelligibility or communication most of the time.
2	The examinee's accent or pronunciation appeared to inhibit intelligibility or communication some of the time.
3	The examinee's accent or pronunciation did not appear to inhibit intelligibility or communication most of the time.
4	The examinee's accent or pronunciation did not appear to inhibit intelligibility or communication at any time.

Scale 12 - Ability to close a conversation appropriately

Theoretical construct definition: the ability to close a conversation in a sociolinguistically acceptable way.

Operational construct definition: the level of participation in the closing adjacency pair ritual, through the provision of a reason for closure and bidding farewell.

Notes:

#1 - Conversations usually close using adjacency pairs (e.g. “I have to be going.” [This constitutes the reason for closing], “OK. See you later”, “Yeah, see you later.”), without which closures might be considered sociolinguistically inappropriate.

#2 - Suitable phrases for bidding farewell include: “Goodbye”, “See you (later)” and “Thank you”.

Paul Moritoshi, The University of Birmingham, 2002.

0	The examinee does not give a reason for closure or bid farewell in any way.
1	The examinee fails to give a reason for closure but bids farewell.
2	The examinee gives a reason for closure but does not bid farewell.
3	The examinee gives a reason for closure and bids farewell, but the process is faltering or does not conform to the adjacency pair pattern.
4	The examinee gives a reason for closure and bids farewell smoothly and conforms to the adjacency pair pattern.

Scale 13 - Ability to contribute pro-actively to a conversation

Theoretical construct definition: pro-active contribution.

Operational construct definition: the level to which an examinee contributes to a conversation through acts that are pro-active rather than reactive, i.e. the extent to which the examinee works to bring the examiner into the conversation. This is considered in terms of greeting behaviour (scale 1a), conversation initiation behaviour (scale 2b), the number of student-initiated topic changes (scale 3b) and questioning behaviour (scales 4a and 4c).

Note - This scale is a composite score based on the ratings for the scales listed above. It is calculated and automatically entered onto the spreadsheet by the software itself, using the following Excel (V.7.0) software formula:

=SUM(IF(J*n*>0,SUM(J*n*-1),0)+IF(M*n*>0,SUM(M*n*-1),0)+IF(P*n*>0,SUM(P*n*-1),0)+O*n*+R*n*)*0.75

where *n* is the row number for each subject. It brings the weighting for the pro-active behaviours in these five scales up to 1.75, while keeping reactive behaviours at a weighting of 1.

Paul Moritoshi, The University of Birmingham, 2002.

APPENDIX VI
THE TECP CODING SHEET

3a	3b	4a			4b				5a	5b	6	7		8	9	10a	10b				11
Range of techs	No. of chngs	Question type			Question's purpose				NVC used	B-C used	Cohesion problems	Turn length		Turn-taking problems	Vocabulary problems	Gram. errors	Grammatical complexity				Pronunciation problems
		Tr	Closed	Open	Tr	Init conv	Chng topic	Mntn topic				1 sent	ext				1 word	unstru	s-c	m-c	
Que	1								1	1											
NVC	2								2	2											
DM#1*	3								3	3											
DM#2	4								4	4											
DM#3									5	5											
DM#4									6	6											
unmarked									7	7											
									8	8											
									9	9											
									10	10											

* DM = Discourse markers which indicate a change of topic (e.g. “By the way”, “Anyway...”, “That reminds me...”, “So....” etc.).

Paul Moritoshi, The University of Birmingham, 2002.

APPENDIX VII
THE TECP RATING SPREADSHEET

The following data are samples taken from actual test results.

			Scales																					
			1a	1b	2a	2b	3a	3b	4a	4b	4c	5a	5b	6	7	8	9	10a	10b	11	12	13			
Test no.	Stud Lvl.	VT Ref.	Greet Quan	Greet Qual	Start hes	Start tech	Chng topic	No. chngs	Que type	Que purp	No. of que	NVC	B-C	Cohesion	Turn lgth	Turn-taking	Vocab	Gram acc	Gram comp	Pron	Close	Contrib	Tot.	%	Grade
1	4	1/0:00:00	3	2	2	2	2	2	4	4	6	4	4	3	3	3	4	3	5	4	4	10.5	74.5	87.39	A
11	3	1/1:07:34	1	2	3	1	1	1	2	4	2	4	4	3	2	3	3	3	5	3	4	3	54	63.34	C
21	2	1/2:16:24	3	2	3	2	0	0	2	4	2	3	4	2	1	3	2	2	2	4	3	4.5	48.5	56.89	B

Paul Moritoshi, The University of Birmingham, 2002.

APPENDIX VIII
THE FACE VALIDITY QUESTIONNAIRE (ENGLISH)

English conversation test questionnaire

Please check (✓) your answers in the boxes. If necessary, please write your comments in Japanese in the spaces provided.

1 **What do you think is the most accurate way to check someone's English conversation ability?**

Please choose *only one* of the answers below.

a Write a script of a conversation. ❑
b Read a script of a conversation, then answer comprehension questions. ❑
c Listen to a recorded conversation, then answer comprehension questions. ❑
d A written test of vocabulary and grammar useful during conversation. ❑
e Have a conversation in English with a native speaker. ❑
f Have a conversation in English with a non-native speaker. ❑
g Another way (please write): _____
h I'm not sure. ❑

2 **Did the conversation test you've just done simulate a normal conversation?**

Please choose *only one* of the answers below.

a Yes, it simulated a normal conversation very well. (Go to question 3) ❑
b Yes, it simulated a normal conversation quite well. (Go to question 3) ❑
c It's OK. (Go to question 3) ❑
d It didn't simulate a normal conversation very well. ❑ (Miss question 3. Go to question 4)
e It didn't simulate a normal conversation at all. ❑ (Miss question 3. Go to question 4)
f I'm not sure. ❑ (Miss questions 3 and 4 and go to question 5)

Paul Moritoshi, The University of Birmingham, 2002.

3 **Why do you think the test simulated a normal conversation?**

Please choose any of the answers below. You can choose more than one answer. After answering this question, miss question 4 and go to question 5.

a It had the parts of a normal conversation (for example, a start, questions, topic changes and a finish). ❑

b I had a roughly equal chance to speak. ❑

c I could ask questions freely. ❑

d I could choose the topic. ❑

e I could change the topic. ❑

f The teacher didn't tell me whether my opinion or answer was right or wrong. ❑

g It was mostly spontaneous and I couldn't write a script for what I would say. ❑

h Other reason(s) (please write): _____

_____

_____

i I'm not sure. ❑

4 **Why do you think the test didn't simulate a normal conversation?**

Please choose any of the answers below. You can choose more than one answer.

a It didn't have the parts of a normal conversation (for example, a start, questions, topic changes and a finish). ❑

b I didn't have a roughly equal chance to speak. ❑

c I couldn't ask questions freely. ❑

d I couldn't choose the topic. ❑

e I couldn't change topics. ❑

f The teacher told me whether my opinion or answer was right or wrong. ❑

g It wasn't spontaneous and I could write a script for what I would say. ❑

h Other reason(s) (please write): _____

_____

i I'm not sure. ❑

Paul Moritoshi, The University of Birmingham, 2002.

5 **Did you have enough chances to speak?**

Please choose *only one* of the answers below.

a No. ❑

b Yes. ❑

c I'm not sure. ❑

6 **Was it difficult to remember the test procedure during the test?**

Please choose *only one* of the answers below.

a Yes, it was very difficult to remember the test procedure during the test. ❑

b Yes, it was quite difficult to remember the test procedure during the test. ❑

c It was neither easy or difficult to remember the test procedure during the test. ❑

d No, it was quite easy to remember the test procedure during the test. ❑

e No, it was very easy to remember the test procedure during the test. ❑

f I'm not sure. ❑

7 **If you have any comments about the conversation test's procedure, please write them below.** Please write your comments in Japanese.

_____

_____

_____

8 **What did you like about the test?** Please write your comments in Japanese.

_____

_____

Paul Moritoshi, The University of Birmingham, 2002.

9 **What did you dislike about the test?** Please write your comments in Japanese.

_____

_____

10 **How could the test be improved?** Please write your comments in Japanese.

_____

_____

11 **Overall, how effective do you think the test was as a test of your conversation ability?** Please choose *only one* of the answers below.

a Excellent. ❑
b Very good. ❑
c Good. ❑
d So-so. ❑
e Not very good. ❑
f Very poor. ❑

12 **How old are you?**

____ years and ____ months.

13 **How long have you been studying English?**

____ years ____ months.

Paul Moritoshi, The University of Birmingham, 2002.

APPENDIX IX
THE FACE VALIDITY QUESTIONNAIRE (JAPANESE)

The original version of this dissertation included the Japanese form of the face validity questionnaire at this point, but it is not included here due to problems with using Japanese fonts on some computers. If you would like a copy of the Japanese version of the face validity questionnaire used in this study, you can contact Paul Moritoshi via e-mail at moritosh@po.harenet.ne.jp.

Paul Moritoshi, The University of Birmingham, 2002.

APPENDIX X
'TRAINING NOTES' TO THE SECOND RATER

Notes to raters

Please read these notes carefully before rating.

The scales

1. This rating scheme is intended for use in scoring English conversation ability. Each scale rates a component part of this construct.
2. Each scale is supplied with a theoretical construct definition and an operational construct definition. Please read these carefully before using the scales.
3. Some scales have one or more notes which clarify the scale's intended purpose, terms used or provide examples. Please read these notes carefully. Adhering to them will help increase inter-rater reliability, score accuracy and interpretability.
4. Each scale takes the form of levels (indicated in the left column) and their corresponding description (indicated in the right column). Please select the description that most accurately assesses the examinee's ability for that component and then write the level number on the spreadsheet provided, under the corresponding scale number, on the row corresponding to that test subject.
5. When scoring each scale, select the highest level the examinee demonstrates during the test.

The coding sheet

Though some scales (1a - 2b and 12) can be rated very quickly as students greet, start or close the conversation, other scales must be rated as an overall performance across the full duration of the test. The coding sheet is intended to help the rater keep track of a subject's performance in these behaviours as the test progresses. When a subject performs a particular behaviour, the rater can make a check mark under the corresponding scale description(s). Note that even only a brief utterance may need to be coded under various scales. For example, "What

Paul Moritoshi, The University of Birmingham, 2002.

are you doing this weekend?" is an open question (scale 4a), used perhaps to change the topic (scale 4b). It is also a single sentence turn (scale 7) and a single-clause sentence (scale 10b). At the end of the test, these check marks can be tallied to help the rater select the most appropriate rating for each of those performance criteria. Reading the coding sheet in conjunction with the rating scheme before you commence rating will help you to become familiar with how the coding sheet works during observation and rating.

The spreadsheet.

The spreadsheet has been designed to make raw data entry and data analysis easier. Raters can use the 'VT ref' column information to find any test quickly and easily. To find the required test:

1. Ensure that the tape is rewound to the beginning.
2. Eject the videotape from the player.
3. Re-insert the tape.
4. Fast-forward the tape until the video player's tape counter displays the required time reference, expressed in hours : minutes : seconds.

Since the tests take up two videotapes, these are also numbered and indicated in the 'VT ref' column. For example '1 / 4:29:13' refers to the test on tape number 1, which starts 4 hours, 29 minutes and 13 seconds into the tape, i.e. test subject #41. These time references are quite accurate, but you may need to search forward or backward a little to find the start of the test.

If entering ratings on the paper copy of the spreadsheet, please be careful to use both the correct column (scale) and row (test subject). A long ruler or edge of paper helps, as does the occasional double check. Also, please ensure that no transcription errors occur when inputting the data into the computer.

Finally, to maintain rating standards, please don't try to mark too many tests in one go. Take breaks when you need them.

Paul Moritoshi, The University of Birmingham, 2002.

APPENDIX XI
A SAMPLE OF TOPICS USED IN THE TECP

* Indicates those topics initiated only by the examiner.

Very commonly occurring topics included:

What we did last weekend.

What we will do this coming weekend.

What we did during the winter vacation.

What we will do during the spring vacation.

Students' part-time jobs.

How we spend our free time.

Hobbies.

Quite commonly occurring topics included:

Family in general.

My children in particular.

Boyfriends.

*The end-of-semester tests.

Differences between England, Japan and China.

The cinema and videos.

Travel experiences.

More obscure topics included:

Differences between Korea and Japan.

Birthdays.

The up-coming birth of one student's baby.

One student's new car.

*Future careers.

The English conversation course itself.

One student's lacerated thumb caused by an accident with a knife at work.

Paul Moritoshi, The University of Birmingham, 2002.

REFERENCES

ACTFL (American Council on the Teaching of Foreign Languages) (2001) Proficiency testing. The ACTFL Oral Proficiency Interview (OPI). About the ACTFL OPI. (www) http://www.actfl.org/public/articles/details.cfm?id=17 (21st March 2002).

Al-Arishi, A. Y. (1994) 'Role-play, real-play and surreal-play in the ESOL classroom.' *ELT Journal* 48/4: 337-346.

Bachman, L. (1988) 'Problems in examining the validity of the ACTFL Oral Proficiency Interview.' *Studies in Second Language Acquisition* 2/10: 149-164.

Bachman, L. (1990) *Fundamental Considerations in Language Testing.* OUP.

Bachman, L. and A. Palmer (1996) *Language Testing in Practice.* OUP.

Badovi-Harlig, K. et al. (1991) 'Developing pragmatic awareness: closing the conversation.' *ELT Journal* 45/1: 4-15.

Baker, D. (1989) *Language testing: a critical survey and practical guide*. Edward Arnold.

Brazil, D. (1992) 'Speaking English or talking to people.' Adapted from a lecture given at Sophia University, Tokyo in January 1992.

Brazil, D. (1995) *Classroom and Spoken Discours*e. Centre for English Language Studies, Birmingham University.

Brown, H.D. (1994) *Principles of Language Learning and Teaching* (3rd. ed.) Prentice Hall.

Canale, M. (1983) 'From communicative competence to communicative language pedagogy.' In Richards, J.C. and R. Schmidt (eds.) *Language and communication*. Longman.

Carroll, B. (1980) *Testing communicative performance*. Pergamon.

Chalmers, A. (1982) *What is this thing called science?* University of Queensland Press.

Cook, G. (1989) *Discourse*. OUP.

Coulthard, M. (1985) *An Introduction to Discourse Analysis.* (2nd. ed.) Longman.

Cronbach, L.J. (1980) 'Validity on parole: how can we go straight?' *New Directions for Testing and Measurement* 5: 99-108.

Cronbach, L.J. and P.E. Meehl (1955) 'Construct validity in psychological tests.' *Psychological Bulletin* 52: 281-302.

Dornyei, Z. and S. Thurrell (1994) 'Teaching conversational skills intensively: course content and rationale.' *ELT Journal* 48/1: 40-49.

Educational Testing Service (1970) *Manual for Peace Corps language testers*. ETS.

Paul Moritoshi, The University of Birmingham, 2002.

Egyud, G. and P. Glover (2001) 'Oral testing in pairs - a secondary school perspective.' *ELT Journal* 55/1: 70-76.

Fairclough, N. (1992) *Discourse and social change.* Polity.

Foot, M. (1999) 'Relaxing in pairs.' *ELT Journal* 53/1: 36-41.

Fulcher, G. (1987) 'Tests of oral performance: the need for data-based criteria.' *ELT Journal* 41/4: 287-291.

Genesee, F. and J. Upshur (1996) *Classroom-based evaluation in second language education*. CUP.

Goffman, E. (1981) *Forms of talk*. Blackwell.

Hatch, E. (1992) *Discourse and language education*. CUP.

Hattori, T. (1987) 'A study of nonverbal intercultural communication between Japanese and Americans - focusing on the use of the eyes.' *JALT Journal* 8/2: 109-118.

Hayashi, R. (1988) 'Simultaneous talk - from the perspective of floor management of English and Japanese speakers.' *World Englishes* 7/3: 269-288.

Holmes, J. (1992) *An introduction to sociolinguistics*. Longman.

Hughes, A. (1989) *Testing for Language Teachers*. CUP.

Jafarpur, A. (1987) 'The short-context technique: an alternative for testing reading comprehension.' *Language Testing Journal* 4/2: 195-220.

Jennings, M. et al. (1999) 'The test-takers' choice: an investigation of the effect of topic on language-test performance.' *Language Testing Journal* 16/4: 426-456.

Jones, R.L. (1985) 'Some basic considerations in testing oral proficiency.' In Lee, Y.P. et al. (eds.) *New directions in language testing*. Pergamon.

Kormos, J. (1999) 'Simulating conversations in oral-proficiency assessment: a conversation analysis of role plays and non-scripted interviews in language exams.' *Language Testing Journal* 16/2: 163-188.

Lado, R. (1961) *Language testing*. Longman.

LoCastro, V. (1987) 'Aizuchi: A Japanese conversational routine.' In L. Smith (ed.) *Discourse across cultures*. Prentice Hall.

Lynch, B. and F. Davidson (1994) 'Criterion-referenced language test development: linking curricula, teachers, and tests.' *TESOL Quarterly* 28/4: 727-743.

Matthews, M. (1990) 'The measurement of productive skills: doubts concerning the assessment criteria of certain public examinations.' *ELT Journal* 44/2: 117-121.

McCarthy, M. (1991) *Discourse Analysis for Language Teachers*. CUP.

McNamara, T. (2000) *Language Testing*. OUP.

Messick, S.A. (1980) 'Test validity and the ethics of assessment.' *American Psychologist* 35: 1012-1027.

Montgomery, M. (1995) *An introduction to language and society*. (2nd ed.) Routledge.

Moritoshi, P. (2001) 'A four-step approach for improved syllabus design coherence.' *The Language Teacher Journal* 25/12: 15-19.

North, B. and G. Schneider (1998) 'Scaling descriptors for language proficiency scales.' *Language Testing Journal* 15/2: 217-263.

Nunan, D. (1987) 'Communicative language teaching: Making it work.' *ELT Journal* 41/2: 136-145.

Nunan, D. (1989) *Understanding language classrooms*. Prentice Hall.

Nunan, D. (1992) *Research methods in language learning*. CUP.

Nunan, D. (1993) *Introducing Discourse Analysis.* Penguin.

Oller, J.W. (1979) *Language Tests at School*. Longman.

Owen, C. et al. (1997) *Testing*. Centre for English Language Studies, Birmingham University.

Piper, T. (1984) 'Putting reality into role play.' *TESL Canadian Journal* 1: 29-34.

Popper, K. (1972) *Objective knowledge*. OUP.

Richards, J.C. (1980) 'Conversation.' *TESOL Quarterly* 14/4: 413-432.

Richards, J.C. and C. Lockhart (1996) *Reflective Teaching in Second Language Classrooms*. CUP.

Richards, J.C. and T.S. Rodgers (1986) *Approaches and methods in language teaching*. CUP.

Richards, J.C. and R. Schmidt (1983) 'Conversational analysis.' In Richards, J.C. and R. Schmidt (eds.) *Language and communication*. Longman.

Robb, T. (1980) 'Towards a theory of English conversation.' *JALT Journal* 4/2: 9-11.

Sacks, H. (1968) *Mimeo Lecture notes.*

Scollon, R. and S. Scollon (1983) 'Face in interethnic communication.' In Richards, J.C. and R. Schmidt (eds.) *Language and communication*. Longman.

Seedhouse, P. (1996) 'Classroom interaction: possibilities and impossibilities.' *ELT Journal* 50/1: 16-24.

Sheal, P. (1989) 'Classroom observation: training the observers.' *ELT Journal* 43/2: 92-104.

Shohamy, E. et al. (1986) 'Introducing a new comprehensive test of oral proficiency.' *ELT Journal* 40/3: 212-220.

Sinclair, J. and D. Brazil (1982) *Teacher talk.* OUP.

Paul Moritoshi, The University of Birmingham, 2002.

Soudek, M. and L. Soudek (1985) 'Non-verbal channels in language learning.' *ELT Journal* 39/2: 109-114.

Spolsky, B. (1995) *Measured Words*. OUP.

Stern, H.H. (1992) *Issues and options in Language Teaching*. OUP.

Stevenson, D.K. (1985) 'Pop validity and performance testing.' In Lee, Y.P. et al. (eds.) *New directions in language testing*. Pergamon.

Swan, M. (1985) 'A critical look at the communicative approach (2).' *ELT Journal* 39/2: 76-87.

Takahashi, T. (1989) 'The influence of the listener on L2 speech.' In S. Gass et al. (eds.) *Variation in Second Language Acquisition: Discourse and Pragmatics*. Multilingual Matters.

Tannen, D. (1994) *Gender and Discourse.* OUP.

Taylor, B. (1982) 'In search of real reality.' *TESOL Quarterly* 16/1: 29-42.

Underhill, N. (1987) *Testing Spoken Language: A handbook of oral testing techniques*. CUP.

Upshur, J. and C. Turner (1995) 'Constructing rating scales for second language tests.' *ELT Journal* 49/1: 3-12.

Valdman, A. (1989) 'The elaboration of pedagogical norms for second language learners in a conflictual diglossia situation.' In Gass, S. et al. (eds.) *Variation in second language acquisition*. Multilingual Matters.

van Lier, L. (1989) 'Reeling, writhing, drawling, stretching, and fainting in coils: Oral Proficiency Interviews as conversation.' *TESOL Quarterly* 23/3: 489-508.

vom Saal, D.R. (1983) 'Condensation and second language acquisition.' *TESOL Quarterly* : 17/3: 494-495.

Wardhaugh, R. (1998) *An introduction to sociolinguistics*. (3rd ed.) Blackwell.

Weir, C. (1993) *Understanding and developing language tests*. Prentice Hall.

Weir, C. and J. Roberts (1994) *Evaluation in ELT*. Blackwell.

Paul Moritoshi, The University of Birmingham, 2002.