

**A partial validation of the contextual validity of
the *Centre listening test* in Japan**

Kozo YANAGAWA

PhD

2012

University of Bedfordshire

**A partial validation of the contextual validity of
the *Centre listening test* in Japan**

by

Kozo YANAGAWA

**A thesis submitted in partial fulfilment of the
requirements for the degree of Doctor of Philosophy
Of the University of Bedfordshire**

July 2012

A partial validation of the contextual validity of the *Centre listening test* in Japan

Kozo YANAGAWA

ABSTRACT

The purpose of this study was to validate the listening comprehension component of *the Centre Test* in Japan (henceforth, JNCTL) in relation to contextual parameters and cognitive processing. For the purpose of this study, a comprehensive framework of contextual parameters and a L2 listening processing model was established. This provided a solid theoretical framework for this study, whereby empirical evidence was elicited in relation to contextual parameters and cognitive processing. The elicitation was made through document analysis, focus group interviews, and a large-scale questionnaire administered to stakeholders including 110 high school English teachers and 391 third year students of high schools. The elicited data was subjected to descriptive, quantitative and qualitative analysis.

The results of Preliminary studies identified ten possible key parameters to help the JNCTL achieve greater validity. They included the number of opportunities to listen to the input, a lack of hesitations, a lack of overlapping turns, a lack of multi-participant discussions, a lack of variety in the English accents used, a lack of L2 speakers, a lack of inference questions, a lack of non-linear texts, a lack of sandhi-variations, and a lack of natural speech rate. The results of the questionnaire revealed that sandhi-variation was the key parameter to help the current JNCTL achieve greater validity in a direction that would be accepted by the stakeholders, and it was further explored in Main Study in attempt to investigate the effect of sandhi-variation on listening comprehension test performance and the level of cognitive load imposed on the test takers.

A series of experiments was conducted involving the manipulation of sandhi-variation. The results revealed that although no statistical difference was found in item difficulty estimates between the

sandhi-variation and non-sandhi-variation versions, sandhi-variation may involve double effects on listening comprehension for the test takers. The positive effects could involve providing more prominent phonological difference between accented and unaccented words in connected speech which are produced by sandhi-variation, and this difference may reduce the cognitive load imposed on the test takers. The negative effects may involve increasing the cognitive load imposed on the test takers by obscuring sounds through elision or unclear pronunciation, and disturbing speech perception or word recognition.

Recommendations are provided for improving the validity of the current JNCTL and for the development of listening comprehension tests more generally. Implications are also suggested for the teaching of listening at secondary schools in Japan. Lastly, the limitations of the study are outlined and suggestions for further research are proposed.

Acknowledgements

This thesis was only possible due to the kindness and generous assistance of many friends, colleagues, and family. Although I cannot truly convey my thanks in a few paragraphs, I would like to take this opportunity to express my gratitude.

First of all, I would like to show my appreciation to my supervisors. In the order that met them, I would like to thank Cyril Weir for first introducing me to the Ph.D. programme at CRELLA (Centre for Research on English Language Learning and Assessment) of University of Bedfordshire. I still remember the day when I first saw him in a cosy room at Putteridge Burry campus. He was kind enough to jot down 'With all good wishes for your Ph.D.' on the back of the front cover of his distinguished book *Language Testing and Validation*, which inspired me to this research. I am very grateful for his encouragement, patience, invaluable suggestions, and a good sense of humour throughout this study.

I would also like to thank Tony Green. It was he who recommended CRELLA, which was just new then. Throughout my studies, he has provided unfailing advice and support despite his extremely busy schedule. Hundreds of emails exchanged with him proves this. He is a lighthouse who guides me on the right track whenever I was laidback, lost, and lazy. I would also like to thank Miyoko Kobayashi. She gave the earlier version of this manuscript invaluable comments, and throughout this research she was very generous and supportive. I would also like to express my appreciation to Fumiyo Nakatsuhara for her recommending a strong pair of Cyril and Tony and for her statistical support. She looks to me like an 'iron woman' with strong will and tenacious heart. Every time I see her, I feel firmly determined to carry out this thesis. I would also like to thank Gerald Sharpling for his constructive comments on the earlier version of this manuscript.

My thanks also go to Hiroshi Matsusaka for his generous support throughout this study. Every time I asked him various things – to be a reviewer or a rater, he did not lose his time in accepting my requests. My thanks also go to Kumi Suzuki. Being a good friend, she has been supportive of this thesis. Her expertise about listening provided me with different points of views. I would also like to express my appreciation to

Yo In'nami, Rie Kozisumi, and Mike Linacre for their statistical supports. Without their supports it would have been impossible to run the Rasch analysis.

This thesis would not have been possible without the help of many reviewers, respondents to the questionnaires, interviewees in group interviews, data collectors, participants in the experiment, and the speakers who took part in the recordings. The reviewers include Tomoko Fujita, Kahoko Matsumoto, Nao Okubo, Atsuko Fukuoka, and Mari Aoki, the interviewees include Yukinori Watanabe, Takeshi Oyama, Reiko Takahashi, Maho Mochizuki, Yuki Ando, and Tomohiko Oba, the data collectors include Motoyuki Takeda, Akiko Kojima, Hikaru Olji, Junko Kobayashi, Masumi Hirata, Mikiko Takanashi, Yoichi Meguro, Yoshito Sampei, Reiko Kai, Jun Kai, and Yumiko Ito, the speakers who joined the recordings include William Sievers, Anna Fujimoto, and Chris Sillio, the respondents to questionnaires and the participants in the experiments include the students at Yamato West High school, Odawara High school, Shonan High school, Yamato High school, Kawawa High school, Hosei university, and many anonymous English teachers in Japan.

I would also like to thank Rotary Foundation and Rotarians for their financial and humanitarian support, which enabled me to stay in England as an honourable Ambassadorial scholar. My special thanks go to three distinguished Rotarians, Geoffrey Farr, John Goodge, and Taira Hasegawa.

Last but not the least, I am deeply indebted to my wife, Harumi, for putting up with me being away in England and looking after our kids, Ryoza and Yoko. Without her generous support and patience this dissertation would not be possible.

Contents

LIST OF CONTENTS

Abstract	I
Acknowledgements	III
List of contents	V
List of Tables	XI
List of Figures	XIV
List of Abbreviations	XV

Chapter 1: Introduction 1

1.1 Introduction	1
1.2 The <i>Centre Test</i> in Japanese context	1
1.3 Statement of the problem	5
1.4 Purpose of this study	7
1.5 Aims of this study	7
1.6 Validity and validation	8
1.7 Cognitive validity	9
1.8 Contextual validity	10
1.9 Overview of the thesis	12

Chapter 2: Literature Review 13

2.1 Introduction	13
2.2 L2 listening processing model	14
2.3 Contextual parameters	23
2.3.1 Rubric	23
2.3.1.1 Structure	23
2.3.1.2 Instructions	24
2.3.1.3 Time allotment	25
2.3.1.4 Scoring method	25
2.3.2 Input	26
2.3.2.1 Format	26
Channel	
Speaker	

Gender of the speaker	
The number of the speakers	
Acquaintance with speaker's voice	
Text length	
The number of opportunity to listen to the input	
2.3.2.2 Phonology	32
Hesitations	
Sandhi-variation	
Tempo and rhythm	
Speech rate	
Accents/L2 speakers	
2.3.2.3 Syntax	37
Syntactic complexity	
Syntactic deviation	
2.3.2.4 Vocabulary	39
Word frequency level	
Lexical diversity	
Lexical density	
Culturally specific words	
2.3.2.5 Discourse features	42
Rhetorical type	
Discourse markers	
Redundancy	
Propositional (text) linearity	
Degree of planning	
2.3.2.6 Pragmatic features	47
Overlapping and backchanneling	
Turn taking	
2.3.3 Expected response	48
2.3.3.1 Question	48
Mode of presentation	
Language	
Provision	
Scope	
2.3.3.2 Response	52
Type	
Mode of presentation	

2.3.3.3 Time constraints	54
2.3.4 Relationship between input and expected response	55
2.3.4.1 Topical knowledge	55
Background knowledge	
Domain knowledge	
2.3.4.2 Text-item interaction	57
Inference question	
Position of relevant information in the text	
Lexical attractiveness of options	
2.4 Research Questions	62
<u>Chapter 3: Preliminary studies: Elicitation and prioritisation of a key parameter</u>	64
3.1 Introduction	64
3.2 Preliminary study 1: Elicitation of discrepancies	65
3.2.1 Introduction	65
3.2.2 Method: Overall research design	65
3.2.2.1 Research Question 1	66
3.2.2.2 Research Question 2	66
Document analysis	
Quantitative analysis	
Qualitative analysis	
3.2.2.3 Research Question 3	78
Vocabulary size	
Practical communication ability	
Language use situations	
Functions of language	
3.2.3 Results and analyses	83
3.2.3.1 Introduction	83
3.2.3.2 Research Question 1	83
3.2.3.3 Research Question 2	83
Document analysis	83
Quantitative analysis	88
Qualitative analysis	96

3.2.3.4 Research Question 3	102
Vocabulary size	
Practical communication ability	
Language use situations	
Functions of language	
3.2.4 Findings	106
3.2.4.1 Research Question 1—Real-life parameters vs. the CS	106
3.2.4.2 Research Question 2—Real-life parameters vs. the JNCTL	107
3.2.4.3 Research Question 3—the JNCTL vs. the CS	108
3.2.5 Discussions	110
3.3. Preliminary study 2: Prioritisation of a key parameter	116
3.3.1 Introduction	116
3.3.2 Method	116
3.3.2.1 Group interviews	116
Purpose	
Setting	
Interviewees	
Procedures	
Questions	
3.3.2.2 A large-scale questionnaire	120
3.3.3 Results and analyses	124
3.3.3.1 Group interviews	124
Introduction	
General impression of the JNCTL	
Discrepancies identified through Preliminary study 1	
Limitations of the group interviews	
Findings	
3.3.3.2 A large scale questionnaire	133
3.4 Preliminary study 3: Confirmation of a key parameter	140
3.4.1 Introduction	140
3.4.2 Method	140
3.4.3 Results	142
3.4.4 Findings and pathway to Main study	144

Chapter 4: Main study: Methodology 146

4.1 Introduction	146
4.2 Overview of the research design	146
4.3 Materials development	149
4.3.1 Listening comprehension test	149
4.3.1.1 Source	149
4.3.1.2 Recording	151
4.3.1.3 Evaluation of the recordings	153
4.3.1.4 Piloting	156
4.3.2 Questionnaire	159
4.4 Participants	160
4.5 Procedures	161
4.6 Analyses	163
4.6.1 Analysis for Research Question 5	163
4.6.2 Analysis for Research Question 6	165

Chapter 5: Main study: Results and analyses 166

5.1 Introduction	166
5.2 Research Question 5	166
5.2.1 Introduction	166
5.2.2 Quality of Rasch analysis	167
5.2.2.1 Introduction	167
5.2.2.2 Reliability	167
5.2.2.3 Fit statistics	171
5.2.2.4 Anchor items	174
5.2.3 Answer to Research Question 5	175
5.2.4 Discussion	179
5.3 Research Question 6	182
5.3.1 Introduction	182
5.3.2 Quality of Rasch analysis	184
5.3.2.1 Introduction	184
5.3.2.2 Reliability	184
5.3.2.3 Fit statistics	187
5.3.3 Answer to Research Question 6 and discussions	188
5.4 Findings	195

List of Tables

Chapter 1

Table 1.1 The JNCTL	5
---------------------------	---

Chapter 2

Table 2.1 Rubric	23
Table 2.2 Format for input	27
Table 2.3 Phonology for input.....	32
Table 2.4 Discourse features for input.....	43
Table 2.5 Question	49
Table 2.6 Response	53
Table 2.7 Framework of contextual parameters	61

Chapter 3

Table 3.1 Research instruments summary	66
Table 3.2 Method for each contextual parameter	68
Table 3.3 Results of coding for the second training session.....	75
Table 3.4 Coding sheet for qualitative parameters	77
Table 3.5 Method for exploring the components of the CS.....	79
Table 3.6 Qualitative analysis for the components of the CS	79
Table 3.7 Structure	84
Table 3.8 Instruction, time allotment, and scoring method	85
Table 3.9 Format	86
Table 3.10 Expected response.....	87
Table 3.11 Text length (# of words)	88
Table 3.12 Sandhi-variations.....	90
Table 3.13 Speech rate (wpm)	91
Table 3.14 Syntactic complexity	92
Table 3.15 Vocabulary	93
Table 3.16 Discourse markers	94
Table 3.17 Pragmatic features (only for dialogues)	95
Table 3.18 Rhetorical type	97
Table 3.19 Topical knowledge	98

Table 3.20	The rate of exact agreement among the raters (%)	99
Table 3.21	Coverage by word frequency level by JACET 8000 list ...	103
Table 3.22	Language use situations	105
Table 3.23	Functions of language.....	106
Table 3.24	Discrepancies identified through Preliminary study 1.....	109
Table 3.25	Backgrounds of the English teachers	118
Table 3.26	Possible changes to the JNCTL	122
Table 3.27	Summary of the interviewees' comments in relation to the discrepancies	132
Table 3.28	Perceived importance of the discrepancies by stakeholders (%)	134
Table 3.29	Ratings of 'naturalness' of sandhi-variation in the JNCTL	143

Chapter 4

Table 4.1	Research instruments summary for Main study	146
Table 4.2	Research design	148
Table 4.3	Source and the number of items for each item set	149
Table 4.4	The number and ratio of items in relation to item types through comparison with the JNCTL	150
Table 4.5	Evaluation of 'naturalness' of the sandhi plus version ...	155
Table 4.6	Evaluation of 'naturalness' of the sandhi minus version	155
Table 4.7	Results of the pilot study	157
Table 4.8	Participants' proficiency measured by TOEIC IP	161
Table 4.9	Administration of the questionnaire	163

Chapter 5

Table 5.1	Descriptive statistics for each administration	167
Table 5.2	Measures for participants and items, and reliabilities	168
Table 5.3	Means and SDs of item difficulty estimates for each version	175
Table 5.4	Item difficulty estimates and standard errors for each item in relation to version	178
Table 5.5	Means and SDs for perceptual difficulty in noticing sandhi-variation features	184
Table 5.6	Measures for participants and items, and reliabilities on the questionnaire	185
Table 5.7	Infit mean square by Zstd	188
Table 5.8	Measures and standard errors of perceptual difficulties of sandhi-variation features in relation to version.....	189

List of Figures

Chapter 2

Figure 2.1 L2 listening processing model	17
--	----

Chapter 3

Figure 3.1 Hesitations	89
Figure 3.2 Required processing level	99
Figure 3.3 Text linearity	100
Figure 3.4 Position of relevant information	101
Figure 3.5 Inference questions	101
Figure 3.6 Practical communication ability	104
Figure 3.7 Perceived importance by stakeholders	135

Chapter 4

Figure 4.1 Common item/person linking	148
---	-----

Chapter 5

Figure 5.1 Map showing the relationship between test takers' ability and item difficulty estimate	170
Figure 5.2 Infit mean square value by Zstd and item difficulty estimate for each item	171
Figure 5.3 Item difficulty estimates in relation to version	176
Figure 5.4 Map showing the relationship between the test takers' sensitivity to sandhi-variation features and their perceptual difficulty to be noticed	186
Figure 5.5 Perceptual difficulty in noticing sandhi-variation features	190

List of Abbreviations

CS	The Course of Study for foreign languages for upper secondary school (implemented since 2003)
EFL	English as a foreign language
EIL	English as an international language
ESL	English as a second language
ESOL	English for Speakers of Other Languages
ETS	Educational Testing Service
IIBC	International Institution of Business Communication
JNCTL	Japan National Centre Test, Listening comprehension component
L1	first language
L2	second language
MCQ	Multiple-choice question
MEXT	Ministry of Education, Culture, Sports, Science, and Technology
NCUEE	National Centre for University Entrance Examination
PET	Preliminary English Test
RQ	Research Question
SAQ	Short answer question
SD	Standard deviation
SEM	Standard error of measurement
STEP	Society for Testing English Proficiency
sv	Sandhi-variation
TEFL	Teaching of English as a Foreign Language
TESOL	Teaching English as a Second Language
TLU	Target language use
TOEFL	Test of English as a Foreign Language
TOEIC	Test of English for International Communication

Chapter 1 Introduction

1.1 Introduction

This chapter describes the background of this study, focusing on English testing for university entrance purposes in the Japanese context to introduce the *Centre Test* in Japan. Next, an important decision to incorporate a listening comprehension component into the test is described in accordance with the reform of the Course of Study (henceforth, CS), a national syllabus for formal education at secondary schools in Japan. Then, the purpose of the present study is outlined, and the major aims of this study are set out. Also, the theoretical framework for the present study is provided. Lastly, an overview of this thesis is presented.

1.2 The *Centre Test* in Japanese context

In Japan, the *Centre Test*, developed and administered by the *Daigaku Nyushi Centre* (the National Centre for University Entrance Examination, henceforth, NCUEE), is a high-stakes national test designed to be used for university entrance purposes (Nishigori & Kuramoto, 2007). A growing number of universities and colleges are using the test for gate keeping purposes. In January, 2011, 665 (85 per cent) out of 780 universities and colleges across Japan, including all 82 national and 79 municipal universities, adopted the *Centre Test* for this purpose. This figure has increased from 60.6 per cent (403 universities or colleges) in 2000 (Ministry of Education, Culture, Sports, Science, and Technology, henceforth, MEXT, 2011a) by more than 20 per cent. Thus, the total number of the test takers

reached 520,680 in 2011 (NCUEE, 2011a).

The growing popularity of the *Centre Test* can primarily be attributable to the higher ratio of high school graduates attending higher education, and participation by more private universities or colleges. Recently more high school students in Japan go up to higher education, partly because of the Japanese belief that admission to a first-reputation university or college guarantees a high ranking position in society (Nishigori & Kuramoto, 2007; Shea, 2009), and partly because a declining birth rate in Japan makes it easier for high school graduates to be accepted to universities or colleges (for example, Watanabe, 2004). Further possible explanation can be that job opportunities are now less available immediately after their high school graduation due to the economic recession in Japan and the concomitant lower job-opening-to-application ratio. As a result, the ratio of high school graduates who attend higher education increased to 54.4 per cent in 2011 (MEXT, 2011a). This figure is relatively higher than that of other developed countries. It is, for example, 24.8 per cent for Germany in 2005, 32.5 per cent for France in 2005, 52.1 per cent for U.S in 2004, 62.6 per cent for U.K in 2005 (MEXT, 2008). Accordingly, out of 548,886 who graduated from high schools in 2011 and were admitted to tertiary institutions, 442,421 (80.6 per cent) took the national test (The rest of them were admitted through other routes such as interviews, recommendations, or writing a short essay).

Second, more and more private universities or colleges are taking advantage of the *Centre Test* for their admissions. 504 (84 per cent) out of 599 private universities or colleges across Japan used the *Centre Test* in 2011 (MEXT, 2011a). The figure more than doubled from 242 in the year of 2000. This shows that now high school graduates have more options to be admitted to private universities—either by taking the original test

developed by the private university, or alternatively by taking the *Centre Test*. Thus, the *Centre test* attracts numerous test takers and is very influential not only for high school students who want to go up to tertiary institutions but also for formal English education conducted at high schools.

These accounts draw our attention to how the entrance exam affects the language teaching or language learning at secondary schools in Japan. In Japan, English has not been learnt so much for communication purposes as for the test and being admitted to good schools. It is fair to assume, then, that the entrance exam has its tremendous influence 'at those high schools where students hope to enter a competitive university near the top of the pyramid of prominence and rank' (Shea, 2009, p. 99). Admittedly, English at secondary schools in Japan should be considered as an academic subject, where the primary goal of studying English is to obtain good scores on the entrance examinations, particularly for the students who intend to go to tertiary institutions (Kuroda, 2012). The existence of the word '*Juken-Eigo*' (entrance examination English) indicates a role for English at secondary schools in Japan. The word makes English into a subject to be learnt for the particular purposes of entrance examinations distinct from English for communication purposes.

Faced with the advent of globalization, however, where English is considered a common language in the world, bridging the huge gap between the '*Juken-Eigo*' and practical English for communication purposes has been identified as an immediate national priority (for example, Watanabe, 2004). It is widely believed that Japanese learners of EFL are not as good at speaking and listening as at reading and writing (for example, Uchida, Kikuchi, Nakaune, Naekawa, & Ishizuka, 2002). The MEXT, which governs educational policy in Japan, therefore, launched its *Strategic Plan to Cultivate Japanese with English abilities* in 2002 in an

attempt to carry out this task. In accordance with the CS for enhancing 'practical communicative competence in English' (MEXT, 2007), the *Strategic plan* urged NCUEE to introduce a listening comprehension test as soon as possible on the assumption that the introduction of a listening comprehension test would promote the development of listening ability in Japanese learners of EFL. In 2006 NCUEE added a 50-point listening comprehension component (henceforth, JNCTL) to the original 200-point paper-pencil section for the first time in its 26-year history. Perhaps this was the most significant change in the entrance exam system over the decades (Murphey, 2006).

The JNCTL is claimed to be an achievement test which purports to measure to what extent high school graduates in Japan have learnt the core knowledge they are supposed to have acquired by the time of their senior high school graduation' (NCUEE, 2006, 2007, 2008, 2009a, 2010). The JNCTL is also intended to discriminate between candidates according to the listening proficiency level as long as it is used for gate-keeping purposes for the universities or colleges (NCUEE, 2007; Ito, Kawamura, Shimada, Nishihara, & Funato, 2007; Negishi, Matsuzawa, Sato, Toyoda, & Nakano, 2010). The JNCTL is composed of four parts or twenty-five items (nineteen for dialogues and six for monologues) as Table 1.1 illustrates (see Section 3.2.3.3 for more details).

Table 1.1 The JNCTL

Part	Task	D or M	# of items
1	Match the picture or number	D	6
2	Complete the conversation	D	7
3	Choose an option according to the content of dialogues	D	6
4	Choose an option according to the content of monologues	M	6
Total			25

Note. D: Dialogue, M=Monologue

Many people, including educational policy makers, language teachers, test developers, and businesses were positive about the introduction of this listening comprehension test into the *Centre Test*, hoping that it would lead to improvements in the listening ability of Japanese learners of EFL at high schools and thus their communicative competence in English (Kougo, 2006; *Zen-eiren*, 2007). *Zen-eiren* or the Association of High School English Teachers, for example, states that ‘teaching listening has been introduced at secondary schools in Japan to enhance listening abilities of the students [since the introduction of the JNCTL into the Centre Test]’ (*Zen-eiren*, 2007, p. 378). Matsuzawa agrees with this, saying that ‘more and more English teachers are focusing on listening as well as on other skills, inspired by the introduction of the listening comprehension section into the *Centre Test*’ (Suzuki, Takashima, Matsuzawa, Kadonaga, Oh, Yoshino, & Namai, 2007, p. 31).

1.3 Statement of the problem

The validity of the listening comprehension component of the *Centre Test* in Japan (henceforth, JNCTL), however, is still open to question (Uchida, Sugisawa, & Ito, 2010) as no empirical research has been published by NCUEE and to date very little evidence of its validity has yet been provided

(Sage & Tanaka, 2006; Ito et al., 2007). Sage and Tanaka (2006) carried out psychometric analysis of the 2006 form of the JNCTL and found low construct validity. This is a key issue: if the test is not measuring listening ability appropriately, then how could we expect the listening ability of Japanese learners of English as a foreign language at high schools to be enhanced? Neither could we expect their practical communication ability in English to be improved (Shea, 2009). Positive washback, which may encourage forms of teaching or learning judged to be appropriate (Green, 2007), will not take place if less valid tests are introduced and used (Messick, 1996). This is where the researcher has been inspired and motivated to carry out the present study.

Another claim by NCUEE (2006) is open to question:

The principle of developing the *Centre Test* lies in accordance with the overall objective of foreign language (English) education at high schools in Japan, which is specified in the CS, that is to develop students' practical communication abilities (MEXT, 2007). Based on this principle, every effort was made to reflect language use situations and functions of language which are operationalised in the CS when developing the JNCTL.

(NCUEE, 2006, p. 446) [translated by the author]

The basis for this claim needs to be established. Given the potential impact on the national interest of Japan, education and business, and washback on English language teaching, there is a pressing need to investigate the validity of this high-stakes national test.

The present study does not envisage, however, that the development of a

valid listening comprehension test will directly lead to enhancing the listening ability of Japanese EFL learners at high schools, because although changing national examinations can serve as a vehicle for shaping teachers' instructional practices, 'that success is not assured' (Chapman & Snyder, 2000, p. 462). 'Ideally, the move from learning exercises to test exercises should be seamless' (Messick, 1996, p. 241). Fostering positive washback, however, is not easy to achieve in practice (Watanabe, 2004; Green, 2007). No consistent washback, for example, was found between college entrance exams and the methods English teachers at high schools in Japan employed in their classrooms (Watanabe, 1996, 2004). These results of the earlier studies do not coincide with the relatively optimistic views put forward by Kougo (2006), *Zen-eiren* (2007), or Suzuki et al. (2007) (see Section 1.2).

Nevertheless, the starting point for this study lies in my belief that the development of a JNCTL with optimum validity is a prerequisite if listening ability and practical communicative ability of Japanese high school learners of EFL is to improve, and this belief is underpinned by Messick (1996), arguing that positive washback can be associated with the introduction and use of more valid tests because 'minimizing construct-under-representation and construct-irrelevant variance in a test should facilitate good educational practices' (p.247).

1.4 Purpose of the study

The purpose of this study is, therefore, to investigate the validity of the JNCTL. Specifically, this study explores this high-stakes test in terms of *cognitive validity* and *contextual validity* (Weir, 2005) and aims to highlight aspects of the JNCTL that might be reformed to enhance its validity and so lead to further improvements in the listening ability of Japanese learners of

EFL at high school.

1.5 Aims of this research

The present study has the following major aims:

- 1)** To propose a comprehensive framework for describing the contextual parameters which affect test performance on listening comprehension tests for EFL learners.
- 2)** To develop a practical L2 listening processing model to inform test design purposes.
- 3)** Using the outcomes of (1) and (2), to identify discrepancies between the JNCTL, the CS and real life listening features in terms of contextual parameters and cognitive processing.
- 4)** To establish how the intra-task manipulation of specified contextual parameters affects test performance.
- 5)** To assess how the manipulation of specified contextual parameters affects the level of cognitive demand imposed on the test takers.

1.6 Validity and validation

It is widely accepted that validity is concerned with whether or not a test measures what it is supposed to measure (for example, Lado, 1961). Messick (1989) developed this traditional view of the validity and defined 'test validity' as nothing less than evaluative summary of both the evidence for and the consequences of score interpretation and use. Based on this revised understanding of test validity and validation, the socio-cognitive approach, which the present study employs, was proposed by Weir (2005), where 'test validity is claimed to reside on test scores' (Weir, 2005, p.12). Weir (2005) argues that test scores should be considered as a reflection of a trait or targeted language ability which can be observed through an

interaction between the cognitive processing engaged by the test takers and the contextual parameters relevant to the test task. Accordingly, test validation involves the justification of the interpretation of a test score as a representation of a construct we want to measure. In order to justify our interpretation, we should provide empirical evidence (Messick, 1989; Bachman, 1990; Bachman & Palmer, 1996; Weir, 2005). This study, therefore, aims to provide the empirical evidence in terms of the contextual validity and cognitive validity proposed by Weir (2005), in an attempt to validate the JNCTL. Each of these forms or aspects of validity is discussed in the sections which follow.

1.7 Cognitive validity

Cognitive Validity is concerned with the extent to which a test or a test task reflects cognitive and meta-cognitive processing (goal setting, planning, and monitoring) in real life language use (as distinct from the uses of language that may be limited to the classroom or to tests). What is important for the present study is that the cognitive processing involved in real life language use should be reflected as far as possible in language test situations if claims for validity are to be supported. This study, therefore, first proposes a L2 listening model based in the literature, and then examines both the JNCTL and the Course of Study on which the JNCTL is based in terms of cognitive validity.

Two conventional L2 listening processing models – the interactive processing model (Marslen-Wilson & Tyler, 1980; Voss, 1984; Lynch, 1988; Anderson & Lynch, 1988; Richards, 1990; Peterson, 1991; Buck, 1990, 2001; Matsusaka, 1995) and Anderson's model (2005) – do not suffice for the development of valid listening comprehension tests, primarily because these models do not allow test developers to operationalise listening ability

well enough. These models do not specify the constructs of a listening comprehension test to the extent that test validation can be conducted. Furthermore, these models do not give sufficient attention to the relationship between cognitive processing and contextual parameters (Weir, 2005).

The issue of how the conditions under which a task is carried out (*context*) relate to the cognitive processes engaged by the listener is central to listening test validation, the concern of this study. It is for this reason that the two conventional models are not sufficiently suitable for test validation purposes, and therefore I propose a practical L2 listening processing model (see Figure 2.1) based on Khalifa and Weir (2009). The L2 listening model will be provided and discussed in detail in Chapter 2.

1.8 Contextual validity

Contextual validity, traditionally described as content validity, is concerned with the extent to which a test reflects linguistic context but also the social and physical context in which a task is performed (Weir, 2005). Since language processing does not take place in a vacuum, language testers need to specify the context in which this processing takes place (Weir, 2005). It is for this reason that the word 'context' rather than 'content' is used in the present study. That is, since a language use situation reflects a socio-cultural dimension, 'context' better accounts for the dimension of a language use than does 'content'. A contextual parameter refers to a linguistic, social, or situational variable, which is likely to affect language test performance, and is used as a criterion for my validation process in this study. Shaw and Weir argue (2007) that it is important to be able to describe target activities in terms of their contextual parameters and to operationalise as many of them as possible in the test task(s).

There is a pressing need for a full description of contextual parameters for listening comprehension tests. Weir's framework (2005, p.45) is helpful, but lacks the detail required for a full description. For example, it does not provide full descriptions of phonological features of the input texts, which are considered elsewhere to be crucial contextual parameters in listening comprehension tests (Buck, 2001). Similarly, alternative frameworks such as Bachman and Palmer's (1996, 2010) framework of test task characteristics overlook some features that are likely to be of particular relevance to listening tests such as phonological features like hesitations, sandhi-variation, or accents. It is for this reason that a preliminary aim of the present study must be to compile, through the literature review, a more comprehensive framework of contextual parameters for listening comprehension tests for EFL learners.

Table 2.7 supplies a comprehensive framework of contextual parameters for L2 listening comprehension tests. Each parameter is mapped out under the four headings proposed by Bachman and Palmer (1996, 2010): *Rubric, Input, Expected response, and Relationship between Input and Expected response*. Although the framework is not intended only for the development or validation of listening comprehension tests, it does provide a very useful taxonomical framework by which contextual parameters can be categorized. This categorization helps to make it easier not only to identify as many contextual parameters as possible but also to operationalise them as appropriately as possible. Alongside the preceding L2 listening processing model, the proposed framework of contextual parameters will provide us with a theoretical framework for the validation of the JNCTL (and the Course of Study on which it is based).

1.9 Overview of the thesis

The present study explores the validity of the JNCTL. This thesis consists of six chapters. Chapter 1 describes *the Centre Test* in the Japanese context and states the purpose for undertaking this study, followed by the intended contribution to the wider field of language testing and teaching. The socio-cognitive approach (Weir, 2005) to test validation informing this study is introduced at the end of the chapter following a general discussion of test validity and validation.

Chapter 2 first aims to establish the socio-cognitive framework for validation purposes through a thorough, extensive review of related research and then research questions are proposed. Chapter 3 sets out the research design, elaborates on the methodology, and presents the results of Preliminary studies (which answers research questions 1 through 4). Methods employed include expert reviews, focus group interviews, and a large-scale questionnaire. It is through Preliminary studies that a key parameter is identified to improve the validity of the JNCTL in a direction that would be accepted by the stakeholders.

Chapter 4 describes the methodology for the Main study, which addresses research questions 5 and 6. Methods included a small case study using experimental design and a questionnaire. Chapter 5 reports on the results, which shed light on the validity of the JNCTL by providing empirical evidence about the key parameter.

Chapter 6 revisits the research questions, and discusses the key findings. Recommendations for the future JNCTL, for the development of listening comprehension tests more generally, and for English teaching and testing are proposed. These recommendations will contribute to enhancing students' listening ability, and thereby 'their practical communication ability' in English in the real world, which is a major goal of this study.

Chapter 2 Literature review

2.1 Introduction

To inform the validation of the listening comprehension component of the *Centre Test* in Japan (henceforth, JNCTL), this chapter sets out the components of

- a) a second language (L2) listening processing model for operational testing purposes and
- b) a comprehensive framework of contextual parameters.

In the socio-cognitive approach to test validation advocated by Weir (2005), there is a need for an L2 listening processing model that will allow test developers and researchers to investigate the cognitive validity of tests of listening. The model employed here is based in the first instance on the reading model developed for test validation purposes by Khalifa and Weir (2009), adapted to apply to listening skills on the basis of a review of the relevant literature.

Khalifa and Weir (2009) also point to the importance of context to listening comprehension and the processes engaged by the listener. The framework of contextual parameters employed in the present study was also established through a thorough and extensive review of the literature. Elements were mapped out under the four headings proposed by Bachman and Palmer (1996, 2010) for the investigation of task characteristics. The possible effects of each parameter on the test performance and/or the cognitive processes of Japanese (high school) learners of EFL are

discussed.

2.2 L2 listening processing model

Cognitive validity, as described by Weir (2005), is similar to *interactiveness* as conceived by Bachman and Palmer (1996). This concerns the extent and type of involvement of a test taker's language ability (language knowledge and metacognitive strategies), topical knowledge, and affective schemata in performing a language task (Bachman & Palmer, 1996). What is important for the present study is that the cognitive processing involved in real life language use should be reflected as far as possible in language test situations if claims for validity are to be supported. Test validation must therefore start from an understanding of L2 listening processes. This study first proposes a L2 listening processing model based in the literature, and attempts to examine the JNCTL (and the Course of Study which requires the JNCTL to reflect itself) in terms of cognitive validity.

It is generally accepted that comprehension is comprised of the complementary use of top-down (knowledge-based, conceptual based) processing and bottom-up (data-driven) processing (Marslen-Wilson & Tyler, 1980; Voss, 1984; Lynch, 1988; Anderson & Lynch, 1988; Richards, 1990; Peterson, 1991; Buck, 1990; Rubin, 1994; Matsusaka, 1995). Top-down processing means making use of prior knowledge – or 'higher-level knowledge' (Ericsson & Kintsch, 1995) such as background knowledge or domain knowledge – in analysing and processing information that is received (words, sentences, etc.). Bottom-up processing, on the other hand, means making use principally of information that is already present in the data (that is, the words, sentences, etc.). Thus, listeners exploit their prior knowledge, expectations, experience, scripts (sequences

of expected behaviours in a given situation), and schemas (mental structures that represent some aspect of the world) while analysing the words and sentences in the text (Richards, Platt, & Platt, 1985).

This leads to the widely accepted view that comprehension is a form of cognitive processing: the input perceived is first decoded from the auditory stream and is then further processed on a general cognitive level (Wolff, 1987). This theory of comprehension accounts for two things: (a) perceptual processing, that is the selection and decoding of perceptual stimuli: sound and word recognition, and (b) the subsequent treatment of these linguistic units: conceptual and propositional processing (Wolff, 1987). Perceptual processing corresponds to bottom-up processing, and conceptual and propositional processing to top-down processing.

Anderson (2005) conceives conceptual and propositional processing as two different stages: parsing and utilisation. Parsing refers to segmentation of an utterance, by which words are transformed into a mental representation of the combined meaning of the words. The segments are then recombined to generate a meaningful representation of the original sequence. Utilization, on the other hand, refers to the stage at which the listener may draw different types of inferences to complete the interpretation. Thus, Anderson's model is comprised of three different levels of processing: perception, parsing, and utilization. It is noteworthy that all three stages are interrelated and recursive and can happen concurrently during a single event (Goh, 2000; Anderson, 2005). This is consistent with the first complementary model, or parallel interactive processing model (Buck, 1990).

Neither the interactive processing model nor Anderson's model, however, suffices for the development of valid listening comprehension tests, primarily because these models do not allow test developers to

operationalise listening ability well enough. Those models do not give sufficient attention to the relationship between cognitive processing and contextual parameters (Weir, 2005). The issue of how the conditions under which a task is carried out (*context*) relate to the cognitive processes engaged by the listener is central to listening test validation, the concern of this study. Furthermore, these models do not specify constructs of a listening comprehension test in sufficient detail for test validation to be conducted. For example, some crucial phonological contextual parameters such as speech rate, hesitations, or reduced forms in connected speech, are not described in the conventional models. It is for this reason that I propose a practical L2 listening processing model for test validation purposes (see Figure 2.1).

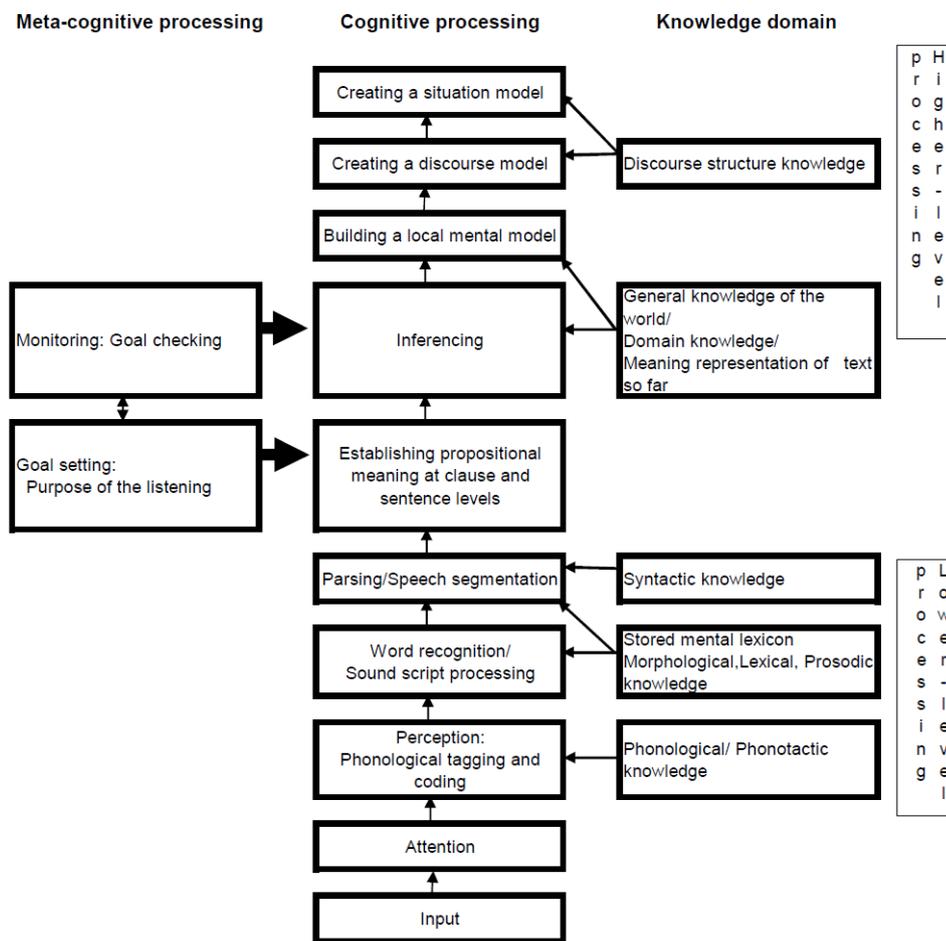


Fig. 2.1 L2 listening processing model (adopted from Khalifa and Weir, 2009, p.43)

The model provides a clear framework for exploring listening comprehension test constructs and helps to explain why the choice of contextual parameters may affect cognitive processing. We accept that, as mentioned earlier, listening comprehension involves parallel processing, although the simplified flow chart model offered may not fully reflect the extent to which the operations described are simultaneous and interactive. We also understand that the model does not incorporate the roles of echoic, working and long term memory. However, these are not seen to be shortcomings given that the model is intended to prioritise practicality for

test validation purposes rather than to provide a fully comprehensive description of L2 listening processes.

This model is based on a similar practical test validation model for reading developed by Khalifa and Weir (2009), which provides a clear framework for exploring reading comprehension test constructs and why the choice of contextual parameters may affect cognitive processing during reading. Major changes were made to Khalifa and Weir's model (2009), particularly at the lower level of processing, reflecting the differences between processing spoken and written texts. It was considered that the higher level of processing makes no difference between the input modes once a propositional meaning at clause and sentence levels is established. It should also be noted that this model is based on what Buck (2001, p. 27) terms 'controlled processing' rather than 'automatic processing' which he claims is necessary for the most efficient language use, since controlled processing is more common when performing unfamiliar skills such as L2 listening. A brief explanation of the features of the model follows.

Column A in Figure 2.1, represents a central executive processing module, involving goal setting (for example, Weir, 2005; Khalifa & Weir, 2009), or the purpose for listening. The purpose of a listening activity will determine the processing necessary for its completion. That is, the purpose governs and influences how language users or test takers approach attention-management during an actual task. In language test situations, test takers have a clear purpose of answering the questions put to them as successfully as possible. Skehan (1996) agrees that post-task activities can change the way in which learners direct their attention during the task.

In column A, metacognitive processing, or monitoring (Bacon, 1992; Rubin, 1994; Bachman & Palmer 1996, Skehan, 1996; Thompson & Rubin 1996; Weir, 2005, Goh & Taib, 2006; Goh, 2008) is also included.

Monitoring refers to the process of checking comprehension as listening is taking place (O'Malley, Chamot, & Küpper, 1989). O'Malley et al. (1989) argue that in listening comprehension monitoring is a key process that distinguishes better learners from poorer learners. They add that monitoring consists of maintaining awareness of task demands and information content and propose that two metacognitive strategies can support monitoring. These are selective attention, or focusing on specific information anticipated in the message, and directed attention, or focusing more generally on the task demands and content. This view of monitoring in listening comprehension is consistent with the concept of selecting between different purposes of listening.

Column B in Figure 2.1 represents cognitive processes. The bottom of the column represents lower (bottom-up) and the top higher level (top-down) processes. As soon as our auditory system receives a signal, the signal is processed through the several levels as long as attention or redirected attention is paid to this signal. These levels include perception, including phonological tagging and coding (Lund, 1991; Matsusaka, 1995; Rost, 2002; Wilson, 2002), word recognition or sound-script processing (Cutler & Butterfield, 1992; Rost, 2002, 2005; Field, 2003; Vandergrift, 2006), parsing or speech segmentation (Richards, 1983; Goh, 2000; Rost, 2002, 2005; Field, 2003), establishing propositional meaning at clause and sentence levels from input, building a local mental model (Rost, 2002; Anderson, 2005), constructing a discourse model (Buck, 2001; Rost, 2002), and situation models (Van Dijk & Kintsch, 1983; Khalifa & Weir 2009). Constructing a discourse model refers to constructing an organised representation of a discourse, whereas constructing a situation model can be conceived as arriving at a pragmatic interpretation. The construction, therefore, involves understanding of implicitly intended message.

It is through phonological tagging or coding of the language that a phonological representation can be achieved and conveyed to the next processing level of word recognition, though it should again be emphasized that this process interacts with all other levels. Word recognition is an important step toward meaning but may be a difficult one for EFL learners (Tsui & Fullilove, 1998; Goh, 2000; Goh & Taib, 2006; Enomoto, 2007; Field, 2009). This is because even if phonological coding appears successful, many L2 listeners may fail to identify words or locate word boundaries in connected speech. One explanation for this is their less developed mental lexicon (Vandergrift, 2006). Matching the aural form of a word with the word in their mental lexicon (sound-script processing) does not work unless they have a sufficient stored lexicon (Goh, 2000). Another explanation for such problems may be the mismatch between listeners' expectations of what they hear and real phonological realisation. Field (2003, p. 330) warns that 'listeners' expectations are sometimes unduly influenced by exposure to the written language.' We can assume that Japanese L2 listeners' phonological expectations are to a large extent affected by the written language, since formal English education in Japan has long prioritised reading and writing over speaking and listening.

Parsing or speech segmentation refers to establishing the relationship between the meaning of individual words and the meaning of whole utterances. Generally, the listener constructs a meaning of an idea unit or 'chunk of information which is viewed by the speaker or writer cohesively' (Kroll 1977), and then forgets the actual words and the syntax, and so only a summary of the meaning, the gist, remains (Sachs, 1967). Following processing through perception, word recognition, and parsing, input can come at this point to constitute a proposition. A proposition is defined as 'a simple idea that consists of a concept (usually a noun) and something

that describes or relates to that concept (usually a verb or adjective)' (Buck, 2001, p. 28). He argues that 'since storing a large number of propositions in memory is a tremendous burden, we make mental models of the content' (Buck, 2001, p. 28). These mental models are continuously updated and revised as more input reaches listeners, or as the monitoring process indicates that it is necessary to construct a better understanding of the discourse. I have, following Khalifa and Weir (2009), termed this processing building a local mental model (see Figure 2.1, p. 17). It is in this on-going processing that coherence plays a key role in facilitating the integration of discrete, local meaning representations into an enhanced mental model. At the top of column B, listeners, while processing discourse, form two models in their minds; a discourse model of a comprehension, which is a semantic representation of the propositions in the discourse, and a situation model, which is a mental model of what the discourse is about (Van Dijk & Kintsch, 1983). The situation model can be conceived as a pragmatic interpretation.

In column B, inferencing (O'Malley et al., 1989; Buck, 2001; Field, 2004; Anderson, 2005) is included in the series of processes, since inferencing is said to be 'at the core of language processing' (Buck, 2001, p. 147). Anderson (2005) agrees that in understanding a sentence, the listener must make a good number of inferences. Buck (2001, p. 148) points out that 'inferencing is involved at all levels of language processing, even where information is explicitly stated.' For example, we may infer the meaning of a word from the context, or we may infer plausible supporting grounds for logical arguments or plausible interpretations of what a speaker has said (Rost, 1990). In the model we place inferencing prior to the mental model or after establishing propositional meaning from input. This location is intended to stress that no initial mental representation (local

mental model) can be created without inferencing, and that meaning does not lie in the spoken discourse itself but in the listener's own mental representation which is activated by inferencing.

Column C in Figure 2.1, represents a knowledge dimension. Linguistic and other forms of knowledge underpin the cognitive process(es) set out in column B, and are engaged in completing all real language use (or test) tasks. Relevant aspects of knowledge include phonological, morphological and phonotactic knowledge (Rost, 1990), the stored mental lexicon (Rost, 1990; Cutler & Butterfield, 1992; Goh, 2000; Field, 2003; Vandergrift, 2006; Chang & Read, 2006), prosodic knowledge (Vanderplank, 1993), syntactic knowledge (for example, Conrad, 1985; Anderson, 2005; Joyce, 2008), general knowledge of the world (for example, Lund, 1991; Bacon, 1992; Rost, 1990, 2002), domain knowledge, co-text knowledge (Field, 2004), and discourse structure knowledge (Khalifa & Weir, 2009).

Alongside the framework of contextual parameters which follows, the proposed L2 listening processing model provides us with a theoretical framework for the validation both of the JNCTL and the Course of Study promulgated by MEXT, on which it is based (see Section 1.2).

2.3 Contextual parameters

Four major characteristics of test tasks proposed by Bachman and Palmer (1996, 2010) provide the description of the contextual parameters: They include Rubric, Input, Expected response, Relationship between input and expected response.

2.3.1 Rubric

Table 2.1 Rubric

Structure	the number of parts/ items
	sequence of parts/items
	relative importance of parts/tasks
Instruction	language (L1 or L2)
	channel (audio or video)
	specifications of procedures and tasks
Time allotment	minutes(seconds)
Scoring method	criteria and weight

The test rubric includes ‘those characteristics of the test that provide the structure to the test and tasks’ (Buck, 2001, p. 118). It is important in that test developers can increase the situational authenticity of the test or the task if they can structure the test rubric in such a way as to replicate a non-test situation the test takers are going to meet in the future. As shown in Table 2.1, the characteristics of the rubric include the structure of the test (2.3.1.1), instructions (2.3.1.2), time allotment (2.3.1.3), and scoring method (2.3.1.4).

2.3.1.1 Structure

The structure includes the number of parts or items, their sequence, and their relative importance. First, in the case of a listening test, it is said that

'the items should ask for information in the same order in which it occurs in the text' (Weir, 2005, p.64). Otherwise, it may confuse test takers, which could lead to unreliable performance (Buck, 2001). Second, preferably, the fewer items associated with an input text, the better, because of the limitations of listeners' short-term memories. This contrasts with reading comprehension, where readers can read the same input text as many times as they like and the cognitive load of short-term memory is not as heavy as in listening. Third, shorter texts should precede longer texts. This not only helps alleviate test takers' test anxiety but also reduces the level of cognitive demand. For the same reason, items should be laid out in such a way that easier items appear first and then more difficult ones. Preventing construct irrelevant variance (Messick, 1996) from intervening into the test and contaminating the results, these considerations would lead to improve the validity of the listening test. Finally, the relative importance of parts or items has an effect on weighting, which is to be discussed later (see Section 2.3.1.4).

2.3.1.2 Instructions

The instructions should be 'candidate-friendly, intelligible, comprehensive, explicit, brief, simple, and accessible' (Weir, 2005, p. 57). In monolingual situations, where complex instructions are involved it is preferable to give them in the candidates' first language. In Japan, where monolingual situation is generally the case, the rubric does not have to be in the target language but could be in Japanese, if comprehensibility or intelligibility is prioritised. Also, if possible, instructions should be associated with sample items to get the test takers familiar with the format and the procedure before the test begins (Buck, 2001). This should also help test takers to become acquainted with speaker' voice (see Section 2.3.2.1).

These considerations contribute to test validity.

2.3.1.3 Time allotment

Time allotment is concerned with the duration of the test as a whole and of the individual tasks (Bachman & Palmer, 1996). Weir (2005) argues that decisions relating to time allotment clearly impact on the processing and hence on the validity of the test. For this reason it must be made clear to test takers how much time should or can be spent on each part of a test, and it should be clearly indicated on the test paper (Weir, 2005). Given sufficient time, L2 readers in a reading test may read a text again and again in order to create a situational model. This is not the case, however, with a listening comprehension test. L2 listeners cannot replay the text as often as they like, suggesting that even if given sufficient time, it does not necessarily benefit L2 listeners if they do not understand the text. Instead, the number of opportunities to listen to the input or speech rate has more impact on listening comprehension for L2 listeners (see Section 2.3.2.1), since L2 listeners cannot replay the text they have just listened to.

2.3.1.4 Scoring method

Weir (2005, p. 63) states that ‘the criteria by which [the test takers’] answers will be judged need to be made apparent.’ We should note that the criteria also have an effect on both goal setting and monitoring in the test takers’ cognitive processing involved in the listening task completion. This means that the criteria affect the validity of the test since goal setting and monitoring are conceived as the two key cognitive parameters (see Figure 2.1).

The criteria have to be explicit about weighting, which is concerned with the assignment of a different number of maximum points to a test item, task

or component in order to change its relative contribution to other parts of the same test (Weir, 2005). Consequently, weighting reflects relative importance of items, parts or tasks in the test. Ideally, weighting should be made a decision based on, for example, the exact item difficulty estimates calibrated through statistical analysis instead of being dependent on the intuitive, unreliable decision by a test developer. This will surely raise the scoring validity proclaimed by Weir (2005) and thereby the whole validity of the test.

Under the current JNCTL practice, each item is assigned equivalent weighting. Discussions can be made about relative importance of each item or part.

2.3.2 Input

Bachman and Palmer (1996, 2010) define input as consisting of material in a given test task, which the test takers are expected to process in some way and to which they are expected to respond. This material is described in terms of format (2.3.2.1) and linguistic aspects (2.3.2.2).

2.3.2.1 Format

Format has to do with the way in which the input is presented, and includes the following characteristics; channel, speaker, text length, and the number of opportunities to listen to the input. Table 2.2 supplies the contextual parameters in relation to format.

Table 2.2 Format for input

Channel	the way in which the input is presented	video or audio
Speaker	gender of the speaker	male or female
	the number of speakers	one, two, or more
	acquaintance with speaker's voice	familiar or non-familiar
Text length		the number of words
The number of opportunities to listen to the input		once, twice, or more

Channel

Channel is mode of presentation of input. Channel may be by video or by audio. It is recommended that input be presented by video as well as audio because 'students who are familiar with television often find video more interesting, challenging, and motivating than audio recordings' (Thompson & Rubin, 1996, p. 334), and because video allows for the use of a wider range of strategies than audio.

Video, however, may have only a very limited effect on listening comprehension (for example, Mueller, 1980; Coniam, 2001). What is important in terms of language test development is whether or not the use of videos (visuals) improves both situational and interactional authenticity (Buck, 2001). That is, if visual information would be available in the target language use situation (Bachman & Palmer, 1996; Wagner, 2010), the channel should be video or visual, while if the target language use situation would involve understanding a radio news bulletin, the channel should be audio. Target language use situations refer to *target language use domain* defined by Bachman and Palmer (1996, p. 44) as 'a set of specific language tasks that the test taker is likely to encounter outside of the test itself, and to which we want our inferences about language ability to

generalize.’ In the case of the JNCTL, the target language use situations are specified in the CS, on which the JNCTL is based. This specification helps test developers make a decision about which channel to employ, and this should also contribute to validity (see Section 3.4.3).

Speaker

The category of speaker includes characteristics of the speaker or speakers such as gender, the number of speakers, and familiarity with the speaker’s voice on the part of the listener.

Gender of the speaker

The gender of the speaker refers to whether the speaker(s) in a text in a test task is/are female or male. Markham (1988) investigated the effect of the gender of the speaker on students’ recall on listening comprehension. The result revealed that students recalled more idea units from the male speaker than from the female speaker, suggesting that students listened more attentively to the male speaker. Another finding of Markham’s is that a brief introduction of the female speaker greatly enhanced the students’ perception of her, while that of the male speaker did not appear to be as important. This may suggest that the introduction of female speakers may neutralise the gender distinction.

Markham’s findings, however, cannot easily be generalized due to the paucity of the related research. Of particular relevance to the gender of the speaker in terms of language test development is a reflection of target language use situations. For example, if the language use situation is concerned with the talk between a professional baseball player and his boss, then the speaker should be male rather than female. We should also note that more and more jobs are becoming gender-free, so that the

gender of the speaker does not matter so much as it used to.

The number of speakers

Brown (1995, p.62) has suggested that 'the fewer individuals and objects, the easier the text is.' We can assume that test takers are likely to find texts easier to follow when they involve fewer speakers. The more speakers in a text, the more test takers have to discriminate between them and to process different accents or idiosyncrasies of each speaker. This may increase their cognitive load and lead to listening difficulties.

Acquaintance with speaker's voice

Acquaintance with speakers' voice refers to whether test takers are familiar with the voice of speaker in the discourse they are exposed to. The familiarity may interact with test anxiety and the number of opportunity to listen to the input (see Section 2.3.2.1) to affect test performance. Among language tests, listening comprehension test have been said to be particularly stressful for test takers because of the speakers' unfamiliar accents, lack of clarity of articulation, or 'sloppy' pronunciation (Rost, 2002). In fact, care was taken to accustom participants to the speaker's voice in a Flowerdew and Tauroza's experiment (1995). Flowerdew and Tauroza (1995) intentionally retained a segment of lecture which was not relevant to the experiment to ensure that the segment could serve as lead-in section during which the participants could get used to the speaker's voice and mannerisms without having to focus on the content presented. Acquaintance with speaker's voice, therefore, can be thought of as an important contextual parameter which is likely to affect test performance.

Some standardized tests (for example, TOEFL CBT, Test of English as a Foreign Language, Computer Based Test) allow test takers to become

familiar with speaker's voice by providing sample questions or instructions spoken in the same voice or voices as the test tasks before the test begins, but others (The *Eiken Test in Practical English Proficiency or the JNCTL*) do not. Taking into account the potential for test taker anxiety resulting from the high-stakes of the test and the role of hearing the speaker's voice in advance in helping them to become attuned to it, sample questions should probably be provided.

Text Length

Text length can be measured either by duration or by the number of words. It affects test takers' attention and cognitive load, and thereby test performance. The longer the text that test takers are required to listen to, the longer they will need to maintain their attention and the greater the cognitive load they will have to endure. It is reasonable to believe that as O'Malley, Chamot, and Küpper (1989) shows, 'the longer the text, the more likely it is that listeners will stop attending to it' (p. 428). For this reason the test developers of the JNCTL need to take into account limitations on the test takers' concentration and their proficiency level.

The number of opportunities to listen to the input

The number of the opportunities to listen to the input refers to how many times the test takers are provided the text. When making decision on this issue, test developers, as Buck (2001, p. 171) suggests, 'need to take into account the construct they want to measure, the characteristics of the testing situation, and the quality of the audio the test takers will hear.' Most listening comprehension tests currently in widespread involve

listening once (TOEFL, TOEIC¹, the *Eiken test*) or twice (for example, the JNCTL or Cambridge First Certificate in English²).

The argument for providing input once is that automaticity and the ability to make inference is an important part of the listening construct (Buck, 2001). He adds that hearing and processing a text a second time may involve different or additional comprehension skills from the first time (Buck, 2001). Another argument for providing the input once is that test tasks should reflect target language use domains or situations (Bachman & Palmer, 1996). Although we can ask confirmation or repetition in everyday life communication, it is often the case that we are in most cases supposed to understand the text by listening once. Further argument for providing the input once is limitation of the time to be spent on listening comprehension. Since many language tests involve not only listening comprehension but also reading comprehension or grammar section, providing the input more than once would end up requiring more time for the administration of the whole test.

On the other hand, arguments in favour of providing the input twice include that it may serve to alleviate psychological stress (Buck, 2001; Filed, 2012) and that it may compensate for a lack of interactivity when listening to recordings as test takers have no opportunity to ask for repetitions or to confirm, or clarify what the speaker means as they would in many non-test language use situations (Anderson, 2005; Geranpayeh & Taylor, 2008; Field, 2012).

The current JNCTL allows the test takers to listen to each text twice. NCUEE should explain its current practice in terms of either the construct the JNCTL is intended to measure or administrative reasons.

¹ Test of English for International Communication, developed by ETS (Educational Testing Service) and administered by TOEIC Steering Committee in Japan.

² Developed by Cambridge ESOL (English for Speakers of Other Languages)

Next, linguistic factors of input are examined. They involve phonology (2.3.2.2), syntax (2.3.2.3), vocabulary (2.3.2.4), discourse features (2.3.2.5), and pragmatic features (2.3.2.6).

2.3.2.2 Phonology

As Table 2.3 supplies, phonology involves hesitations, sandhi-variation, tempo and rhythm, speech rate, and accents, each of which is discussed below.

Table 2.3 Phonology for input

Hesitations	filled pauses(<i>uh, um ,er</i>), repetitions, false starts/repairs, fillers (<i>well, I mean, you know</i>)
Sandhi-variation	formulaic expression, weak form, assimilation, elision
Tempo	speed at which stressed words are spoken per minute
Rhythm	proportion of stressed words to total number of words on average
Speech rate	word per minute (wpm) or syllable per second (sps)
Accents	standard (north American), standard (UK), standard (others), non-standard, L2 speaker

Hesitations

Hesitations involve filled pauses (non-lexical sounds such as *uh, uhm, er*, etc), fillers (*you know, well, I mean*, etc), repetitions, and false starts and subsequent repairs.

Underwood (1989) stresses that L2 learners need to get used to hearing and dealing with spontaneous speech that includes hesitations. If listening tests fail to reflect these, then they will under-represent the listening construct (Buck, 2001; Weir, 2005).

Research, which has investigated the effects of hesitations on language learners' listening comprehension, however, has provided conflicting evidence. Voss (1979) found that about one third of perception errors resulted from hesitations. Also, Voss (1984) found two typical errors in handling hesitations: (i) a hesitation feature is misinterpreted as (part of) a word (ii) (part of) a word is misinterpreted as a hesitation feature and left out accordingly. Griffiths's (1991) findings are consistent with Voss's. Griffiths found that among hesitations, filled pauses (*uh, uhm, er, etc.*) cause perception difficulties for EFL listeners. In disagreement with Voss (1974, 1984) and Griffiths (1991), however, Blau (1991) provides counter-evidence that filled pauses and fillers are actually beneficial for EFL learners because the listeners can take advantage of the processing time provided. Thus, the question of whether hesitations hinder or help listening comprehension must still be considered to be open.

Sandhi-variation

Sandhi-variation is defined as 'phonological modification of grammatical forms which have been juxtaposed' (Henrichsen, 1984, p. 311). It refers to formulaic expressions or contraction (for example, *gonna, wanna, hasta* in place of *going to, want to* and *has to*), reduction (for example, [əglɑsə] for 'a glass of'), assimilation (for example, [dídʒu] for 'Did you'), elision, and weak forms. Thus, it can be conceived as reduced forms in connected speech (Gimson & Cruttenden, 1994; Brown & Brown, 2006).

Sandhi-variation is considered to be problematic for L2 learners (for example, Matsusaka, 1995) as it reduces the saliency of phonological features (Henrichsen, 1984) and leads to the disappearance of word boundaries and to the omission of certain vowels and consonants. As a result a sequence will sometimes be susceptible to different interpretations

(Matsusaka, 1995; Suzuki, 2008). This may be further problematic for L2 listeners who do not have as much knowledge of phonetic reduction as native English speakers (Matsusaka, 1995) and are as a consequence more signal dependent, or more dependent on phonological cues for their comprehension. We should note that 'in some situations, a particular microlinguistic element is crucial to the meaning of the entire sequence, and if the listener fails to comprehend it, he in turn fails to comprehend the sequence' (Matsusaka, 1995, pp. 58-59).

Yamauchi (2002) provides empirical support for the arguments. The study attempted to identify 'saturated point' which he defines as the stage where further repetition of input cannot enhance listeners' comprehension any more. The study revealed that overall, the saturation point overlapped with sandhi phenomena in a stream of speech. This finding confirms that sandhi-variation renders listening comprehension more difficult for L2 listeners.

One issue regarding sandhi-variation is the way in which the rate of delivery affects the occurrence of sandhi-variation as sandhi-variation is implicated in the speech rate (Gimson & Cruttenden, 1994). Gimson and Cruttenden claim that 'it may be said that the more rapid the delivery the greater the tendency to reduction and obscuration of unaccented words' (Gimson & Cruttenden, 1994, p. 230), though sandhi-variation is not limited to fast speech (Ito, 2006). This complementary relationship between sandhi-variation and speech rate makes it so difficult for researchers to isolate and identify the effects of sandhi-variation on a listening comprehension test performance that little research has been available to date about this. Consequently, the effects of sandhi-variation only on multiple-choice question (henceforth, MCQ) listening comprehension test are still open to question, despite the predominant use of MCQ format for

listening comprehension tests. Thus, further research to examine the effect of sandhi-variation on MCQ listening comprehension test performance is necessary.

Tempo and rhythm

Tempo refers to the speed at which stressed words are spoken, while rhythm refers to the proportion of stressed words to total number of words on average. Vanderplank (1993) calls the tempo 'pacing', for example, sixty syllables per minute, and the rhythm 'spacing.' He had a group of advanced level learners of English transcribe and mimic Margaret Thatcher, a former British prime minister, being interviewed, and found that 'pacing' and 'spacing' may be a more useful indicator of difficulty of listening comprehension for non-native speakers of English. He, therefore, argues that simple counting of words or syllables per minute does not bring out the real difficulties in processing a speech for learners, even though they are widely accepted (see p.36). Rather, he adds that tempo and rhythm may well determine the difficulty.

This finding is very informative for Japanese learners of EFL since Japanese, which is the first language of the JNCTL's candidates and French, which is the first language of the participants in the study by Vanderplank (1993), are syllable-timed languages unlike English, which is a stress-timed language. The difference in the rhythm and stress patterns between Japanese and English implies that the 'pacing' and 'spacing' may be no less appropriate an indicator of difficulty in listening comprehension for Japanese learners of EFL than the number of words or syllables per minute.

Speech rate

Speech rate is the speed at which a person speaks. It is an important predictable variable for listening comprehension (Buck, 2001). Words per minute or syllables per minute has been conventional uses of measuring the speech rate (for example, Tauroza & Allison, 1990). The former tends to be more widely used than the latter as counting syllables is more complex and time-consuming whereas counting words is relatively easy (Buck, 2001).

One possible explanation for the effect of speech rate on listening comprehension is automaticity of processing. If the speech rate is too fast for listeners, they do not have sufficient time to decode the input and their processing is overloaded, and automaticity will break down. Another explanation is the interaction with sandhi-variation (see p.33). Normal speed of speech involves a good deal of sandhi-variation (Gimson & Cruttenden, 1994), which may make the text more difficult for L2 listeners to understand. Speech rate no doubt affects listening comprehension test performance (for example, Griffiths, 1990). The developers of the test should note that the compromise of speech rate is most likely to threaten the validity of the test.

Accents/ L2 speakers

Accent refers to a particular way of speaking which tells the listener whether or not the speaker is a native speaker of English, or the region or country they come from (Richards et al., 1985). It is potentially a very important variable in listening comprehension. An unfamiliar accent can make comprehension almost impossible for the listener (Buck, 2001). Buck offers the example of listeners hearing an unfamiliar accent — perhaps hearing an Australian for the first time after studying with American

teachers — this can cause problems and may disrupt the whole comprehension process.

From a World Englishes perspective, Kim (2006, p.37) argues that 'English educators and researchers should be aware of the importance of a range of Englishes in international contexts while test developers should reflect this practical issue in the context of language assessment.' In the era where English is acknowledged as an international language (EIL) or as a 'lingua franca' or common language in the world, Kachru (1992) supports the importance of the variety of English or non-native speakers in international contexts, arguing that communication more often involves non-native speakers of English than native-speakers or a mix of non-native and native speakers. Also, 'native speakers' are becoming more difficult to define due to globalization. These arguments lead us to assume that use of L2 speakers can be acceptable as long as they are intelligible.

By contrast, Taylor and Geranpayeh (2011) argue that test developers should be cautious about the introduction of multiple varieties of accents into the testing of lower levels since it 'deprives listeners of a major set of phonetic cues' (p. 98).

Either way, test developers should account for the choice of a particular accent(s), and this would contribute to test validity.

2.3.2.3 Syntax

Syntax includes syntactic complexity and syntactic deviation.

Syntactic complexity

Syntactic simplification refers to simplifying a text by the use of shorter, simpler propositional syntax. 'Syntactic simplification is the process of reducing the grammatical complexity of a text, while retaining its

information content and meaning' (Siddharthan, 2006, p. 2). It involves replacing particular syntactic constructs like relative clauses, apposition or conjunction in sentence so that some target group of people (for example, EFL learners) can find the text easier to understand or process.

Syntactic complexity can be indicated by the C-unit to T-unit ratio for dialogue, while it can be indicated by the S-node to T-unit ratio for talk. The higher the ratio is, the more complex the syntactic structure is. While we can assume that the less complex the syntactic structure is, the easier the text is for the listeners, research provides conflicting evidence about the effect of simplifying the syntax of aural text on learner comprehension (for example, Mecartty, 2000). Cervantes and Gainer (1992) and Hasan (2000) show that syntactic simplification assists listening comprehension. By contrast, Speidel, Tharp, and Kobayashi (1985), Blau (1990), and Mecartty (2000) do not find beneficial effects from syntactic simplification. One possible reason is that syntactic simplification might have resulted in simplified text losing coherence, making the text harder to comprehend.

The inconsistency of the effect of syntactic simplification may result from the difficulty with isolating the effects. It is often the case that the effects of syntactic simplification are confounded with the effects of other linguistic modifications such as lexical or discoursal modification or slower rate of speech (Blau, 1990). Alternatively, the inconsistency can be attributed to different definitions of different researchers. Thus, the effect of strictly simplifying syntactic structure of aural passages on comprehension remains to be shown.

Syntactic deviation

Syntactic deviation refers to the violation of spoken grammar defined by Carter and McCarthy (2006). The reason why the present study adopts

the framework for accuracy judgments instead of conventional written English grammar lies in the widely accepted view that different grammar should be adopted into different mode of production (that is, writing and speaking). Hughes (2003) agrees that different criteria should be adopted for spoken text, arguing that criteria for evaluating spoken mode remain largely unchanged from the conventional corpus, which is based on written text. She contends that a static text-oriented view of language 'does not do justice to the spoken form and dynamic interactional features which make it the most sophisticated means of human communication' (Hughes, 2003, p. 296).

Richards (1983) says that syntactic deviation is frequent due to the effort speakers put into planning and organizing the content of their utterances in real time. Of particular relevance is that syntactic deviation could have negative effects on listening comprehension for EFL learners as the formal instruction has long placed its emphasis on accuracy based on written English grammar rather than on spoken grammar. It is assumed, therefore, that most learners are likely to be susceptible to syntactic deviations from written English grammar, even when acceptable in most forms of spoken English grammar, and have difficulties processing the syntactic deviation.

2.3.2.4 Vocabulary

Vocabulary includes word frequency level, lexical diversity, lexical density, and culturally specific words.

Word frequency level

It is self-evident that vocabulary is important to L2 listening. If L2 learners do not have access to enough of the vocabulary of an input text,

they will be unable to engage the lowest levels of the listening process. A high proportion of unfamiliar words is therefore likely to hinder comprehension. Nissan, DeVincenzi, and Tang (1996) reveal that the occurrence of infrequent words is, as we might expect, an important predictor of item difficulty for dialogue items in TOEFL listening comprehension tests.

Yanagawa and Green (2008) confirmed the effect of infrequent words in dialogue items but in an opposite direction to Nissan et al. (1996). Infrequent words defined as 2000 word frequency level or more by JACET 8000 (JACET, 2003) lowered item difficulty of the dialogue times (Yanagawa & Green, 2008). This is opposite to a widely accepted assumption that the higher word frequency level is in the text, the more difficult the text is. The directional discrepancy can primarily be explained by the hypothesis that less frequent words may provoke a narrower range of relevant mental, stored lexicon of listeners than more frequent words. That is, less frequent words, some of which can be used in particular language use situations, may provoke particular mental stored lexicon of listeners. This may have helped test takers establish meaning of proposition more easily and quickly. We should note that the words over no more than 2000 word frequency level (for example, advertise, airport, or department) were defined as infrequent words in Yanagawa and Green (2008). We can speculate that these seemingly frequent words but defined as infrequent words in their study may account for the directional discrepancy in the result between Nissan et al. (1996) and Yanagawa and Green (2008). Kasahara (2009) did not confirm the effect of infrequent words on item difficulty of the JNCTL administered in 2008, either. Instead, he revealed that lexical density affects the item difficulty of the JNCTL. Thus, while in general infrequent words may make listening

comprehension more difficult, the effect varies with other factors.

Lexical diversity

Lexical diversity is an indicator of the degree to which a text is constructed by different variety of lexis. It can be expressed as a percentage and calculated by the formula: Number of different words (type) is divided by the total number of words in the text (token). This is called type-token ratio, which is in wider use for indicating the degree.

It seems that interactional discourse may show less lexical diversity because it may involve more redundancy, whereas monologue or talk representing transactional discourse may show more lexical diversity. The inclusion of both types of discourse, therefore, seems important to reflect as many language use situations as possible, especially when it comes to the JNCTL, which is based on the CS which operationalise a wider variety of language use situations including both types of discourse.

Lexical density

Lexical density is an indicator of the degree to which a text is constructed by content words, rather than functional words (for example, prepositions or auxiliary verbs). It is expressed as a percentage of the content words in the total number of a text.

Kasahara (2009) showed that it was lexical density instead of lexical frequency or lexical diversity (see the previous section) that affected the item difficulty of the JNCTL administered in 2008. We can assume that the higher lexical density, the more packed the information in the text, and the heavier the cognitive load on the listeners. Furthermore, item difficulty is greater.

Culturally specific words

Cultural specific words refer to words which are specific to particular cultures and involve high-contextual schemata. For example, proper nouns (names for specific people, places, and products) or words that reflect unfamiliar cultural practices (for example, Civil War) are defined as culturally specific words. Sasaki (2000) demonstrated that those who read the culturally familiar cloze text tried to solve more items and generally understood the text better, which resulted in better performances than those of the students who read the original text. This leads us to assume that a text involving culturally familiar words will aid listening comprehension for L2 listeners while a text involving culturally unfamiliar words will not. This shows that the developer of the JNCTL should make every effort to keep the test fair by reducing to minimum culturally specific words which would favour or unfavour particular test takers.

2.3.2.5 Discourse Features

Discourse features are defined as those that relate to 'the nature and structure of text as a whole, including rhetorical type and textual organization' (Jamieson, Jones, Kirsch, Mosenthal, & Taylor, 2000, p. 16). They include rhetorical types, discourse markers, redundancy, propositional linearity, and degree of planning as shown in Table 2.4.

Table 2.4 Discourse features for input

Rhetorical type (Jamieson et al., 2000)	logical structure by which a text progress	definition, description, classification, illustration, cause and effect, problem and solution, comparison/contrast, analysis, regulatory
Discourse markers (Halliday & Hassan, 1976, etc.)	Signposting by which text progress is explicitly shown	<i>but, however, so, because of, as a result,</i> etc.
Redundancy (Derwing, 1996; Bejar et al., 2000)	ratio of new propositions to the total propositions of the text and the number of elements	
Propositional linearity (Tsui & Fullilove, 1998)	whether a discourse contradicts the initial schema, and the listeners have to revise it	linear or non-linear
Degree of planning (Bejar et al., 2000)	planned, somewhat planned, unplanned	

Rhetorical type (only for monologue or talk)

Rhetorical types can be generally classified as definition, description, classification, illustration, cause and effect, problem and solution, comparison and contrast, analysis, or regulatory. Although there is no definitive list of rhetorical types, this set seems reasonably representative (Weir, 2005). Also, we should note that most texts involve two and more than two tasks. A text, for example, that is primarily a description of a mechanism might also include some definitions or short descriptions of a process (Jamieson et al., 2000).

Much reading research (for example, Goh, 1990) shows that rhetorical

types affect the difficulty of reading comprehension. Goh (1990) revealed that problem-solution, comparison and contrast, and cause and effect properties produced better recalls than classification or description properties. This leads us to assume that rhetorical types affect difficulty of listening comprehension since reading and listening are both information processing intended to establish a coherent situational model of a text (Van Dijk & Kintsch, 1983). If different rhetorical types are employed, then the validity of JCNTL may be raised as different cognitive processing is likely to be engaged.

Discourse markers

Discourse markers refer to the signposting by which text progression is explicitly shown to the comprehenders. They include conjunctions (additive, adversative, causal, temporal) proposed by Halliday and Hassan (1976) as well as rhetorical signaling (Dunkel & Davis, 1994), macro-markers (Chaudron & Richards, 1986), and micro-markers (Chaudron & Richards, 1986; Flowerdew & Tauroza, 1995). Rhetorical signalings include brief pointer words such as '*first*' or '*in contrast*'. While macro-markers include '*to begin with*' or '*this is why*', micro-markers include '*well*', '*OK*', '*all right*', '*you know*', '*so*', '*but*', or '*you see*' and so on. Since the more coherent the text, the easier it will be to process (Brown, 1995), it is hypothesized that the discourse markers can aid listening comprehension. Thus, many studies (for example, Chaudron & Richards, 1986) have been conducted to investigate whether such discourse markers help EFL or ESL learners. This research, however, has produced conflicting evidence, suggesting that the effect of discourse markers on L2 listening comprehension may vary with other linguistic features or the listening proficiency of L2 listeners.

Redundancy

Redundancy can be defined as a repetition of words or phrases, paraphrase, or an extension of a concept already discussed (Derwing, 1996). Several studies have found that redundancy aids listening comprehension for L2 listeners. Redundancy can be indicated by the ratio of new propositions to total propositions in the whole text (Bejar, Douglas, Jamieson, Nissan, & Turner, 2000). In general, a lower percentage means extensive redundancy, indicating that the text is expected to be easier for L2 listeners, while a higher percentage means less redundancy and indicates that the text is expected to be more difficult, though the definition of the lower or higher has not been established.

Propositional (text) linearity

Propositional (text) linearity refers to how much the text derives from a linear presentation of ideas (Bejar et al., 2000). We can assume that a linear text is easier for listeners than a non-linear text as the listeners could guess the meaning of the linear text on the basis of the schema activated by linguistic input that they manage to process even if they failed to process some linguistic input. This is not the case, however, if the text is non-linear. The listeners of a non-linear text are supposed to process linguistic input that 'contradicts the initial schema activated and revise it accordingly' (Tsui & Fullilove, 1998, p. 446). Tsui and Fullilove investigated the kind of processing skill that is more important in discriminating the performance of L2 learners on listening test items and found that bottom-up processing is more important than top-down processing (see Section 2.3.4.1).

Since the JNCTL claims to prioritise interactional discourse, which may involve both types of discourse (linear and non-linear texts), it cannot

achieve optimum validity in relation to contextual parameters and cognitive processing unless it involves both types so as to reflect language use situations in the real world or operationalised in the CS.

Degree of planning

A spoken text can be placed on a continuum according to the degree of planning. Spontaneous speech usually found in everyday conversation is located at one end of the continuum, while a radio news or a prepared formal speech is at the other end. Although little research has investigated how degree of planning itself might affect item difficulty (Bejar et al., 2000), some of phonological (see Section 2.3.2.2, for example, hesitations or sandhi-variation), syntactic (see Section 2.3.2.3, for example, syntactic complexity), lexical (see Section 2.3.2.4, for example, word frequency level or lexical diversity), or discoursal (see Section 2.3.2.5, for example, redundancy) features vary according to whether a speech is planned or not (Buck, 2001). The degree of planning of a text should be one of the organizing features of test construction if a listening comprehension test intends to be valid. However, no matter how unplanned a text intends to be in a language test, the text cannot be unplanned any more once it is developed and used for the testing purpose, since language testing situation itself is planned. We cannot be free from this dilemma. Realising that our test development is under the constraints of the dilemma, it is possible to define the degree of planning variable as planned, somewhat planned, and unplanned (Bejar et al., 2000). This definition would help operationalise, for example, the quality or quantity of hesitations to be involved or speech rate, suggesting that contextual parameters are interrelated with each other and add to a language use situation a text is intended to reflect.

2.3.2.6 Pragmatic features

Pragmatic features include overlapping and backchanneling, and turn taking.

Overlapping and backchanneling

Overlapping is defined as the situation in which both speakers are trying to speak at the same time. Backchanneling, on the other hand, involves 'yeah', 'Oh, yes', or 'hmm' (Rost, 2005). Overlapping and backchanneling are both essential elements of interactional discourse as interlocutors (listeners) do not necessarily wait until the speaker terminates his or her turn. A brief review of, however, listening section of TOEIC or the *Eiken* test (STEP test) shows that these large-scale tests do not include overlappings or backchannelings. This can partially be attributed to reducing cognitive load on the test takers as processing the overlapping utterance will incur serious cognitive load on EFL learners. It is easier to understand any text involving individuals and objects which are clearly distinct from one another (Brown, 1995). It depends on the construct of the JNCTL whether or to what extent the JNCTL should reflect overlappings or backchannelings. If the construct includes the listening ability to comprehend spontaneous, overlapping interactional discourse, then the lack of those features suggests the compromise of contextual validity, and hence cognitive validity.

Turn taking

Turn taking refers to the change of speaker during conversation. It is assumed that the more turn taking is involved, the more difficult the text is, not only because more turn takings may mean longer text, but also

because it may lead to an unpredictable development of the conversation (see Section 2.3.2.5). Also, listeners might be supposed to make more inferences to relate each utterance to the preceding utterances (Brown, 1995) if more turn takings are involved in the dialogue.

Of particular relevance is the way in which the interlocutors are engaged in turn takings. Some turn takings do not take place in vacuum but occur with overlappings or/and backchannelings. Careful examination of interactional discourse, therefore, is necessary to raise the validity of listening comprehension test for EFL learners such as the JNCTL.

2.3.3 Expected response

Expected response refers to the physical response the test developers are attempting to elicit by the way the instructions have been written, by the task designed, and by the kind of input provided (Bachman & Palmer, 1996). It determines the processing necessary for the completion of a listening activity, that is to answer the question put to them as successfully as possible (see Section 2.2).

Expected response is composed of three elements; question (2.3.3.1), response (2.3.3.2), and time constraints (2.3.3.3).

2.3.3.1 Question

Question involves mode of presentation, language, provision, and scope as set out in Table 2.5.

Table 2.5 Question

Mode of presentation	spoken (audio) or written (text)
Language	L1 or L2
Provision	either before or after the text is presented
Scope	broad or narrow

Mode of presentation

Mode of presentation of a question is concerned with how the question is presented to the test takers, spoken or written. Some large-scale English tests like the *Eiken* test suite employs spoken mode of presentation while others (for example, Part 3 and 4 of TOEIC) does written mode only. If the question is presented in spoken mode, the test takers are required to listen to the question and understand the question before arriving at a correct answer. This suggests that the question is another important part of the input as they cannot arrive at a correct answer even if they understand the input itself. As a matter of fact, listeners performed better when questions were presented in written mode rather than in spoken mode (Iimura, 2011). This would be more often the case if the response type is a constructed one (for example, summary) rather than a selected one (for example, multiple-choice question, see Section 2.3.3.2).

If the question is presented in written mode, on the other hand, test takers recognise what the question is before the listening. One concern with the presentation of the question in written mode is that the items may include a measure of reading ability as well as the listening ability of test takers, especially if the question is provided in L2 and linguistically complicated. It is expected, therefore, that the questions should be as short as possible to avoid the contamination of the other confounding factors (for example, reading ability).

Zen-eiren (Association of High School English Teachers in Japan) argue that the options are too lengthy and so they should be less than seven or eight words and reduce other confounding variables as much as possible since the JNCTL is supposed to focus on measuring listening proficiency (Zenn-eiren, 2007). This argument comes from a premise that test takers should arrive at correct answers unless they fail to build up a successful mental model of a text. In other words, it threatens the validity of the listening comprehension test if the test takers cannot get an item correct just because they do not understand the options presented in L2, or just because they do not have the time enough to read the options while reaching an understanding of the text itself. This is where other contextual parameter such as time constraints (see Section 2.3.3.3), provision of question and/or response (see Section 2.3.3.1 or 2.3.3.2), or language (see Section 2.3.3.1) interact with each other in a test. Language test developers, therefore, should integrate all of the relevant contextual parameters into an item.

Language

The language of the question is concerned with whether the test taker's L1 or L2 is employed. In monolingual situations, it is possible to employ the test takers' L1 in the questions, but if the population of test takers includes different L1 groups, only the L2 or target language can be appropriate.

Decisions about which language to use in the test questions should take into account that where different languages are used in the input and in the questions, this requires a mode shift for the test takers, which might be distracting for them. As the JNCTL currently employs L2 as the language of the questions, the JNCTL is free from this kind of distraction.

Provision

Provision of question is concerned with when item stems or questions are provided to the test takers (that is, either before or after the text is presented to them). It is often the case that if the question is presented in written mode, it is presented to the test takers before listening. If the question, on the other hand, is provided in spoken mode, it is often presented to the test takers after listening to the text.

Berne (1995) examined relative effectiveness of different pre-listening activities. The result shows that scores for participants completing the question preview activity were significantly higher than scores for participants who were not given the activity. Berne (1995) attributes this result to activated schemata relating to passage content by previewing the questions. This suggests that question preview may affect listening comprehension test performance.

In theory, since our communication takes place in a particular context, it seems sensible that listeners or test takers know beforehand what the text is going to be about by referring to the questions. Therefore, the current practice of the JNCTL of providing the question and options before listening seems also sensible as they provide the test takers with activated schemata relating to the input text.

Scope

A broad scope question refers to tasks that require the test taker to process a lot of input. An example of a broad scope question is a 'main idea' listening comprehension question that deals with the content of an entire text. A narrow scope question refers to tasks that require the processing of only a limited amount of input. An example of a narrow

scope question is a listening comprehension question that focuses on a specific detail or a limited part of the listening text (Bachman & Palmer, 1996). Thus, a broad scope question and a narrow scope question can be referred to as a higher level processing question and a lower level processing question respectively. Of particular relevance is the L2 processing model established earlier (see Figure 2.1). A broad question, in other words, may require the construction of up to a situation model or a discourse model whereas a narrow question may require the construction of a local mental model or of propositional meaning at clause or sentence levels.

Shohamy and Inbar (1991) shows that listeners performed better on narrow scope items than on broad scope items. This was observed across topics as well as across text types and across students' levels. This may provide a useful insight into the development of listening comprehension test that broad scope questions may discriminate the test takers better than narrow scope questions. Since the JNCTL is used for gate-keeping purposes, the question scope, broad or narrow, can be a relevant contextual parameter not only for achieving greater validity but also for the actual use.

2.3.3.2 Response

As Table 2.6 shows, response includes type, mode of presentation, and time constraints.

Table 2.6 Response

Type	selected (e.g. MCQ) provision or non-provision , limited (e.g. SAQ), extended (e.g. summary)
Mode of presentation	spoken (audio), written (text) L1 or L2
Time constraint	minutes and/or seconds

Note. MCQ=multiple-choice question, SAQ=short answer question

Type

The type of the response can vary considerably (Buck, 2001). A selected response does not require the test takers to construct an answer on their own. Instead, they only have to choose an answer among the alternatives or options provided to them. Multiple-choice questions (henceforth, MCQ) is a good example of this. If the response type is selective, a decision to make is when the options are provided to the test takers (that is, either before or after they listen to the text). Research (for example, Yanagawa & Green, 2008; Iimura, 2011) shows that the choice of when to provide the options in MCQ listening comprehension test affects test performance for EFL learners.

A more significant issue, however, would be how the choice of provision of questions and/or responses would affect the ongoing cognitive processing of test takers facing a MCQ listening comprehension test, since this would be more relevant to cognitive validity (see Section 1.7) and hence provide insights into developing more valid listening comprehension tests. Little research (Wu, 1998) has explored to date the ongoing cognitive processing of the test takers on a MCQ listening comprehension test. Given that the JNCTL employs the MCQ format, this kind of research is a pressing need.

Limited and extended response, by contrast, require the test takers to

construct an answer on their own. One word answer, for example, which is often used in short answer questions is one of the limited response, whereas short summary of a text either in L1 or L2 is an example of the extended response. Large-scale tests such as the JNCTL, which attracts hundreds of thousands of test takers every year (NCUEE, 2011b), employ selected, usually MCQ type. Given the huge number of test takers, it is inevitable and reasonable to employ this response type.

Mode of presentation

Mode of presentation takes shapes of either written or spoken. Written response, for example, refers to writing one word, a sentence, or even a short summary either in L1 or L2, whereas spoken response refers to, for example, sentence-repetition tasks or summarizing what they have understood by speaking in their native language (Buck, 2001). Written response is currently in wider use than the spoken response, especially in large-scale tests.

In the case of constructed, written responses the language of the response is important (Buck, 2001). These could be given in either L1 or L2. In the case of L2 responses, the main issue is whether the response will be evaluated for correctness, appropriacy, and so forth, and whether those judgments will affect the score. In a listening test, 'it seems reasonable not to penalise mistakes if the response is intelligible and clear' (Buck, 2001, p. 126).

2.3.3.3 Time constraints

Time constraints seem to have impact on the validity of listening comprehension test as it may induce test strategy or test wiseness, which is irrelevant to the construct of listening comprehension. Provided

sufficient time but unable to build up an appropriate mental model of a text (see Section 1.2), the test takers are likely to rely on guessing to find out a correct answer by chance, especially when it comes to MCQ format. For this reason, the test developers should make careful decisions about how much time should be allotted to each section or each item of MCQ format. This can only be done after piloting a test in question and examining the empirical data (Weir, 2005). This would help raise the test validity.

2.3.4 Relationship between input and expected response

2.3.4.1 Topical knowledge

Topical knowledge includes background knowledge and domain knowledge.

Background knowledge

The ease with which discourse is understood depends in part on the listener's own background knowledge and familiarity with the scenario presented. 'Because comprehension involves constructing meaning by relating information in the input to information stored in long-term memory, the process is facilitated if the content of the input is familiar' (Leeser, 2004, p. 590) to the listener. Schmidt-Rinehart (1994) provided empirical support for the argument by showing that the more prior knowledge the listener has about the topic of a lecture, the easier it is for that listener to comprehend the lecture and retain general points of information. Also he showed that there was no interaction between topic familiarity and proficiency level of the participants. Thus, Schmidt-Rinehart concludes that topic familiarity emerged as a powerful factor at all levels of proficiency.

Some researchers, however, showed that background knowledge may

have limited or even negative effects on listening comprehension. Chiang and Dunkel (1992) found that a significant effect for prior knowledge only appeared on participants' performance on the passage-independent items of the postlecture comprehension test: If EFL learners listen to an unfamiliar topic, question type does not affect item difficulty, but that if they listen to a familiar topic, passage-independent items are easier than passage-dependent items. In addition, Jensen and Hansen (1995) showed that the size of prior knowledge effect was trivial. Furthermore, O'Malley et al. (1989) warn of the negative impact on comprehension from overextended use or misapplication of listeners' background knowledge. Tsui and Fullilove (1998) provided empirical support for the argument. They provided two types of texts for the participants in their experiment; a schema matching text and a schema non-matching text. In a schema matching text, the schema activated by the initial incoming linguistic input is congruent with the subsequent input, while in a schema non-matching text, the initial schema is refuted by subsequent input and hence test takers need to be able to continuously process the subsequent input and revise the initial schema to reach the correct answer. The result showed that less-skilled listeners were more likely to have difficulties revising their initial schema formed by their background knowledge in accordance with the incoming linguistic data than skilled listeners. This suggests that background knowledge may have negative effects on listening comprehension for less-skilled listeners, especially if the discourse does not follow what they expect it to be. Thus, the effect of background knowledge on listening comprehension may have something to do with the predictability of the discourse or the propositional linearity (see Section 2.3.2.5) and with the proficiency level of the listeners.

Domain Knowledge

Domain knowledge refers to knowledge which is shared by the people who specialise in a particular domain but is not shared by the people outside the domain. This suggests that domain knowledge is narrower than general knowledge of the world or background knowledge (see Fig. 2.1). It is thought to have a facilitative effect on listening comprehension. For example, those students who are specialising in the study of applied linguistics may know what 'comprehensible input' means, and how it affects second language acquisition, and consequently they will understand more easily what a text about relationship between input and language acquisition means than the people who are not interested in the academic field. Thus, it is reasonable to believe that domain knowledge facilitates the listening comprehension of the test takers.

Care should be taken to make sure that the JNCTL should not favour particular test takers who are interested in particular domains in order to keep the test fair.

2.3.4.2 Text-item interaction

Text-item interaction includes inference question, position of relevant information in the text, and lexical attractiveness of options

Inference question

Inference question refers to the questions which require the test takers to interpret the meaning beyond what is explicitly stated in the text. Nissan et al. (1996) investigated 283 listening comprehension dialogue items, most of which were taken from fifteen TOEFL forms administered before 1996 and identified inference questions as one of five significant predictor

variables³ affecting the item difficulty of TOEFL listening comprehension dialogue items. Yanagawa and Green (2008) confirmed that the inference questions were a predictable variable of the item difficulty of dialogue items in a test of listening comprehension. Besides, Limura (2011) identified a significant correlation between inference question and item difficulty of multiple-choice questions of listening comprehension test where the questions were presented in spoken mode (see Section 2.3.3.1). The finding that inference questions make items more difficult can be explained by the fact that inference questions require the listeners to construct a situational model and then to draw pragmatic meanings from the utterance. There must be, therefore, doubts about the validity of a listening comprehension test if it fails to include sufficient inference questions as our intended message is often implicit.

Position of relevant information in the text

Freedle and Kostin (1999) identified the position of relevant information as a strong predictor variable affecting item difficulty of short talk items on the TOEFL listening comprehension test. Relevant information – or ‘necessary information’ (Buck & Tatsuoka, 1998) – is the information that test takers have to understand to arrive at a correct answer. Freedle and Kostin (1999) found that overall, items dealing with information presented early or in the last sentence in a mini-talk were associated with easier items, whereas if the relevant information for correctly answering an item was found in the middle of a minitalk this pattern was associated with harder items. This is consistent with the so-called primacy effect or recency effect (Deese & Kaufman, 1957; Murdock, 1962). Kostin (2004),

³ Other predictable variables include utterance pattern (statement-end items or question-end items), negatives in the text, speaker’s role (specific role the speakers play in a dialogue), and infrequent vocabulary.

Yanagawa and Green (2008), and Imura (2011) confirmed the positional effect on dialogue item difficulties. It is for this reason that test developers should be aware of where relevant information is found in the text—at the beginning, in the middle, at the end, or distributed—so that the language test can achieve greater validity.

Lexical attractiveness of options

Lexical attractiveness of options is an indicator of how attractive test takers find answer options based on the occurrence of lexical overlaps and lexical inferences between the listening text and the test items. A lexical overlap in an option refers to any word appearing both in the text and in one of the answer options, whereas a lexical inference refers to a word in an item being associated with words appearing in the text. Thus, this parameter does not apply to limited or extended type of response (see Section 2.3.3.2) but only to selected type of response, which is the most popular for a language test that attracts a large number of candidates such as TOEFL or the JNCTL.

Freedle and Fellbaum (1987) found that test items can be made more difficult by increasing the number of lexical overlaps among the incorrect options and/or by decreasing the number of lexical overlaps among the correct options. These findings suggest that relationship between input (that is, text) and expected response (that is, options) would lead to differences in the ways in which test takers process the text and so are likely to affect listening comprehension test performance. Freedle and Kostin (1996, 1999) and Kostin (2004) also revealed the significant effect of lexical attractiveness of options on the item difficulty of talk items or dialogue items respectively.

Lexical attractiveness of options affects test taking strategies and test

wiseness (Cohen, 1984, 1998). Cohen (1984, 1998) found that when taking MCQ reading comprehension tests, test takers tended to adopt a test taking strategy of matching the options with the text and selecting an option because (a) it had a word/words that also appeared in the text; (b) it had words similar in sound, or meaning, to words in the text; (c) it had a word which belonged to the same word family; or (d) it just seemed somehow to be related to word(s) in the text (Cohen, 1998). This suggests that when taking listening comprehension tests, test takers will employ similar strategies, and the strategy use will be affected by the difference in when the item stems and answer options are provided to the test takers (see Section 2.3.3.1). This leads us to assume that the relationship between input and expected response will affect test takers' test taking strategy use and hence their cognitive processing during the test. Although the use of test taking strategies by the test takers during the test under-represents the construct of a language test, it is inevitable that a test invites test taking strategy use as long as a test employs a particular test format such as MCQ. This is where test validity cannot help being compromised or approximated.

These are a full description of contextual parameters that the present study identified through the extended review of literature. Thus, having established an L2 listening processing model (Fig. 2.1) and the framework of contextual parameters (Table 2.7), it is now appropriate to move on to the validation of the JNCTL.

Table 2.7 Framework of contextual parameters

Rubric	Structure	The number of parts/items, Sequence of parts/items, Relative importance of parts/items, Number of tasks/items per part				
	Instruction	Language (L1 or L2), Channel (audio or video), Specifications of procedures and tasks				
	Time allotment	Minutes				
	Scoring method	Criteria and weight				
Input	Format	1) Channel		Video or audio		
		2) Speakers	Gender, number, and acquaintance			
		3) Text length			The number of words	
		4) The number of opportunities to listen to the input			Once, twice, or more	
	Linguistic	Phonology	1) Hesitation/ pause	Filled pause, Repetition, False starts /repair, Filler (e.g. <i>I mean, well</i>)		
			2) Sandhi-variation	Formulaic expression, Weak form, Assimilation, Elision		
			3) Tempo	Speed with which stressed words are spoken per minute		
			4) Rhythm	Proportion of stressed words to total # of words on average		
			4) Speech rate	Wpm or SPS (syllable per second)		
		5) Accents	Standard, non-standard, L2 speakers'			
		Syntax	1) Syntactic complexity	S-node/T-unit ratio or C-unit /T-unit ratio		
		Vocabulary	1) Word frequency level	JACET 8000		
			2) Lexical diversity	Type-token ratio		
			3) Lexical density	Percentage of the content words in a text		
			4) Culturally specific words	The number of words		
		Discourse	1) Rhetorical type (only for talk)	definition, description, classification, illustration, cause and effect, problem/solution, contrast /comparison, analysis, regulatory		
			2) Discourse markers	The number of words		
			3) Redundancy			
			4) Propositional (text) linearity	Linear or non-linear		
			5) Degree of planning	Planned, somewhat planned, or unplanned		
		Pragmatic	1) Overlapping/ back channelling	Occurrence of the feature		
			2) Turn taking	The number of the turn takings		
		Expected Response	Question	1) Mode of presentation (audio or text), 2) Language (L1 or L2), 3) Provision (before or after), 4) Scope (Broad or Narrow)		
	Response		1) Type	Selected (provision), limited (e.g. SAQ), extended (e.g. summary)		
			2) Mode of presentation	Audio, visual, or text (L1 or L2)		
	Time constraints	Minutes and seconds				
	Relationship between Input and expected response	Topical knowledge	1) Background knowledge	2) Domain knowledge		
Text-item interaction		1) Inference question	Do the test takers need to infer beyond what is explicitly stated?			
		2) Position of relevant information in the text	Partial (beginning, in the middle, or end) or distributed			
		3) lexical attractiveness of options	overlappings between a key and the text against the overlappings between the distractors and the text			

Test validation involves justification of our interpretations or hypotheses between a test score and ability or a construct a particular test is meant to measure. In order to justify our interpretation, we should provide empirical evidence (Messick, 1989; Bachman, 1990; Bachman & Palmer 1996; Weir, 2005). This study, therefore, provides empirical evidence based on the L2 listening processing model and the framework of contextual parameters in an attempt to validate the JNCTL. To carry out the purpose of this study, six research questions were set out below;

2.4 Research Questions

Four research questions were set out for Preliminary study which precedes the Main study, while two research questions were set out with possible modifications once the outcome of Preliminary study is clear.

Preliminary study

- RQ 1 To what extent does the current Course of Study regulated by the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) achieve optimum validity in terms of contextual parameters and cognitive processing?
- RQ 2 To what extent does the current format of the JNCTL achieve optimum validity in terms of contextual parameters and cognitive processing?
- RQ 3 To what extent does the current format of the JNCTL reflect the Course of Study?

RQ 4 Which of the discrepancies identified through RQ 1 to RQ 3 are perceived to be most important to the validity of the JNCTL by the following groups of stakeholders: high school students and high school English teachers?

Main study

RQ 5 How does the intra-task manipulation of specified contextual parameters affect test performance?

RQ 6 How does the manipulation of specified contextual parameters affect the level of cognitive load imposed on the test takers?

Chapter 3 Preliminary studies: Elicitation and prioritisation of a key parameter

3.1 Introduction

This chapter describes the methodology, results, and findings for a series of Preliminary studies to answer the four research questions set out at the end of Chapter 2.

Preliminary study 1 is intended to answer RQ 1, RQ 2, and RQ 3 to identify the discrepancies between real life listening features and the JNCTL/the CS, and between the JNCTL and the CS.

Preliminary study 2 is intended to answer RQ 4 to prioritise a key parameter among the discrepancies found through Preliminary study 1 to improve the validity of the JNCTL in a direction that would be accepted by stakeholders.

Preliminary study 3 is intended to explore whether or to what extent the current JNCTL reflects a realistic level of the key parameter to address the question in greater depth than Preliminary study 1.

3.2 Preliminary study 1

3.2.1 Introduction

This section describes the methodology, results, and findings for Preliminary study 1 to answer the three research questions set out at the end of Chapter 2. After providing the overall design, methods are elaborated for each research question.

RQ 1 (contextual parameters and cognitive processing in the CS) is investigated through document analysis. The methodology for answering RQ 2 (contextual parameters and cognitive processing in the JNCTL) is presented in three different parts: a) documentary analysis, b) quantitative analysis and c) qualitative analysis. The methodology for answering RQ 3, (the JNCTL compared with the CS), is divided into four parts in line with the four components of the CS; vocabulary, practical communication ability, language use situations and functions of language.

Results and analyses for each research question are presented, followed by the findings to each research question. Discussions are briefly made about the findings.

3.2.2 Method

The proposed L2 listening processing model (see Figure 2.1) and the framework of contextual parameters (see Table 2.7) informed the collection of empirical evidence relating to the three research questions. Table 3.1 shows the research instruments used in relation to each research question. Document analysis of the Course of Study was conducted for the purpose of investigating RQ 1 through RQ 3. Additionally, to address RQ 2 and RQ 3, qualitative and quantitative analyses were carried out. Expert raters, who are specialists in L2 listening or language testing, were asked to review the forms of the JNCTL administered in the past three years (2007,

2008, and 2009) and the CS on which the JNCTL is based.

Table 3.1 Research instruments summary

RQ	Target	Document analysis	Quantitative analysis	Qualitative analysis
1	Real-life parameters vs. the CS	✓		
2	Real-life parameters vs. the JNCTL	✓	✓	✓
3	The CS vs. the JNCTL	✓	✓	✓

3.2.2.1 Research Question 1 (Real-life parameters vs. the CS)

Guided by the proposed L2 listening processing model and the framework of contextual parameters, the researcher conducted a document analysis of the CS (MEXT, 2007) to pick out the statements which were relevant to the construct of listening comprehension, and described them for the purpose of exploring into to what extent the contextual parameters listed in the framework and the L2 listening processing are reflected in the CS.

3.2.2.2 Research Question 2 (Real-life parameters vs. the JNCTL)

Document, quantitative, and qualitative analysis were carried out using the L2 listening processing model and the proposed framework of contextual parameters. The document and quantitative analysis was conducted by the researcher whereas the qualitative analysis was conducted by raters or/and by the researcher. Table 3.2 indicates what types of analysis was employed to address each contextual parameter listed in the framework. *Tempo/rhythm, redundancy, and lexical attractiveness* of options are not included in this analysis due to the absence of a clear baseline against which we might judge these features as to whether those parameters are sufficiently reflected in the JNCTL.

Also, *degree of planning* was not included in this analysis since it is self-evident that all JNCTL texts are planned. Each analysis is described below.

Table 3.2 Method for each contextual parameter

Parameter		Docu.	Qua.	Quali.	
Rubric		✓			
Input	Format	channel	✓		
		speaker	✓		
		text length		✓	
		#of opportunities to listen to the input	✓		
	Phonology	hesitations		✓	
		sandhi-variations		✓	(✓) ¹
		speech rate		✓	(✓) ²
		accent/L2 speaker			✓
	Syntactic	complexity		✓	
		deviation		✓	
	Vocabulary	frequency level		✓	
		density		✓	
		diversity (type-token ratio)		✓	
		culturally specific word		✓	
	Discourse	rhetorical type			✓
		discourse marker		✓	
		propositional (text) linearity			✓
	Pragmatic	overlapping		✓	
		turn taking		✓	
	Expected response	Question	mode/language/provision	✓	
scope				(✓) ³	
Response		type/mode	✓		
	Time constraints	✓			
Relationship between Input and Expected response	Topical knowledge			✓	
	inference question			✓	
Text-item interaction	position of relevant information			✓	
Required processing level				✓	

¹ Qualitative analysis of sandhi-variation is conducted in Preliminary study 3.

² Qualitative analysis of speech rate is conducted in Preliminary study 3.

³ Question scope was collapsed into the position of relevant information or required processing level.

Document analysis

As explained in Table 3.2, document analysis was conducted for *Rubric*, *Format* with the exception of *text length*, and *Expected response*. The researcher carried out the analysis, using the 2007 form of the JNCTL. The forms of 2008 and 2009 were not included since no difference was found in relation to these parameters across the three forms.

Quantitative analyses

The quantitative analyses were conducted by the researcher, addressing a larger part of the contextual parameters as shown in Table 3.2. The features addressed included text length, hesitations, sandhi-variations, speech rate⁴, syntactic complexity, syntactic deviation, word frequency level, lexical density, lexical diversity, culturally specific words, discourse markers, overlapping turns, and turn takings.

Text length for each text was calculated by the researcher. Hesitations referred to filled pauses (non-lexical sound such as *uh*, *uhm*, *er*, etc), fillers (*you know*, *well*, *I mean*, etc), repetitions or repeats, and false starts and subsequent repairs, based on Maclay and Osgood (1956), Blau (1991), Foster, Tonkyn, and Wigglesworth (2000), and Rost (2002). Repetitions which were intended for restating speakers' intentions or for rhetorical effect were not considered hesitations. False starts were defined as an utterance which is begun and then either abandoned altogether or reformulated in some way (Foster et al., 2000). The researcher listened to the CD recordings of the three forms while checking with the scripts available on the website of the NCUEE (2009b).

Sandhi-variations (Henrichsen, 1984) or reduced forms in connected

⁴ Sandhi-variation and speech rate also undertook qualitative analysis in Preliminary study 3.

speech (Brown & Brown, 2006) involved weak form or unaccented form (Gimson & Cruttenden, 1994), flap, assimilation (for example, [dídʒu] for 'Did you'), elision, and formulaic expressions (*gonna, wanna, hasta* in place of *going to, want to* and *has to*). The researcher listened to the CD recordings to identify as many sandhi-variations as possible.

Words per minute was employed for the analysis of the speech rate of the input texts since it tends to be more widely used (for example, Griffiths, 1990; Zhao, 1997; Brindley & Slatyer, 2002) and easily compared to other studies than articulation rate, which excludes the time spent for hesitations or pauses in its duration calculation (Koreman, 2006; Jacewicz, Fox, O'Neill, & Salmons, 2009). Although syllables per second or minute could have been a more appropriate indicator of this parameter, words per minute was employed since counting syllables is more complex and time-consuming whereas counting words is relatively easy (Buck, 2001).

For the analysis of syntactic complexity, different measures were used for different text types (dialogues and monologues) to reflect better the syntactic complexity of each text type (for example, Foster et al., 2000). Syntactic complexity for dialog was expressed by the proportion of the number of C-units to the total number of T-units, and for monologues by the proportion of the number of S-nodes to the total number of T-units (see Section 2.3.2.3). C-unit refers to a shortest unit which has a communicative value in dialogues, whether it is a word, a phrase, a sentence, or grammatical or ungrammatical (Brock, 1986), while S-node refers to sentence-nodes. T-unit refers to the shortest unit which a sentence can be reduced, and consisting of one independent clause together with whatever dependent clauses are attached to it (Richard et al., 1996). The higher the proportion is, the higher syntactic complexity is. Syntactic deviation referred to the violation of the spoken grammar defined

by Carter and McCarthy (2006), and was judged by the researcher.

Lexical density was expressed as a percentage of content words to the total number of words in the texts (Rose, 2008). Lexical diversity was also expressed as a percentage of different words to the total number of words in the text using the formula: Number of different words (type) is divided by the total number of words in the text (token). This is called type-token ratio, which is in wider use for indicating the degree. Both parameters were calculated using the Complete Lexical Tutor (<http://www.lexutor.ca/vp/eng/>). Word frequency level was calculated using JACET list of 8000 basic words (JACET, 2003). The use of the list for this analysis was primarily due to the fact that the list was developed for Japanese learners of EFL, and it fits with the purpose of this analysis: it can show the coverage by every 1000 word frequency level. Culturally specific words were defined as words which are specific to particular cultures and involve high-contextual schemata (Sasaki, 2000). For example, proper nouns (names for specific people, places, and products) or words that reflect unfamiliar cultural practices were defined as culturally specific words.

Discourse markers referred to the signposting by which text progression is explicitly shown to the comprehenders. They included rhetorical signalling showing additive, adversative, causal, and temporal relationships between the texts (Halliday & Hassan, 1976; Dunkel & Davis, 1994), and macro-markers (Chaudron & Richards, 1986) including '*this is why*' or '*to begin with*'. Accordingly, discourse markers included *so, and, in fact, in addition, furthermore, but, to the contrary, by contrast, on the other hand, nevertheless, this is why, therefore, thus, consequently, for this reason, as a result, to begin with, first, lastly, before long, after a while, after all, eventually*, and so on. The researcher checked with the scripts and

counted the number of discourse markers.

The analyses of overlapping turns and turn takings were conducted only for the dialogues from Part 1 to Part 3 by the researcher, who listened to the CD recordings and referred to the scripts where necessary.

Qualitative analyses

Three sets of qualitative analyses are described: Raters, Training sessions, and Procedures. These analyses, as supplied in Table 3.2, addressed accent/L2 speaker, rhetorical type, text linearity, inference questions, topical knowledge, position of relevant information, and the required processing level.

Raters

The analysis for accent/L2 speaker, rhetorical type, and topical knowledge was conducted by the researcher. The analysis of rhetorical type and topical knowledge addressed only talks or monologues in Part 4. Rhetorical types included, as presented in Table 2.7, definition, description, classification, illustration, cause and effect, problem and solution, contrast/comparison, analysis, and regulatory. The rater (i.e. researcher) attempted to identify, if any, multiple rhetorical types for a monologue.

In the meantime, the analysis for text (propositional) linearity, position of relevant information, inference questions, and required processing was conducted by the three raters, who are specialists in L2 listening or language testing. One is a university professor who specialises in English phonetics and phonology, another is an associate professor with a PhD in the field of language testing, and the other is an experienced high school teacher with an MA degree in the field of L2 listening. They were asked to review the JNCTL administered in 2007, 2008 and 2009 in relation

to the four qualitative parameters.

Training sessions of the raters and the development of a coding sheet

The first training session

Before their review of the JNCTL, a series of training sessions was held in order to get the raters familiar with the qualitative parameters but also to raise the validity of the coding sheet. The listening comprehension component for Preliminary English Test (henceforth, PET) developed by Cambridge ESOL (English for Speakers of Other Languages) and for Second Language English Proficiency Test (henceforth, SLEP Test) developed by Educational Testing Service were used for their practice as they were considered to be at an equivalent proficiency level of the JNCTL. One item each from four different parts of the listening comprehension component of the PET was chosen, while three items were chosen from the listening comprehension component of the SLEP Test, resulting in seven items in total.

The raters were asked to rate the seven items from two to four nominal (categorical) scales at home in relation to the following six qualitative parameters: question scope, (rhetorical) organization, propositional complexity, position of relevant information, inference questions, and required processing level. The group convened in Tokyo in 2007 to discuss any difficulties they had experienced in their coding, and the discrepancies between the coding. The discussions led to three major changes to the coding sheet:

- 1) Definitions or descriptions were given to each parameter for clarification.
- 2) Organisation, which concerned with the text structure (that is, either explicit or implicit), was collapsed as it was considered overlapping with

an inference question.

- 3) Propositional complexity was renamed propositional predictability as the new name reflected more aptly the parameter which was concerned more with text predictability than with text complexity. Thus, the first training session, the subsequent discussions, and the resulting amendments to the coding sheet were hoped to raise the validity of this method.

The second training session

The second training session aimed to further improve the validity of the ratings. In the second session, the seven items in the part 1 of the listening comprehension component of the PET were used. The same three raters were given the material and asked to code the seven items in relation to the five parameters. As expressed in the Table 3.3, the rate of agreement among the raters was fairly high. Unable to arrange another meeting due to the availability of the raters, the researcher informed the raters of the model codings.

Table 3.3 Results of coding for the second training session

Parameters	Rater	Item						
		1	2	3	4	5	6	7
Required processing level	A	2	2	2	1	2	1	1
	B	2	2	1	2	2	2	1
	C	2	2	1	1	2	1	2
Propositional predictability	A	3	3	3	3	3	3	3
	B	2	2	2	2	2	2	2
	C	2	2	2	3	3	2	3
Question scope	A	2	2	2	1	1	1	1
	B	2	2	2	1	2	2	1
	C	2	1	1	1	2	2	1
Position of relevant information	A	4	4	4	2	4	2	2
	B	4	4	2	2	4	2	2
	C	4	4	2	3	4	4	2
Inference questions	A	1	1	1	1	2	1	1
	B	1	1	1	1	1	1	1
	C	1	1	1	1	2	1	1

Note. Required processing level: 3=situation model, 2=discourse model, 1= propositional model
 Propositional predictability; 3= unpredictable, 2= somewhat predictable, 1= predictable
 Question scope; 2=broad, 1= narrow,
 Position of relevant information; 4=distributed, 3=at the 1st utterance, 2= in the middle, 1= at the last utterance
 Inference question; 2=inference question, 1=non-inference question

Based on the results of the second training session, further amendments were made to the rating sheet:

- 1) Question scope was eliminated because it was found to be redundant. This was because it was closely associated with position of relevant information and required processing level. If the question scope is broad, for example, then the item requires a processing level higher than constructing a propositional meaning at clause or sentence level (see Figure 2.1) and understanding two or more than two utterances by finding coherence. The position of relevant information is likely to be distributed in the text. By contrast, if the scope is narrow, the item will require the processing level of constructing a propositional meaning at clause and sentence level and understanding word(s) or short utterances. Relevant information will be found locally (such as in the middle or at the end of the text). This simplification could have contributed to the practicality of the review.

- 2) A non-linear-text was redefined as a text where listeners should revise their initial schema as the on-going text contradicts the initial schema: It was initially defined as 'a text which can be predictable by using background knowledge.' Accordingly, 'propositional predictability' was newly termed 'text linearity'. Three options were collapsed into dichotomous options: 'Yes' or 'No', which seemed to be much easier for the coding.

- 3) The four variables for position of relevant information were collapsed into three variables; 'in the beginning or the end', 'in the middle', or 'distributed' as no relevant information was found in the final utterances of the source texts in the second session.

The collapse of question scope, the simplification of variables, and

redefinition of non-linear texts was intended to make the coding exercise easier for the raters. It was through this process that the validity of the coding was improved. Thus, the final coding sheet was prepared (see Table 3.4).

Table 3. 4 Coding sheet for qualitative parameters

Parameters	Question	Option
Processing level	What processing level is required to arrive at a correct answer?	(1) A level to understand word(s) or a utterance (2) A level to understand two or more than two utterances (3) A level to understand the whole text (including integrating information or using background knowledge)
Text linearity	Is the text linear?	(1) Yes (2) No
Position of relevant Information	Where is the relevant information found in the text?	(1) either at the beginning or the end of the text (2) in the middle (3) distributed
Inference question	Does the item require the test takers to interpret the implicit meaning beyond what is explicitly stated in the text?	(1) Yes (2) No (Chose this if the item requires simple calculation or understanding anaphora)

Procedures

A packet of coding materials was prepared. This included

- the test brochures for each form of the JNCTL administered in 2007 to 2009 taken from NCUEE website (NCUEE, 2009b)
- the answer keys
- the tape scripts
- a set of CD recordings for each form

The coding sheets were sent out by post to the raters in August, 2009. For the purpose of refreshing their memories and reconfirming their understandings of the rating, a coding sheet was attached with five practice items taken from the 2006 form of the JNCTL. The practice items were two short dialogues, two longer dialogues, and one longer monologue, each accompanied by an expected coding which they were not allowed to consult until they had completed their own coding. If they chose an option different from the expected coding proposed by the researcher, then they were strongly recommended to review their coding and reconfirm their understanding of the relevant parameter or parameters. The raters were asked to return the coding sheet filled out with their coding to the researcher within a few weeks (see Table 3.20 for the rate of agreement between/among the raters).

3.2.2.3 Research Question 3 (the CS vs. the JNCTL)

Document analysis was conducted of the CS (RQ 1) by the researcher, and the analysis found that the CS operationalises four components – vocabulary size, practical communication abilities, language use situations, and functions of language. Accordingly, the research question was specified as follows:

Whether or to what extent does the JNCTL reflect the CS in terms of vocabulary, practical communication ability, language use situations, and functions of language?

Table 3.5 Method for exploring the components of the CS

Components of the CS	Quantitative	Qualitative (rater judgment)
vocabulary size	✓	
Practical communication ability		✓
Language use situations		✓
Functions of language		✓

To answer the specified RQ 3, quantitative analysis was conducted to explore vocabulary size, while rater judgement was used to investigate practical communication ability, language use situations, and functions of language, as presented in Table 3.5. Details of the raters and how final decisions about each component were made are given in Table 3.6.

Table 3.6 Qualitative analysis for the components of the CS

Components of the CS	the number of raters	How to make a final decision
Practical communication ability	2 (including researcher)	discussion
Language use situations	1 (researcher)	NA
Functions of language	3 for the 2007 form, 2 for the 2008 and 2009 forms	In case of no agreement, the researcher decided.

Practical communication ability was coded by two raters including the researcher, while the other one was one of the same raters that conducted the qualitative analysis to answer the research question 2 (see Section 3.2.2.2). In case of any discrepancy, the final decision was to be made

through discussion between the raters. Language use situations were categorised by the researcher himself. Functions of language for the 2007 form were judged by the same three raters that conducted the qualitative analysis of RQ 2, while functions of language for the 2008 and 2009 forms were judged by two different raters: One was a university professor who specialises at TEFL (Teaching of English as a Foreign Language) or TESOL (Teaching of English as a Second Language), and the other was an expert with MA degree in the field of language testing. The researcher would make the final decision about this component if no agreement could be reached among the raters.

It was not intended that there should be differences in the number of raters and it was not planned that the researcher should act as one of the raters, but the availability of the raters and the demanding nature of the coding task made these compromises unavoidable.

The method for analysing each component is discussed below.

Vocabulary size

JACET list (JACET, 2003) of 8000 basic words was selected for investigating the vocabulary used in the JNCTL since it is based on the vocabulary of Japanese learners of English. All of the texts for each form of the JNCTL administered from 2007 to 2009 were put into v8an.pl, an analytical programme attached with the JACET 8000 list. This programme reports the level of coverage of each 1000 word frequency level in a text up to the 8000 word level.

Practical communication ability

Practical communication abilities operationalised in the CS are set out below.

- a) Understanding information, content, and the speaker's or writer's intentions
- b) Grasping the outline and the main points
- c) Expressing or conveying their own ideas
- d) Responding in a way appropriate to the situation and the purpose
- e) Deepening the understanding of language and culture

(MEXT, 2007)

The qualitative analysis of the components of practical communication ability covered by the CS confined itself to ability b) – grasping the outline and the main points – as it is self-evident that a) applies throughout the JNCTL while ability d) is applied in Part 2 of the JNCTL. Abilities c) and e) are clearly not tested by the JNCTL.

Two raters as shown in Table 3.6 reviewed each form of the JNCTL administered from 2007 to 2009. That is, each rater judged seventy-five items in total according to whether each item requires the listeners to grasp the main point or not. Discussions were made to resolve the discrepancies between the raters until agreement was reached (see p. 103 for the rate of agreement).

Language use situations

The CS operationalises four different language use situations (MEXT, 2007): a) Individual uses such as phone calls, travelling, or shopping. b) Use in groups such as speeches, presentations, or discussions. c) Uses involving larger number of people such as newspapers, advertisements, or radio. d) Creative communication such as skits or dramas based on a group discussion programmes. The researcher attempted to identify a

language use situation for each item in the JNCTL administered from 2007 to 2009.

Functions of language

The CS specifies five different functions to be engaged (MEXT, 2007); smoothing human relationships, transmitting feelings, transmitting information, transmitting ideas and intentions, and instigating action. Three raters as in RQ 2 were asked to identify the most salient function of language out of the five functions specified in the CS for the items in the 2007 form of the JNCTL, while two raters were asked for the 2008 and 2009 forms (see Table 3.6). Although it was argued that isolating one function from the multiple functions involved in any stretch of discourse was difficult, the raters were encouraged to single out the most salient function for each item (see p. 105 for the rate of agreement and Appendix 2).

3.2.3 Results and analyses

3.2.3.1 Introduction

This section describes the results of the analyses for the three research questions for Preliminary study 1. Following the results and analyses for RQ 1, the section for RQ 2 is divided into three parts, documentary, quantitative, and qualitative analysis, while the section for RQ 3 is divided into four parts, vocabulary, language use situations, practical communication ability, and functions of language.

3.2.3.2 Research Question 1

The document analysis showed that the CS specifies 1,800 different words, 'basic' collocations, use of 'standard' English as a model, particular sentence patterns and grammatical elements, and places importance on grasping the main points of a text. The CS does not provide any specifications concerning the format, rubric, response format, and other linguistic elements such as discoursal or phonological (excluding accent) features. The CS, therefore, does not provide a basis for achieving optimum validity in terms of contextual parameters and cognitive processing.

3.2.3.3 Research Question 2

Document analysis

Rubric

The results for structure are presented in Table 3.7. The JNCTL is composed of four major parts and six different tasks. The six tasks are

- 1) match the correct picture or number,
- 2) complete the conversation,

- 3A) choose the correct options according to the content of a dialogue,
 3B) information gap,
 4A) choose the correct option according to the content of a monologue,
 and
 4B) choose the correct options according to the content of a longer
 monologue.

Part 1 through 3 employ dialogues and Part 4 employs monologues.
 While each text is associated with only one item in Parts 1, 2, 3A, and 4A,
 the longer texts in Part 3B and 4B are each associated with three items.
 The total number of items was twenty-five, with twenty-one texts.

Table 3.7 Structure

Part	Task	D or M *	# of text	# of items for a text	# of items
1	Match the correct picture or number	D	6	1	6
2	Complete the conversation	D	7	1	7
3A	Choose a correct option according to the content of a dialogue	D	3	1	3
3B	Information gap	D	1	3	3
4A	Choose a correct option according to the content of a monologue	M	3	1	3
4B	Choose correct options according to the content of a monologue	M	1	3	3
Total			21		25

Note. D refers to dialogues while M refers to monologues.

The results for the other contextual parameters grouped into the Rubric— instruction, time allotment, and scoring method are supplied in Table 3.8. The instructions are not only provided in L1 on the test brochure but also through audio. Time allotment is thirty minutes after the reading or grammar sections is over. Scoring method employ machine-scanning with two marks for each item, totalling fifty points. No additional weighting is given to any of the sections or the tasks; each item is worth two points.

Table 3.8 Instruction, time allotment, and scoring method

Parameters	
Instruction	L1(Japanese), printed + audio
Time allotment	30 minutes, no flexibility
Scoring method	machine scanned
	no relative importance according to sections or tasks
	2 marks for each item, totalling 50 points (25 items)

Format

The results for *format* are summarised in Table 3.9; Format includes channel, speakers, text length⁵, and the number of opportunities to listen to the input.

⁵ The results of text length are presented in Table 3.11 (p. 88).

Table 3.9 Format

Channel	audio through individual IC players
Speakers (the number, gender, acquaintance)	four native speakers (two male and two female) no practice items to get the test takers acquainted with speakers' voice
The number of opportunities to listen to the input	twice

The channel was audio through individual IC players given to teach test taker. Only the volume can be controlled by the individual test taker. Regarding the speakers, four speakers (two male and two female) were employed to record all of the texts. The number of speakers, however, used in a text was either one or two. Nineteen out of twenty-five items (76 per cent) involved two participants whereas six items (24 per cent) involved individual language use. That is, the texts used were limited to either one or two speakers. This shows that the JNCTL is devoid of multi-participant discussions and this may under-represent the construct of listening comprehension.

Care was not taken to get the test takers acquainted with the speaker's voice as practice items are not provided. This can partly be attributed to NCUEE's assumption that test takers are familiar with the procedure or the types of tasks since the practice tests or the real tests of the JNCTL are widely available for the test takers. Also instructions are all provided in L1 (Japanese) instead of L2 (English), which helps the test takers understand how the test is going and what they are supposed to do. Another possible explanation for the absence of practice items can be time constraints. With practice items, the duration of the test would be longer and this would affect the entire administration of the *Centre Test*. Thus, test takers do not have a chance to get acquainted with the speaker's voice before listening.

Texts are played twice: test takers have two opportunities to listen to the input. There is a brief interval between the two playings of the text. This does not necessarily reflect listening in the real world, where language users would generally have only one opportunity to listen. The number of opportunities to listen to the input, therefore, may be one of the discrepancies between real life listening features and the JNCTL.

Expected response

Table 3.10 summarises the results for expected response, which includes question, response, and time constraints.

Table 3.10 Expected response

Question	printed presentation L2 preview
Response	multiple-choice questions printed presentation L2 preview
Time constraints	fixed

Printed in the test brochure, the question is provided in L2 (English) before listening. The response employs multiple-choice questions with a single correct answer and three distractors, and the options are also provided in L2 before listening to the input text. Time constraints are fixed. Test takers should answer each question at fixed intervals.

Thus, the document analysis of the CS and JNCTL so far elicited two discrepancies between real life listening features and the CS/JNCTL: They are:

- 1) a lack of multi-participant discussions
- 2) the number of opportunities to listen to the input

Quantitative analysis

The quantitative analysis included text length, hesitations, sandhi-variations, speech rate⁶, syntactic complexity, syntactic deviation, word frequency level, lexical density, lexical diversity, culturally specific words, discourse markers, overlapping turns, and turn takings. The result for each parameter is reported.

Table 3.11 Text length (# of words)

Part	2007	2008	2009
1 (short dialogues)	15 to 30	24 to 31	29 to 32
2 (short dialogues)	15 to 30	14 to 24	17 to 32
3A (medium-length dialogues)	45 to 50	47 to 51	43 to 54
3B (a longer dialogue)	146	151	149
4A (medium-length monologues)	66 to 80	91 to 98	89 to 98
4B (a longer monologue)	184	204	189
Total number of words	1,010	1,087	1,110

Text length

As shown in the Table 3.11, the text length for each section does not vary much across the forms. For example, it was 184, 204, 189 words for Part 4B for the 2007, 2008, and 2009 forms respectively. Also, as the test is proceeding, the text length becomes longer. The JNCTL starts with dialogues of 20 to 30 words in Part 1 and ends with a longer monologue of

⁶ Sandhi-variation and speech rate also undertook qualitative analysis in Preliminary study 3.

about 200 words in Part 4B. This suggests that the JNCTL is well controlled not only within an administration but also over the forms.

Phonology

The results for three phonological parameters are presented in this section. They include hesitations, sandhi-variation and speech rate.

Few *hesitations* were found across the forms as illustrated in Figure 3.1. Six fillers (*well (4), you know (1), kind of (1)*) and one filled pause (*uhm*) appeared out of a total of 1,010 words in the 2007 form whereas five fillers and five filled pauses were found out of a total of 1,087 words in the 2008 form. In the 2009 form, five fillers, one filled pause, and two false starts were found. No repetition for repairs occurred across the three forms. The lack of hesitations can be considered one of the discrepancies between listening in the real world and the JNCTL since hesitations represent interactional discourse in the real world.

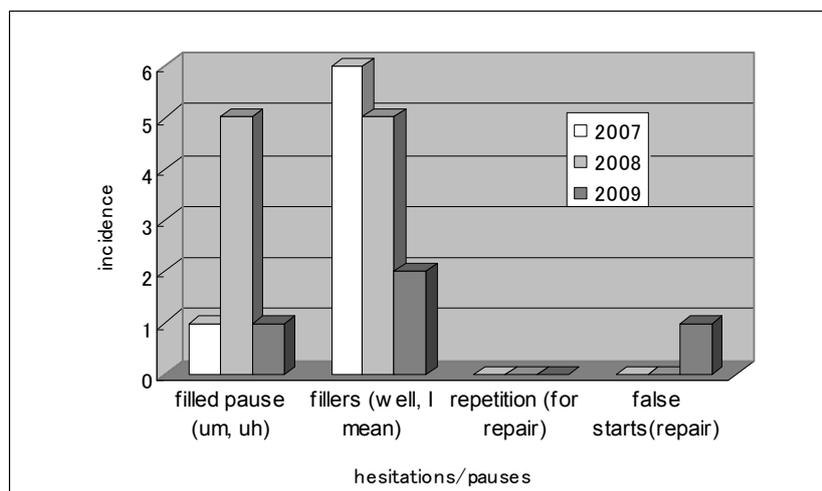


Fig.3.1 Hesitations

Table 3.12 supplies *sandhi-variations* included in the current JNCTL. We should note, however, that no baseline against which we can judge the

validity of sandhi-variation was established. For this reason, a qualitative analysis of sandhi-variation was planned in Preliminary study 3 to clarify the validity of the sandhi-variation in the JCNTL.

Table 3.12 Sandhi-variations

Sub-category	2007	2008	2009
Weak form	<i>have been(1), and(3), them(1), of(1), had(1),</i>	<i>and, could I, been(1), will(1),them(2)</i>	<i>has been, they, We're, and it, that'll (2), them (3), I'll, you can</i>
Flap [r]	<i>get out of(1), pair of(1), writing(1), letters(1), But I(1), get it(1), How's(1), but it(1), little(1), put it(1)</i>	<i>letters, got(3),couldn't have done it, right, in it(1)</i>	<i>putting, get a, did I, saw it in, should I, pretty, When I, started</i>
Assimilation (reduction)	<i>could you(2), did you(2), don't you(1), join us(1)</i>	<i>with us, an eye on my, couldn't have done it, got your, when I</i>	<i>but your, in an hour, should we, on it</i>
Elision	<i>take her(1), have him(1), it(3), would(2), wanted(1), centre(1), asked(1), that(2)</i>	<i>didn't(4), them, them, wha(t) part, don't you, couldn't have done it, wanted, tha(t)(2), couldn'(t),ca(p),cen(t)i-meter</i>	<i>aren't, put, get to, centre</i>
Formulaic expression	<i>kind of(1)</i>	<i>kind of(1)</i>	<i>(0)</i>

Note. () indicates the number of incidences.

Table 3.13 Speech rate (wpm)

part	2007	2008	2009
1 (short dialogues)	167	140	175
2 (short dialogues)	189	165	175
3A (medium-length dialogues)	159	170	160
3B (a longer dialogue)	179	175	150
4A (medium-length monologues)	149	175	160
4B (a longer monologue)	151	155	140

The results for *speech rate* are supplied in Table 3.13. Overall, the speech rate for interactional discourse in Part 1 through 3 in the JNCTL seems to be 'slightly' slower than that in listening in the real world, whereas the speech rate for transactional discourse in Part 4 is considered similar to that in the real world (see Section 2.3.2.2, Tauroza & Allison, 1990). However, given that typical speech rate varies according to the speaker, the vehicle, types of discourse, or communication events, it is in practice almost impossible to establish a baseline against which we can judge the validity of the speech rate of the current JNCTL. It is for this reason that at this point the speech rate has not been determined as one of the discrepancies between real life listening features and the JNCTL. Speech rate is to be further addressed in Preliminary study 3.

Syntax

Table 3.14 presents the results for syntactic complexity for each form.

Table 3.14 Syntactic complexity

Part	2007	2008	2009
1 (short dialogues)	1.08	1.67	1.32
2 (short dialogues)	1.23	1.18	1.2
3A (medium-length dialogues)	1.13	1.14	1.04
3B (a longer dialogue)	1	1.25	1.12
4A (medium-length monologues)	1.08	1.32	1.62
4B (a longer monologue)	2.5	1.73	1.46

Note. Part 1 through 3B is based on C-unit/T-unit

while Part 4A and 4B is based on S-node/T-unit.

The results show that syntactic complexity looks slightly larger than 1. It ranges between 1 and 1.73 with the exception of Part 4B in 2007 (2.5). This may reflect the JNCTL's prioritising the interactional nature of discourse (NCUEE, 2006, 2007, 2008, 2009a, 2010) rather than transactional discourse (Richards, 1990).

Of particular relevance to this study is the difficulty of claiming validity about syntactic complexity with those figures as no established baseline is available for the syntactic complexity.

No syntactic deviations were found across the three forms. This may be consistent with the lack of sandhi-variations such as false starts and repetitions for repairs by the speakers (see Figure 3.1). This suggests that the JNCTL prioritises accuracy of the input text to the 'naturalness'. The result, that no syntactic deviations was found in the monologues, is sensible in that monologue texts should originally be in written mode and should not allow for any syntactic deviations. Dialogues, on the other hand, representing interactional discourse in our everyday life communication could have involved some syntactic deviations because this could have made the texts simulate listening in the real world to a

greater extent. Nevertheless, lack of syntactic deviations in the text cannot claim to be a discrepancy between real life listening features and the JNCTL since syntactic deviations is not a construct of the JNCTL.

Vocabulary

This section reports the results for lexical diversity, lexical density, and culturally specific words⁷.

Table 3.15 Vocabulary

Parameter	2007	2008	2009
Lexical diversity (type-token)	42 %	42 %	41 %
Lexical density	50 %	51 %	51 %
Culturally specific words	<i>Mr. Richard Clayton,</i> <i>Ms. Tucker,</i> <i>Thompson, Joe, Pat</i>	<i>A.B Brothers, Billy,</i> <i>George, Jim, Kathy,</i> <i>Lee, Leonardo da</i> <i>Vinch, M, Maria,</i> <i>May, Redwood,</i> <i>Smokey, Tayler,</i> <i>Timothy, Tommy,</i> <i>Wavery</i>	<i>Centerville,</i> <i>Wakefield,</i> <i>Margaret</i>

Note. See Table 3.21 for the results of word frequency level.

As shown in the Table 3.15, lexical diversity (type-token ratio) and lexical density (the proportion of content words to the total number of words in a text) were consistent across the three forms, suggesting that both parameters are well controlled across the forms. By contrast, this may not be necessarily the case with culturally specific words. More culturally specific words were used in the 2008 form than in the 2007 and the 2009

⁷ The results for word frequency level are to be presented in Section 3.2.3.4.

forms. Also, most of the culturally specific words were proper nouns such as the name of a person or a place (a city or a station).

Discourse

This section reports the results for discourse markers.

Table 3.16 Discourse markers

	2007	2008	2009
Discourse markers	<i>but(7), so(2), because of(1), nevertheless(1)</i>	<i>but(6)</i>	<i>but (11), so (3), in fact (1), therefore (1)</i>
Total	11	6	16

Note. The total number of words were 1,010, 1087, 1,110 for the 2007, 2008, and 2009 forms respectively.

As Table 3.16 illustrates, the number of discourse markers was eleven for the 2007 form, six for the 2008 form and sixteen for the 2009 form. This seems relatively small, given that the total number of words used for each form was more than 1,000 words. Besides, a relatively smaller variety of discourse markers was found. Out of thirty-three markers identified across the forms, twenty-five markers (76 per cent) was '*but*' and five markers (15 per cent) was '*so*', and the rest was *because of*, *nevertheless*, *in fact*, and *therefore*. This shows paucity and a smaller variety of discourse markers in the JNCTL.

The paucity and a smaller variety of discourse markers, however, cannot be claimed to be a discrepancy between real life listening feature and the JNCTL since NCUEE stresses that the JNCTL prioritises the interactional nature of discourse (NCUEE, 2006, 2007, 2008, 2009a, 2010), where the test structure tends not to be explicitly signposted by discourse markers.

Pragmatic features

The results for two pragmatic parameters are reported in this section. They are turn-takings and overlapping turns. As Table 3.17 shows, the number of turn-takings is consistent across the forms. Also, the number of turn-takings is increasing from a few in Part 1 to about ten in Part 3B. This suggests that the number of turn-takings is well controlled for, and that heavier cognitive load is incurred in the later sections.

No overlapping turns were found, suggesting that optimum contextual validity may not be achieved since overlapping turns may be one of the salient features which characterise interactional discourse in our everyday communication. We should also note NCUEE stresses that the interactional discourse is the JNCTL's primary target (NCUEE, 2006, 2007, 2008, 2009a, 2010). Lack of overlaps, therefore, may be considered one of the discrepancies between listening in the real world and the current JNCTL.

Table 3.17 Pragmatic features (only for dialogues)

Parameter	Part	2007	2008	2009
Turn takings	1	2 or 3	4	4
	2	2 or 3	2 or 3	3
	3A	4 to 6	3 to 6	4 to 6
	3B	12	9	13
Overlapping turns	1 to 3B	0	0	0

Note. This analysis was not applied to Part 4, which employs monologues.

Thus, the quantitative analysis elicited two further discrepancies in addition to the two discrepancies identified by the document analysis (see Section 3.3.2.1). They are a lack of hesitations, and a lack of overlapping turns, both of which characterises interactional discourse the JNCTL claims

to prioritise (NCUEE, 2006, 2007, 2008, 2009a, 2010). Next section elaborates on the results of the qualitative analysis of RQ 2.

Qualitative analyses

Qualitative analysis was conducted either by the researcher or by a panel of three raters. This section first reports the results for accent/L2 speaker, rhetorical types, and topical knowledge, which were analysed by the researcher, and then the results for required processing level, text linearity, position of relevant information, and inference questions analysed by the panel.

Accent

The JNCTL employs standard North American accent. It does not use other accents at all, neither does it L2 speakers. This reveals 'standard model' of English in the CS (MEXT, 2007) can be conceived as North American accent. This may coincide with a widely accepted view that 'standard model of English' to be learnt at secondary schools in Japan should be North American.

The exclusive use of North American accent, however, may under-represent listening in the real world. The test takers, who are high school students or high school graduates, are likely to encounter a wider range of language use situations in the future, where more communications are likely to be engaged through the media of other accents than North American accent. This is more often the case when English is established as a common language and used as an international language (EIL) throughout the world (McKay, 2002). The lack of variety of English accents used, therefore, can be considered one of the discrepancies between listening in the real world and the JNCTL.

Rhetorical type

The analysis of rhetorical type addressed only monologues in Part 4B. The result was that the JNCTL employed descriptions across the forms. The constant use of descriptions suggests that the JNCTL is more interested in the narratives, which may not need as many varieties of rhetorical types as expositions where logical connections should be explicitly made by them. This result may be consistent with the paucity and a small variety of discourse markers (see Table 3.16) in that the JNCTL prioritises interactional discourse over transactional discourse where the text structure tend to be explicit.

We should note, however, that rhetorical type is not a discrepancy between listening in the real world and the JNCTL as other rhetorical types such as contrast/comparison or cause/effect were also employed in 2007 and 2009 forms as shown in Table 3.18.

Table 3.18 Rhetorical type

Parameter	2007	2008	2009
Rhetorical type	description, contrast/comparison, cause and effect	description	description, cause and effect

Note. This analysis was only applicable for Part 4B.

Topical knowledge

As shown in Table 3.19, the JNCTL used texts on a range of topics—a prize given to a woman novelist, a shrimp as a souvenir from Hawaii, and a new on-line journal—which look unique and interesting. None of the topics and contents are likely to favour or disfavour particular test takers. That is, the JNCTL does not presume particular topical or domain

knowledge the test takers are supposed to have to arrive at a correct answer. Thus, topical knowledge is not a discrepancy between listening in the real world and the JNCTL.

Table 3.19 Topical knowledge

	2007	2008	2009
Topical knowledge	a prize given to a woman novelist	a shrimp as a souvenir from Hawaii	a new on-line journal

Note. These topics were named by the researcher.

Four contextual parameters, *required processing level, text linearity, position of relevant information, and inference questions*, were analysed by three raters. Their coding, however, diverged in spite of a series of training sessions (see Section 3.2.2.3) as the rate of exact agreement shows in Table 3.20. Consequently, the final decision about each coding was made on the agreement of the codings: where two or three raters agreed, this decision was taken as final. The researcher’s coding was referred only when the three raters’ codings were different with one other, which occurred at nine codings across the three forms—six (four items for required processing level and two items for inference questions) for the 2007 form, two (one for required processing level and one for position of relevant information) for the 2008 form, and one (for position of relevant information) for the 2009 form.

Table 3.20 The rate of exact agreement among the raters (%)

Parameters	2007	2008	2009
Required processing level	52	44	16
Text linearity	76	86	67
Position of relevant information	36	60	24
Inference question	50	61	67

Required processing level

Figure 3.2 illustrates the results for required processing level, suggesting that three hierarchical processing levels are required across the forms. In addition, relatively more items which require constructing discourse (a level to understand two or more than two utterances) were found across the forms than the items which required constructing propositional meanings. Furthermore, relatively more items which require constructing situation models (a level to understand the whole text including integrating information or using background knowledge, see Table 3.4) were found in the 2008 form than the items which required constructing discourse or propositional meanings. This suggests that the JNCTL involves items calling for a wider range of cognitive processing. Thus, required processing level is not a discrepancy between listening in the real world and the JNCTL.

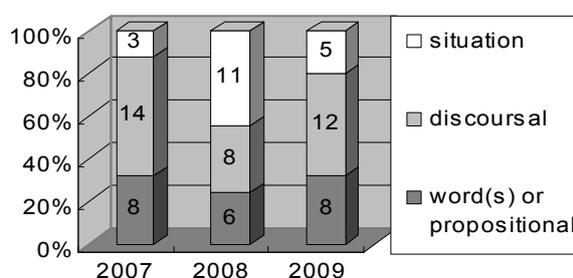


Fig. 3.2 Required processing level

Text linearity

The results for text linearity are displayed in Figure 3.3, suggesting that few non-linear texts were employed in the 2008 and 2009 forms. The number of non-linear text reached seven (28 per cent), two (8 per cent), three (12 per cent) for the 2007, 2008, and 2009 forms respectively. This result shows that text linearity is one of the discrepancies between listening in the real world and the JNCTL.

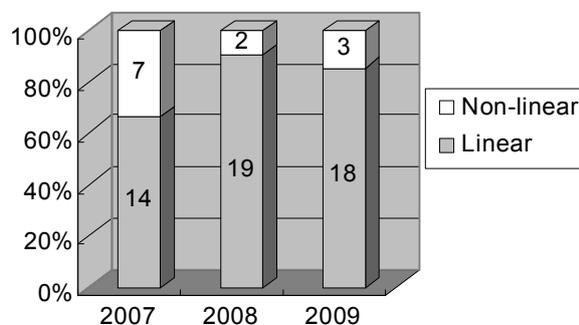


Fig. 3.3 Text linearity

Position of relevant information

Figure 3.4 presents the results for the position of relevant information. The relevant information for thirteen items in the 2007 and 2008 forms (52 per cent) was distributed across the texts while the relevant information for twelve items (48 per cent) was found locally (that is, either in the middle or at the beginning or the end). In the 2009 form, the result is different from the 2007 and 2008 forms in that more relevant information was found in the middle. Overall, found in the different places, the position of relevant information cannot be claimed to be a discrepancy between real life listening features and the JNCTL.

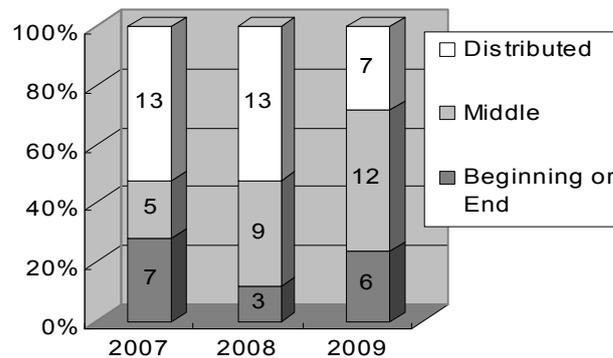


Fig. 3.4 Position of relevant information

Inference question

The number of inference questions reached no more than three (12 per cent), four (16 per cent), and two (8 per cent) for the 2007, 2008, and 2009 forms respectively as displayed in Figure 3.5. This relatively small number of inference questions suggests that this parameter is a discrepancy between listening in the real world and the JNCTL.

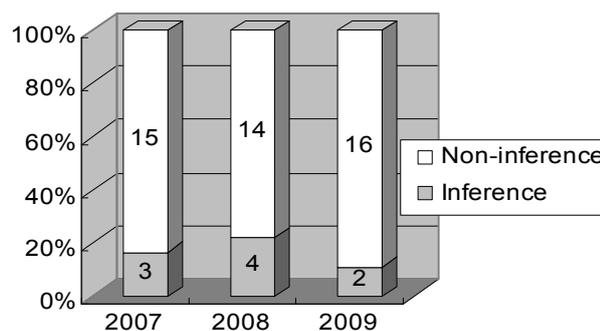


Fig. 3.5 Inference question

Note. This analysis was not applicable to the seven items in Part 2 as they are not intended to test inferencing skills (see Appendix 1).

These results elicited two further discrepancies between the JNCTL and the listening in the real world — a lack of non-linear texts and a lack of inference questions. Thus, the results of RQ 1 and RQ 2 elicited eight discrepancies in total.

They are:

- a lack of multi-participant discussions
- the number of opportunities to listen to the input (listening twice—rather than once as in most real life situations)
- a lack of hesitations
- a lack of variety in the English accents used
- a lack of L2 speakers
- a lack of non-linear texts
- a lack of overlapping turns
- a lack of inference questions

The identification of these discrepancies shows that the JNCTL does not achieve optimum validity in terms of contextual parameters and cognitive processing. Next section draws on the results of RQ 3.

3.2.3.4 Research Question 3

Vocabulary size

Table 3.21 supplies the ratio of the words used in the JNCTL, which is covered by the every 1000 word frequency level of JACET 8000 list (JACET, 2003). The ratio is indicated by both by type or by token for each form of the JNCTL. Token refers to the total number of words in a text, while type refers to the number of different words (Richards et al., 1985). The CS allows for the use of 1800 different words (MEXT, 2007). In the analysis, the 1800 words was approximated as 2000 word frequency level.

The result, that, as supplied in Table 3.21, 73 to 87 per cent of the words used in the JNCTL are covered by 2000 word frequency level, shows that overall, the JNCTL reflects the CS in terms of vocabulary use⁸.

Table 3.21 Coverage by word frequency level by JACET 8000 list

Word frequency level	Type			Token		
	2007	2008	2009	2007	2008	2009
1,000 word	67%	61%	63%	80%	79%	80%
2,000 word	82%	73%	75%	88%	85%	87%
3,000 word	85%	78%	80%	89%	87%	89%
4,000 word	87%	81%	83%	91%	91%	91%

Practical communication ability

Two raters were asked to identify whether the items require grasping the outline or the main point of the discourse. Rate of agreement was 84 per cent for the 2007 and 2009 forms and 68 per cent for the 2008 form. All discrepancies between the two raters were resolved by discussion.

As Figure 3.6 shows, fourteen (56 per cent) or sixteen (64 per cent) items out of twenty-five were identified as requiring an understanding of the main points of the texts whereas eleven (44 per cent) or nine (36 per cent) items were focusing on local points of the texts. This result suggests that the JNCTL reflects the CS well in terms of the component of practical communication ability in English.

⁸ The words beyond 4000 word frequency level are listed in Appendix 4.

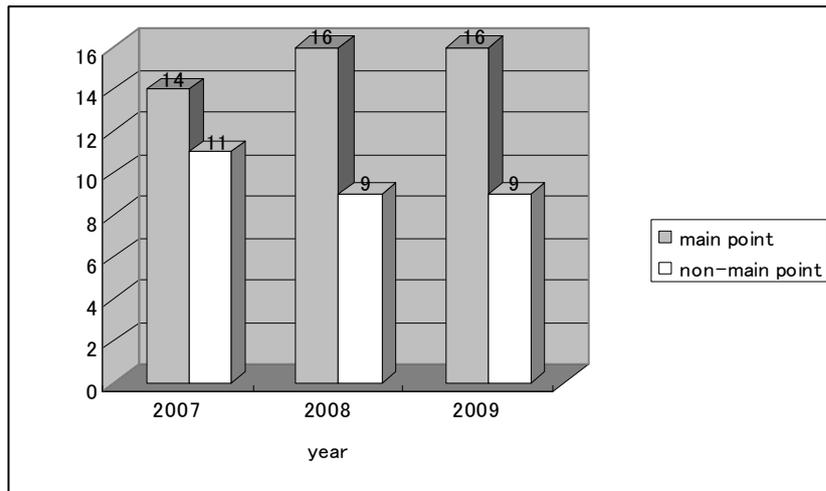


Fig. 3.6 Practical communication ability

Language use situations

As Table 3.22 shows, nineteen (71 per cent) or twenty (80 per cent) items concerned individual language use situations whereas three to five items were addressed to an unspecified broadcast audience. Few presentations to a group and no creative communication were found. This is consistent with the finding from RQ 2 that the JNCTL involves either one or two speakers and is lacking multi-participant discussions (see Section 3.2.3.3).

This result shows that language use situations in the current JNCTL were predominantly covered by individual language use situations and thereby the JNCTL does not reflect all of the language use situations operationalised in the CS. Thus, the result confirms that lack of multi-participant discussions is a discrepancy not only between real life listening features and the JNCTL (RQ 2), but also between the JNCTL and the CS.

Table 3.22 Language use situations

Language use situations	2007	2008	2009
(a) Individual language use	20 (80%)	19 (71%)	20 (80%)
(b) Presentation to a group	0	3 (12%)	1(4%)
(c) Unspecified audience	5 (20%)	3 (12%)	4(16%)
(d) Creative communication	0	0	0

Functions of language

Out of the twenty-five items in the 2007 form of JNCTL, the three raters agreed on the coding for thirteen items (52 per cent), two raters agreed on eleven items (44 per cent), and no agreement was reached on one item (4 per cent). Where two or three raters agreed, this decision was taken as final. On the item where no agreement was reached, the researcher took the final decision.

As for the 2008 and 2009 forms, the two raters agreed on the coding for twenty-four items (96 per cent), and no agreement was reached on one item (4 per cent) for each form where the researcher took the final decision. That is, on only three items across the three forms (4 per cent, 3/75) did the researcher take the final decision. This low ratio was considered not to reduce the validity of this analysis.

As shown in Table 3.23, eighteen, fourteen, or eleven items out of twenty-five items were identified as involving transmitting information in the 2007, 2008, 2009 forms respectively. Three, four, or six items were found transmitting ideas and intentions in the 2007, 2008, 2009 forms respectively. Three, six, or seven items were identified instigating action in the 2007, 2008, 2009 forms respectively. Few items were found across the forms for transmitting feelings and smoothing human relationships.

The result shows that the JNCTL most appreciates transmitting information.

This does not necessarily mean, however, that the JNCTL do not appreciate other language functions. We should note that the reviewers were asked to single out the most salient function among the five provided, even though they found out multiple functions of a text (see Section 3.2.2.2). For this reason, it is impossible to say that the JNCTL is lacking the functions specified in the CS.

Table 3.23 Functions of language

Function	2007	2008	2009
Smoothing human relationships	0	0	1
Transmitting feelings	1	1	0
Transmitting information	18	14	11
Transmitting ideas and intentions	3	4	6
Instigating action	3	6	7
Total	25	25	25

Thus, the results of RQ 3 revealed that the JNCTL does not fully reflect the CS, since it does not involve full situations of language use as specified in the CS, though the JNCTL reflects the CS in terms of vocabulary and practical communication ability (the ability to grasp the main points of a text).

3.2.4 Findings

3.2.4.1 Research Question 1; Real-life parameters vs. the CS

The results of this Preliminary study showed that the CS is not sufficiently detailed to inform the development of the JNCTL. The answer to the first research question, therefore, was that the CS does not achieve optimum

validity in terms of contextual parameters and cognitive processing. Since the CS is not intended for the development of any language test, it is not surprising that the CS does not achieve optimum validity. The CS is no more than a set of guidelines, which shows general educational policy at secondary schools in Japan (see Section 1.1). Therefore, NCUEE should specify what the JNCTL is meant to measure rather than just repeating 'It is intended to measure the core knowledge the high school graduates are supposed to have acquired by the time of their senior high school graduation' (NCUEE, 2006, 2007, 2008, 2009a, 2010). A clearer specification of the constructs of the JNCTL and the availability may potentially bring about positive backwash effects on teaching listening at high schools but also on English learning for the students, because they can inform English teachers and high school students better about the test.

3.2.4.2 Research Question 2; Real-life parameters vs. the JNCTL

The second research question concerned the extent to which the current form of the JNCTL achieves optimum validity in terms of contextual parameters and cognitive processing. The answer was that the current JNCTL does not since several discrepancies were found between real life listening features and the JNCTL. These include providing the input twice, a lack of hesitations, a lack of overlapping turns, a lack of multi-participant discussions, a lack of variety in the English accents used, a lack of L2 speakers, a lack of non-linear texts, and a lack of inference questions. Each discrepancy is discussed in Section 3.2.5. We should also note that the JNCTL achieves validity to the extent which it involves the items calling for a wider range of cognitive processing.

3.2.4.3 Research Question 3; the JNCTL vs. the CS

The third research question concerned the extent to which the current JNCTL reflects the CS. The answer to this question was that it does not fully reflect the CS. The JNCTL reflects the CS in terms of vocabulary and the ability to grasp the main point of the discourse, but it lacks a variety of language use situations specified in the CS. This result is consistent with that of RQ 2 that the JNCTL is lacking multi-participant discussions.

In conclusion, the results of Preliminary study 1 elicited eight major discrepancies as shown in Table 3.24 between real life listening features and the JNCTL, and the JNCTL and the CS.

Table 3.24 Discrepancies identified through Preliminary study 1

Parameter		Discrepancy	
Input	Format	channel	
		speaker (lack of multi-participant discussions)	✓
		text length	
		#of opportunities to listen to the input	✓
	Phonology	hesitations	✓
		sandhi-variations	
		speech rate	
		accent	✓
		L2 speaker	✓
	Syntactic	complexity	
		deviation	
	Vocabulary	frequency level	
		density	
		diversity	
		culturally specific word	
	Discourse	rhetorical type	
		discourse marker	
		propositional (text) linearity	✓
Pragmatic	overlapping	✓	
	turn taking		
Expected response	Question	mode/language/provision	
		scope	
	Response	type/mode	
		Time constraints	
Relationship between Input and Expected response	Topical knowledge		
		inference question	✓
	Text-item interaction	position of relevant information	
Required processing level			

They include:

- a lack of multi-participant discussions (RQ 2 and RQ 3)
- providing the input text twice (RQ 2)
- a lack of hesitations (RQ 2)
- a lack of variety in the English accents used (RQ 2)
- a lack of L2 speakers (RQ 2)
- a lack of non-linear texts (RQ 2)
- a lack of overlapping turns (RQ 2)
- a lack of inference questions (RQ 2)

Each discrepancy is discussed in the next section.

3.2.5 Discussions

A lack of multi-participant discussions

The result, that the current JNCTL only employs dialogues and monologues, shows that the JNCTL does not achieve optimum validity, and that it does not fully reflect the CS. The JNCTL may be advised to include multi-participant discussions because the inclusion of multi-participant discussions would make the JNCTL more reflective of contextual parameters (RQ 2) and of language use situations operationalised in the CS (RQ 3).

In the meantime, it may be true that real life listening does not necessarily include multi-participant discussions. Furthermore, the inclusion of the multi-participant discussions will increase the level of cognitive load for the test takers (Brown, 1995). Besides, a potential problem with the multi-participant discussions may involve the difficulty for test takers to distinguish one speaker from another. Those things considered, further discussions are made in Section 6.3.3.

Providing the input text twice

If a listening test is intended to reflect most language use situations, it is assumed that there should be just one opportunity to listen to the input. Given that the *Centre Test*, however, is a high-stakes test in Japan, this parameter cannot help being compromised to the extent which NCUEE is able to be responsible for the unexpected disturbances such as even small noises which might disturb test takers' listening. In addition, double play of the input texts can invite three more advantages. First, it can compensate for the absence of those paralinguistic features or contextual cues which are available in face-to-face interactions. Second, double play can compensate for a lack of interactivity to confirm or clarify what the speaker says. Third, it can alleviate the tension in the test takers who are afraid of missing a point in the high-stakes test (Field, 2012). Thus, the current practice of providing the input twice may be justified.

A lack of hesitations

The result, that few hesitations occurred in the forms of the JNCTL, suggests that the JNCTL does not achieve optimum validity in relation to contextual parameters and cognitive processing because input texts devoid of hesitations under-represent the construct of listening comprehension. Since NCUEE claims to place priority on interactional discourse, which is characterised by spontaneous features such as hesitations (NCUEE, 2006, 2007, 2008, 2009a, 2010), it may consider including more hesitations into the dialogues.

The current practice, a lack of hesitations, seems to be based on an assumption that they would disturb the listening comprehension for L2 listeners. No conclusive evidence, however, has been obtained yet,

leaving behind conflicting evidence (see Section 2.3.2; Voss, 1979; Blau, 1991; Griffiths, 1991). Hesitations such as repetitions or unfilled pauses might facilitate listening comprehension by providing more time for the listeners to decode and process the input text. Thus, there is no reason why the JNCTL should *hesitate* to include hesitations. The inclusion will surely make the input text more authentic, and hence raise the contextual validity.

A lack of variety in the English accents used

The introduction of more varieties of accents or English as an International language (EIL) into the JNCTL may accord with the objective of English education declared in the CS, which is to enhance students' practical communication ability in English or to deepen the understanding of language and culture (MEXT, 2007). McKay (2002) argues that EIL assumes that the learning of English is for communication with people from different backgrounds whereas EFL education has been designed to offer linguistic knowledge with the learners as accurately as possible. Now that English is acknowledged as an international language in the world (see for example, Kachru, 1992), Kim's (2006) contention that, 'test developers should reflect this practical issue in the context of language assessment' (p. 37) may appear sensible.

NCUEE, however, should be cautious about the inclusion of a variety of English accents since the CS claims to prioritise the use of 'standard' English as a model (MEXT, 2007), which can be considered North American accent. Furthermore, taking into account the proficiency level of the test takers who are still at lower level, priority may be put on exposing them first to a 'standard' accent of English. Besides, it should be examined to what extent a variety of English accents have been introduced

into the textbooks or other teaching materials in current use before the decision is made about the parameter.

Another issue concerns the choice of a variety(ies). What accent(s) should be adopted if multiple accents are employed? British, Australian, Canadian accent, others, or EIL? The choice should be explained in terms of the construct of the JNCTL. Thus, the decision being made to the variety in the English accents used should wait for the stakeholders' views, which are to emerge in Preliminary study 2.

A lack of L2 speakers

The result, that no L2 speakers are used in the current JNCTL, shows that the JNCTL prioritises the use of native speakers of English.

The inclusion of L2 speakers into the JNCTL may potentially bring about a backwash effect to language learning and teaching where English teachers and high school students in Japan seem to assume that the English they are supposed to speak should be like native speakers'. The inclusion of L2 speakers into the JNCTL may free them from the biased assumption and encourage them to use English more in their classrooms. This is in accordance with the revised CS to be in effect in 2013, where English classes are supposed to be conducted in the target language, that is English (MEXT, 2010), and thereby aims to enhance students' practical communication ability in English.

The CS, however, advocates the use of 'standard' English as a model (MEXT, 2007). Thus, the decision being made to this parameter should wait for the stakeholders' views, which are to emerge in Preliminary study 2.

A lack of non-linear texts

The result, that the JNCTL are lacking non-linear texts, shows that the JNCTL does not achieve optimum validity in relation to contextual parameters and cognitive processing. The JNCTL should employ more non-linear texts because listening in the real world contains as many non-linear texts as linear texts, and so comprehension of the non-linear texts is considered as one of the important constructs of listening comprehension in the real world and of the JNCTL.

Another reason is that the items with non-linear texts discriminate better between successful listeners and less successful ones than items with linear-texts (see Section 2.3.2.5, Tsui & Fullilove, 1998). Admittedly, the JNCTL is intended to discriminate between candidates according to the listening proficiency level as long as it is used for gate-keeping purposes for the universities or colleges (see Chapter 1). In sum, the JNCTL should involve more unpredictable texts to achieve its greater validity without heavily relying on predictable texts. This also would discriminate the test takers better, which would also serve to help the JNCTL achieve the purpose.

A lack of overlapping turns

The result, that no overlapping turns were found in the JNCTL, may suggest that the JNCTL does not achieve optimum validity because overlapping turns is one of the features which characterise interactional discourse. It is the case that overlapping turns occur when a turn taking does not take place successfully between the interlocutors (Sacks, Schegloff, & Jefferson, 1974) or when a strong solidarity between them is indicated in some cultures (Wierzbicka, 2003).

It is assumed, however, that overlapping turns will pose a considerable

level of cognitive load on the test takers (Brown, 1995). Furthermore, the understanding of the input texts including overlapping turns seems beyond the 'core knowledge', which the JNCTL is claimed to measure (NCUEE, 2007). For these two reasons, it seems sensible that the JNCTL does not include overlapping turns, even if it helps the JNCTL achieve greater validity.

A lack of inference questions

The result, that few inference questions were found across the three forms, shows that the JNCTL does not achieve optimum validity, primarily in terms of cognitive processing. Speakers do not necessarily convey their messages explicitly but implicitly (Widdowson, 1978). It is more often the case with everyday communication the JNCTL claims to place priority on (NCUEE 2006, 2007, 2008, 2009a, 2010). For this reason, the JNCTL should include more inference questions and thereby achieve greater validity in relation to contextual parameters and cognitive processing. The more inference questions the JNCTL includes, the higher and wider level of processing the test takers are required to be engaged in (see Figure 2.1).

The next Section reports on which one among the eight discrepancies is perceived to be the most important parameter by the stakeholders to make the JNCTL achieve higher validity.

3.3 Preliminary study 2: Prioritisation of a key parameter

3.3.1 Introduction

Preliminary study 2 was intended to answer RQ 4: that is to identify among the discrepancies found through RQ 1, RQ 2 and RQ 3 those parameters that the stakeholders perceived to be the most important to the validity of the JNCTL. This Section first describes the methods; a) focus group interviews and b) a large-scale questionnaire. Then, general impressions by the interviewees of the JNCTL and of the discrepancies are reported by revealing their verbatim verbal protocols in the group interviews. Summaries of questionnaire responses are presented alongside the open-ended free comments made by respondents. One key parameter is identified for further investigation.

3.3.2 Method

3.3.2.1 Group Interviews

Purpose

The purpose of the focus group interviews was twofold; first to elicit how stakeholders perceived the discrepancies identified through Preliminary study 1 (and possible changes to the JNCTL that could be made as a result). The other was to elicit more general views of the JNCTL from the perspectives of the different stakeholders because this was also considered useful in thinking about how the JNCTL might be revised in the future. Three stakeholder groups were targeted: college lecturers involved in English language instruction, high school English teachers, and high school students preparing for the *Centre Test*. These groups were chosen since they were considered as being most affected by the JNCTL.

Setting

Group interviews were held between October 2008 and January 2009. Each group comprised three participants in addition to the researcher as an interviewer. The interviewer (the researcher) avoided giving any impression of approval or disapproval based on his opinions and feelings about the answers he received. This was intended not to cause interviewer bias where the interviewees try to give the answers they think the interviewer wants to hear and would approve of (Browne, 2011). All interviewees voluntarily participated in these group sessions.

All the interviewees had been asked to take the 2008 form of the JNCTL at home and to mark their scores by referring to an attached sheet of correct answers before the interviews. This was intended to ensure that they would all be familiar with the JNCTL at the time of the interview.

Interviewees

Table 3.25 supplies background information about the high school English teachers and college lecturers. All of the high school English teachers were familiar with one another. The researcher assumed that this acquaintanceship could make the participants feel more relaxed and so would encourage them to share more candid comments about the JNCTL than would have been the case if they were talking to strangers. This, it was felt, would enhance the validity of the group session. The three teachers each work for different high schools where there are candidates for the *Centre Test* every year.

The high school teachers all had Masters degree in the field of Applied Linguistics or Teaching English as a Foreign Language (TEFL). This could be considered an advantage as their expertise in TEFL could make the group interview more in-depth and productive. On the other hand, they

might not be representative of the wider population of high school English teachers in Japan, where only 14.1 per cent of high school teachers are estimated to have Masters degrees (MEXT, 2011b).

The three college lecturers each came from different colleges. They all teach English at their colleges as a full-time (Lecturer A) or a part-time lecturer (Lecturer B and C). Compared with the high school teachers, they had relatively short experience of teaching as shown in Table 3.25.

The three high school students (henceforth, Miho, Asami, and Yoko as pseudonyms) were all third-year students preparing for the *Centre Test*. They came from the same high school. Since they were familiar with one another, this was expected to make them feel relaxed and it was considered that this would make their group interview more productive.

Table 3.25 Backgrounds of the English teachers

	High school teachers			College lecturers		
	A	B	C	A	B	C
Sex	m	f	m	m	f	m
Teaching experience (years)	25	20	2	6	2	6
Experience of developing high-stakes English tests	*No information			✓		
Experience of developing a listening test	*No information				✓	✓

Note. m=male f=female, *The Interviewer did not ask for the information.

Procedures

The interviews were conducted in Japanese and were recorded with permission from all the participants, who agreed to the recording with their signature (see Appendix 5). The researcher acted as the interviewer

across the group interviews. The participants were informed that no interviewees would be identified in any subsequent report.

In order to elicit as many candid comments as possible from the participants, the interview had been intended to work as 'focus group interview', where the interviewer plays a limited role and the interviewees freely discuss issues, but in fact, the interviews had more of the character of 'group interviews' as the interviewer had to take a leading role, and the interviewees took turns to answer the interviewer's questions. This happened perhaps because of a social norm unique to Japanese society where it is considered rude to cut in or overlap with others' talks, especially when it comes to formal talking like this.

All group interviews followed common procedures: phase 1 aimed to elicit how stakeholders perceived the JNCTL; phase 2 introduced the aim of the research and the process of identifying the discrepancies between the CS and the JNCTL, and between them and real life listening features in terms of contextual parameters and cognitive processing (this phase was omitted for the interview with high school students because the researcher was afraid that it would be difficult for the high school students to understand the theoretical background of this research and this would keep their attention away from this interview). Phase 3 attempted to elicit what stakeholders thought of the discrepancies identified. The duration of the interviews was approximately one hour for the students, one and a half hours for the high school English teachers, and two hours for the college lecturers.

Questions

The development of the interview questions was based on the results of Preliminary study 1. That is, the interview questions came primarily from discrepancies (see Table 3.24) between the CS and the JNCTL, and

between them and real life listening features in terms of contextual parameters and cognitive processing⁹.

The following were pre-set questions which were presented across the group interviews.

Preset questions

Phase 1: How do you perceive the JNCTL?

Phase 2: (Interviewer's explanation about the aim of this research and about the method of identifying the discrepancies)

Phase 3: What do you think of providing input text twice/a lack of hesitations/a lack of multi-participant discussions/a lack of varieties in the English accents used /a lack of L2 speakers/ a lack of overlapping turns/a lack of inference questions?

Modifications and elaborations of the questions for the students

What differentiated the group interview for the high school students was the interviewer's modifications to the questions to make them appear less technical so that the students could raise their awareness of their own feelings and thoughts and express their views more easily and clearly. For example, when the topic was about a variety of English accents, the interviewer referred to Malaysia where the interviewees had been for their school trip. The interviewer asked, 'Did you find any difference between the English you usually listen to in the classroom and the English you heard in Malaysia?'

3.3.2.2 A large-scale questionnaire

The results of RQ 1 through RQ 3 and the small group interview sessions

⁹ Non-linear texts were not identified a discrepancy before the group interviews.

informed the development of the questionnaire which followed. The purpose of the questionnaire was to elicit the stakeholders' views of the most important parameters that might help to make the JNCTL more reflective of listening in the real world.

The lead question was 'What do you think about the following changes to the JNCTL to make it better¹⁰? Choose an option which best fits your opinions.' The possible changes referred to the eight discrepancies identified through the initial investigation of RQ 1, RQ 2, and RQ 3 (see Table 3.24), and to two further parameters, which were speech rate and sandhi-variation. Even though speech rate and sandhi-variation did not emerge as discrepancies through Preliminary study 1 (perhaps due to the absence of an established baseline -Section 3.3.3.2), they were added because the literature (for example, Gimson & Cruttenden, 1994) suggests that they are important to the construct of listening comprehension. Thus, possible changes to the JNCTL, which were included as the questionnaire items, were summarised in Table 3.26.

¹⁰ 'Better' was used in the questionnaire instead of 'more valid' as the researcher was afraid that the respondents would have little idea of what 'valid' means.

Table 3.26 Possible changes to the JNCTL

Possible changes (discrepancies)
providing the input text twice
a lack of natural speech rate
a lack of varieties in the English accents used
a lack of L2 speakers
a lack of inference questions
a lack of multi-participant discussions
a lack of hesitations
a lack of sandhi-variations
a lack of overlapping turns
a lack of non-linear texts

Possible changes to the JNCTL summarised in Table 3.26 informed the development of ten questionnaire statements. The statements were:

- 1 At present the JNCTL allows test takers to listen to each text twice. Each text should be played only once.
- 2 At present the JNCTL are delivered at a rate that is a bit slower than natural speed. The speech rate should be natural speed.
- 3 At present all speakers included in the JNCTL speak with a standard North American accent. Other accents (for example, British or Australian) should be used in the JNCTL texts as well as the North American.
- 4 At present all speakers included in the JNCTL are native speakers of English. Non-native speakers of English should be included in the JNCTL texts.

5 At present the JNCTL hardly ask for interpretations on listeners based on the factual information in the text. The JNCTL should include more those items which call for the test takers' interpretation.

6 At present the JNCTL listening texts involve one or two people. The JNCTL should include texts with more than two speakers.

7 In natural speech speakers of English use fillers or hesitations, which are unique to spontaneous speech. They are hardly covered in the JNCTL. More of these features should be included in the JNCTL texts.

8 In natural speech speakers of English use assimilation (for example, *did you, would you*) or elision (ex. *tell her, next station, camera*). More of these features should be included in the JNCTL texts.

9 At present the JNCTL does not involve overlappings where two speakers speak at the same time. Overlappings should be included in the JNCTL texts.

10 At present the JNCTL texts are mostly straightforward and linear. More of non-linear texts should be included.

The respondents of the questionnaire were asked to read each statement and rate the degree of agreement by choosing one of the options which best fits their opinions. The options were constructed on a five-point scale using 'agree' to 'disagree'. Also, a space was provided where the respondents were allowed to freely write their views to make the JNCTL 'better'.

The questionnaire was piloted on a small group of stakeholders, twenty-seven third year high school students and thirteen high school English teachers from January to March, 2010. Based on their responses and feedback, the wording of five questionnaire items was revised to make the items easier to understand. For example, one item was initially 'At present the JNCTL hardly asks for understanding of non-linear or unpredictable texts. More of non-linear texts should be included in the JNCTL texts.' But this was modified to 'At present the JNCTL texts are mostly straightforward and linear. More non-straightforward texts should be included in the JNCTL texts.'

The questionnaire (see Appendix 6A, 6B) was administered to 391 high school students from four different high schools and to 110 English teachers from seventy-five different high schools from July to November, 2010. The following sections report the results and analyses for RQ 4.

3.3.3 Results and analyses

3.3.3.1 Group Interviews

Introduction

Based on the recordings of the group interviews, the researcher attempted to provide a reasonable summary of each interview. The summary was shown to the three high school teachers and college lecturers to make sure that the summary reported their views in a way that they agreed was true and accurate. Unfortunately, this could not be done for the high school students as they had graduated from the high school and proved difficult to track down. The quotes in the following report were translated by the researcher.

General impression of the JNCTL

The three high school English teachers were all positive about the introduction of the listening comprehension component into the *Centre Test* as it is expected that this will contribute to enhancing the listening ability of high school students in Japan. They assumed that the introduction encouraged high school English teachers to incorporate teaching listening into their classes, while it encouraged high school students to listen to spoken English outside their classrooms. Apparently, teaching listening has been better appreciated in their actual classrooms since the listening component was introduced into the *Centre Test* (high school teacher A).

In addition, high school teacher A was positive about the JNCTL as 'It uses a variety of different situations high school students or the graduates are likely to encounter in their everyday lives.' Teacher B commented positively saying that 'The JNCTL appears to use quite natural spoken language, as the texts include many pauses.'

In contrast, the high school students, who were going to take the JNCTL one month after the interview, tended to be more negative about the current JNCTL. Miho argued;

The JNCTL does not contribute to improving students' understanding of spoken English as we do not listen to English. You know, the JNCTL only covers 50 scores out of 250 of the English test, and so it is often the case that we end up checking some useful expressions listed in vocabulary books [which usually accompany CDs], instead of listening to English.

Asami and Yoko agreed with Miho, saying that 'We do not have time to devote ourselves to the listening section as we should use our time for

vocabulary, grammar and reading comprehension', which covers a larger part (that is, four-fifths) of the *Centre Test*.

The college lecturers, on the other hand, seemed positive about the introduction of the listening comprehension component into the *Centre Test*. Lecturer A said; 'This movement is in line with more emphasis on enhancing practical communication ability in English in formal English education in Japan. It is expected that speaking and writing tests should follow the listening comprehension test soon.'

One reservation, however, was proposed. 'What is important', lecturer A claimed, 'is the consistency between what the high school test takers are actually taught in high schools and what the test is meant to measure.' That is, the optimum validity of the JNCTL should be balanced by the need for consistency between teaching and testing.

Discrepancies identified through Preliminary study 1

This Section reports what the interviewees said about each discrepancy identified through Preliminary study 1.

Providing the input text twice

Different interviewees had different views about providing the input text twice. One view was 'listening twice is reasonable since the *Centre Test* is an extraordinarily high stakes test' (lecturer B and high school teacher B). Another affirmative view of providing the input text twice was that since real life communication allows for clarification or repetition, playing a text twice is natural and therefore acceptable' (high school teacher B and lecturer A). Some interviewees, by contrast, were against playing a text twice. High school teacher C and Yoko argued that 'Playing a text twice does not reflect real life listening but invites test-taking strategies.' High school teacher C

argued;

Since automaticity is crucial to cognitive processing in listening comprehension, playing a text twice may prevent the listeners from engaging in the necessary processes, and in the long run this may have negative effects on improving their listening ability.

One solution to the problem, which was proposed by high school teacher B, was that different texts should employ different numbers of listenings. For example, discourse types which do not allow for repetitions of the message, such as news or announcements, should be provided only once, reflecting real life communication. On the other hand, interactional discourse types such as dialogues or discussions, which allow for repetitions of the message, should be provided twice. Another solution proposed by the teacher was to include a clarification turn in the dialogue itself, which might provide a more natural alternative means of repeating relevant utterances without repeating the recording.

A lack of hesitations and overlapping turns

Overall, the interviewees were against involving more hesitations and overlapping turns in the JNCTL. Miho argued that 'the JNCTL should reflect the current situation of English education in high schools in Japan,' where 'Juken-Eigo' (see Section 1.2) is prioritised rather than English for communication. If the JNCTL were to use authentic texts involving hesitations or overlapping turns, it would not match with the current English teaching in the classroom.

Also, test fairness should be considered. Lecturer A argued;

Not all the test takers in Japan can necessarily get access to authentic texts containing spontaneous features such as hesitations or overlapping turns. Since the JNCTL should be fair, priority should be placed on accessibility to text books rather than text authenticity itself.

Miho agreed with lecturer A, saying 'Authenticity can be or should be compromised for test fairness.' High school teacher A agreed, saying that 'Those texts which involve spontaneous features might not be suitable for the students until they reach a certain proficiency level where they can cope with them. Gradual incorporation may be possible, though.'

High school teacher B and lecturer B, however, stressed that 'Since real life speech involves many spontaneous features, the JNCTL should endeavour to make itself as close to real life as possible and thereby positive washback can be expected.' Only lecturer B referred to issues of validity, suggesting that 'It depends on what the *Centre Test* is meant to measure whether the JNCTL should use more authentic texts involving hesitations or not.'

A lack of multi-participant discussions

High school teacher B said, 'I do not think that the number of speakers in the current JNCTL is good enough as the text book for *Oral Communication I*, [one of key subjects in the high school curriculum], employs many multi-participant discussions.' Asami, lecturer B, and lecturer C agreed, 'Multi-participant discussions sound interesting, don't they?' Asami added, 'I would be happy if Part 4B could be replaced with a multi-participant discussion 'cause Part 4B, employing a longer monologue, is one of the most difficult sections for me, and the replacement would

make the JNCTL easier.’

One of the problems to emerge, however, is ‘how the test takers can distinguish one speaker from another’ (Asami and high school teacher A) in multi-participant discussions’. One possible solution, proposed by lecturer A, was to provide the test takers with visual information in addition to textual information so that they might be able to identify the speakers. Another solution was not to develop items calling for the test takers to identify speakers in order to arrive at correct answers. Instead, the correct answer should be based on the main point or the conclusion of the discourse.

A lack of variety in the English accents used and L2 speakers

Most of the interviewees agreed with the exclusive use of North American accents in the JNCTL, because of test fairness and the proficiency level of the test takers. Test fairness is concerned with the accessibility of high school test takers to the varieties of English. We can assume that North American English accents are the most accessible in Japan now as it is the dominant accent there, while other varieties may be more difficult to access in some areas in Japan. High school teacher A, Miho, and Asami insisted that ‘Accessibility of the varieties of English used and consistency with high school text books should be prioritised.’ Five interviewees out of nine stated that priority should be given to understanding North American English accents first, as this can be conceived as a ‘standard model’ of English in Japan and so is in the widest use in the available English textbooks. Then, other varieties of English can be incorporated into teaching or testing to meet particular needs or the interests of particular students.

Miho argued;

I don't want varieties to be involved in the test as different varieties represent not only different phonological realisations but also lexical differences, which is demanding. Can you imagine if a Japanese proficiency test involves *Kansai-ben* [a dialect or accent in use in western parts of Japan] or other accents, what do you think? It's true that TOEIC adopts varieties of English, but it is not as high-stakes as the *Centre Test*. You can take TOEIC as many times as you like whereas you cannot take the JNCTL more than once a year.

It was only high school teacher B who partly agreed with the introduction of non-native accents of English such as Japanese or Asian. 'This', she argues, 'will help Japanese learners of EFL be aware of intelligibility and acceptability of English by non-native speakers of English in the era of World Englishes.'

A lack of inference questions

Different interviewees had different views on inference questions. Some interviewees were against the inclusion of inference questions. High school teacher B argued that 'Inference questions presume particular cultural or societal conventional norms associated with particular speech communities, which are irrelevant to the construct of listening comprehension.' She was concerned that different cultures may have different communicative values, so that an item may have different correct answers. Lecturer A called into question the validity of inference questions, arguing that 'Whether an item is an inference question or not is relative and so the definition is too ambiguous and elusive.' Yoko claims

that taking into account the proficiency level of test takers, priority should be placed on whether they can successfully build up textual meaning or not, rather than reaching pragmatic interpretations of the texts.

Miho, lecturer B, and lecturer C, in contrast, were positive about the inclusion of inference questions. If the JNCTL is meant to measure listening ability in the real world, then the JNCTL should involve those items which focus on the implicit message conveyed by the speaker. They argued 'It is often the case that our message is indirect or implicit.'

Table 3.27 summarises comments made by the interviewees in relation to each discrepancy. Although the small group interviews themselves did not contribute to the immediate elicitation of important parameters and also involved some limitations mentioned earlier (see Section 3.3.2.1), it elicited different views of the discrepancies and the possible changes to be made to the current JNCTL from different stakeholders.

**Table 3.27 Summary of the interviewees' comments
in relation to the discrepancies**

Discrepancies	Supporting current practice	Opposing current practice
Providing the input text twice	High-stakes test	Not reflecting real life listening Preventing automaticity Inviting test taking strategies
A lack of hesitations	The proficiency level of the test takers	Not reflecting real life listening
A lack of overlappings	Consistency with the textbooks	
A lack of multi-participant discussions	Heavy cognitive load Difficulty to differentiate one speaker from another	Not reflecting real life listening Consistency with textbooks
A lack of varieties in the English accents used	The proficiency level of the test takers	Raising awareness of varieties of English accents
A lack of L2 speakers		Not reflecting real life listening
A lack of inference questions	Inference questions may presume particular cultural norms.	Message is often indirect or implicit

Note. Lack of non-linear texts was not discussed in this group interview session as it was not identified as a discrepancy between real life listening features and the JNCTL before the group interview.

Limitations of the group interviews

Clearly, there are quite a few limitations to the group interviews: First, the admittedly small sample of three cannot reflect the views of all students, teachers, or lecturers in Japan, and this indicates the need for caution about any claims that are made. Priority was placed on feasibility and

manageability of the interview and in-depth elicitation of the stakeholders' perceptions of the JNCTL rather than the representativeness of the sample. Second, each 'focus group interview' ended up merely as a group interview, where the interviewer took a leading role. Had the groups been able to discuss issues more freely among themselves, this could have provided more insights into the JNCTL and the discrepancies identified, and different or wider views of the JNCTL might have been elicited. Third, non-linearity of text was not addressed as a topic in the group interviews because it had not been identified by the researcher as a discrepancy before the group interviews were undertaken.

Findings

Two major findings emerged from the group interviews:

- 1) The English teachers were more positive than the students about possible changes which can make the JNCTL 'better'.

- 2) English teachers and students tended to object to the possible changes to the current JNCTL because of the wide range of proficiency levels among test takers and the lack of consistency that would emerge between teaching and testing.

The next Section describes whether or to what extent the views elicited through group interviews were reflective of the wider population of stakeholders by reporting the results of a large-scale questionnaire.

3.3.3.2 A large-scale questionnaire

The key result from this questionnaire was that overall, both high school English teachers and students were not positive about possible changes (see Table. 3.26) which can be made to the current JNCTL. Table 3.28

and Figure 3.7 show the number and/or the ratio of those respondents who chose either 'agree' or 'agree to some extent' with the possible changes described in the earlier section. With the exception of sandhi-variation and variety of accents for English teachers, none of the items obtained more than 50 per cent of agreement from the respondents, which seems a cut-point. This seems more often the case with the students.

Table 3.28 Perceived importance of the discrepancies by stakeholders (%)

Possible changes to the JNCTL	Teachers (N=110)	Students (N=391)
Providing the input text once	35 (31.8 %)	18 (4.6 %)
More natural speed	48 (44.0 %)	64 (16.5 %)
Variety of accents	59 (54.1 %)	54 (13.9 %)
Inclusion of L2 speakers	36 (33.0 %)	46 (11.8 %)
More inference questions	53 (48.2 %)	77 (19.9 %)
Inclusion of discussions	52 (47.3 %)	85 (22.1 %)
More hesitations	47 (42.7 %)	53 (13.7 %)
More sandhi-variations	85 (78.0 %)	110 (28.4 %)
Inclusion of overlapping turns	26 (23.6 %)	12 (3.1 %)
More non-linear texts	45(41.0 %)	46 (12.0 %)

Note. The figure shows the number and/or the ratio of those respondents who chose either 'agree' or 'agree to some extent' with the possible changes.

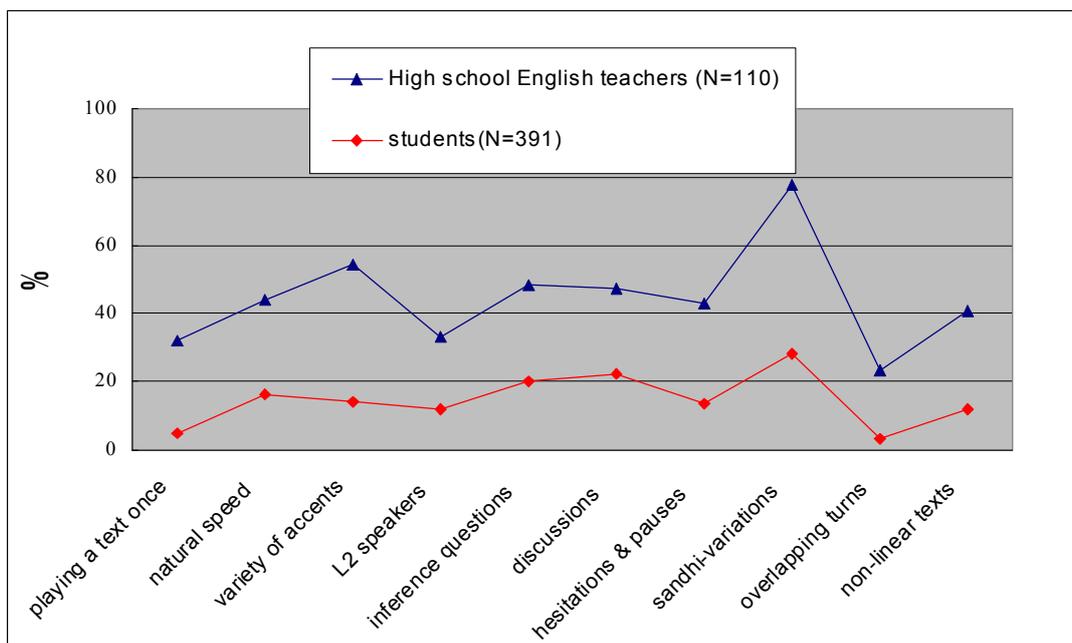


Fig. 3.7 Perceived importance by stakeholders

Note. The figure shows the ratio of those respondents who chose

either 'agree' or 'agree to some extent' with the possible changes.

The reasons for this can be seen through the students' comments in the space provided in the questionnaire.

One student argues, saying 'The *Centre Test* should be kept as it is now since it should take into account the wide range of listening proficiency among the test takers.' The JNCTL seems to the third year high school students no more than an additional task. One student expressed disagreement with the introduction of the JNCTL itself. 'A listening comprehension test does not make any sense as we cannot learn to speak English even if we can get higher scores on the test.' The student seems to sense that getting higher scores on the JNCTL does not necessarily mean that he or she can demonstrate his or her higher practical communication ability in the real world. He may assume that he can get higher scores by utilising test-taking strategies as the JNCTL does not

reflect listening in the real world. If justified, this would be a very damaging criticism. If scores on the JNCTL do not indicate practical ability in English (the key aim of the introduction of the JNCTL), then the introduction of the JNCTL would have to be considered a failure.

Another student's comment coincides with the widely-accepted perception (see Section 1.2) that there are two kinds of English in Japan—'*Juken Eigo*' (entrance examination English) and English for communication.

Occupied by this assumption, the student comments: 'I do not want English for communicative purpose to mix with 'Juken-Eigo'. Another student even makes complaints like this: 'It is not fair because returnees are favoured by including a listening comprehension component.' These comments indicate how strongly the students disagree with even the introduction of the JNCTL or with the possible amendments to make the current JNCTL 'better'.

Having said this, only one parameter was supported by 78 per cent of the English teachers and 28.4 per cent of the students. That was sandhi-variation, suggesting that sandhi-variation is a key parameter, and more inclusion of sandhi-variations would make the JNCTL achieve greater validity in the direction that would be accepted by stakeholders.

The priority on the sandhi-variation over the other parameters can partly be attributable to the pervasiveness of sandhi-variation. Sandhi-variation is observed so often in real life listening that they might have sensed that the feature should be included in the JNCTL to the extent which reflects realistic level. In fact, the examples used in the questionnaire (see Appendix 6A, 6B) such as *did you* [didʒu:] or *tell her* [telə] must have been so popular and typical to the stakeholders that they might have perceived the feature to be important.

Another reason might be the stakeholders' understandings that

sandhi-variation discriminates a spoken text from a written text and so it should be included in a listening comprehension test of the JNCTL, which should be composed of spoken texts.

Thus, the pervasiveness of sandhi-variation among the stakeholders, and their understandings of the importance of sandhi-variation in listening comprehension tests might have elicited sandhi-variation as the key parameter.

Other findings include:

- 1) shared perceptions between the English teachers and the students of the relative importance of the items. For example;
 - Overlapping turns is perceived to be the least important.
 - Playing a text once is perceived to be relatively less important than the other parameters.

- 2) shared perceptions between the English teachers and the students of the importance of the consistency between teaching and testing.

A student wrote, 'Changes to the current JNCTL would not be acceptable unless the current English teaching in the classroom stops serving only to give students grammatical knowledge.' This comment implies that the changes to make the JNCTL closer to real life listening would only be acceptable if the teaching changed. One teacher also stressed the relevance between teaching and testing, saying 'If the *Centre Test* is intended to measure the core knowledge represented in the textbooks authorised by MEXT, then it should focus on that core knowledge instead of on the listening in the real world.' We should note here that this coincides with the findings of group interviews (see Section 3.3.3.1). Thus, it seems

that stakeholders believe it is more necessary to make teaching and testing consistent than to make the JNCTL closer to listening in the real world.

In sum, on one hand, both English teachers and high school students do not favour changes to the current JNCTL. In particular, the students strongly hope to keep the JNCTL as it is as students are most likely to be affected by the JNCTL. Students' negative responses (free comments on open-ended space) about changes to the current JNCTL may suggest that listening is still a neglected area at educational practices in Japan, even six years after the introduction of the JNCTL. This does not necessarily coincide with the optimistic views provided by *Zen-eiren* (2007) or Kougo (2006) described in Chapter 1, but coincides with Watanabe (1996, 2004), which showed lack of linear washback between testing and teaching.

On the other hand, sandhi-variation was identified as the only parameter widely accepted by both teachers and students that might make the JNCTL achieve greater validity. We should remind ourselves at this point that sandhi-variation emerged as a key parameter originally from literature review not from the review of the current JNCTL: The results of RQ 2 have not clarified yet that sandhi-variation is one of the discrepancies between real life listening features and the JNCTL (see Table 3.24) due to the absence of an established baseline. Therefore, we first need to demonstrate in Preliminary study 3 which follows that sandhi-variation is a discrepancy, and then we can or should further explore the effect of the sandhi-variation on the test performance and the level of cognitive load imposed on the test takers. Of particular relevance to the sandhi-variation is that the occurrence of sandhi-variation is implicated in the speech rate. The more rapid the delivery is, the greater the tendency to reduce and

obscure unaccented words (Gimson & Cruttenden 1994) is. The results of RQ 2 have not clarified yet either that speech rate is one of the discrepancies between real life listening features and the JNCTL due to the absence of an established baseline about valid speech rate. Preliminary study 3, which follows, therefore, addresses both sandhi-variation and speech rate in the current JNCTL to relate Preliminary studies 1 and 2 to Main study.

3.4 Preliminary study 3: Confirmation of a key parameter

3.4.1 Introduction

Preliminary studies 1 and 2 identified several discrepancies between the JNCTL and real life listening features. Responses from stakeholders suggested that one parameter – namely sandhi-variation – should be prioritised for investigation as a means of improving the validity of the JNCTL in a direction that would be accepted by stakeholders.

Preliminary study 3 is intended, before Main study investigation, to explore whether or to what extent the current JNCTL reflects a realistic level of sandhi-variation. This study addresses the question in greater depth than the Preliminary study 1, which did not reach any conclusion about the validity of the extent of sandhi-variation in the test. Rater judgements will be employed to answer the question.

This section is divided into three parts; methods, results, and findings, followed by the reformulation of the research questions set out earlier.

3.4.2 Method

Material

Four different excerpts from four different parts (see Table 3.7) for each of the forms of the JNCTL administered in 2008 and 2009 were selected and used as the material. Eight texts were used as the material to be rated. For each of the two forms, this comprised two short dialogues (one each from Part 1 and 2), one longer dialogue (from Part 3), and one longer monologue (from Part 4).

Rating sheet

For the purpose of this study, a rating sheet was developed. The rating sheet addresses eight features of sandhi-variation and a holistic judgement.

The features are assimilation, weak forms, elision, formulaic expressions, flap [r] sound, intrusion, elision of voiceless plosives, and speech rate.

Of these, assimilation, weak forms, elision, and formulaic expressions were derived from Henrichsen (1984, see Section 2.3.2.2), flap [r] and intrusion were from Gimson and Cruttenden (1994), Brown and Brown (2006), or Ito (2006), and elision of voiceless plosives was derived from the researcher's experience that Japanese high school learners of EFL, particularly those at lower levels of proficiency, may have difficulty in recognising voiceless plosives which are elided at the end of sentences. Speech rate is strongly implicated in the occurrence of sandhi-variation (Gimson & Cruttenden, 1994). A holistic judgment of 'naturalness' was included as an item since it was considered that the addition of a holistic rating could provide a more appropriate indication of the 'naturalness' of the sandhi-variation than an aggregate of judgements of each distinct feature of sandhi-variation.

The rating sheet employed a seven-point Likert scale, with seven indicating that the text 'very much' reflected naturalness and one indicating 'not at all'. Six (indicating 'fairly') on the seven-point Likert scale was taken as the cut point for adequate 'naturalness'. This was because as five indicated only 'to some extent', it was considered insufficient to reflect 'naturalness' while seven indicating 'very much' was considered too strict as a minimum level of 'adequacy'.

The rating sheet had been presented to rater B, an expert in English phonetics, prior to the coding to make sure that the sheet was valid for the purpose of judging sandhi-variation. Based on his comments, a few minor changes were made to the examples of each feature to make it clearer for the raters.

A limitation of the rating sheet should be pointed out. First, not all

important sandhi-variation features were included. Linking of [n], for example, with a vowel which follows such as *on it* or *in it*, or glottal stop as a variation of /t/ which usually occurs before a semi-vowel such as *It was* (Matsusaka, 1986), were not included in the item lists. This was because the linking was considered a form of assimilation while glottal stop was not as noticeable as other features. For this reason it was considered that this limitation would not affect the reliability of the method.

Procedures

Two experts, an expert in language testing (rater A) and an expert in English phonetics (rater B), were asked to be involved in the rating. Both of them, whose L1 was Japanese, were regularly exposed to 'natural' English in their everyday lives at home or on business. In January, 2011, they were asked to rate the degree of naturalness according to the rating sheet (see Appendix 3) attached with the CD recordings.

3.4.3 Results

The two raters agreed on the rating for; weak form, elision, intrusion, elision of voiceless plosive, and the holistic impression, where their rating was taken as final. Where the two raters diverged in their ratings, the researcher's rating was referred to. If the researcher's rating was identical with either rater A or B, then that rating was taken as final¹¹. This occurred for assimilation, flap [ɾ], and speech rate. Where no agreement was reached among the three raters, the average of the three ratings was used in the subsequent analysis. This was applied only to formulaic expressions. Table 3.29 shows the ratings of the three raters in relation to

¹¹ Agreement among the raters was considered to reflect reality of sandhi-variation features more than the average rate among the raters.

each feature of sandhi-variation and holistic impression.

It is important to note that the difference in the judgement between raters A and B was within one scale point for seven items among the nine as shown in Table 3.29, suggesting that this judgement was overall reliable. On the other hand, the difference for flap [r] was three scale points, and it was two scale points for speech rate. Rater A later stated that she felt that she might have been more lenient in the judgement of the JNCTL than she had been. The researcher made a final decision about the ratings for flap [r] and speech rate.

Table 3.29 Ratings of ‘naturalness’ of sandhi-variation in the JNCTL

	very much (7)	fairly (6)	to some extent (5)	uncer- tain (4)	not suffici- ently (3)	little (2)	not at all (1)
Assimilation		A	BR				
Weak form		AB	R				
Elision			AB	R			
Formulaic expressions			A	B	R		
Flap [r]		A			BR		
Intrusion				ABR			
Elision of voiceless plosive			ABR				
Speech rate			A		BR		
Holistic impression			ABR				

Note. The question was ‘To what extent does the current JNCTL reflect listening in the real world in relation to the sandhi-variation in connected speeches? Check the box, please.’

A and B: raters, R: the researcher. Bold indicates the final decision for each item.

Since six on the seven-point Likert scale was taken as the cut point for

adequate 'naturalness', the results show that the current JNCTL appears not to reflect a fully realistic level of sandhi-variation. Assimilation, elision, elision of voiceless plosives, and holistic impression were all judged to reflect real life listening only to some extent. Speech rate and flap [ɾ] were judged as clearly under-representing real life listening. Formulaic expressions and intrusion were rated as 'uncertain'. It was only weak forms that reached the cut point of six, indicating a realistic level.

3.4.4 Findings and pathway to Main study

Although weak forms were judged to be fairly realistically reflected on the current JNCTL, the results suggest that the current JNCTL does not more generally reflect a realistic level of sandhi-variation and therefore sandhi-variation is a discrepancy between real life listening features and the JNCTL. Additionally, the results suggest that speech rate is another discrepancy between real life listening features and the JNCTL.

These two findings not only justified the inclusion of sandhi-variation and speech rate in the questionnaire employed in Preliminary study 2 but also confirmed sandhi-variation as a key parameter. Accordingly, in Main study, the effect of sandhi-variation on the JNCTL is to be further explored. The two research questions set out earlier were reformulated as follows:

- RQ 5) How does the manipulation of sandhi-variation affect test performance?

- RQ 6) How does the manipulation of sandhi-variation affect the level of cognitive load imposed on the test takers?

The next chapters (Chapter 4 & 5) examine the effect of sandhi-variation on the validity of the JNCTL.

Chapter 4 Main study: Methodology

4.1 Introduction

This chapter sets out the methodologies used to address research questions 5 and 6 for Main study. The chapter is divided into the following sections:

1. overview of the research design,
2. materials development,
 - a. listening comprehension test
 - i. source
 - ii. recording
 - iii. evaluation
 - iv. piloting
 - b. questionnaire
3. participants,
4. procedures, and
5. analyses.

4.2 Overview of the research design

Table 4.1 Research instruments summary for Main study

RQ	Target	Questionnaire	Case study
5	test performance		✓
6	level of cognitive load	✓	✓

As Table 4.1 shows, the case study was used to answer both RQ 5 and RQ 6 while the questionnaire addressed only RQ 6. To carry out the case

study, a series of experiments was conducted involving the manipulation of sandhi-variation. Two different sets of test materials (henceforth, item set A and B) were developed, and each item set had two versions. In the first version of each item set, the texts involved realistic level of sandhi-variation (henceforth, sandhi plus version) while in the second, they involved minimal level of sandhi-variation (henceforth, sandhi minus version). That is, sandhi-variation was manipulated while keeping the other parameters as close as possible to the original. The items for these tests were obtained from administrations of the JNCTL within the past four years, from 2007 to 2010 (NCUEE, 2009b, 2010).

Pilot instruments were developed and modified before the administration of the experiment. Participants in the pilot study included seventy-five second year high school students intending to take the JNCTL the following year. The piloting helped inform the refinements to the instruments (see Section 4.3.1.4).

The experiment employed a cross-over design as presented in Table 4.2 to control for a possible test effect and order effect. The first administration was conducted in April, 2011, and it was two months after the first administration (that is, June, 2011) that the second administration was conducted. In addition, it employed a common-item and common-person design to treat different test forms as one set for analysis as presented in Figure 4.1 and thereby allowed for the use of Rasch analysis (Linacre, 2007).

Table 4.2 Research design

Group	April 2011	June 2011
1 (23)	A+	B-
2 (19)	A-	B+
3 (31)	B+	A-
4 (38)	B-	A+
5 (43)	B-	A+

Note. 'A+' indicates sandhi plus version of item set A while 'A-' indicates the sandhi minus version.

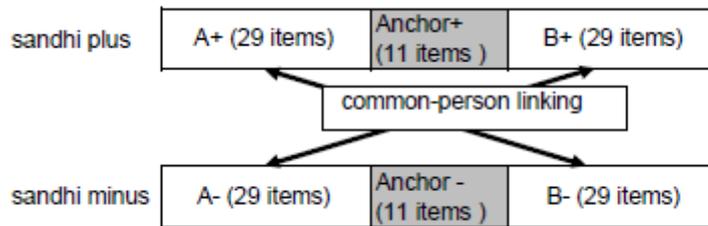


Fig. 4.1 Common-item/person linking

A total of 154 first and second year university students majoring in science and engineering participated in the experiment. These students came from five intact classes at a university in Japan. The researcher was concerned that the total number of participants across the groups could be less than 100, a minimum requirement for Rasch analysis (McNamara, 1996; Otomo, 1996), due to unexpected absences. Group 5 was intended to follow Group 4, which seemed most likely to reduce its participants.

All participants took a listening comprehension test and then responded to a brief questionnaire composed of five items all in the form of five-point

Likert scale.

4.3 Materials development

4.3.1 Listening comprehension test

4.3.1.1 Source

Two item sets (A and B) of the test were developed: Each item set had both a sandhi-variation version where the texts involved realistic level of sandhi-variations (sandhi plus version) and a non-sandhi-variation version, where the texts involved minimal sandhi-variation (sandhi minus version). The test source was the 2007, 2008, 2009, and 2010 forms of the JNCTL as supplied in Table 4.3. Because the purpose of this study was to investigate the effect of sandhi-variation on the validity of the JNCTL, the intention was to reflect the current JNCTL as closely as possible.

Table 4.3 Source and the number of items for each item set

Item set	# of items	Independent items	Anchor items
A	40	25 items from the 2009 form	
		4 items from the 2007 form	11 items from the
B		25 items from the 2008 form	2010 form
		4 items from the 2007 form	

Out of the forty items in each set, twenty-nine items were **independent** items to be used in either item set A or B, while eleven were **anchors** to be shared in both item sets. Out of the twenty-nine independent items, twenty-five items from the 2009 form were used in item set A while twenty-five items from the 2008 form were used in item set B. The four items making up the twenty-nine independent items for each item set came from the 2007 form.

The eleven anchor items were all taken from the 2010 form. It was considered that eleven items (27.5 per cent, 11/40) were necessary to enable a common-item linking between item sets, following Saida and Yanagawa (2011) where a common-item design was developed to compare the test scores among forms. Saida and Yanagawa (2011) included thirteen anchor items (28.9 per cent, 13/45) between two different test forms. As presented in Table 4.4 in the present study, two anchor items were allotted to different six item types of the current JNCTL with the exception of only one item allotted to Part 3A (medium-length dialogue). The anchor items were distributed in the same locations across the four test forms (A+, A-, B+, B-).

Table 4.4 Listening comprehension test: the number and ratio of items in relation to item types through comparison with the JNCTL

Items	# of items	Dialogue				Monologue	
		Comprehension			Response	Comprehension	
		Short	Medium-length	Long*		Short	Long*
Independent	25	6	3	3	7	3	3
	4	1	0	0	2	1	0
Anchor	11	2	2	2	2	1	2
# of items(%)	40	9 (22.5%)	5 (12.5%)	5 (12.5%)	11 (27.5%)	5 (12.5%)	5 (12.5%)
The JNCTL (%)	25	6 (24%)	3 (12%)	3 (12%)	7 (28%)	3 (12%)	3 (12%)

Note. *One long text is associated with two or three items.

Care was taken to keep the ratio of different types of items in the material similar to that of the current JNCTL. The current JNCTL is composed of six different types of items;

- 1) short dialogues with one comprehension question (6 items or 24%),
- 2) short dialogues with one response question (7 items or 28%),
- 3A) medium-length dialogues with one comprehension question (3 items or 12 %),
- 3B) a longer dialogue associated with three comprehension questions (3 items or 12%),
- 4A) short monologues with a comprehension question (3 items or 12%),
and
- 4B) a longer monologue associated with three comprehension questions (3 items or 12%).

To closely parallel the ratio of item types in the current JNCTL, each experimental form was composed as shown in Table 4.4.

- 1) nine short dialogues with one comprehension question for each (22.5%)
- 2) eleven short dialogues with one response question for each (27.5%)
- 3A) five medium-length dialogues with one comprehension question for each (12.5 %)
- 3B) two longer dialogues associated with three or two comprehension questions (5 items or 12.5%)
- 4A) five short monologues with a comprehension question for each (12.5%)
- 4B) two longer monologues associated with three or two comprehension questions (5 items 12.5%).

4.3.1.2 Recording

It was important that while the degree of sandhi-variation should vary

across versions, other features should be held as constant as possible. The same voices, for example, had to be used in recording the different versions of the input texts. Accordingly, rather than using the recordings from past administrations of the JNCTL (which are available on NCUEE's website), fresh recordings were prepared.

There were two differences between these new recordings from the recording of the current JNCTL that should be noted. First, while the current JNCTL only employs North American accents to reflect the 'standard model' of English in Japan (see Section 3.2.3), two native speakers, an American man and a woman from New Zealand, participated in the recording for the present study. The inclusion of the New Zealander was necessary because no North American woman was available. The inclusion of a non-North American accent, however, was not considered to threaten the validity of the experiment as the same voice was used across the versions. Furthermore, she seemed to try to use North American accent, taking into account the dominant use of the accent in the current, formal English education at schools in Japan.

Second, only two native speakers were used for the recording, unlike the current JNCTL where four different native speakers are used (see Section 3.2.3). This was again a matter of availability. However, the question numbers were recorded by another native speaker with a North American accent.

The native speakers were informed about the purpose of the recording and the experiment, and were asked to rehearse with the scripts, which were made available to them prior to the recording. Roles in the dialogues were played by the two native speakers, a male and a female: the same gender balance as on the current JNCTL, while the monologues were each randomly assigned to one of them. The recording was transferred to

computer, and was edited using 'Sound It', a sound editing programme operating at 16 bits and 44,100KHz.

One major challenge with the recording was keeping the speech rate similar between versions since speech rate is strongly implicated in the occurrence of sandhi-variation (Gimson & Cruttenden, 1994). Although the two native speakers were asked to keep their speed as comparable as possible while making the occurrence of sandhi-variation distinctive, in practice it proved very difficult to make the speech rate identical.

4.3.1.3 Evaluation of the recordings

The possible confounding variable of disparity in the speech rate was addressed by asking two language experts to evaluate the two versions for each item set. They were asked to listen to five different texts selected from the texts (two short dialogues, a longer dialogue, a short monologue, and a longer monologue) to assess how far they felt that the speech rate was similar across the two versions. The same Likert scale as was used in Preliminary study 3 (Appendix 3) to assess the 'naturalness' of the sandhi-variation in the current JNCTL was employed again here.

The result was that in the evaluation of the sandhi plus version, both raters agreed that the speech rate would 'fairly reflect' the 'naturalness' of real life listening. In the evaluation of the sandhi minus version, on the other hand, while rater B evaluated the speech rate as reflecting natural language use 'to some extent', rater A felt that it did so 'little'. The results suggest that the relatively slow delivery of the sandhi minus version was noticeable to the raters and so might affect test taker performance. Admittedly, the average speech rate of item set A was 146.6 wpm for the sandhi minus and 162.7 wpm for the sandhi plus version, while it was 152.2 wpm and 167.6 wpm for item set B, respectively (see p. 70 for the use of

speech rate). In short, speech rate could not be controlled at this moment, and the comparison might have been between a relatively fast sandhi plus and relatively slow sandhi minus version. Taken together with the piloting of the material, which followed, this finding led to further amendment of the speech rate (see Section 4.3.1.4).

The two raters were also asked to assess the 'naturalness' of the sandhi-variation in each item set. The results for holistic impression, as shown in Table 4.5 for sandhi plus version and in Table 4.6 for sandhi minus version, was that rater B rated the sandhi plus version as reflecting 'very much' and rater A did as reflecting 'fairly' natural speech. With respect to the different features of sandhi-variations both raters agreed on assimilation, elision, and intrusion, rating them as 'fairly reflect' the 'naturalness' of sandhi-variation. Besides, rater A considered that the weak form, formulaic expression, elision of voiceless plosives, and flap [ɾ] 'fairly reflect'. Although rater B rates these four features 'reflect to some extent', the discrepancy was consistently one scale point on a seven-point Likert scale. This suggests that the sandhi plus version could be considered suitable for the experiment.

In the evaluation of the sandhi minus version, on the other hand, the two raters did not agree on any of the features or on their holistic impression as Table 4.6 shows. Less sandhi-variation, however, occurred in the sandhi minus version than in the sandhi plus version. This suggests that the recording intended for use in the sandhi minus version could be considered suitable for the experiment as it was considered by both raters to involve considerably less sandhi-variation than the sandhi plus version.

Table 4.5 Evaluation of ‘naturalness’ of the sandhi plus version

	very much	fairly	to some extent	uncer- tain	not suffici- ently	little	not at all
Assimilation		AB					
Elision		AB					
Weak form		A	B				
Formulaic expressions		A	B				
Intrusion		AB					
Elision of voiceless plosive		A	B				
Flap [ɾ]		A	B				
Speech rate		AB					
Holistic impression	B	A					

Note. A=rater A, B=rater B

Table 4.6 Evaluation of ‘naturalness’ of the sandhi minus version

	very much	fairly reflect	to some extent	uncer- tain	not suffici- ently	little	not at all
Assimilation				B		A	
Elision				B		A	
Weak form				B		A	
Formulaic expressions				B		A	
Intrusion	B					A	
Elision of voiceless plosive				B		A	
Flap [ɾ]				B		A	
Speech rate			B			A	
Holistic impression		B				A	

Note. A=rater A, B=rater B

4.3.1.4 Piloting

For the purpose of ensuring the validity of the experimental material, a pilot study was conducted. Seventy-five second year high school students who intended to take the JNCTL the next year participated in the piloting. It was considered that the discrepancy in participants between piloting (high school students) and the experiment (the first or second year university students) would not threaten the validity of the pilot, as it was assumed that these relatively able high school students would have a listening proficiency level comparable of that of the likely participants in the experiment.

Two versions for item set A and the sandhi plus version for item set B were piloted. The sandhi minus version for item set B could not be piloted due to a shortage of participants. It was considered that this would not adversely affect the value of the piloting, as the tryout of the three forms (A+, A-, and B+) would be sufficient to identify any relevant problems.

As Table 4.7 shows, the piloting found the two item sets (A and B) to be comparable in difficulty. The mean score for the sandhi plus version of item set A was 12.75 (32 per cent) and it was 13.77 for item set B (37 per cent because the three items associated with one dialogue was excluded from the analysis. This was because of an editing error on the test form: a picture which should have been printed on the test brochure to accompany the longer dialogue was missing. Accordingly, full marks for B+ version was reduced from forty to thirty-seven.).

Table 4.7 Results of the pilot study

Item set	Version	<i>N</i>	<i>M</i>	<i>SD</i>
A	+	20	12.75 (32%)	3.91
	-	20	14.55 (36%)	3.0
B	+	35	13.77 (37%)	5.04

Note. Sandhi minus version of item set B was not piloted.

The results of this piloting informed three major refinements of the material. First, the test needed to be made easier since the mean scores across the forms (32 to 37 per cent) were only around one third out of the total. It was considered that the low scores might be attributable to the fact that during the pilot, each input text was played only once. Although the current JNCTL provides the input twice for each text, the number of opportunities to listen was kept to one in the piloting as there could have been repetition effects on the test takers' listening comprehension, contaminating the results as an intervening variable. Furthermore, test takers' possible fatigue was also taken into account. Providing the input text twice must have taken forty-five minutes, and this was considered too long for them to keep their concentration on the test, which might have resulted in reducing the validity of the test.

However, for Main study, the decision was taken to employ two listenings. This would have the twin benefits of keeping the tests as similar as possible to the current JNCTL and making them a little easier. The time allowed to answer each item after the first and second listenings was kept identical with the time allowed in the current JNCTL.

A second amendment involved replacing two items which were identified as being too difficult for the participants. The item difficulty values of these two items were 0.17 and 0.03 for B-, 0.1 and 0 for A+, and 0.1 and 0.05 for

A-. The replacement items were taken from an EFL textbook for Japanese high school learners which focuses on listening skills.

The third amendment with the piloted material was making the speech rate identical between the two versions for each item set. This was intended to keep the other parameters than the speech rate as closely as possible in order to raise the validity of the experiment. Although in theory the speech rate adjustment should have been made before the pilot study, it proved very difficult that the native speakers kept the speech rate as comparable as possible while making the occurrence of sandhi-variation distinctive (see Section 4.3.1.2). Consequently, taking advantage of this opportunity to modify the whole instrument, the speech rate was technically adjusted. The recording was edited using the 'Time compressor' function of the sound editing software 'Sound it'. This function allows the user to manipulate the rate of the delivery while keeping the same tempo and pitch as in the original. The principle for the manipulation was to adjust the speech rate of the sandhi minus version to match the rate of the sandhi plus version. Accordingly, the speech rate of the sandhi minus version was raised by 11 per cent for item set A and by 10 per cent for item set B.

The only concern with the technical manipulation was that it could have affected sound quality to the extent that the test takers would find the recording unnatural. The researcher confirmed that the material sounded natural, and was not distorted.

Thus, the pilot instruments were modified before the administration of the experiment. The three amendments, providing the input text twice to the test takers, replacing the two most difficult items with easier ones, and making the speech rate identical across versions, would contribute to the validity of the experiment.

4.3.2 Questionnaire

A questionnaire was developed to explore how the manipulation of the sandhi-variation would affect the cognitive load of the test takers. The questionnaire items were derived from sandhi-variation features and speech rate, as operationalised in Preliminary study 3 (see Table 4.5 and Table 4.6)—assimilation, elision, weak forms, formulaic expressions, intrusion, flap [ɾ] sound, and speech rate. Both expert review and piloting were conducted before the administration. The expert review was intended to raise the validity of the items, while piloting was intended to reaffirm that the items would be understood by the respondents in the way intended. Reviewed by the four experts—one in language testing, one at English phonetics, one at second language listening, and one a high school English teacher—two items were collapsed because of lack of relevance to the purpose of the questionnaire or redundancy with other items. The items included ‘I could have understood the texts better if I had looked at the scripts while I was listening’, and ‘I felt the words were pronounced differently from when they are spoken more slowly or more clearly’.

Piloting with two university students resulted in changing the wording of the items, so that the respondents could be able to easily understand the items. ‘Rate of delivery’, for example, was replaced with ‘speech rate’. Thus, the questionnaire composed of five items of a five-point Likert scale (see Appendix 11A, 11B) was ready. The question items included:

- I felt that overall the sounds ran together. [linking]
- I found it difficult to recognise when one word finished and the next word began. [boundary]
- I felt many of the words were unclearly and ambiguously pronounced. [unclearly]

- I felt some words were elided or dropped. [elided]
- I felt the speech rate was fast. [fast]

The respondents were asked to read each statement presented in their L1 (Japanese) and to rate the degree of agreement by choosing an option which best fits their feelings. The options were 5=agree, 4=agree to some extent, 3=uncertain (cannot decide), 2=disagree to some extent, and 1=disagree.

4.4 Participants

A total of 154 university students participated in the experiment from five different intact classes at a university in Tokyo. Their English proficiency had been measured by TOEIC IP (Test of English as International Communication, Institutional Programme), which is a popular standardised test composed of two sections: listening comprehension and reading comprehension, each worth a possible 495 points: the maximum possible score is 990 (International Institution of Business Communication, henceforth, IIBC, 2010). The use of TOEIC IP was intended to place all of the first and second year students into appropriate classes.

Groups 1 and 2 (see Table 4.2), who were first year students, had taken the test three weeks before the first administration of the experiment while Groups 3, 4, and 5, who were second year students, had taken the test five months before. Despite the difference in the time at which the participants took the test, the result was considered to be a general indicator of their English proficiency when the present study was conducted. Given the total score of the TOEIC IP supplied in Table 4.8, overall the participants' proficiency was estimated to be of a lower-intermediate level according to IIBC, which develops and administers TOEIC IP in Japan. IIBC reports

that the mean scores for the third-year high school students who had taken the test in 2009 ($N=14,045$) was 233 for the listening component and 402 in total (IIBC, 2010), while the equivalent figures were 201.9 ($SD=40.3$) and 361.9 ($SD=63.7$) respectively for the participants in the experiment (see Table 4.8). The difference in mean scores for the listening portion, 31.1 ($233 - 201.9$), was within $2SE$ ($2 \times 25=50$, ETS, 2007), and was not considered statistically significant. In addition, the difference in mean scores for the total including listening and reading components, 40.1 ($402 - 361.9$), was within $2SE$ ($2 \times 50=100$, ETS, 2007) and was not considered statistically significant, either. Thus, the English proficiency of the participants in the experiment is similar to that of the third-year high school students and the results can reasonably be generalised to that population.

Table 4.8 Participants' proficiency measured by TOEIC IP

Group	<i>N</i>	<i>M</i> for listening (<i>SD</i>)	<i>M</i> for total (<i>SD</i>)
1	23	221.3 (24.5)	404.8 (11.1)
2	19	256.6 (27.8)	475.8 (42.2)
3	31	212.3 (26.6)	357.8 (29.5)
4	38	182.0 (33.3)	334.2 (21.9)
5	43	176.3 (32.6)	317.0 (55.8)
	154	201.9 (40.3)	361.9 (63.7)

4.5 Procedures

The experiment employed a cross-over design to control for any possible test order effects (see Table 4.2). The first administration of the experiment was conducted in April, followed by the second administration two months later in June, 2011. The interval of two months was intended to control for any possible memory effects, given that this experiment

employed a common item design as supplied in Figure 4.1.

Taking advantage of the interval, the instructor (the researcher) taught listening to the participants. Three classes were scheduled prior to the second administration, each with 90-minute long, including instructions and practice for improving students' listening abilities. The instruction was intended to encourage the students to be motivated in the second administration and to guide them to be successful L2 listeners. One possible intervening variable from this instruction, which is teaching effect, must have been compensated for by the same teaching procedure that was consistently made across the groups.

In the administration care was taken to follow consistent procedures across the groups. First, the same portable CD player was used in both administrations to control for the sound quality and volume. Second, each listening comprehension test was conducted at the very beginning of the scheduled class. Third, participants were informed that the result of the listening comprehension test in the second administration would affect their grading in the semester.

The questionnaire, which had been developed in an attempt to explore the level of cognitive load imposed on the test takers, was administered to Groups 1 through 4 once the listening comprehension test in the second administration was over, whereas it was administered to Group 5 both in the first and second administration in order to employ a common-person linking and thereby allow for the use of Rasch analysis. Table 4.9 presents how the questionnaire was administered in the present study.

Table 4.9 Administration of the questionnaire

Group (N)	1 st (April, 2011)	2 nd (June, 2011)
1 (23)		-
2 (19)		+
3 (31)		-
4 (38)		+
5 (43)	-	+

Note. '+' and '-' indicates sandhi plus and minus version respectively.

4.6 Analyses

4.6.1 Analysis for Research Question 5

Rasch analysis was undertaken after conducting basic descriptive analysis of the results for each group in relation to each administration. The use of Rasch analysis allowed for converting the item difficulty estimates and the person abilities on the four different forms (A+, A-, B+, and B-) into scores on a logit scale, the measure unit common to both person ability and item difficulty, where 'the differences between persons, between items, and between persons and items can directly be read' (Bond & Fox, 2007, p.48) from the logit scale and compared with one another. That is, the use of Rasch analysis in the present study was intended to consolidate the 138 item difficulty estimates, sixty-nine from each version into one set for analysis, so that each item difficulty estimate can be compared with each other to answer RQ 5: how does the manipulation of sandhi-variations affect listening comprehension test performance?

The quality of the Rasch analysis was investigated before answering RQ 5 through exploring Rasch person/item reliability, fit statistics, and the quality of the anchor items. Rasch person ability is an indicator of whether there are enough items spreading along the continuum, and enough spread of

ability among persons, whereas Rasch item reliability indices indicate ‘the replicability of item placements along the pathway if these same items were given to another sample of the same size that behaved the same way’ (Bond & Fox, 2007, pp. 41-42). Item misfit and overfit was analysed using Winsteps ver. 3.72.0 (Linacre, 2011) to determine if the elicited data (item and person estimates) fit well with the Rasch model. Fit refers to ‘the degree of match ... between the expectations of the model and the actual data for that candidate on each item’ (McNamara, 1996, p. 137). There are two types of fit statistics: misfit and overfit. Misfitting items or persons differ from common response patterns, whereas overfitting ones behave too similarly to the predictions of the Rasch model (McNamara, 1996).

The quality of the anchor items was explored as to whether there would be a good coverage of the difficulty levels. In addition, each anchor item was explored as to whether the average ability of those test takers who answered correctly would be higher than that of those who answered incorrectly.

Based on these analyses the answer to RQ 5 was explored both holistically and discretely. Holistic analysis aimed to find whether there would be any statistical difference in the item difficulty estimates between the versions. This analysis was carried out through running analysis of variance (ANOVA) using SPSS (version 20.0 for Windows) Analysis of discrete items was conducted as to whether each item would have statistical difference in the item difficulty estimates between the versions. The following formula was to be employed as the test of difference: If the proposition was true for an item, then the difference was considered to be statistically significant.

$$|ID_1-ID_2| > SE_1+SE_2 \text{ (Formula 1)}$$

ID₁, ID₂ = item difficulty estimates for a pair of items on both versions

SE₁, SE₂ = standard errors for a pair of items on both versions

It was through this process that the answer to RQ 5 was identified.

4.6.2 Analysis for Research Question 6

The descriptive statistics including mean scores and the standard deviations for each item difficulty estimate, in the case of this study, perceptual difficulty in noticing sandhi-variation features were presented. Then, Rasch analysis was carried out, treating the ten questionnaire item difficulty estimates, five from each version, as one set for analysis, so that perceptual difficulty in noticing sandhi-variation features could be compared with each other to answer RQ 6: how does the manipulation of sandhi-variations affect the level of cognitive load imposed on the test takers? Formula 1 (see Section 4.6.1) was employed to explore whether there would be any statistical difference in the item difficulty estimates between the versions. To determine the quality of the results, Rasch item/person reliability and fit statistics was explored using Winsteps ver. 3.72.0 (Linacre, 2011). Also, the appropriacy of category structure (Bond & Fox, 2007) was investigated to reaffirm that the five-point Likert scale employed in the questionnaire functioned in the way intended.

This chapter described the methodology to address the research questions 5 and 6 for Main study. The next chapter reports on the results of manipulation of sandhi-variation on listening comprehension test performance and on the level of cognitive load imposed on the test takers.

Chapter 5 Main study: Results and analyses

5.1 Introduction

This chapter reports the results and analyses for research questions 5 and 6. First, the results for descriptive analysis are reported and the quality of Rasch analysis is explored. Then, through the in-depth analysis of the actual items the answers to RQs are provided. Discussions are provided on the effects of manipulation of sandhi-variation features on test performance and the cognitive load imposed on the test takers.

5.2 Research Question 5

5.2.1 Introduction

Research question 5 addressed how the manipulation of sandhi-variation affects listening comprehension test performance. Table 5.1 shows mean scores and standard deviations for each administration. The mean scores on the listening comprehension tests generally ranged from 15.5 (38.8 per cent) to 20.7 (51.8 per cent) across groups and administrations with the exception of the comparatively high mean scores for Group B (24.3 and 26.6). This suggests that overall the test forms were difficult for the test takers of the present study. The relatively higher mean scores for Group B can be attributed to the higher listening proficiency of the participants as indicated by the TOEIC IP scores (see Table 4.8).

Table 5.1 Descriptive statistics for each administration

Group (<i>N</i>)	1st administration		2nd administration	
	item set &	<i>M (SD)</i>	item set &	<i>M (SD)</i>
	version		version	
Group 1(23)	A+	16.4 (3.5)	B-	20.7(3.1)
Group 2(19)	A-	24.3 (5.3)	B+	26.6 (5.2)
Group 3 (31)	B+	17.3 (3.9)	A-	18.6 (4.7)
Group 4 (38)	B-	19.0 (4.4)	A+	16.7 (4.4)
Group 5 (43)	B-	17.5 (4.4)	A+	15.5 (4.8)
(154)				

Note. 'A+' indicates sandhi plus version of item set A. Full mark is forty.

5.2.2 Quality of Rasch analysis

5.2.2.1 Introduction

Rasch analysis was conducted to consolidate the 138 item difficulty estimates, sixty-nine from each version, into one set for analysis, so that each item difficulty estimate can be compared with each other to answer RQ 5: how does the manipulation of sandhi-variations affect listening comprehension test performance?

Out of 154 students who were registered in any of the five intact classes (Group 1 to 5), the six students who did not take either of the tests were excluded from this analysis. Accordingly, the number of test takers to be analysed was reduced to 148 from 154.

5.2.2.2 Reliability

The quality of Rasch analysis was in the first place investigated through the Rasch person/item reliability statistics. Rasch person reliability index is an indicator of whether there are enough items spreading along the

continuum, and enough spread of ability among persons, whereas Rasch item reliability index indicates ‘the replicability of item placements along the pathway if these same items were given to another sample of the same size that behaved the same way’ (Bond & Fox, 2007, pp. 41- 42). The result for Rasch person reliability and Rasch item reliability were 0.79 and 0.92 respectively, as shown in Table 5.2. This shows that there are enough items spread according to the range of person ability, and that item placements and person ordering are replicable, suggesting that the results are substantially reliable.

Table 5.2 Measures for participants and items, and reliabilities

Measure	<i>N</i>	Measure		<i>R</i>
		<i>M</i>	<i>SD</i>	
Person	148	-0.1	-0.6	0.79
Item	138	-0.04	1.18	0.92

Figure 5.1, which shows the relation between estimation of participants and items, displays the linear relationship between the Rasch calibration for the 148 test takers and the 138 items. On the far left side of the Figure is the Rasch logit scale, a true interval scale on which zero indicates the average, higher numbers represent greater person ability and greater item difficulty. The second column shows person ability, and in the case of this study, person ability represents the participants’ listening ability. Each ‘#’ represents two test takers and a dot (.) represents one. The higher the marks appear in the column, the higher the level of listening ability the test taker is estimated to have. The third column displays item difficulty. Here, the higher the items appear in the column, the more difficult they are estimated to be. The numbers in this column indicate item number and a

'-' before the number indicates that the item comes from the sandhi minus version. For example, '-31' indicates item number 31 in the sandhi minus version while '31' is item 31 in the sandhi plus version. The logit scale represents the probability that a person of given ability will succeed on an item of given difficulty. Where the difference between the person ability value and the item difficulty value is 0 logits, the person is estimated to have 50 per cent probability of success on that item. When the person ability is higher than item difficulty, the person is more likely to give a correct response, when item difficulty is higher than person ability, an incorrect response is the likely outcome.

Figure 5.1 reveals that the items are spread more widely than the test takers' abilities. Besides, there are not only no gaps in the empirical item hierarchy but there is considerable redundancy, as items with similar difficulty estimates are found along a wider range of measurement (between -2 and 2). This is consistent with the higher Rasch person/item reliability.

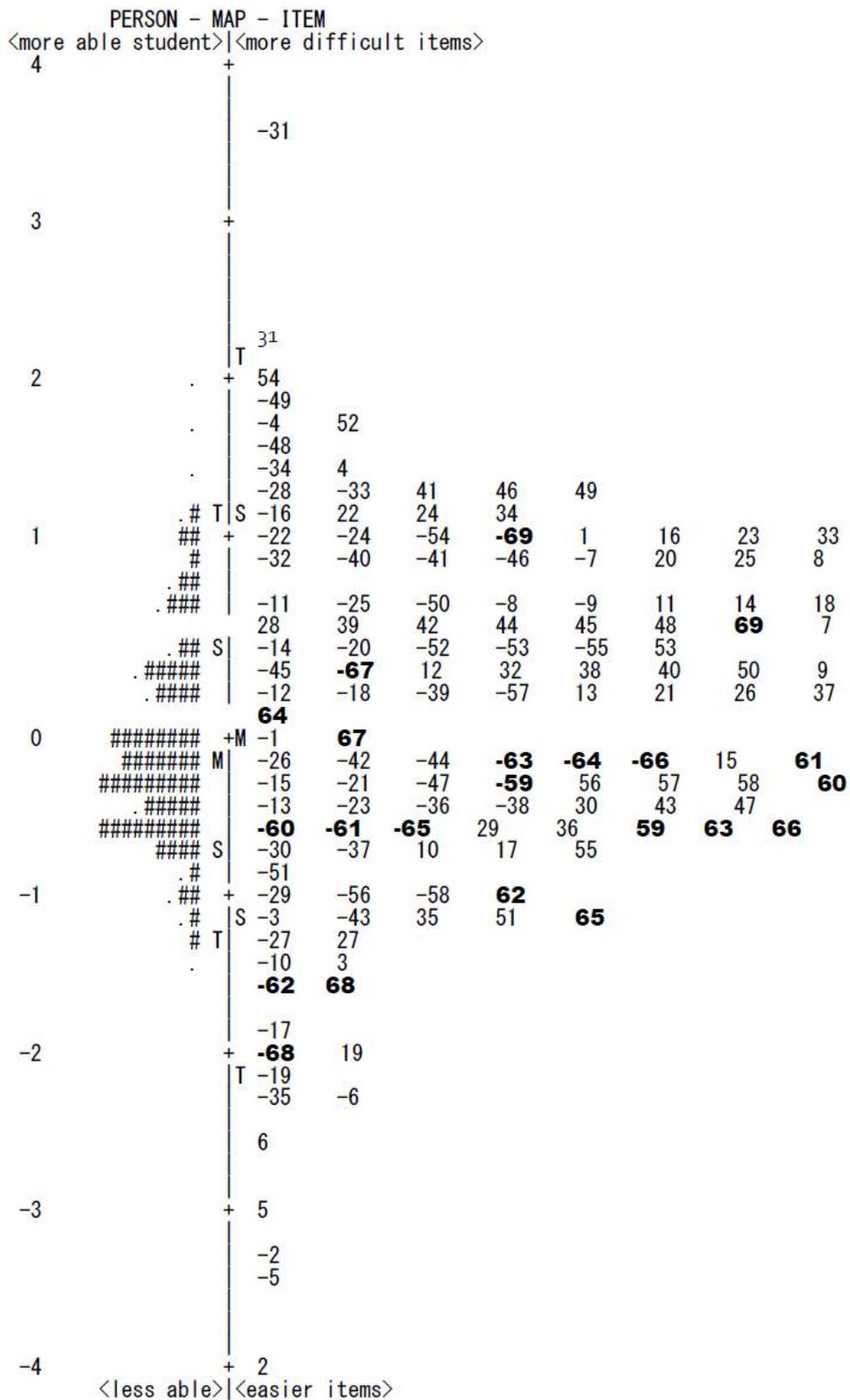


Fig. 5.1 Map showing the relationship between test takers' ability and item

difficulty estimate Note. M=mean; S=1SD; T=2SD; Bold=anchor items

5.2.2.3 Fit statistics

Second, to determine if the elicited data (item and person estimates) fit well with the Rasch model, item misfit and overfit were analysed using Winsteps ver. 3.72.0 (Linacre, 2011). Based on Bond and Fox (2007), infit mean square values were considered acceptable in the present study when they ranged from -2 to 2 by its Zstd. Since infit mean square value by Zstd refers to a standardised infit value, a value of less than -2 was considered to show overfit, whereas a value greater than 2 was considered to exhibit misfit (Bond & Fox, 2007).

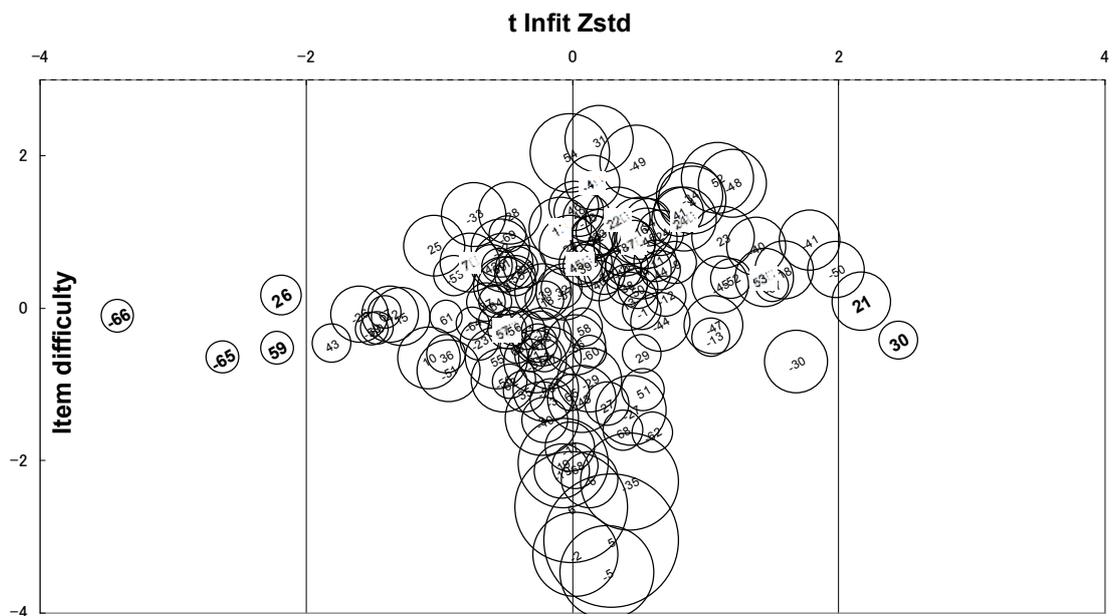


Fig. 5.2 Infit mean square value by Zstd and item difficulty estimate for each item

Note. Two items (# 2 and # -31), which were located beyond the logit scale of -4 and 3, were omitted from this figure. The logit was -5 and 3.55 for # 2 and # -31 respectively.

Figure 5.2 displays the infit mean square values by Zstd and item difficulty estimates for the 136 items (Two items, # 2 and # -31 were omitted from the figure). The vertical line indicates item difficulty as a logit score, a true interval scale, while the horizontal line indicates infit mean square by Zstd. In the figure, the higher the items appear, the more difficult the items are. The lower the items appear, the easier they are. The size of the circle for each item indicates the degree of imprecision or the standard error of the measurement. The majority of the items (130 items) fell in the acceptable range of between -2 and 2 of infit value by Zstd while six items fell outside the range. Two items were found beyond the acceptable range, on the right outside the range, to have a lack of fit to the Rasch model, while four items were on the left side. This reveals that 132 items including # 2 and # -31 omitted from Figure 5.2 contribute to the measurement of only one construct—in the case of the present study, listening ability, which is an underlying assumption to be met for Rasch analysis.

The two items which did not match the expectations of the Rasch model were # 21 and # 30. The item # 21 (Infit Mnsq=1.22, Infit Zstd=2.2) had been answered correctly by relatively lower-ability test takers ($Mean=0.24$, $SE=0.17$) while incorrectly by relatively higher-ability test takers ($Mean=0.68$, $SE=0.21$). The other item # 30 (Infit Mnsq=1.12, Infit Zstd=2.5) was answered correctly by relatively lower-ability test takers ($Mean=-0.26$, $SE=0.06$) and incorrectly by relatively higher-ability test takers ($Mean=-0.22$, $SE=0.07$). These two items were found not to meet the assumption of the Rasch model that the higher the person ability, the more likely an item is to be answered correctly.

The two items, however, were not excluded from the subsequent analysis, because the primary purpose of this study was to compare item difficulty estimates under the two different conditions where one text involves a

natural level of sandhi-variation and the other involves minimal sandhi-variation. Also, the measures constructed from Rasch model are claimed to be robust to minor deviations from the model's requirements (Henning, Hudson, & Turner, 1985). The two items only cover 1.45 per cent (2/138), which is far lower than the ratio of the percentage—five percent of misfitting items that Belgar (2010) considered problematic. Thus, having such a small number of misfitting items was not considered to reduce the validity of the test itself.

Overfit to the Rasch model was defined in the present study as items displaying standardised infit value less than -2. As illustrated in Figure 5.2, four items (# 26, # 59, # -65, # -66) overfit the Rasch model. The number of overfitting items, however, was not considered problematic, given that the proportion of overfitting items is less than five per cent (Smith, 2005). The four items only covered 2.9 per cent of the 138 item data set. Besides, Shizuka (2007) claims that overfitting items do not degrade the model but are no more than redundant. The four items, therefore, were included in the subsequent analysis to make a comparison with their counterparts, which is the aim of RQ 5.

Thus, the analysis of fit statistics shows that the test made up of those items which are for most part located acceptable range of between -2 and 2 by Infit value by Zstd, fits well with the expectations of the Rasch model. This suggests that the test forms were measuring one latent trait (construct)—listening ability of the test takers, which is an underlying assumption to be met for the use of Rasch analysis.

The analysis of fit statistics also showed that five participants (3.4 per cent, 5/148) did not fit the Rasch model (infit value greater than 2) and five participants (3.4 per cent, 5/148) overfitted the model. None of them, however, was excluded from the subsequent analysis due to the lower ratio

to the total number of participants. Thus, all items and all participants were included in the subsequent analyses.

5.2.2.4 Anchor items

Lastly, the quality of the Rasch analysis was explored through inspecting the results of eleven anchor items for each version (twenty-two items in total), since the anchor items affect the quality of the Rasch analysis. As Figure 5.1 displays, the item difficulty estimates for the anchor items (# 59 to # 69 indicated by bold) spread widely along the logit scale and fit with the range of the participants, suggesting that there was a good coverage of the difficulty levels. Besides, the standard errors of the measurement (*SE*) were relatively small compared with those for the independent items (# 1 to # 58) (see Table 5.4).

In addition, infit statistics show that although three items (# 59, # -65, # -66) overfit the Rasch model, none of the items were found underfitting the model. Furthermore, average ability of those test takers who answered correctly was higher than that of those who answered incorrectly for the twenty-one anchor items with the exception of # -62, even where no statistically difference ($p < .01$) was found in the average ability (-0.07) of those who answered correctly and that of those who answered incorrectly (0.01). This is consistent with higher Rasch item reliability (0.92, see Table 5.2). Lastly, distributed at the same places across the test forms, the eleven anchor items had been allotted to different six parts of the test (see Section 3.2.3.3) to closely parallel the ratio of the current JNCTL. All taken together, the anchor items functioned in the way intended to connect the different data sets and thereby treat the 138 items onto a same scale to allow for the use of Rasch analysis.

5.2.3 Answer to Research Question 5

Table 5.3 Means and SDs of item difficulty estimates

for each version

version	<i>N</i>	<i>M</i>	<i>SD</i>
sv+	69	0.02	1.16
sv-	69	-0.09	1.2

Note. Sv+ and sv- indicates the sandhi plus and minus version respectively.

RQ 5 addressed how the manipulation of sandhi-variation affects listening comprehension test performance. Mean scores and standard deviations for the item difficulty indices for each version were presented in Table 5.3. The mean score for the sandhi plus version and the sandhi minus version was 0.02 and -0.09 respectively. This result appears to show that overall the sandhi plus version was slightly more difficult than the sandhi minus version (by 0.11 logits). Analysis of variance (ANOVA) was run to explore this observed difference. It was found that there was no statistically significant ($p < .01$) difference in the item difficulty estimates between the two versions ($F=(1, 136)=0.307, p=0.58$). This suggests that manipulation of sandhi-variation does not substantially affect listening comprehension test performance.

A further analysis was conducted into possible differences in the individual item difficulty estimates. A scatter plot was displayed to illustrate comparative difference in item difficulty estimates for each item in relation to the occurrence of sandhi-variation. Figure 5.3 shows that the manipulation of sandhi-variation may affect item difficulty. Forty items, which are displayed under the line in the figure, were more difficult on the sandhi plus version than on the sandhi minus version ($sv+ > sv-$), whereas twenty-eight items, which are displayed above the line, were more difficult on the sandhi minus version than on the sandhi plus version ($sv+ < sv-$).

Only one pair of item (# 14 and # -14) did not yield any difference between the versions.

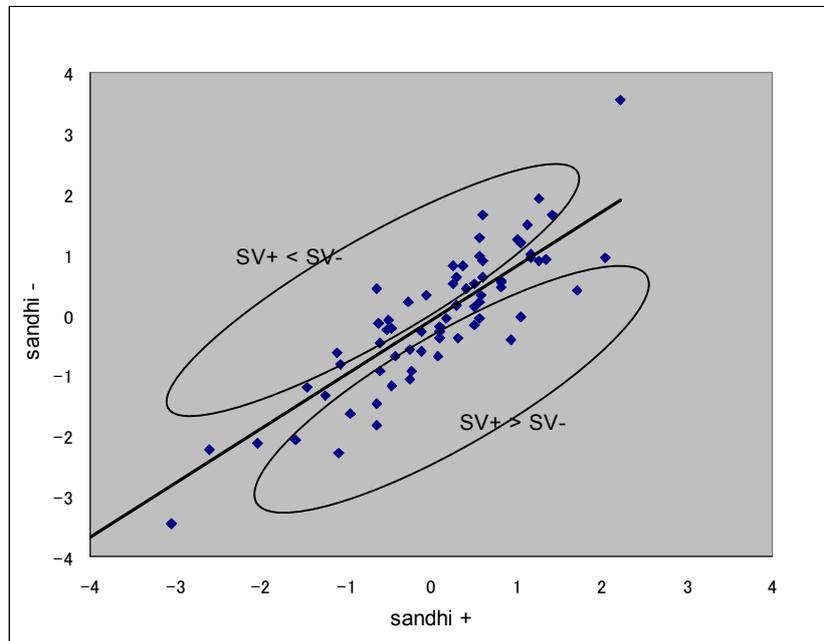


Fig. 5.3 Item difficulty estimates in relation to version

Note. Sv+ and sv- indicates sandhi plus and minus version respectively.

For the purposes of this study, each item was examined as to whether it would have statistical difference in item difficulty estimates between the versions. The following formula was employed as the test of difference between the versions (see Section 4.6):

$$| ID_1 - ID_2 | > SE_1 + SE_2 \text{ (Formula 1)}$$

ID_1, ID_2 = item difficulty estimates for a pair of items on both versions

SE_1, SE_2 = standard errors for a pair of items on both versions

If the proposition was true for an item, then the difference was considered to be statistically significant ($p < .01$). Table 5.4 supplies items difficulty estimates and standard errors for each item in relation to version, and the

statistically difference between the versions.

Table 5. 4 Item difficulty estimates and standard errors for each item in relation to version

item	sv+		sv-		difference	item	sv+		sv-		difference
	ID	S.E.	ID	S.E.			ID	S.E.	ID	S.E.	
1	1.05	0.34	-0.03	0.21	sv+ > sv-	36	-0.6	0.21	-0.48	0.33	
2	-5	1.84	-3.23	0.46		37	0.08	0.21	-0.7	0.34	sv+ > sv-
3	-1.5	0.4	-1.2	0.23		38	0.31	0.22	-0.38	0.32	sv+ > sv-
4	1.42	0.37	1.65	0.3		39	0.56	0.23	0.21	0.31	
5	-3.05	0.73	-3.47	0.51		40	0.36	0.22	0.81	0.32	
6	-2.61	0.61	-2.25	0.31		41	1.27	0.27	0.89	0.33	
7	0.6	0.32	0.87	0.24		42	0.56	0.23	-0.08	0.31	sv+ > sv-
8	0.82	0.33	0.53	0.23		43	-0.46	0.21	-1.19	0.37	sv+ > sv-
9	0.29	0.32	0.6	0.23		44	0.5	0.22	-0.18	0.32	sv+ > sv-
10	-0.65	0.34	-1.47	0.25	sv+ > sv-	45	0.59	0.23	0.31	0.31	
11	0.6	0.32	0.62	0.23		46	1.34	0.27	0.92	0.33	
12	0.29	0.32	0.16	0.22		47	-0.46	0.21	-0.22	0.32	
13	0.09	0.32	-0.38	0.21		48	0.61	0.23	1.64	0.37	sv+ < sv-
14	0.5	0.32	0.5	0.22		49	1.27	0.27	1.92	0.4	
15	-0.11	0.32	-0.29	0.21		50	0.26	0.22	0.51	0.31	
16	1.05	0.34	1.19	0.26		51	-1.07	0.23	-0.82	0.34	
17	-0.65	0.34	-1.82	0.27	sv+ > sv-	52	1.71	0.39	0.39	0.22	sv+ > sv-
18	0.5	0.32	0.11	0.21		53	0.4	0.32	0.43	0.22	
19	-2.03	0.49	-2.14	0.3		54	2.04	0.43	0.93	0.24	sv+ > sv-
20	0.82	0.32	0.44	0.22		55	-0.65	0.34	0.43	0.22	sv+ < sv-
21	0.09	0.32	-0.29	0.21		56	-0.24	0.21	-0.94	0.35	sv+ > sv-
22	1.17	0.35	0.93	0.24		57	-0.28	0.21	0.21	0.31	
23	0.93	0.34	-0.43	0.21	sv+ > sv-	58	-0.25	0.21	-1.06	0.36	sv+ > sv-
24	1.17	0.35	0.99	0.25		59	-0.52	0.18	-0.27	0.18	
25	0.82	0.33	0.55	0.23		60	-0.26	0.18	-0.57	0.18	
26	0.17	0.22	-0.08	0.31		61	-0.11	0.18	-0.62	0.18	sv+ > sv-
27	-1.25	0.24	-1.34	0.38		62	-0.96	0.19	-1.63	0.22	sv+ > sv-
28	0.56	0.23	1.25	0.34	sv+ < sv-	63	-0.62	0.18	-0.16	0.18	sv+ < sv-
29	-0.6	0.21	-0.94	0.35		64	0.09	0.18	-0.2	0.18	
30	-0.42	0.21	-0.7	0.34		65	-1.11	0.2	-0.64	0.18	sv+ < sv-
31	2.22	0.37	3.55	0.73	sv+ < sv-	66	-0.51	0.18	-0.1	0.18	sv+ < sv-
32	0.26	0.22	0.79	0.32		67	-0.05	0.18	0.3	0.18	
33	1	0.25	1.23	0.35		68	-1.6	0.22	-2.07	0.25	
34	1.13	0.26	1.48	0.36		69	0.56	0.19	0.97	0.2	sv+ < sv-
35	-1.09	0.23	-2.28	0.53	sv+ > sv-						

Note. Eleven items from # 59 through # 69 were anchor items. Sv+ and sv- indicates sandhi plus and minus version respectively.

The result was that twenty-four items were found to have statistically significant differences in the item difficulty estimates between the versions. Sixteen items were found to be more difficult on the sandhi plus version whereas eight items were found more difficult on the sandhi minus version. The rest of the items (forty-five items) were found not to have statistically significant differences in item difficulty estimates. This result has raised further questions: Why does sandhi-variation disturb listening comprehension in some cases, and why does sandhi-variation facilitate listening comprehension in others? Discussions are made about this question in the next section.

5.2.4 Discussion

The result of RQ 5, that sixteen pair of items out of sixty-nine (23 per cent, 16/69) were found more difficult on sandhi plus versions, can be attributed to the possibility that sandhi-variation disturbed test takers' listening comprehension. One of the examples can be # 35, a short dialogue from Part 2 where the test takers are supposed to choose an option which best fits to complete the conversation. The item difficulty estimate of the # 35 was -1.09 ($SE=0.23$) and it was -2.28 ($SE=0.53$) for # -35.

(text of # 35)

M: Hurry up! We're running late!

W: Hold on. Where did I put my purse?

M: I saw it in the kitchen.

(options)

- 1) Let's see. Don't let it go.
- 2) Oh, how much was it?
- 3) OK. *I'll take a look.* (key)
- 4) Why? Are you happy?

The recording of the short dialogue involved a considerable number of sandhi-variation features. The [d] in the *hold on* in the second utterance was elided, and the [d] in *did I* turned into a flap [ɾ], [t] in *put* was either elided or turned into a glottal stop [ʔ], an intruding [j] was found between *saw* and *it*, [t] in *saw it in the kitchen* was either elided or turned into a glottal stop [ʔ], and [ðə] in *the kitchen* was reduced to schwa [ə] as a result of linking with [n] which precedes. The occurrence of these sandhi-variation features might have caused the text to become more difficult than the sandhi minus version where [d] or [t] were not so much reduced.

On the other hand, the result of RQ 5, that eight items (12 per cent, 8/69) were found more difficult on sandhi minus version than on sandhi plus version, may partly be attributable to the test takers having less cognitive resource to spare when faced with the sandhi minus version. The test takers might have been driven to pay more attention to individual words, which were clearly and distinctively pronounced and thereby more noticeable to them. This might have drawn their attention and helped exhaust their cognitive resources, which would have been better reserved for higher level processing which builds a mental model (Goh, 2000). Because short-term memory or working memory can hold only about seven 'items' (Miller, 1956), clearly and distinctively pronounced words might have

filled up the limited processing capacity more rapidly. This could explain why the eight items might have displayed more difficult item estimates on the sandhi minus version than on its counterpart.

An example is # 28 from Part 1 where the test takers were asked to choose the correct option according to the content of the conversation. The item difficulty estimate of the # -28 was 1.25 ($SE=0.34$) and it was 0.56 ($SE=0.23$) for # 28.

(text of # 28)

M: Isn't the soccer game starting now?

W: Yeah, but my favourite drama comes on *in an hour on another channel*.

M: So, should I record the game?

W: That's OK. I'll record my program.

(the question and options)

What will the man do?

- 1) Record the drama.
- 2) Record the game.
- 3) Watch the drama.
- 4) *Watch the game.* (key)

The phonological difference between the versions was mostly derived from the sequence of '*in an hour on another channel*' in the second utterance by a woman. In the sandhi plus version, the sequence '*in an hour on another*' was reduced as follows [inən'auə-anə'nʌðə'tʃænl] as a

result of the linking of [n] with the following vowels, while on the sandhi minus version the same sequence was pronounced as follows [inænauə'ʔənʔə'nʌðə'tʃænl]. It was possible that the test takers of the sandhi plus version were able to commit less cognitive resource because the sequence on the sandhi plus version was perceived as two distinctive chunks 'in an hour on' and 'another channel'. By contrast, on the sandhi minus version, where individual words were more phonologically distinctive than on the sandhi plus version, the same sequence might have been perceived as discrete words rather than as two chunks. This heavier cognitive load on the sandhi minus version might have overloaded the test takers' cognitive resources, kept them away from paying attention to the following utterance including the necessary information (see Section 2.3.4.2), and led to the failure to construct a successful mental model.

This explanation appears, however, very speculative and cannot readily account for why some items are more difficult on the sandhi plus version while others are more difficult on the sandhi minus version. Further discussion of this point will be presented in Section 5.3.3.

5.3 Research Question 6

5.3.1 Introduction

RQ 6 was concerned with whether manipulation of sandhi-variation would affect the level of cognitive load imposed on the test takers. The result of RQ 5, as illustrated in the section above, revealed that while overall the manipulation of sandhi-variation did not affect listening comprehension test performance, the item difficulties of particular items were affected by the occurrence of sandhi-variation. This implied that manipulation of sandhi-variation would also affect the cognitive load imposed on the test takers.

To answer RQ 6, a questionnaire made up of five items was developed and administered to the participants. They were asked to rate the degree of agreement by a five-point Likert scale. The actual items were as follows (see Chapter 4):

- I felt that overall the sounds ran together. [linking]
- I found it difficult to recognise when one word finished and the next word began. [boundary]
- I felt many of the words were unclearly and ambiguously pronounced. [unclearly]
- I felt some words were elided or dropped. [elided]
- I felt the speech rate was fast. [fast]

The number of respondents to be analysed was 140 as fourteen students were absent out of 154 students who were registered in the five intact classes. The results including mean scores and the standard deviations for each sandhi-variation are presented in Table 5.5. The higher the mean score is, the more noticeable the sandhi-variation feature is for the test takers.

Table 5. 5 Means and SDs for perceptual difficulty in noticing sandhi-variation features

Group (N)	version	linking	boundary	unclearly	elided	fast
1 (22)	-	4.23(0.58)	4 (0.78)	3.73(0.89)	3.27(0.94)	3.95(0.91)
2 (18)	+	4 (0.56)	3.78(1.0)	4 (0.73)	3.61(0.93)	4.06(1.00)
3 (27)	-	3.7 (1.03)	3.78 (1.01)	3.37(0.89)	3.56(1.14)	3.93(0.88)
4 (30)	+	4.03(0.82)	3.97(0.82)	3.57(0.90)	3.87(0.75)	4.07(0.88)
5 (43)	- (1 st)	3.88(0.77)	3.73(0.91)	3.29(0.86)	3.41(0.99)	4.32(0.87)
	+ (2 nd)	3.61(1.01)	3.63(0.98)	3.54(0.97)	3.56(1.01)	3.63(1.01)

Note. linking=sounds running together, boundary=unclear word boundaries,

unclearly=unclearly pronounced sounds, elided=elided sounds, fast=faster speech rate

Overall, the descriptive statistics shows that linking (sounds running together) and fast speech rate were more noticeable to the test takers irrespective of version than unclear word boundaries, unclearly pronounced sounds, and elided sounds as the mean scores for the linking and fast speech rate was four indicating 'Agree to some extent' or over four for the three groups.

5.3.2 Quality of Rasch analyses

5.3.2.1 Introduction

Rasch analysis was carried out, treating the ten questionnaire item difficulty estimates, five from each version, as one set for analysis, so that each item difficulty estimate, in the case of this study, perceptual difficulty in noticing sandhi-variation features could be compared with each other to answer RQ 6: how does the manipulation of sandhi-variations affect the level of cognitive load imposed on the test takers? To determine the quality of the results, Rasch item/person reliability and fit statistics were

explored before attempting to find the answer to RQ 6.

5.3.2.2 Reliability

Preliminary analysis showed that Rasch person reliability was 0.62 and Rasch item reliability was 0.80, and that eight respondents did not fit the Rasch model as their values of *Zstd* were larger than 2, indicating underfit. When the eight respondents were excluded from the subsequent analysis, this raised the Rasch person and item reliabilities to 0.66 and 0.81 respectively, as shown in Table 5.6. Thus, the number of the respondents to be analysed was reduced to 132 (140 - 8).

**Table 5.6 Measures for participants and items,
and reliabilities on the questionnaire**

Measure	<i>N</i>	Measure	<i>R</i>	
		<i>M</i>	<i>SD</i>	
Person	132	1.88	1.21	0.66
Item	10	0	0.38	0.81

The result shows that item placements are considerably replicable, while person ordering is not as reliable. Also, the mean score for person measure was considerably larger (1.88) than that (0) for the item measure, suggesting that the participants had higher perceptual sensitivity to sandhi-variations than the items could have covered. This disparity between the participants and the items can easily be observed in Figure 5.4.

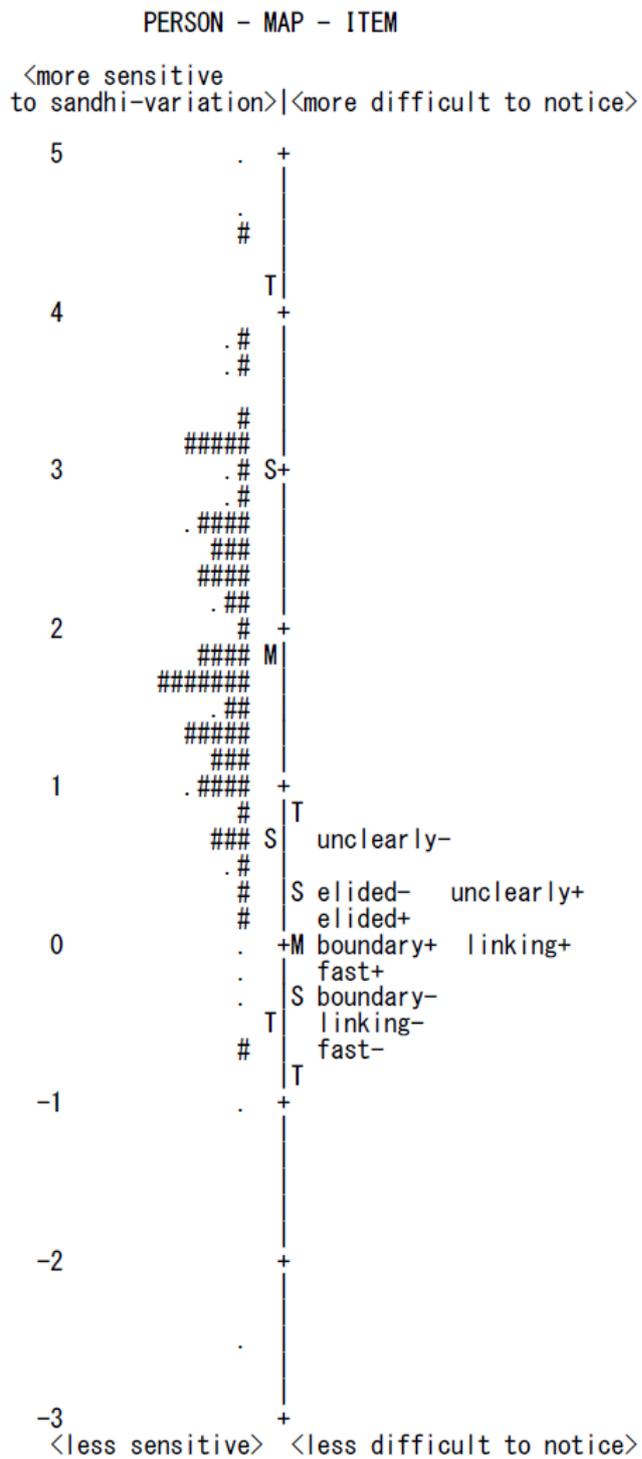


Fig. 5.4 Map showing the relationship between the test takers' sensitivity to sandhi-variation features and their perceptual difficulty to be noticed

Note. For example, 'elided+' and 'elided-' indicates elided sounds in the sandhi plus version and in the sandhi minus version respectively.

In the map the first column from the left is a logit scale. The second column shows person ability, and in the case of this study, the person ability is the participants' perceptual sensitivity to sandhi-variation features. Each '#' represents two test takers and a dot (.) represents one. If the marks appear higher in the column, it means the test takers are more sensitive to sandhi-variation. The third column displays perceptual difficulty in noticing sandhi-variation features. The more difficult-to-perceive features appear higher in the column. '-' indicates sandhi-variation features involved in the sandhi minus version.

The result shows that no sandhi-variation features were located higher on the scale than the majority of the respondents, suggesting that the majority of the participants found all of the features rather easy to perceive, whether they had taken the sandhi plus or minus version. The relatively lower Rasch person reliability (0.66) is consistent with this result. More items to measure the perceptual difficulties of the test takers in noticing sandhi-variations should have been provided (see Section 6.7).

5.3.2.3 Fit statistics

The fit statistics were investigated using Winsteps ver. 3.72.0 (Linacre, 2011). One item, speech rate for the sandhi minus version, was identified as not fitting the Rasch model as the value of Zstd (2.4) was larger than 2 as presented in Table 5.7. The average person ability of the test takers who perceived the speech rate to be fast ('agreed' with the questionnaire item) was 2.3 ($SE=0.17$). This was lower than that of the test takers who agreed with other sandhi-variation features, suggesting that even the test takers who are less sensitive to speech rate perceived it in the sandhi minus version to be fast.

Speech rate in the sandhi minus version (fast-), however, was not excluded from the subsequent analysis because it is a key focus of this study.

Table 5.7 Infit mean square by Zstd

Questionnaire items	sv+	sv-
sounds running together [linking]	-2.1	0.6
unclear word boundaries [boundary]	-0.9	-0.3
unclear pronounced sounds[unclearly]	-0.4	-0.2
elided sounds [elided]	-0.4	1.4
faster speech rate [fast]	-0.1	2.4

Lastly, no disordered data was found in terms of the appropriacy of category structure (Bond & Fox, 2007), suggesting that the five-point Likert scale employed in the questionnaire functioned in the way intended.

5.3.3 Answer to Research Question 6 and discussions

Given the clarified limitations on the quality of the Rasch analysis, further analysis was conducted to answer RQ 6.

Table 5.8 Measures and standard errors of perceptual difficulties of sandhi-variation features in relation to version

sandhi-variation features	sandhi+		sandhi -		difference <i>p</i> <.01
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	
sounds running together [linking]	-0.02	0.16	-0.51	0.16	sv+ > sv-
unclear word boundaries [boundary]	0.08	0.16	-0.3	0.16	n.s.
unclear pronounced sounds[unclearly]	0.3	0.16	0.58	0.15	n.s.
elided sounds [elided]	0.23	0.16	0.4	0.15	n.s.
faster speech rate [fast]	-0.12	0.16	-0.65	0.16	sv+ > sv-

Table 5.8 shows the item difficulty estimates, in the case of this study, the perceptual difficulty in noticing sandhi-variation features, and standard errors. The higher the measure is, the more difficult the sandhi-variation feature is to be noticed by the test takers. In other words, the lower the measure is, the more noticeable the sandhi-variation feature is. For example, in the sandhi plus version, sounds running together (-0.02) and fast speech rate (-0.12) were relatively more easily perceived by the test takers than unclear word boundaries (0.08), unclear pronounced sounds (0.3), and elided sounds (0.23). In the sandhi minus version, on the other hand, sounds running together (-0.51), unclear word boundaries (-0.3), and speech rate (-0.65) were more easily perceived by the test takers than unclear pronounced sounds (0.58) and elided sounds (0.4).

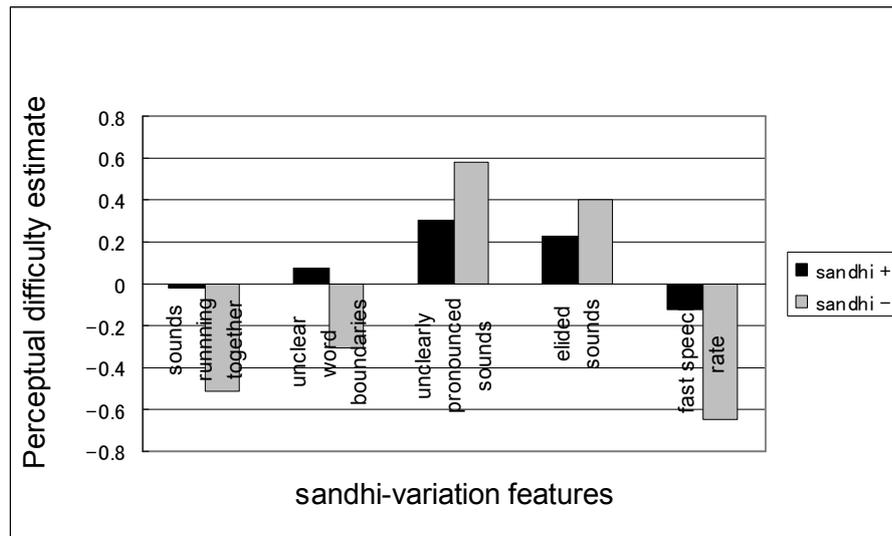


Fig. 5.5 Perceptual difficulty in noticing sandhi-variation features

Figure 5.5 displays perceptual difficulties in noticing sandhi-variation features in relation to version. Relevant to RQ 6 was, as Figure 5.5 shows, that the manipulation of sandhi-variation had different effects on different sandhi-variation features between the versions. Although the difference in the measure (item difficulty estimates) between the versions was not statistically significant by using the same formula (see Formula 1 in Section 4.6.1), unclearly pronounced sounds or elided sounds were relatively easily perceived by the test takers on the sandhi plus version, suggesting that unclearly pronounced sounds or elided sounds may have increased the cognitive load for the test takers on the sandhi plus version. This is consistent with the nature of the texts, and could account for the result of RQ 5, that sixteen items were found more difficult on the sandhi plus version than on its counterpart.

One of examples is # 54 from Part 3 where the test takers were asked to choose the correct option according to the content of the talk. The item difficulty (standard errors) was 2.04 (0.43) and 0.93 (0.24) for the sandhi plus version and the minus version respectively (see Table 5.4).

(text of # 54)

OK, Pat. *Here's* the medicine for your cold. *There* are two kinds. These yellow tablets are for your sore throat. Take two *of* them three times a day, after each meal. Take one *of* these green pills after meals only *when* you have a very high fever. *They'll* bring your temperature down to normal. *They're* strong, so take no more than three a day.

(the question and options)

What should Pat take after eating if she has a sore throat and a slight fever?

- 1) One yellow tablet and two green pills.
- 2) Three green pills.
- 3) *Two yellow tablets.* (key)
- 4) Two yellow tablets and one green pill.

The recording of the short talk involves several unclearly pronounced sounds or elided sounds: Among them *Here's*, *there are*, *they'll* might have been the most problematic for the test takers. Unaccented weak forms, these functional words involved reduction of length of sounds and the elision of vowels and consonants, and realised such as [iəz], [ðə], and [ðəl] for *Here's*, *there are*, and *they'll* respectively. Appearing at the very beginning of each sentence, those three unaccented words might have disturbed word recognition or speech perception, and have increased the cognitive load for the test takers on the sandhi plus version. It is noteworthy that the necessary information of this item (see Section 2.3.4.2) is distributed across the text, and this item requires a discourse-level of understanding. Taken together, unaccented and unclearly pronounced

words attributed to sandhi-variations might have increased the level of cognitive load for the test takers so much that more of them might have failed to answer correctly.

In the meantime, as Figure 5.5 displays, sound running together ($p < .01$), unclear word boundaries (n.s.), and fast speech rate ($p < .01$) were relatively easily perceived by the test takers of the sandhi minus version, suggesting that sounds running together with unclear word boundaries and speech rate perceived to be fast may have increased the cognitive load of the test takers on the sandhi minus version.

One good example of this can be # -55 from Part 1 where the test takers were asked to choose an option which best fits according to the content of a short conversation. The item difficulty (standard errors) was -0.65 (0.34) and 0.43 (0.22) for the sandhi plus and minus version, respectively.

(text of # 55)

W: Where have you **been**? I've been waiting here **so long**.

M: Sorry. I couldn't get out of the office until 5:30.

W: But that was **half an hour ago**.

(the question and options)

What time is it now?

1) 5:00 2) 5:30 3) 6:00 (key) 4) 6:30

This example of # -55 can potentially provide an empirical support for the speculative interpretation (see p.182) why the eight texts were found more difficult on the sandhi minus version than on its counterpart. The recording of the short dialogue makes a difference between the versions,

particularly in the first and the third utterance by the female speaker. In the sandhi minus version individual words tended to be clearly and distinctively articulated with relatively clear words boundaries. Even the functional words including *where, have, 've been, here, but, that,* were considered being pronounced in accented strong forms (Gimson & Cruttenden, 1994). By contrast, on the sandhi plus version, phonologically salient words (indicated by bold in the text above) involved only *been* and *so long* in the first utterance and *half an hour ago* in the third utterance. As a result, it was possible that the test takers of the sandhi minus version were perhaps able to catch more individual words than those of the sandhi plus version, and this might have prevented the test takers of the sandhi minus version to construct a successful mental model. The individual words evenly given to accents might have increased the level of cognitive load for the test takers and committed their cognitive resources, and they might have had little or no processing capacity available for higher level processing which forms a mental model of the text. On the sandhi plus version, by contrast, the test takers had no choice to focus on a few words which were clearly pronounced and reserved in their short term memory, and this could have kept themselves away from overloaded or 'engulfed' in the stream of sounds and helped them arrive at a destination—a correct answer.

The higher the processing level an item requires, the more likely this account seems. The required level of processing for # -55 was rated to be a discourse level of processing instead of a lower level of processing or propositional level or word level. If the item # -55 had required such a low level of processing, then the test takers of the sandhi minus version could have managed to arrive at a correct answer even if they did not construct a mental model of the text, because they only had to catch a few

semantically salient key words or phrases to answer correctly in such an item.

This account does not seem consistent with the result of RQ 6, that sound running together with unclear word boundaries may have disturbed test takers' cognitive processing on the sandhi minus version, if the assumption is true that clearly pronounced words on the sandhi minus version make word boundaries clear and prevent sounds from running together. This assumption, however, might not be necessarily true to the test takers of the present study because it is possible that the listening difficulty might have resulted in having them 'perceive' the sound to run together with fewer word boundaries and with faster speech rate. That is, the 'perception' of sound running together and fast speech rate can be considered to be a by-products of their unsuccessful listening rather than a cause. Bloomfield, Wayland, Rhoades, Blodgett, Linck, and Ross (2011) argue that 'listeners are more likely to perceive speech as fast when other features challenge comprehension' (p. 67).

Another example similar to # -55 is # 63 from Part 2 where the test takers were asked to choose an option according to the content of the conversation. The item difficulty estimate (standard errors) was -0.62 (0.18) and -0.16 (0.18) for the sandhi plus and minus version, respectively. In the sandhi plus version the limited number of the content words (indicated by bold in the text) were given phonological prominence. In the sandhi minus version, by contrast, individual words were perceived to have equivalent phonologically prominence. This might have increased the level of cognitive load and exhausted more rapidly and ineffectively the cognitive resources of the test takers of the sandhi minus version than those of its counterpart, and have kept them away from building a successful mental model of the text.

(text of # 63)

M: What's **wrong**, Mary?

W: I have a **terrible** cold. Today my doctor told me to **stay in bed** for a few **days** and **see** her again if I'm still **sick**.

M: So, you won't be going to **school** tomorrow. Do you want me to **turn in** your **homework** for **you**?

W: That'd be **great**.

(the question and options)

What is the woman likely to do tomorrow?

- 1) Go to school.
- 2) Hand in the man's homework.
- 3) See the doctor again.
- 4) *Stay at home.* (key)

In sum, sandhi-variations, contrary to our anticipation, may help listening comprehension for EFL test takers. This can partly be attributed to the perceptual, phonological differences between accented and unaccented words in a text, which can be produced by sandhi-variations. The differences may help the accented words, usually with highly communicative value, more noticeable to the test takers, and thereby facilitate their listening comprehension.

5. 4 Findings

These results of RQ 5 and RQ 6 lead us to conclude that although overall sandhi-variation does not affect listening comprehension test performance, sandhi-variation features may have both positive and negative effects on test takers' listening comprehension. The positive effects could involve

providing more prominent phonological difference between accented and unaccented words in connected speech which are produced by sandhi-variation. This difference may reduce the level of cognitive load imposed on test takers, and thereby save their cognitive resources for higher level of processing, which is necessary to form a mental model of a text. This facilitative role of sandhi-variation identified through the present study is not consistent with the literature, and new. Negative effects may involve increasing the level of cognitive load imposed on the test takers by obscuring sounds through elision or unclear pronunciation, and disturbing speech perception or word recognition. This is consistent with the literature. These findings may reflect overall differences in the level of cognitive load imposed on the test takers in relation to the characteristic of a text (that is, including realistic level of sandhi-variation or minimal level of sandhi-variation).

Chapter 6 Discussions and conclusions

6.1 Introduction

The purpose of this study was to investigate the validity of the JNCTL. Specifically, this study explored this high-stakes test in terms of *cognitive validity* and *contextual validity* (Weir, 2005) and aimed to highlight aspects of the JNCTL that might be reformed to enhance its validity and so lead to further improvements in the listening abilities of Japanese high school learners of EFL.

In this chapter, a summary of the study is presented, and the research questions are revisited and answered in turn, and then the key findings are discussed. Recommendations are provided for improving the validity of the current JNCTL and for the development of listening comprehension tests more generally. Implications are also suggested for the teaching of listening at secondary schools in Japan. Lastly, limitations of the study are outlined and suggestions for further research are proposed.

6.2 Summary

The purpose of this study was to validate the listening comprehension component of *the Centre Test* in Japan in relation to contextual parameters and cognitive processing. The present study in the first place established a comprehensive framework of contextual parameters and a L2 listening processing model for validation purposes. This provided a theoretical framework for this study, whereby empirical evidence was elicited through document analysis, focus group interviews, and a large-scale

questionnaire administered to stakeholders. The elicited data was subjected to descriptive, quantitative, and qualitative analysis.

Preliminary studies identified ten discrepancies between real life listening features and the current JNCTL. They included the number of opportunities to listen to the input, a lack of hesitations, a lack of overlapping turns, a lack of multi-participant discussions, a lack of variety in the English accents used, a lack of L2 speakers, a lack of inference questions, a lack of non-linear texts, a lack of sandhi-variations, and a lack of natural speech rate. The results of the questionnaire revealed that sandhi-variation was perceived to be the most important parameter by stakeholders to help the current JNCTL more reflective of real life listening features. Then sandhi-variation was further explored in the main study in attempt to explore the effect on listening comprehension test performance and the level of cognitive load imposed on the test takers.

A series of experiments was conducted involving the manipulation of sandhi-variations. The results suggested that although statistically significant differences were not found in the item difficulty estimates between the sandhi plus and minus versions, sandhi-variations may involve double effects on listening comprehension test performance and the level of cognitive load for the test takers. The positive effects may lubricate word boundaries and help the listeners to parse a stream of connected speech more efficiently and thereby decrease the level of their cognitive load, whereas the negative effects may obscure the sounds through elision or unclear pronunciation and increase the level of their cognitive load for the test takers.

6.3 Answers to Research Questions

6.3.1 Research Question 1

The research question 1 was:

To what extent does the current Course of Study regulated by the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) achieve optimum validity in terms of contextual parameters and cognitive processing?

This question was investigated through a review of the Course of Study in relation to frameworks derived from the literature review. The answer to RQ 1 was that the CS does not achieve optimum validity in terms of contextual parameters and cognitive processing. The document analysis showed that the CS specifies 1800 different words, 'basic' collocations, the use of 'standard' English as a model, particular sentence patterns and grammatical elements, and places importance on grasping the main points of a text in reading and listening comprehension. The CS does not provide any specifications about discoursal, pragmatic or phonological features (excluding accents), or features connected with test format, rubric, or response format.

It is perhaps not to be expected that the CS should achieve optimum validity in relation to contextual parameters and cognitive processing since the CS is no more than a set of broad guidelines, which is promulgated by MEXT and regulates the English education at secondary schools in Japan. It is not intended to provide a level of detail that could fully inform the development of a language proficiency test (or a listening comprehension test).

For this reason English teaching at secondary schools in Japan needs a more detailed operationalisation of listening comprehension, one that will contribute to the greater contextual and cognitive validity of the JNCTL. This operationalisation will help bridge the gap between the CS/the JNCTL and listening in the real world. Further discussions are made in Section 6.4.1.

6.3.2 Research Question 2

The research question 2 was:

To what extent does the current format of the JNCTL in Japan achieve optimum validity in terms of contextual parameters and cognitive processing?

This research question was investigated through a review of the 2007-2009 forms of the JNCTL in relation to frameworks derived from the literature review. The answer to RQ 2 was that although the current JNCTL involves items calling for a wider range of cognitive processing, it does not achieve optimum validity as ten discrepancies were found between the JNCTL and real life listening features. These were:

- providing the input text twice
- a lack of hesitations
- a lack of overlapping turns
- a lack of multi-participant discussions
- a lack of variety in the English accents used
- a lack of L2 speakers
- a lack of non-linear texts

- a lack of inference questions
- a lack of sandhi-variations
- a lack of natural speech rate

Of these, five discrepancies were concerned with phonological features of the input text. They are a lack of hesitations, a lack of variety in the English accents used, a lack of L2 speakers, a lack of sandhi-variations, and a lack of natural speech rate. Since hesitations and sandhi-variations differentiate a spoken text from a written text (see for example, Rost, 2002), the shortfall in those features in the JNCTL under-represents the construct of listening comprehension. Speech rate should be prioritised as speech rate is the most predictable parameter to affect test takers' cognitive processing and thereby test performance (see for example, Buck, 2001). In addition, speech rate is strongly implicated in the occurrence of sandhi-variation (Gimson & Cruttenden, 1994) and is associated with automaticity. For this reason the compromise on speech rate in the JNCTL is most likely to threaten the validity of the listening comprehension test.

In the meantime, a variety of English accents and/or L2 speakers should not be introduced, even though test developers should in theory reflect the importance of a range of Englishes in international contexts (Kim, 2006). This is primarily because the CS acknowledges the use of 'standard' English as a model as the result of RQ 1 reveals, (see Section 6.3.1), and partly because a variety of English accents or L2 speakers are likely to impose a considerable level of cognitive load on the test takers, who are at lower proficiency level. Thus, the inclusion of a variety of English accents and L2 speakers cannot be considered achievable. Further discussions are made in Section 6.5.1.

Non-linear texts and inference questions may be crucial for improving the cognitive validity of the current JNCTL. It is often the case that in the real world listeners have to interpret beyond what is explicitly stated in the text. So it is the case that listeners have to be flexible enough to revise their initial schemata, activated by the earlier input, to build a successful mental model of a text. The JNCTL, therefore, should include a sufficient number of inference questions or non-linear texts so that it can claim to achieve greater validity than is currently the case in relation to cognitive processing.

Multi-participant discussions and overlapping turns may be considered one of the unachievable parameters to be included into the JNCTL, since they can be considerably disturbing to the test takers who are at lower proficiency level of listening and are afraid of missing a point in the high-stakes test. The decision being made to these two parameters should wait for further discussions (see section 6.3.3) and for the elicitation of stakeholders' views (see section 6.3.4).

Although the possible changes outlined above are justified from the theoretical perspective of the literature on L2 listening, we should remind ourselves at this point that the JNCTL is an achievement test and should be developed in accordance with the CS (NCUEE, 2006, 2007, 2008, 2009a, 2010), and that the CS itself does not provide any detailed operationalisation of listening comprehension (see Section 6.3.1).

We should also note that those stakeholders who are most affected by this high-stakes test (teachers and students at high schools and receiving institutions) have the right to make decisions about what the JNCTL should be like, since the JNCTL cannot itself be disembodied from a social context where the consequential use of the scores should be responsible for

resultant teaching and learning (Messick, 1996). That is, any possible changes to the JNCTL should be in a direction that is acceptable to these key stakeholders and so feasible in the social context of the test. Because of this consideration, possible changes to the current JNCTL for the purposes of this study were limited in line with stakeholders' views, which were addressed in RQ 4 (see Section 6.3.4).

6.3.3 Research Question 3

The research question 3 was

To what extent does the current format of the JNCTL reflect the Course of Study?

This question was investigated through a review of 2007 to 2009 forms of the JNCTL in relation to the CS. The answer to RQ 3 was that the JNCTL does not fully reflect the CS. The JNCTL reflects the CS in terms of vocabulary and the ability to grasp the main point of the discourse, but it is lacking a variety of language use situations specified in the CS. This result is consistent with that of RQ 2 in that the JNCTL is lacking multi-participant discussions.

The inclusion of multi-participant discussions will not only help the JNCTL reflect the CS better but also help the JNCTL achieve greater validity in relation to contextual parameters. For this reason, although it can pose a higher level of cognitive load for the test takers, the JNCTL may consider introducing multi-participant discussions.

6.3.4 Research Question 4

The research question 4 was:

Which of the discrepancies identified through RQ 1 to RQ 3 are perceived to be most important to the validity of the JNCTL by the following groups of stakeholders: high school students and high school English teachers?

This question was investigated through a series of group interviews participated by stakeholders and a large-scale questionnaire which followed. A total of ten contextual parameters were examined to answer RQ 4. Eight parameters emerged through RQ 1 and RQ 3 as discrepancies between real life listening features and the JNCTL, while two parameters (that is, sandhi-variation and natural speech rate) were chosen from the literature review.

Across stakeholder groups, sandhi-variation clearly emerged as the single most important discrepancy. This can primarily be attributable to the pervasiveness of sandhi-variation in spoken English (see Chapter 2, for example, Gimson & Cruttenden, 1994; Brown & Brown, 2006). Sandhi-variation is observed so often in real life listening that the stakeholders might have sensed that the feature should be included in the JNCTL to an extent which better reflects realistic levels. In fact, the examples used in the questionnaire (see Appendix 6A, 6B) such as *did you* [dɪdʒu:] or *tell her* [tɛlə] must have been so familiar to the stakeholders that they may have perceived as the feature to be the most important.

For another reason, the stakeholders, particularly English teachers, may have felt that a listening comprehension test cannot be suitable if it does

not include a realistic level of sandhi-variation as sandhi-variation characterises a spoken text (Brown & Yule, 1983). They may have sensed that an input text cannot be acceptable as a spoken text for listening unless it includes a realistic level of sandhi-variations.

Another finding through answering RQ 4 was that stakeholders prioritise the consistency between teaching and testing as to what changes can be made to the current JNCTL. This provides a key point of view by which to further address what recommendations can be feasible and acceptable for the future JNCTL. The recommendations are outlined in Section 6.5.1.

6.3.5 Research Question 5

Research question 5 concerned how the intra-task manipulation of specified contextual parameters would affect test performance. Research question 5 was specified into the following through Preliminary studies since sandhi-variation emerged as the single most important parameter to help make the JNCTL achieve optimum validity in the direction that would be acceptable by stakeholders.

How does the intra-task manipulation of sandhi-variation affect listening comprehension test performance for the test takers?

This question was investigated through a case study including a series of experiments. The answer to RQ 5 was that the manipulation of sandhi-variation did not substantially affect listening comprehension test performance. This result was not consistent with much of the literature, which shows that sandhi-variation disturbs listening comprehension for L2

listeners (for example, Matsusaka, 1995; Field, 2003; Brown & Brown, 2006; Enomoto, 2007; Joyce, 2008, see Section 2.3.2.2).

One possible explanation for the discrepancy between this result and the earlier research is the difficulty with isolating, if any, the effect of sandhi-variation on a multiple-choice listening comprehension test for L2 listeners, where many predictable variables affect their test performance (Freedle & Kostin, 1996; Brindley & Slatyer, 2002; Kostin, 2004; Yanagawa & Green, 2008; Iimura, 2011). The possible significant effect of sandhi-variation may have been compromised by these or other factors.

Another possible explanation may be 'trade off' effect. As the result of RQ 6 shows, sandhi-variation may have both positive and negative effects on the listening comprehension. The positive effects may involve lubricating word boundaries and provide more prominent phonological difference between accented and unaccented words in connected speech, and this difference may help the test takers to parse a stream of connected speech more efficiently. The negative effects may disturb speech perception or word recognition by obscuring the sounds through elision or unclear pronunciation. The positive effects may have been compromised by the negative effects, and vice versa.

Furthermore, even the sandhi minus versions (a version without sandhi-variations) could not help involving minimal sandhi-variations. It was by no means easy in the recording to completely control for sandhi-variations between the versions. For instance, in the utterance of '*Could you ...?*', the native speaker could not help producing [kədʒu:] instead of [kudju:] even for the recording of the sandhi minus version. Accordingly, the unavoidable inclusion of minimal sandhi-variation in the sandhi minus version may have limited the differentiation between the versions.

Lastly, the adjustment of speech rate after the piloting may have further limited the differentiation between the versions, which had been more noticeable before the piloting.

6.3.6 Research Question 6

Research question 6 was concerned with the manipulation of specified contextual parameters and how these would affect the level of cognitive load imposed on the test takers. This research question was reformulated into the following:

How does the manipulation of sandhi-variation affect the level of cognitive load imposed on the test takers?

This question was investigated through a questionnaire consisting of five items. The answer to RQ 6 was that the manipulation of sandhi-variation had different effects on the cognitive load imposed on the test takers. The questionnaire responses revealed that sounds obscured by unclear pronunciation or elision were relatively easily perceived on the sandhi plus version, suggesting that obscured sounds may have increased the cognitive load of the test takers, and disturbed word recognition and speech perception on the sandhi plus version. The result of RQ 5, namely that sixteen items were found more difficult on the sandhi plus version than on its counterpart, may be attributable to this effect.

In the meantime, sounds running together ($p < .01$) and fast speech rate ($p < .01$) were more easily perceived on the sandhi minus version, suggesting that the lack of sandhi-variation may have increased test takers' cognitive load. The result of RQ 5, namely that eight items were found more difficult on the sandhi minus version than on its counterpart, may be attributable to

this. This was not, however, consistent with the hypothesis that lack of sandhi-variation would facilitate listening comprehension.

One possible interpretation for this could be that the lack of sandhi-variation may make a chunk (a sense unit composed of words) less noticeable for test takers because the lack of sandhi-variation prevents the lubrication of word boundaries with which the learners are familiar and makes each word more distinctive than usual. We should note that 'chunking ability is essential for accurate comprehension of extended texts' (Rost, 2005, p. 517). The more distinctive words may have required more of their cognitive resources and have resulted in little or no mental capacity available for higher level processing which forms meaningful associations with existing knowledge. As a consequence, the test takers may have experienced the sounds as running together or the speech rate may have appeared relatively faster. That is, it is considered that the test takers may have perceived the sounds running together or the speech as being faster because of difficulties in chunking and following the stream of speech attributable to the lack of sandhi-variation. This interpretation implies that sandhi-variation may support listening comprehension if it lubricates word boundaries, promotes chunking more effectively, and helps speech segmentation, while it may not if it merely obscures sounds through elision or unclear pronunciation.

Thus, contrary to much of the literature and our general perception of sandhi-variation, this result shows that sandhi-variation may sometimes have a positive effect on listening comprehension in addition to the negative effects that earlier research has shown (see for example, Yamauchi, 2002).

6.4 Discussions

6.4.1 Research Questions 1 to 3

The answers to RQ 1 through 3 revealed that the JNCTL does not achieve optimum validity as ten discrepancies were found between the CS, the JNCTL and real life listening features. This is perhaps not surprising since the JNCTL was not originally designed to reflect the constructs of 'real life' listening but to reflect the ones operationalised in the CS. What the results of RQ 1 clarified, however, was that the CS does not provide any detailed specifications about the construct of listening comprehension. A question thus arises: How could NCUEE develop the JNCTL without such detailed specifications? It might be assumed that NCUEE has the specifications but it does not make them public. If this assumption is true, the specifications should be made available to the stakeholders. If NCUEE does not have them, then it should first develop some and then make them available to stakeholders, because the specifications may bridge the gap between the CS/the JNCTL and English teaching or learning at secondary schools in Japan.

Even if detailed specifications were available, however, the question would still be open as to whether it is actually possible to achieve 'optimum' validity in a language test. Weir (2005) and Shaw and Weir (2007) admit that full authenticity is unattainable in test development. Nevertheless, every effort should be made to improve on the validity of any language test, because it is our collective responsibility as language educators to maximise validity.

Of course, even if 'optimum' validity were to be achieved for the JNCTL, then that might not actually lead to any immediate improvement in high school students' listening or their practical communication abilities in English: the major goal of the introduction of the listening component into

the *Centre Test* by NCUEE in 2006. Watanabe (1996, 2004) (see Chapter 1) found no consistent washback between college entrance exams and the methods English teachers at high schools in Japan employed in their classrooms. Nevertheless, high school students' listening abilities and their practical communication abilities will be difficult to improve unless the teaching methods shift from '*Juken Eigo*' (entrance examination English) to English for communication.

One factor that may reduce the possibility of positive washback effects is that NCUEE allots only a small ratio of the overall *Centre Test* score to the JNCTL (50/250). Because of this, listening may not be considered as important as other areas such as reading or grammar by high school students intending to take the high-stakes test (see section 3.3.3.1). It is for this reason that NCUEE needs to consider allotting a higher ratio to the JNCTL than it does at present if they aim to realise the intended benefits of the introduction of the JNCTL.

Thus, there is a pressing need to make detailed specifications about the constructs of listening comprehension available, and to develop the JNCTL which achieves 'optimum' validity and is allotted a higher ratio to the overall *Centre Test* score.

6.4.2 Research Questions 4 to 6

The result of RQ 4 identified sandhi-variation as the most important parameter to help the current JNCTL achieve greater validity in a direction that would be acceptable to stakeholders, and the answers to RQ 5 and 6 showed that sandhi-variation may involve both positive and negative effects on listening comprehension for L2 listeners.

Negative effects may involve obscuring the sounds through elision or unclear pronunciation, disturbing lower level processing of speech

perception or word recognition. This may be more often the case with Japanese learners of EFL because their phonological expectations are sometimes 'unduly influenced by exposure to the written language' (Field, 2003, p. 330): Formal English education in Japan has long prioritised passive knowledge of grammar and vocabulary, and intensive word-for-word reading translation over speaking and listening (Yoshida, 2010). This is where optimum balance should be made between teaching written English for entrance examinations and teaching spoken English for communication, keeping in accordance with NCUEE's aims in introducing the JNCTL into the *Centre Test*.

Positive effects of sandhi-variation may involve lubricating word boundaries, providing more prominent phonological difference between accented forms and unaccented forms, and helping L2 listeners parsing the stream of connected speech more efficiently. This finding is new and somewhat unexpected as it is not consistent with earlier research findings and general views of sandhi-variation. English teachers and language test developers can be more positive about introducing sandhi-variation into their classrooms and into the test development. The specific approaches to teaching sandhi-variation are proposed in Section 6.6.

6.5 Recommendations from this study

6.5.1 Recommendations for the JNCTL

The present study proposes three practical recommendations for the JNCTL in relation to contextual parameters and cognitive processing. First, the JNCTL should improve its contextual validity by including more hesitations, multi-participant discussions, sandhi-variations, and a more natural speech rate, while inclusion of a variety of English accents, L2 speakers of English, and overlapping turns should wait for further

discussion and for the greater convergence between teaching and L2 listening theory. Of particular importance is the inclusion of more sandhi-variation features since the present study revealed both that it should be prioritised as a means of improving the validity of the JNCTL in a direction that would be accepted by stakeholders, and that it can affect the cognitive processing of the test takers.

Overlapping turns, a variety of English accents used, and L2 speakers are advised not to be introduced into the JNCTL until English teachers introduce these features into their classrooms, even though it would help the JNCTL achieve greater validity. This is because optimum validity should be balanced by the need for consistency between teaching and testing, especially when it comes to an achievement test based on the high school curriculum, the JNCTL. We should also note that the CS states that the language elements should be 'standard' English in principle (MEXT, 2007).

The number of opportunities to listen to the input will not be as easy to change as others, and impossible to achieve under current conditions. This is primarily because providing an input text twice or double play can compensate for unexpected disturbances such as levels of noise within and around test centres. One possible change might be to include providing different numbers of opportunities to listen to the input for different parts of the JNCTL. For instance, interactional discourse such as the dialogues in Parts 1 through 3 (see Chapter 3) could be provided twice, whereas the transactional discourse in Part 4, such as radio news broadcasts or university lectures can be provided only once. This change would better reflect listening in the real world more than the current practice of the JNCTL because the interlocutors in the real world can ask for clarifications or confirmations on the spot during the dialogues, whereas

the listeners to radio news broadcasts cannot.

Second, the JNCTL should improve its cognitive validity by including more inference questions and non-linear texts. No matter how many contextual parameters are realised, the JNCTL cannot claim to achieve optimum validity unless it also addresses cognitive validity. Inference questions and non-linear texts are essential and achievable without going against the current educational practice. One caution about the inclusion of more inference questions is that as a high school English teacher argued in the group interview (see Section 3.3.3.1) ‘inference questions presume particular cultural or societal conventional norms associated with particular speech communities, which may be irrelevant to the construct of listening comprehension.’

Third, the JNCTL should maintain the positive features that the present study has identified. The quality of items is generally high: It includes items calling for a wide range of cognitive processes and for the understanding of main points of texts distributed across those texts. These features contribute to the validity of the JNCTL in relation to cognitive processing.

In sum, it is highly recommended for the current JNCTL to achieve greater validity by reflecting contextual parameters and cognitive processing in the real world while maintaining the quality of the items.

6.5.2 Recommendations for testing listening

With the framework of contextual parameters and the L2 listening processing model, both of which have been established in the present study, this study can make a substantial contribution to the future development of listening comprehension tests.

The framework of contextual parameters will serve as a useful checklist in

the development and validation of listening comprehension tests for L2 listeners. The extensive review of the literature from a wider perspective has revealed a range of contextual parameters for listening comprehension. Furthermore, the way in which the contextual parameters are presented is innovative. They are mapped out for practical reasons under the four headings proposed by Bachman and Palmer (1996, 2010). Thus, being both comprehensive and practical, the framework can be very useful in testing listening. One good example of the use of a checklist is that the test developers should first identify a target language use situation that the test takers are expected to encounter outside the test situation, and then operationalise as many contextual parameters as possible by referring to each parameter in the checklist. This will help raise the situational authenticity of the test, and thereby contribute to the greater contextual validity of the listening test.

The established L2 listening processing model can also serve as a useful reference against which test developers can check which level of processing is required for particular language use situations. A different focus on a text allows for the development of different items requiring different processing levels. This will help raise the interactional authenticity of the test, and thereby contribute to the enhancement of the cognitive validity of the test.

Full authenticity is unattainable in test development (Weir, 2005; Shaw & Weir, 2007) insofar as it is a test: Some contextual parameters have to be compromised and others have to be at least approximated. Even so, test developers, including language teachers, can and should attempt to provide more valid listening comprehension tests by referring to the framework of contextual parameters and the L2 listening processing model that the present study has established.

The researcher's experience as a director of a prefecture-wide test in Japan for a few years suggests that few high school English teachers have the fundamental knowledge or skills to develop a listening comprehension test. In the meantime, the importance of listening comprehension tests is increasing as English for communication is strongly advocated at the policy level. The framework of contextual parameters and the L2 listening processing model will serve as a platform on which novice English teachers can embark on the development of more valid listening comprehension tests.

6.6 Implications for teaching listening

The present study provides three major pedagogical implications for teaching listening at secondary schools in Japan.

First, English teachers should be aware how complex and fragile L2 listening processing is. As Figure 2.1 illustrates, it takes an integrated process with the help of many forms of linguistic and para-linguistic knowledge to arrive at a situation model of a text. For this reason L2 listening comprehension can potentially involve many pitfalls where L2 listeners' listening comprehension may be impeded or fail. Listening comprehension, for example, may easily fail if listeners lose their attention to the input. Even if their attention is being sustained, listening comprehension may fail if word recognition is not successful (see for example, Field, 2003). Even if word recognition is successful, learners cannot form a mental model of a text if they do not have a sufficient mental lexicon (Goh, 2000; Vandergrift, 2006). That is, L2 listening processing is very fragile. English teachers should in the first place realise this.

Second, English teachers should incorporate and integrate all of the relevant parameters listed in the framework (see Table 2.7) into their

teaching of listening to spoken English. It is assumed that since high school English teachers pay little attention to teaching listening, they do not know exactly what listening difficulties the students may have, where the difficulties come from, or how the students can overcome these difficulties. Keeping the framework of contextual parameters and the L2 listening processing model at hand, and referring to them when needed, English teachers can be more analytical, informative, and instructive in guiding their students to become successful L2 listeners.

Third, given that sandhi-variations entail both positive and negative effects on test takers' listening comprehension, English teachers should introduce sandhi-variations into their teaching of listening. For example, they could support explaining in what circumstances sounds such as elision of [t] occur (for example, [wɪnə-] for *winter*), or how pronouns such as *his*, *him*, *her*, *them* tend to be pronounced as weak forms in connected speech rather than as citation forms. This knowledge may keep learners away from being disturbed by obscured sounds — presumably this is what disturbed the cognitive processing of the test takers taking the sandhi plus version in this study. Another example of a possible approach to teaching learners about sandhi-variations involves a series of exercises including introducing a dictation, identifying instances of sandhi-variations, and letting the students read the text aloud. This kind of exercise has been shown to raise their awareness of phonological features of sandhi-variation, reinforce the consistency between actual aural forms and their mental lexicon (Goh, 2000), and help them to parse the stream of the connected speech more efficiently (Shiki, Mori, Kadota, & Yoshida, 2010). Explicit knowledge about sandhi-variations and enhanced awareness of phonological features may help students tune in to a variety of sandhi-variation features in connected speech.

In sum, by realising how complex and fragile L2 listening processing is for L2 listeners, and by incorporating relevant contextual parameters including sandhi-variation into their teaching, high school English teachers are likely to improve their students' English listening ability.

6.7 Limitations of this study and suggestions for further research

In order to both accurately represent the study and to aid future work in the field, it is important to reflect on the weaknesses in the research, and the areas for future investigation. Since these issues are interrelated, they are presented together.

The first limitation concerns employing different raters in judging different contextual parameters, the components of the CS, and the sandhi-variation features in Preliminary studies. Concessions had to be made due to their availability and to the demanding nature of the rating tasks. Future studies should make every effort to use the same raters consistently throughout the research.

The second limitation concerns some contextual parameters which were not addressed. This was attributable partly to the absence of any baseline against which the raters could judge the material and partly to the difficulty of addressing the parameters effectively. Further studies should aim to conduct a more comprehensive analysis of as many contextual parameters as possible, and then investigate other parameters to see if their manipulations would make any difference to test takers' performance and the cognitive load imposed on the test takers.

The third limitation is to do with interpretations of the results of the group interviews and large-scale questionnaire. Although the responses led us to assume that listening has been still a neglected area in educational practices at secondary schools in Japan, and that real life listening features

have not been substantially introduced into their classrooms yet, this assumption has not been empirically proved yet. Further studies should explore whether or to what extent listening is taught in actual classrooms in Japan, and also whether or to what extent real life listening features including overlapping turns, hesitations, and multi-participant discussions are introduced into textbooks or other teaching materials. These studies will lend empirical support for what changes are to be made to the JNCTL.

The fourth limitation concerns the listening comprehension tests which were used in the experiment in Main study. Since even the sandhi minus version could not help involving minimal sandhi-variation as Gimson and Cruttenden (1994) suggests, the differentiation between the versions was limited to an opposition between minimal and more extensive sandhi-variation rather than between its absence and presence. Further studies should employ two different texts which are mutually exclusive with each other in relation to sandhi-variation, and compare the results with those of the present study.

The fifth limitation is to do with the questionnaire used in the present study to explore the cognitive load imposed on the test takers. Most of the respondents were so susceptible to sandhi-variation features that a ceiling effect may have occurred. In addition, the questionnaire was not originally designed to elicit differences in test takers' cognitive demands for each item, but to elicit differences in relation to the different versions of the test. As a result, it was not fine-grained enough to reveal exactly what types of sandhi-variation features may have disturbed or facilitated listening comprehension for the test takers. Further research should consider employing a wider range of methods, perhaps including interview, think-aloud protocols, or stimulated recall to elicit more accurate and extensive data on cognitive demands.

If future studies are able to overcome the limitations above, they may be in a position to present more significant results.

6.8 Conclusion

This study first established two key platforms from which to investigate the validity of the listening comprehension component of the *Centre Test* in Japan. They are:

1. a comprehensive framework of contextual parameters
2. a cognitive L2 listening processing model for operational validation purposes

Based on these platforms, this research was carried out. The results revealed that the JNCTL does not achieve optimum validity in relation to contextual parameters or cognitive processing. The discrepancies between the CS/the JNCTL and real life listening features included the number of opportunities to listen to the input, a lack of hesitations, a lack of overlapping turns, a lack of multi-participant discussions, a lack of variety in the English accents used, a lack of L2 speakers, a lack of inference questions, a lack of non-linear texts, a lack of sandhi-variations, and a lack of natural speech rate.

These results informed specific recommendations for improving the validity of the JNCTL. Priority should be placed on the introduction of more sandhi-variation to improve the contextual validity of the JNCTL. More inference questions and non-linear texts should also be employed to improve the validity in relation to cognitive processing. With the two key areas of validity improved in tandem with each other, the JNCTL could achieve greater validity overall.

Recommendations were also proposed for testing and teaching English.

English teachers should use the framework of contextual parameters and the L2 listening processing model as a useful check list and reference in an effort to develop a more valid listening test. In the teaching of listening, English teachers should first be aware how complex and fragile L2 listening comprehension is, and then should integrate relevant parameters including sandhi-variation into their teaching of English.

It is hoped that the recommendations from this study will serve to improve the validity of the JNCTL and inform and inspire English teachers, helping to trigger the intended breakthrough in English teaching and testing at secondary schools in Japan. This must contribute to enhancing the ability of high school students to listen to spoken English and so improve their practical communication abilities in English in the real world.

Appendices

Appendices: Contents

Printed

1	Coding sheet for cognitive parameters for Preliminary study	223
2	Coding sheet for practical communication ability and functions of language	224
3	Rating sheet for sandhi-variation features	225
4	Words beyond 4000 word frequency level for the 2007-2009 forms of the JNCTL	226
5	Permission for recording of the interviews	228
6A	Questionnaire to elicit a key parameter(s)	229
6B	Questionnaire to elicit a key parameter(s)-Translation	230
7	Listening comprehension test: Item set A-Test paper	232
8	Listening comprehension test: Item set A-Transcription	242
9	Listening comprehension test: Item set B-Test paper	253
10	Listening comprehension test: Item set B-Transcription	263
11A	Questionnaire to elicit the level of cognitive load imposed on the test takers	274
11B	Questionnaire to elicit the level of cognitive load imposed on the test takers-Translation	275

Audio (Listening comprehension tests)

1. sandhi plus version of item set A (A+)
2. sandhi minus version of item set A (A-)
3. sandhi plus version of item set B (B+)
4. sandhi minus version of item set B (B-)

**Appendix 1 Coding sheet for cognitive parameters
for Preliminary study**

Choose one option which you think best fits.			Part 1					Part2					Part3				Part4						
Parameter	questions	option	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17-19	20	21	22	23-25
Non-Linear	Is the text linear?	(1) Yes (2) No																					
Inference question	Does the item require the test takers to interpret the implicit meaning beyond what is explicitly stated in the text?	(1) Yes (2) No (Chose this if the item requires simple calculation or understanding anaphora)							/														
Processing level	What processing level is required to arrive at a correct answer?	(1) A level to understand word(s) or a utterance (2) A level to understand two or more than two utterances (3) A level to understand the whole text (including integrating information or using background knowledge)																					
Position of Relevant Information	Where is the relevant information found in the text?	(1) either at the beginning or the end of the text (2) in the middle (3) distributed																					

Appendix 3 Rating sheet for sandhi-variation features

Check the box you think best fits.

	very much	fairly reflect	to some extent	uncer- tain	not suffi- ciently	little	not at all
Assimilation (e.g. <i>As <u>you</u> know</i>)							
Elision (e.g. <i>next <u>station</u>, tell <u>her</u>, in<u>ternational</u>)</i>							
Weak form (e.g. <i>John <u>has</u> seen <u>it</u>.</i>)							
Formulaic expressions(e.g. <i>kina, gonna</i>)							
Intrusion (e.g. <i><u>feel</u>, <u>cooperate</u>, <u>three</u> <u>apples</u>)</i>							
Elision of voiceless plosive (e.g. <i>big ha<u>t</u>., find <u>dogs</u>)</i>							
Flap [ɾ] (e.g. <i>bet<u>ter</u>, bu<u>t</u> I, pu<u>t</u> <u>it</u> up)</i>							
Speech rate							
Holistic impression of sandhi-variation							

**Appendix 4 Words beyond 4000 word frequency level for the
2007-2009 forms of the JNCTL**

Frequency level	2007
5,000 word	closed, continuing, dot, fever, license, pill, sunny, tablet
6,000 word	carrot, mushroom, noon, novelist, ski, sore, swimming
7,000 word	microwave, snack, stripe
8,000 word	curly, racket, refrigerator
Over 8,000 word	beach-goers, bookcase, classmates, cleanup, cloudy, comer, email, one-kilometre, refrain
Proper nouns	Clayton, Grandpa, Joe, Ken, Lisa, Richard, Southwest, Thompson, Tina, Tucker, Pat

Frequency level	2008
5,000 word	appetite, exit, plan, washing
6,000 word	attendant, bloom, farmhouse, noon, pick up, ski, suitcase, swimming
7,000 word	centimetre, cone, congratulation, scoop, spacious
8,000 word	basketball, password, reunion, shrimp, striped
Over 8,000 word	all-time, brackish, cookies, country-style, fishbowl, homemade, modernized, one-dollar, organically-grown, salty, seawater, self-timer, souvenir, square-shaped, tastefully, ten-cent, twenty-dollar, vanilla
Proper nouns	A.B.Brothers, Billy, George, Hawaii, Hawaiian, Jim, Kathy, Lee, Leonardo, Maria, Mary, Mike, Northeast, Redwood, Smokey, Taylor, Timothy, Tommy, U.S., Vinci, Waverly, York

Frequency level	2009
5,000 word	annoy, closed, editorial, eventual, miniature, missing, online, popularity, waist
6,000 word	purse
7,000 word	logo, seriousness, waterfall
8,000 word	animation, basketball, elevator, penguin
Over 8,000 word	cicada, hoops, plaza, subway, website, zoo
Proper nouns	Amy, Centreville, Emily, Emma, Kato, Kyoto, Linda, Lisa, Madison, Maine, Margaret, Melissa, Singapore, Spanish, Susan, Takashi, U.S., Wakefield

Appendix 5 Permission for recording of the interview

さま

このたびは、ご多忙中にもかかわらず、インタビューにご協力を頂ましてありがとうございます。このインタビューは現在、柳川が University of Bedfordshire において作成中の PhD 論文の重要な一部を成すものです。

尚、インタビュー内容を録音することに同意していただける場合は、お手数ですが、下記の同意書に署名をお願いいたします。

署名 _____

この件に関する質問等は下記までお願いします。

柳川 浩三 (携帯 090-1669-2466, email: kozo@msh.biglobe.ne.jp)

(Translation)

Dear

I am very grateful to your participation in the interview. The interview is an important part of my dissertation to be submitted to University of Bedfordshire.

Could you please sign in the appropriate space below if you agree that the interview is recorded?

Signature: _____

For further questions,
contact with kozo Yanagawa (090-1669-2466 or kozo@msh.biglobe.ne.jp)

Appendix 6A Questionnaire to elicit a key parameter (s)

大学入試センター英語リスニング試験を
より良いものにするために

回答欄の 〇 を鉛筆やボールペンなどで塗りつぶしてください。[可: ●, ○ / 不可: ⑤, ④, ③, ②, ①]

I 大学入試センター英語リスニング試験をより良いものにするために、下の提案(1~10)についてあなたはどのように思いますか。
選択肢から一つだけ選んでマークしてください。

⑤: 賛成 ④: やや賛成 ③: やや反対 ②: 反対 ①: どちらか(どちらともない)

黒く塗りつぶしてください。

- | | | | |
|----|--|-----------|---|
| 1 | 現行では、本文が放送される回数は2回です。
それを1回にすることについてどのように思いますか。 | ⑤ ④ ③ ② ① | ● |
| 2 | 現行では、話される英語のスピードは実際のスピードよりもやや遅くなっています。
センター試験の英語のスピードを実際のスピードと同じ速さにすることについてあなたはどのように思いますか。 | ⑤ ④ ③ ② ① | ● |
| 3 | 現行では、アメリカ英語以外の英語圏の英語(イギリス英語やオーストラリア英語など)は使われていません。
アメリカ英語以外の英語圏の英語を含めることについてどう思いますか。 | ⑤ ④ ③ ② ① | ● |
| 4 | 現行では、ネイティブスピーカー(英語を母国語とする人)の話す英語しか使われていません。
ネイティブスピーカー以外の人が話す英語を含めることについてどのように思いますか。 | ⑤ ④ ③ ② ① | ● |
| 5 | 現行では、話し手の気持ちや考えが間接的・暗示的に示される対話や説明文は多くありません。
それを増やすことについてどのように思いますか。 | ⑤ ④ ③ ② ① | ● |
| 6 | 現行では、3人以上の会話(例 ディスカッション)は使われていません。
それを含めることについてどのように思いますか。 | ⑤ ④ ③ ② ① | ● |
| 7 | 現行では、言いよみや言いまじりがいを含んだ話し言葉はほとんど含まれていません。
それを取り入れることについてどのように思いますか。 | ⑤ ④ ③ ② ① | ● |
| 8 | 英語では多くの音変化が起きます。たとえば、同化(例 did you, would you)や脱落(tell her next station, camera)などの音変化を取り入れることについてどのように思いますか。 | ⑤ ④ ③ ② ① | ● |
| 9 | 現行では、会話が重なる(overlappings; 二人の人が同時に話す)ことはありません。
それを取り入れることについてどのように思いますか。 | ⑤ ④ ③ ② ① | ● |
| 10 | 現行では、話の展開が単純な場合が多いです。
話の展開がやや複雑な対話や説明を増やすことについてどのように思いますか。 | ⑤ ④ ③ ② ① | ● |

II センター英語リスニング試験をより良いものにするために、
何かご意見がありましたら自由にお書きください。
ご協力、ありがとうございました。

【自由記述欄(枠内に記述してください。)]

Appendix 6B Questionnaire to elicit a key parameter (s)-Translation

In an attempt to make the JNCTL better.

What do you think about the following changes to the JNCTL to make it better?

Choose an option which best fits your opinions.

5= agree, 4=agree to some extent, 3=disagree to some extent,

2= disagree, 1= uncertain/ cannot decide

1 At present the JNCTL allows test takers to listen to each text twice. Each text should be played only once.	5 4 3 2 1
2 At present the JNCTL are delivered at a rate that is a bit slower than natural speed. The speech rate should be natural speed.	5 4 3 2 1
3 At present all speakers included in the JNCTL speak with a standard North American accent. Other accents (for example, British or Australian) should be used in the JNCTL texts as well as the North American.	5 4 3 2 1
4 At present all speakers included in the JNCTL are native speakers of English. Non-native speakers should be included in the JNCTL texts.	5 4 3 2 1
5 At present the JNCTL hardly ask for interpretations on listeners based on the factual information in the text. The JNCTL should include more those items which call for the test takers' interpretation.	5 4 3 2 1
6 At present the JNCTL listening texts involve one or two people. The JNCTL should include texts with more than two speakers.	5 4 3 2 1
7 In natural speech speakers of English use fillers or hesitations, which are unique to spontaneous speech. They are hardly covered in the JNCTL. More	5 4 3 2 1

of these features should be included in the JNCTL texts.	
8 In natural speech speakers of English use assimilation (for example, <i>did you, would you</i>) or elision (ex. <i>tell her, next station, camera</i>). More of these features should be included in the JNCTL texts.	5 4 3 2 1
9 At present the JNCTL does not involve overlappings where two speakers speak at the same time. Overlappings should be included in the JNCTL.	5 4 3 2 1
10 At present the JNCTL texts are mostly straightforward and linear. More of non-linear texts should be included.	5 4 3 2 1

Feel free to write your opinions of the JNCTL to make it better.

Thank you for your cooperation.

Appendix 7 Listening comprehension test: Item set A

リスニングテスト A

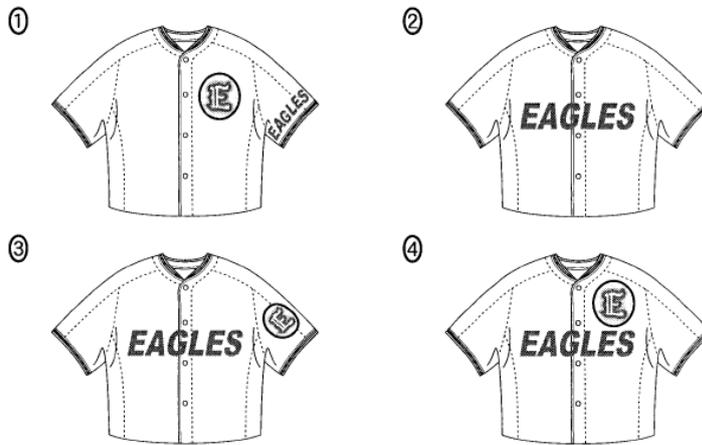
第 1 問

第 1 問は問 1 から問 9 までの 9 問です。それぞれの問について対話を聞き、
答えとして最も適切なものを、四つの選択肢のうちから一つずつ選びなさい。

問1 What time should the man come to the gate?

- 1) 4:45 2) 5:15 3) 6:45 4) 7:15

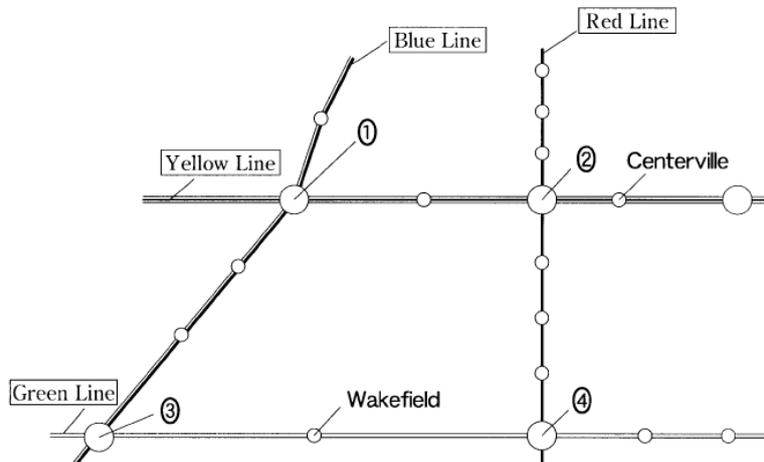
問2 What will the new uniform look like?



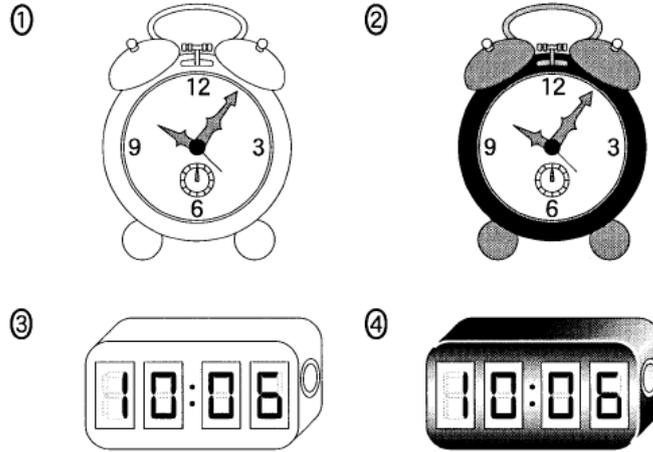
問3 What will the man do?

- 1) Record the drama. 2) Record the game.
3) Watch the drama. 4) Watch the game.

問4 Where does the man and the woman change trains first?



問5 Which clock do the man and the woman order?



問6 How much does the whole group have to pay?

- 1) \$44 2) \$52 3) \$55 4) \$65

問7 What is the woman's correct email address?

- 1) lisa.b@infomail.com
2) lisabrown@infomail.com
3) lisa-brown@infomail.com
4) lisa.brown@infomail.com

問8 What annoys the woman?

- 1) The man's voice.
2) The noise of the bed.
3) The sound of the festival.
4) The TV drama.

問9 How many vans will the group rent?

- 1) 1 2) 2 3) 3 4) 4

第2問

第2問は問10から問20までの11問です。それぞれの問について対話を聞き、最後の発言に対する相手の応答として最も適切なものを四つの選択肢のうちから一つずつ選びなさい。

問10

- 1) Let's ask the waiter to bring some more.
- 2) Let's invite the others to join us.
- 3) Let's look at a menu.
- 4) Let's sit at the table together.

問11

- 1) I see. I hope she'll be here soon.
- 2) I see. I remember her, too.
- 3) OK. She must have forgotten.
- 4) OK. We should go ahead then.

問12

- 1) It's because I have a Spanish friend.
- 2) It's going very well.
- 3) It's not so far from here.
- 4) It's really too bad I had to give it up.

問13

- 1) Let's see. Don't let it go.
- 2) Oh, how much was it?
- 3) OK. I'll take a look.
- 4) Why? Are you happy?

問14

- 1) My pleasure. Have a good day.
- 2) No, problem. I'll find it.
- 3) Not at all. I really appreciate it.
- 4) You're welcome. Try again next time.

問 15

- 1) I said "watch," didn't I.
- 2) I see. Where did you buy it?
- 3) Oh, my tie! No it isn't new.
- 4) Oh, your tie! Yes, it's nice.

問 16

- 1) Don't worry. It's friendly.
- 2) Don't worry. You'll like it.
- 3) You'll be fine. There are several entrances.
- 4) You'll be fine. You can't miss it.

問 17

- 1) Oh, no! That's more than I expected.
- 2) Oh, no! That's not enough.
- 3) OK, I think I'll buy the car.
- 4) OK, you can pay me next week.

問 18

- 1) OK, I'll find something for her to do.
- 2) OK, I'll find something for her to eat.
- 3) OK, I'll take her for a walk tonight.
- 4) OK, I'll take her tomorrow morning.

問 19

- 1) I didn't know new ones were so expensive.
- 2) So why don't we have it repaired?
- 3) Thanks for asking, but I already had some.
- 4) Well, I guess we should buy one then.

問 20

- 1) But I'd actually rather watch TV.
- 2) Don't bother making dinner now.
- 3) Oh, I get it. You must be starving.
- 4) OK. Let's watch it while we eat.

第3問 第3問はA～Cの三つの部分に分かれています。

A第3問Aは、問21から問25までの5問です。それぞれの問について対話を聞き、
答えとして最も適切なものを、四つの選択肢のうちから一つずつ選びなさい。

問21 According to the woman, what would people have to do if they had four fish?

- 1) Pay double.
- 2) Pay for four pets.
- 3) Pay for one pet.
- 4) Pay nothing.

問22 What will the man do?

- 1) Gather his belongings.
- 2) Let the woman get her things.
- 3) Move to seat 19A.
- 4) Stay in seat 9A.

問23 What does the woman say?

- 1) The man can take Saturday off this week.
- 2) The man cannot go to his friend's wedding.
- 3) The man should not work an additional day next week.
- 4) The man should tell her about his absence next week.

問24 What is the woman likely to do tomorrow?

- 1) Go to school.
- 2) Hand in the man's homework.
- 3) See the doctor again.
- 4) Stay at home.

問25 According to the man, which statement is true?

- 1) The new light bulb can be dangerous.
- 2) The new light bulb is unbreakable.
- 3) The new light bulb isn't sold yet.
- 4) The new light bulb never needs to be replaced.

B 第 3 問 B は、問 26 から問 28 までの 3 問です。長めの対話を一つ聞き、問に対する答えとして最も適切なものを、下の 1)~6) の六つの選択肢のうちから一つずつ選びなさい。

対話の場面: 二人の友人が、赤ん坊の名前について話しています。
 問い: 下の表の [26] ~ [28] にあてはまる名前はどれですか。

Top Three Female Baby Names by Decade

Decade	RANK 1	RANK 2	RANK 3
2000s	Emily	Madison	Emma
1990s	Jessica	Ashley	[28]
1980s	Jessica	[27]	Amanda
1970s	Jennifer	Amy	Melissa
1960s	Lisa	Mary	Susan
1950s	Mary	Linda	Patricia
1940s	Mary	Linda	Barbara
1930s	[26]	Betty	Barbara

選択肢

- 1) Emily 2) Emma 3) Jennifer 4) Linda 5) Madison 6) Mary

C 第3問Cは、問29から問30までの2問です。長めの対話を一つ聞き、問に対する答えとして最も適切なものを、下の1)~6)の六つの選択肢のうちから一つずつ選びなさい。

対話の場面; 図書館で学生と司書が文学作品の読者数について話しています。
問い: 下の表の[29] と[30]にあてはまる数値はどれですか。

Age Group	Change in Number of Readers (%)
all adults	+ 7
18-24	[29]
25-34	+ 5
35-44	+ 9
45-54	[30]
55-64	+ 9
65-74	+ 8
75+	+ 15

選択肢 1) -9 2) -3 3) +3 4) +9 5) +15 6) +21

第4問 第4問はA～Cの三つの部分に分かれています。

A第4問Aは、問31から問35までの5問です。それぞれの問について英語を聞き、
答えとして最も適切なものを、四つの選択肢のうちから一つずつ選びなさい。

問31 How does the speaker feel about
the sound of cicadas?

- 1) He feels grown-up.
- 2) He feels surprised.
- 3) He finds it disturbing.
- 4) He finds it relaxing.

問32 What does the speaker want
Takashi to do?

- 1) To apologize for losing her gloves.
- 2) To call her back as soon as possible.
- 3) To look for something she can't find.
- 4) To send her green coat to her in
Kyoto.

問33 What is today's special event?

- 1) The animation characters' show.
- 2) The basketball game with dolphins.
- 3) The Penguin Village opening.
- 4) The train ride to the waterfall.

問34 What can the participants receive
after the cleanup?

- 1) A bag and gloves.
- 2) A movie ticket.
- 3) A thank-you letter.
- 4) Free drinks and snacks.

問35 According to the photographer, which of the following is true?

- 1) Doing background research about the area is essential.
- 2) Freezing condition should be avoided.
- 3) It is best to leave it to chance when taking pictures.
- 4) It is important to stay on the main path.

B第4問Bは、問36から問38までの3問です。これから流れる英語を聞き、それぞれの問いの答えとして最も適切なものを、四つの選択肢のうちから一つずつ選びなさい。

問36 How did the journal *The Global Village* get started?

- 1) Conference organisers wanted to continue to exchange ideas.
- 2) It was one of the major purposes of the conference.
- 3) The department wanted to get articles from students.
- 4) The original members wanted to host another conference.

問37 Why did a particular student become the assistant editor?

- 1) The student attended every day of the conference.
- 2) The student had experience editing on the Web.
- 3) The student knew more about global issues than others.
- 4) The student wrote the best response to the blog.

問38 What is the ultimate goal of the journal?

- 1) To advertise the new program in global journalism.
- 2) To educate people in general about the use of blogs.
- 3) To give training in computer editing to future academics.
- 4) To make young people's voices heard by influential people.

☐第4問Cは、問39と問40の2問です。これから流れる英語を聞き、それぞれの問いの答えとして最も適切なものを、四つの選択肢のうちから一つずつ選びなさい。

問39 Who is most likely to be the speaker?

- 1) A flight attendant.
- 2) A tour guide.
- 3) A shop clerk.
- 4) A photographer.

問40 Why does the castle attract many tourists?

- 1) Because it is the oldest.
- 2) Because it is close to a souvenir shop.
- 3) Because it is well looked after.
- 4) Because it is very famous.

Appendix 8 Listening comprehension test: Item set A-Transcription

() indicates the serial item number across the two item sets (A & B).

59 to # 69 are anchor items.

1 (# 26)

M: Is this the gate for Flight 557 leaving for Singapore at 5:00 pm?

W: Yes, but your flight has been delayed.

M: Really? How long?

W: Two hours. Please come back 15 minutes before departure.

2 (# 27)

W: Let's put the team name on the sleeve of the new uniform.

M: Maybe the name should be on the front.

W: OK. How about putting the logo on the sleeve?

M: Good idea.

3 (# 28)

M: Isn't the soccer game starting now?

W: Yeah, but my favourite drama comes on in an hour on another channel.

M: So, should I record the game?

W: That's OK. I'll record my program.

4 (# 29)

W: Let's take the subway from Centreville to Wakefield.

M: I don't like the Red Line because it only runs every 30 minutes.

W: Then we should take the Blue Line.

M: That's a good idea.

5 (# 30)

M: Let's get a new alarm clock.

W: Yes, this white one looks really easy to read.

M: They have a black one with big hands, too.

W: OK, let's order that one.

6 (# 31)

M: The sign says admission is five dollars each.

W: But since we're a group, we can save a dollar on each ticket.

M: We have eleven students and teachers are free.

W: Sounds good.

7 (# 51)

W: No wonder I didn't get your email! You typed my address wrong.

M: What's wrong with it?

W: You need a dot between 'Lisa' and 'brown.'

8 (# 59)

W: Why are you talking so loudly?

M: I'm practicing my part for the drama festival.

W: OK. Can you finish by nine? I want to go to bed early.

M: Sure.

9 (# 60)

W: We'd like to rent a van next Saturday.

M: Sure, we have some available.

W: We have eighteen people. Is one van enough?

M: Well, each van carries up to twelve.

10 (# 32)

W: Where should we sit?

M: How about that table?

W: But there aren't enough chairs for all of us.

11 (# 33)

M: Is everybody here now?

W: We're still missing Margaret.

M: Oh, now I remember. She said she can't make it.

12 (# 34)

M: I heard you're studying Spanish.

W: Yes, I listen to a radio program every day.

M: Great! How do you like it so far?

13 (# 35)

M: Hurry up! We're running late!

W: Hold on. Where did I put my purse?

M: I saw it in the kitchen.

14 (# 36)

M: Excuse me. How can I get to Mr. Kato's office?

W: Take the elevator to the third floor. It's the first room on your right.

M: Thank you.

15 (# 37)

W: Wow! That's nice. Is it new?

M: Yes, thanks! I just bought it. It fits perfectly around my wrist, and it keeps accurate time.

W: No, no. I meant your tie.

16 (# 38)

W: Where should I meet you?

M: The south entrance. There're three entrances to the zoo. The south entrance has a big picture of a monkey on it.

W: Hope I can find it.

17 (# 56)

W: Well Joe, how much is it going to cost me to fix this old car of mine?

M: Hmn, since the engine needs major repairs, it'll be at least 3,000 dollars.

18 (# 57)

M: I'm worried about the dog.

W: Yeah, she hasn't eaten anything for two days.

M: Maybe we should take her to Dr. Thompson.

19 (# 61)

M: The coffee maker's broken.

W: Can we get it fixed?

M: That would cost more than getting a new one.

20 (# 62)

W: Time for my favorite TV show.

M: Come on, dinner is ready.

W: Oh, but I really don't want to miss it.

21 (# 39)

W: Have you heard about the new law? When you rent an apartment and have a pet, it'll cost 50 dollars extra a month.

M: So, with a cat and a dog you'd pay double?

W: Right.

M: What about four fish in a bowl? Would you have to pay 200 dollars?

W: Well, a pet is a pet.

22 (# 40)

W: Sorry, but isn't this 9A?

M: Yes, it is.

W: Well, I've reserved this seat.

M: But my ticket says... oops, I reserved 19A. Sorry, I'll move, but first let me gather my things.

W: Uh, don't bother. I'll just sit in the seat you reserved.

M: Thank you.

23 (# 41)

M: Could I take a day off tomorrow?

W: Well, we're pretty busy on Saturdays.

M: I know, but I'd like to attend my friend's wedding. I can work an extra day next week instead.

W: I guess that'll work, but next time let me know earlier.

24 (# 63)

M: What's wrong, Mary?

W: I have a terrible cold. Today my doctor told me to stay in bed for a few days and see her again if I'm still sick.

M: So, you won't be going to school tomorrow. Do you want me to turn in your homework for you?

W: That'd be great.

25 (# 64)

M: There's a new type of light bulb for sale that lasts almost forever.

W: You mean you don't need to replace it so often?

M: That's right.

W: Wow, I should get some.

M: Yeah, but you still have to be careful not to break them because they contain poisonous material.

W: I'll keep that in mind.

26-28 (# 42-44)

M: Look at this website. It shows the top three female baby names in the U.S. by decade.

W: Is my name there?

M: Let's see . . . your name was the most popular in the 70s.

W: That's when I was born. Look, Amy and Melissa were also popular then. And I have two friends with those names.

M: And your name was still popular in the 80s. It was ranked second.

W: My grandmother's name was the most popular when she was born in the 30s, and stayed the most popular for another two decades.

M: *Linda* was very popular back then, too.

W: And there are some names that only appear once.

M: Yeah, like Lisa and Susan in the 60s.

W: How about now?

M: The top three are Emily, Madison and Emma.

W: Hmm . . . Emily went from third to the most popular in one decade.

M: It's interesting to see how the popularity of names changes over the years.

29-30 (# 65-66)

W: I'm interested in reading trends in the U.S.

M: Here's a table showing the changing number of American readers between 2002 and 2008.

W: Hmm... Interesting. There was an overall increase of seven per cent.

M: That's right. The number of readers increased in most age groups. Only readers between the ages of 45 and 54 decreased in number by three per cent.

W: I wonder why.

M: It's hard to explain, because readers near that age group increased. Those aged 35 to 44 increased by nine per cent, as did people in their late 50s to early 60s.

W: I noticed my grandmother started reading for pleasure after she retired.

M: I'm not surprised. The elderly aged 75 and older showed the second largest increase of fifteen per cent.

W: What about young people?

M: They tend to read books on the Internet. That explains why people in their late teens and early twenties showed the largest increase twenty-one per cent.

W: That's helpful information. Thanks.

31 (# 45)

When I first came to Japan, I was surprised when people asked if I was annoyed by the sounds of insects such as cicadas. They had heard many western people couldn't stand such sounds. Actually, the sound of cicadas-which are called semi in Japanese-makes me feel at home. When I was growing up, the cicadas would sing in a tree just outside my bedroom window all summer long. I would lie on my bed, listening to their peaceful song. When I heard the cicadas during my first summer in Japan, it brought back happy childhood memories.

32 (# 46)

Hello, Takashi? This is Rose. I'm in Kyoto now. I enjoyed staying with you and your family last week. Sorry to bother you, but I've got a problem. I can't find my gloves. Have you seen them? Maybe I left them on the table in the bedroom, but I'm not sure. They're green and match my coat. If you have them, can you please call me here? The number is, oh, I hardly stay in my hotel room, so I'll contact you again tomorrow. Thanks. Talk to you later.

33 (# 47)

Welcome to Maine Park. Before you begin your marine adventure, we have a few announcements to make. As always, from 12:30, you can see our very popular sea animal show with dolphins jumping through hoops and playing basketball. We are sorry that Penguin Village is closed today, but starting at 3:00, in addition to our usual attractions, you can see your favourite animation characters sing and dance in the plaza in front of the waterfall. You can get there easily if you take the miniature train from the information centre. Don't miss this exciting event!

34 (# 58)

Good morning, everybody. Welcome to Central City's annual community cleanup day. Today we'll pick up litter along a one-kilometre section of the river. Pick up a bag and a pair of gloves before you get started. At 10:30 we'll take a break for cold drinks and snacks. After we finish around noon, stop by the Community Centre to receive a fun ticket for a movie of your choice at any of our local theatres. This is our way of saying thank you for volunteering.

35 (# 67)

As a professional photographer I would like to give you some suggestions for successful landscape photography. In winter, for example, when the days are short, you need to know where you're going and what you want to photograph. You can get familiar with the area you're planning to visit by reading guidebooks and studying maps. Then, you'll know beforehand where the most attractive locations are, rather than leaving it to chance. At the location, you may need to get off the main path, so you should be careful. To take a good photo, it maybe necessary to be in freezing conditions which might be dangerous.

36-38 (# 48-50)

As the head of the International Relations Department I am pleased to announce our new online journal, *The Global Village*. After hosting an international conference on global issues, the department started a biog to stay in touch and communicate about important issues such as the energy crisis and human rights. Soon our biog started to attract exciting ideas and articles from journalists, scholars and even students who had heard about us. In fact, some of the best responses to our blog were from non-academics and students who acknowledged the seriousness of these issues.

Therefore, when we decided to form the editorial group, students were also included. A professor in our department was chosen as the chief editor, but the assistant editor is a student with previous experience editing an online journal.

The Global Village will be a bridge between journalists, scholars and students who share a common interest in global issues. Our eventual objective is to make world leaders aware of our concerns, especially those of young people. They're the ones who will inherit the earth. We would like those leaders to realize this. Thank you for your time.

39-40 (# 68-69)

Ladies and gentleman, this is one of the nine gates to the castle. It's particularly beautiful from this angle. It's the oldest of the gates, but they did some repairs to that door about twenty years ago. There are many reasons this castle attracts a lot of tourists. One of them is that the buildings are always in very good repair. OK. Let's move on. Our next stop will be the souvenir shop.

Appendix 9 Listening comprehension test: Item set B-Test paper

リスニングテスト B

第 1 問

第 1 問は問 1 から問 9 までの 9 問です。それぞれの問について対話を聞き、
答えとして最も適切なものを、四つの選択肢のうちから一つずつ選びなさい。

問 1 When is the book report due?

- 1) Tuesday 2) Wednesday 3) Thursday 4) Friday

問 2 Which pair of glasses is the man looking for?



問 3 Where is the man's hometown?



問 4 How much are the stamps?

- 1) \$2.50 2) \$5.10 3) \$17.50 4) \$20.00

問5 Which one is the man ordering?

①



②



③



④



問6 Which password is correct?

- 1) cB234 2) CB234 3) Cb2345 4) CB2345

問7 What time is it now?

- 1) 5:00 2) 5:30 3) 6:00 4) 6:30

問8 What annoys the woman?

- 1) The man's voice.
2) The noise of the bed.
3) The sound of the festival.
4) The TV drama.

問9 How many vans will the group rent?

- 1) 1 2) 2 3) 3 4) 4

第2問

第2問は問10から問20までの11問です。それぞれの問について対話を聞き、最後の発言に対する相手の応答として最も適切なものを、四つの選択肢のうちから一つずつ選びなさい。

問10

- 1) OK, I hope you can join us next time.
- 2) OK, I'll let the others know you're coming.
- 3) OK, I'm glad you can make it.
- 4) OK, we'll wait for your to pick us up.

問11

- 1) No, I just brought it last week.
- 2) No, I just want to use the restroom.
- 3) No, I'm just going on a long vacation.
- 4) No, I'm sure it belongs to you.

問12

- 1) Oh, I didn't know you liked to take chances.
- 2) Oh, I didn't know your son was a pilot.
- 3) That's great. So, how long have you been working there?
- 4) That's great. So, what countries have you been to?

問13

- 1) No, I don't have someone special.
- 2) No, I just can't find her anywhere.
- 3) Yes, it's my birthday today.
- 4) Yes, it's our 20th wedding anniversary.

問14

- 1) So, what have you learned to make so far?
- 2) So, what should I make at the school tonight?
- 3) So, when are you going to start taking lessons?
- 4) So, why did you miss the lesson tonight?

問 15

- 1) No, I didn't really need it.
- 2) No, I really mean it.
- 3) No, I'm really glad I could help.
- 4) No, it wasn't really a success.

問 16

- 1) All right. I'd appreciate your help.
- 2) OK, give me a few minutes, and I'll have a look.
- 3) OK, give me the map, and I'll show you the way.
- 4) You're right. You can solve them by yourself.

問 17

- 1) All right, I'll ask him to meet you.
- 2) All right, I'll call him at 4:00.
- 3) All right, I'll see him.
- 4) All right, I'll tell him.

問 18

- 1) Yes, but I don't like Chinese food.
- 2) Yes, but I don't want to try a different place.
- 3) Yes, but they've raised their prices.
- 4) Yes, but we've never eaten there before.

問 19

- 1) I didn't know new ones were so expensive.
- 2) So why don't we have it repaired?
- 3) Thanks for asking, but I already had some.
- 4) Well, I guess we should buy one them.

問 20

- 1) But I'd actually rather watch TV.
- 2) Don't bother making dinner now.
- 3) Oh, I get it. You must be starving.
- 4) OK. Let's watch it while we eat.

第3問 第3問はA～Cの三つの部分に分かれています。

A第3問Aは、問21から問25までの5問です。それぞれの問について対話を聞き、
答えとして最も適切なものを、四つの選択肢のうちから一つずつ選びなさい。

問21 What will the woman buy first?

- 1) A suitcase.
- 2) A TV.
- 3) A washing machine.
- 4) Some clothes.

問22 What will Billy do?

- 1) Borrow his mother's cellphone.
- 2) Borrow his mother's watch.
- 3) Buy a new battery.
- 4) Use his own cellphone.

問23 Where will the woman get on the
bus?

- 1) At Central Bus Station.
- 2) At Taylor Hall.
- 3) At the Crown Theatre.
- 4) At the Redwood Hotel.

問24 What is the woman likely to do
tomorrow?

- 1) Go to school.
- 2) Hand in the man's homework.
- 3) See the doctor again.
- 4) Stay at home.

問25 According to the man, which statement is true?

- 1) The new light bulb can be dangerous.
- 2) The new light bulb is unbreakable.
- 3) The new light bulb isn't sold yet.
- 4) The new light bulb never needs to be replaced.

B第3問Bは、問26から問28までの3問です。長めの対話を一つ聞き、問に対する答えとして最も適切なものを、下の写真の六つの選択肢のうちから一つずつ選びなさい。

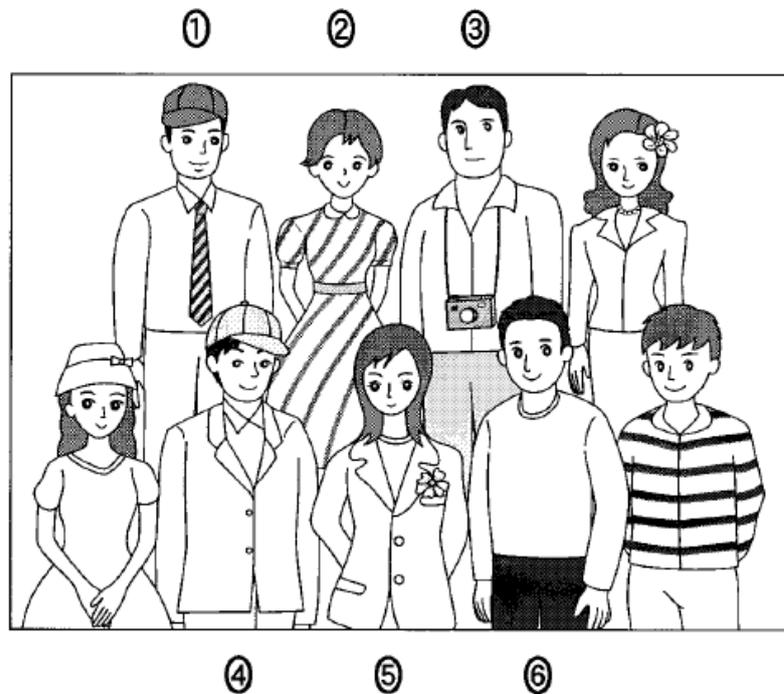
対話の場面: Kathy と父親が、先日撮った写真を見ながら電話で話しています。

問い: 下の三人(Sue, Tommy, George)はそれぞれ写真の何番の人にあたりますか。

問26 Sue

問27 Tommy

問28 George



☐第3問Cは、問29から問30までの2問です。長めの対話を一つ聞き、問に対する答えとして最も適切なものを、下の1)-6)の六つの選択肢のうちから一つずつ選びなさい。

対話の場面; 図書館で学生と司書が文学作品の読者数について話しています。
問い; 下の表の[29] と[30]にあてはまる数値はどれですか。

Age Group	Change in Number of Readers (%)
all adults	+ 7
18-24	[29]
25-34	+ 5
35-44	+ 9
45-54	[30]
55-64	+ 9
65-74	+ 8
75+	+ 15

選択肢 1) -9 2) -3 3) + 3 4) + 9 5) + 15 6) + 21

第4問 第4問はA～Cの三つの部分に分かれています。

A第4問Aは、問31から問35までの5問です。それぞれの問について英語を聞き、
答えとして最も適切なものを、四つの選択肢のうちから一つずつ選びなさい。

問31 What is so special about Maria M.?

- 1) She has just reached the top of the charts.
- 2) She is a popular record-breaking singer.
- 3) She is an Italian who sings country music.
- 4) She will do a recording with the A.B. Brothers.

問32 What is this message
advertising?

- 1) A country home for sale.
- 2) A health food restaurant.
- 3) A sports training centre.
- 4) A vacation resort.

問33 When is a good time to visit the
museum?

- 1) Monday at 1 p.m.
- 2) Tuesday at 6 p.m.
- 3) Thursday at 7 p.m.
- 4) Saturday at 10 a.m.

問34 What should Pat take after eating if she has a sore throat and a slight fever?

- 1) One yellow tablet and two green pills.
- 2) Three green pills.
- 3) Two yellow tablets.
- 4) Two yellow tablets and one green pill.

問35 According to the photographer, which of the following is true?

- 1) Doing background research about the area is essential.
- 2) Freezing condition should be avoided.
- 3) It is best to leave it to chance when taking pictures.
- 4) It is important to stay on the main path.

B第4問Bは、問36から問38までの3問です。これから流れる英語を聞き、それぞれの問いの答えとして最も適切なものを、四つの選択肢のうちから一つずつ選びなさい。

問36 Where does this kind of shrimp usually live?

- 1) In black-colored water on the shore.
- 2) In mineral water from Hawaii.
- 3) In slightly salty water by the sea.
- 4) In the sea near Hawaii.

問37 What did the speaker do when she was given the shrimp as a souvenir?

- 1) She decided to buy a new fishbowl for them and left them on the kitchen table.
- 2) She didn't know what to do and decided to ask her mother the following morning.
- 3) She put them in a clear plastic bottle, which she left on the table by her bed.
- 4) She put them in a glass fishbowl that she found in the kitchen cupboard.

問38 Why did the speaker run to the kitchen the following morning?

- 1) Her mother wanted her to remove the bottle right away.
- 2) Her mother called out to her that coffee was ready.
- 3) She realized that her pets might be in danger.
- 4) She was eager to feed the shrimp.

☐第4問Cは、問39と問40の2問です。これから流れる英語を聞き、それぞれの問いの答えとして最も適切なものを、四つの選択肢のうちから一つずつ選びなさい。

問39 Who is most like to be the speaker?

- 1) A flight attendant.
- 2) A tour guide.
- 3) A shop clerk.
- 4) A photographer

問40 Why does the castle attract many tourists?

- 1) Because it is the oldest.
- 2) Because it is close to a souvenir shop.
- 3) Because it is well looked after.
- 4) Because it is very famous.

Appendix 10 Listening comprehension test: Item set B-Transcription

() indicates the serial item number across the two item sets (A & B).

59 to # 69 are anchor items.

1 (# 1)

W: You didn't finish your book report yet?

M: Well, it's only Tuesday. I still have three days.

W: No, it's due the day after tomorrow.

M: Uh-oh.

2 (# 2)

M: I lost my glasses. Has anyone turned them in?

W: Can you describe them?

M: Sure the lenses are square-shaped, and uh, the parts that fit over the ears are rounded.

W: OK, I'll check.

3 (# 3)

W: Where are you from?

M: I'm from the U.S.

W: What part?

M: The Northeast. I'm from a small town on the coast, south of New York City.

4 (# 4)

W: Hi, I'd like five ten-cent stamps and two one-dollar stamps, please.

M: All right. Anything, else?

W: No, but I only have a twenty-dollar bill.

M: No problem.

5 (# 5)

M: Hi, I'd like a scoop of chocolate ice cream.

W: In a cup or cone?

M: A cone, please. Oh, and could you add a scoop of uh, vanilla?

W: Sure.

6 (# 6)

M: OK. I'm ready to log on.

W: Remember, for your password, you need two letters and three numbers.

M: Right.

W: And make sure both letters are capitals.

7 (# 55)

W: Where have you been? I've been waiting here so long.

M: Sorry. I couldn't get out of the office until 5:30.

W: But that was half an hour ago.

8 (# 59)

W: Why are you talking so loudly?

M: I'm practicing my part for the drama festival.

W: OK. Can you finish by nine? I want to go to bed early.

M: Sure.

9 (# 60)

W: We'd like to rent a van next Saturday.

M: Sure, we have some available.

W: We have eighteen people. Is one van enough?

M: Well, each van carries up to twelve.

10 (# 7)

W: Mike, are you coming with us?

M: I wish I could, but something came up.

11 (# 8)

W: Hey Jim, could you keep an eye on my bag for me?

M: Uh, will you be long?

12 (# 9)

W: You really like to travel abroad, don't you?

M: Well, I have a lot of chances because my son manages a travel agency.

13 (# 10)

W: Could I help you find something sir?

M: Yes, I need something for my wife.

W: Is it a special occasion?

14 (# 11)

W: Oh, I almost forgot. I've got a cooking lesson tonight.

M: You're going to cooking school?

W: Yeah! It's really fun!

15 (# 12)

W: Congratulations! The project was a big success.

M: Thanks, but I couldn't have done it without your help.

W: Oh, you're just being modest.

16 (# 13)

W: Could you help me with something?

M: Uh, I'm kind of busy. What is it?

W: I'm trying to do these math problems, but I'm totally lost.

17 (# 52)

M: Ms Tucker, a Mr. Richard Clayton is on the phone.

W: I'm a little busy right now. Could you have him call me back after four o'clock?

18 (# 53)

M: How about going to the Chinese restaurant for dinner?

W: Let's try a different restaurant tonight.

M: Why? I thought that was your favourite place.

19 (# 61)

M: The coffee maker's broken.

W: Can we get it fixed?

M: That would cost more than getting a new one.

20 (# 62)

W: Time for my favourite TV show.

M: Come on, dinner is ready.

W: Oh, but I really don't want to miss it.

21 (# 14)

W: I was going to buy a new TV this weekend, but my washing machine broke down, and it's too old to fix. I also wanted a new suitcase for my trip next week, but I can't afford everything now.

M: So, what'll you do?

W: Well, clean clothes are the most important thing.

22 (# 15)

W: What's wrong, Billy?

M: My watch stopped, Mom.

W: Maybe you need a new battery. I can get it changed for you tomorrow.

M: OK thanks.

W: Here, use mine for now. I can check the time on my cell phone.

M: Why didn't I think of that? I've got a cellphone, too.

23 (# 16)

W: When does the next city tour start?

M: Let's see. It leaves Central Bus Station at 1:00, and there are several pick-up points along the way.

W: Does it stop at the Redwood Hotel?

M: Well, it stops at Taylor Hall and the Crown Theatre.

W: I think it's easier to start from the beginning.

24 (# 63)

M: What's wrong, Mary?

W: I have a terrible cold. Today my doctor told me to stay in bed for a few days and see her again if I'm still sick.

M: So, you won't be going to school tomorrow. Do you want me to turn in your homework for you?

W: That'd be great.

25 (# 64)

M: There's a new type of light bulb for sale that lasts almost forever.

W: You mean you don't need to replace it so often?

M: That's right.

W: Wow, I should get some.

M: Yeah, but you still have to be careful not to break them because they contain poisonous material.

W: I'll keep that in mind.

26-28 (# 17-19)

M: Thanks for calling, Kathy. I got your letter and the photo of your class reunion last week.

W: Good. I have the same one. Mike took the picture.

M: Oh, that's nice. You look good in your white suit with the flower on the pocket. Who's that in the striped dress? Mary?

W: No, she couldn't make it. She's a flight attendant, and she had to work. That's Sue, who used to live on Pine Street.

M: Who's the handsome young guy next to you?

W: Which one? The one with the cap is Timothy Bloom. He was our star basketball player, remember? And the other one is Tommy Lee. He's in law school now.

M: Too bad Mike's not in the picture with all of you.

W: Yeah, he didn't have a self-timer on his camera. You can see George had a camera, too, but he didn't take any pictures of us.

M: Oh that's just like George.

29-30 (# 65-66)

W: I'm interested in reading trends in the U.S.

M: Here's a table showing the changing number of American readers between 2002 and 2008.

W: Hmm... Interesting. There was an overall increase of seven per cent.

M: That's right. The number of readers increased in most age groups. Only readers between the ages of 45 and 54 decreased in number by three per cent.

W: I wonder why.

M: It's hard to explain, because readers near that age group increased. Those aged 35 to 44 increased by nine per cent, as did people in their late 50s to early 60s.

W: I noticed my grandmother started reading for pleasure after she retired.

M: I'm not surprised. The elderly aged 75 and older showed the second largest increase of fifteen per cent.

W: What about young people?

M: They tend to read books on the Internet. That explains why people in their late teens and early twenties showed the largest increase twenty-one per cent.

W: That's helpful information. Thanks.

31 (# 20)

Good morning! This is Pop Music Top Twenty! Have we got a surprise for you: a special guest right here in our studio! Last week, she broke the record to become the first singer from another country to stay at the top of the charts for seventeen weeks in a row. Plus, she's only two weeks away from the all-time record set by the rock band the A.B. Brothers seven years ago. By now you should know who our guest is. Yes, it's the one and only Maria M. from Italy!

32 (# 21)

Waverly Hills welcomes you to our spacious grounds and facilities right in the middle of the Smokey Mountains. The old farmhouse has been completely modernized with tastefully decorated rooms. There's plenty to do in the area. You can enjoy swimming in the summer, skiing in the winter, or a run through the woods. You can work up a good appetite to enjoy country-style meals prepared in our own kitchen with fresh organically-grown vegetables. Between meals you can join us for a glass of fresh milk and homemade cookies. We're often fully booked, so make your reservations early.

33 (# 22)

Attention please. It is now 5:30 and the museum will close in 30 minutes. When you have finished looking at the exhibits in the room you are in, please make your way to the exit. We hope you have enjoyed the exhibition. For your information the museum is open every day including holidays. Opening hours are from 10 a.m. to 6 p.m. Monday through Thursday. From Friday to Sunday the museum is open from noon to 9 p.m. The coming exhibition on Leonardo da Vinci will start on September 10th. We hope you will visit us again soon.

34 (# 54)

OK, Pat. Here's the medicine for your cold. There are two kinds. These yellow tablets are for your sore throat. Take two of them three times a day, after each meal. Take one of these green pills after meals only when you have a very high fever. They'll bring your temperature down to normal. They're strong, so take no more than three a day.

35 (# 67)

As a professional photographer I would like to give you some suggestions for successful landscape photography. In winter, for example, when the days are short, you need to know where you're going and what you want to photograph. You can get familiar with the area you're planning to visit by reading guidebooks and studying maps. Then, you'll know beforehand where the most attractive locations are, rather than leaving it to chance. At the location, you may need to get off the main path, so you should be careful. To take a good photo, it maybe necessary to be in freezing conditions which might be dangerous.

36-38 (# 23-25)

The other day, a friend of mine came back from Hawaii and gave me some live Hawaiian Red Shrimp. They're really thin and tiny, less than one centimetre long, but if you look closely, you'll see that they're shaped like any other shrimp. They live in pools of "brackish water," which is slightly salty water found along the shore where fresh water from the land and sea water mix. I was given about twenty of them as a souvenir, in a small, clear plastic bottle the kind that mineral water comes in. The shrimp were playing around in the bottle. I planned to go and find a nice glass fishbowl on the weekend so my new pets could swim in it and I could enjoy watching them swim around. I left the bottle on the kitchen table and went to bed.

The next morning when I woke up, I suddenly remembered that my mother makes coffee every morning, using mineral water. I ran to the kitchen, and there she was, holding my bottle in her hand. I was about to say 'Stop!', when she said, "Ugh, something's moving in this water!" So I told her the story and prevented a disaster just in time.

39-40 (# 68-69)

Ladies and gentleman, this is one of the nine gates to the castle. It's particularly beautiful from this angle. It's the oldest of the gates, but they did some repairs to that door about twenty years ago. There are many reasons this castle attracts a lot of tourists. One of them is that the buildings are always in very good repair. OK. Let's move on. Our next stop will be the souvenir shop.

**Appendix 11A Questionnaire to elicit the level of cognitive load
imposed on the test takers**

リスニングテスト後 Jun 13 version *学籍番号は下4桁のみをたてに記入して下さい

回答欄の 0 を鉛筆やボールペンなどで塗りつぶしてください。[可: ●, ○ / 不可: □, ○, ◎, ⊙]

番号	① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨ ⑩
	① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨ ⑩
	① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨ ⑩
	① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨ ⑩
氏名	



I 受験したリスニングテストに関して、以下の1-5の質問項目に対し、あなたの感覚にもっとも近いものを選択肢の中から選び、その番号を塗りつぶしてください。

⑤: そう思う ④: ある程度そう思う ③: どちらともいえない ②: あまりそう思わない ①: そう思わない

		塗りつぶしてください。
1	全体的に、音がつながっている気がした。	⑤ ④ ③ ② ①
2	単語と単語の切れ目がわかりづらかった。	⑤ ④ ③ ② ①
3	弱くあいまいに発音されている部分が多かった。	⑤ ④ ③ ② ①
4	発音がとこどこで省略されている感じがした。	⑤ ④ ③ ② ①
5	話すスピードが速かった。	⑤ ④ ③ ② ①

**Appendix 11B Questionnaire to elicit the level of cognitive load
imposed on the test takers-Translation**

*Question: Choose the option which best matches your perception about
the listening comprehension test you have just taken.*

5=agree, 4= agree to some extent, 3=uncertain (cannot decide),
2=disagree to some extent, 1=disagree

1 I felt that overall the sounds ran together.	5 4 3 2 1
2 I found it difficult to recognise when one word finished and the next word began.	5 4 3 2 1
3 I felt many of the words were unclearly and ambiguously pronounced.	5 4 3 2 1
4 I felt some words were elided or dropped.	5 4 3 2 1
5 I felt the speech rate was fast.	5 4 3 2 1

References

- Anderson, A., & Lynch, T. (1988). *Listening*. Oxford: Oxford University Press.
- Anderson, J. R. (2005). *Cognitive psychology and its implications* (6th ed.). New York: Worth Publishers.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.
- Bacon, S. M. (1992). Authentic listening in Spanish: How learners adjust their strategies to the difficulty of the Input. *Hispania*, 75(2), 398-412.
- Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper* (TOEFL Monograph Series Report No. 19). Princeton: Educational Testing Service.
- Belgar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing*, 27 (1), 101-118.
- Berne, J. (1995). How does varying pre-listening activities affect second language listening comprehension? *Hispania*, 78, 316-29.
- Blau, E. K. (1990). The effect of syntax, speed, and pauses on listening comprehension. *TESOL Quarterly*, 24 (4), 746-753.
- Blau, E. K. (1991). More on comprehensible input: The effect of pauses and hesitation markers on listening comprehension. *ERIC ED 340 234*.
- Bloomfield, A., Wayland, S.C., Rhoades, E., Blodgett, A., Linck, J., & Ross, S. (2011). What makes listening difficult: Factors affecting second language listening comprehension. Retrieved from http://www.casl.umd.edu/our_publications [2012, 19, June] University of

- Maryland, Centre of Advanced Study of Language.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). New York: Routledge.
- Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing*, 19 (4), 369-394.
- Brock, C. (1986). The effects of referential questions on ESL classroom discourse. *TESOL Quarterly*, 20 (1), 47-59.
- Brown, G. (1995). Dimensions of difficulty in listening comprehension. In D. Mendelsohn & J. Rubin (Eds.), *A guide for the teaching of second language listening* (pp. 59-73). San Diego: Dominie Press.
- Brown, G., & Yule, G. (1983). *Teaching the spoken language*. Cambridge: Cambridge University Press.
- Brown, J. D., & Brown, K. K. (2006). Introducing connected speech. In J.D. Brown & K. Brown (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 1-16). Honolulu: National Foreign Language Resource Centre, University of Hawai'i at Manoa.
- Browne, K. (2011). *An introduction to sociology* (4th ed.). Cambridge: Polity Press.
- Buck, G. (1990). *Testing second language listening comprehension* (Unpublished doctoral dissertation). University of Lancaster, Lancaster.
- Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.
- Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: examining attributes of a free response listening test. *Language Testing*, 15 (2), 119-157.
- Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English: A comprehensive guide spoken and written English grammar and usage*. Cambridge: Cambridge University Press.

- Cervantes, R., & Gainer, G. (1992). The effects of syntactic simplification and repetition on listening comprehension. *TESOL Quarterly*, 26 (4), 767-770.
- Chang, A. C., & Read, J. (2006). The effects of listening support on the listening performance of EFL learners. *TESOL Quarterly*, 40 (2), 375-397.
- Chapman, D. W., & Snyder, C.W. Jr. (2000). Can high stakes national testing improve instruction: Reexamining conventional wisdom. *International Journal of Educational Development*, 20, 457-474.
- Chaudron, C., & Richards, J. (1986). The effect of discourse markers on the comprehension of lectures. *Applied Linguistics*, 7 (2), 113-127.
- Chiang, S. C., & Dunkel, P. (1992). The effect of speech modification, prior knowledge, and listening proficiency on EFL lecture learning. *TESOL Quarterly*, 26 (2), 345-374.
- Cohen, A. (1984). On taking language tests: What the students report. *Language Testing*, 1 (1), 70-81.
- Cohen, A. (1998). Strategies and processes in test taking and SLA. In L. Bachman & A. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 90-111). Cambridge: Cambridge University Press.
- Coniam, D. (2001). The use of audio and video comprehension as an assessment instrument in the certification of English language teachers: a case study. *System*, 29, 1-14.
- Conrad, L. (1985). Semantic versus syntactic cues in listening comprehension. *Studies in Second Language Acquisition*, 7, 59-72.
- Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, 31, 218-236.

- Deese, J., & Kaufman, R. A. (1957). Serial effects in recall of unorganized and sequentially organized verbal material. *Journal of Experimental Psychology*, 54 (3), 180-187.
- Derwing, T. M. (1996). Elaborative detail: Help or hindrance to the NNS listener? *Studies of Second Language Acquisition*, 18, 283-297.
- Dunkel, P. A., & Davis, N. J. (1994). The effects of rhetorical signaling cues on the recall of English lecture information by speakers of English as a native or second language. In J. Flowerdew (Ed.), *Academic listening: research perspectives* (pp. 55-74). Cambridge: Cambridge University Press.
- Enomoto, A. (2007). Perception of short English sentences by Japanese Learners of English: A gating study. *Journal of Pan-Pacific Association of Applied Linguistics*, 11 (2), 151-167.
- Ericsson, K., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, 102 (2), 211-245.
- ETS (2007). *User Guide: Listening and reading*. Princeton, NJ: Educational Testing Service.
- Field, J. (2003). Promoting perception: lexical segmentation in L2 listening. *ELT Journal*, 57 (4), 325-333.
- Field, J. (2004). An insight into listeners' problems: too much bottom-up or too much top-down? *System*, 32, 363-377.
- Field, J. (2009). *Listening in the language classroom*. Cambridge: Cambridge University Press.
- Field, J. (2012). The cognitive validity of the lecture-based question in the IELTS listening paper. In L. Taylor & C. Weir (Eds.), *IELTS collected papers 2* (Studies in Language Testing, 34) (pp. 391-453). Cambridge: Cambridge University Press.
- Flowerdew, J., & Tauroza, S. (1995). The effect of discourse markers on

- second language lecture comprehension. *Studies in Second Language Acquisition*, 17, 435-458.
- Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21 (3), 354-375.
- Freedle, R., & Fellbaum, C. (1987). An exploratory study of the relative difficulty of TOEFL's listening comprehension Items. In R. Freedle & R. Duran (Eds.), *Cognitive and linguistic analyses of test performance* (pp. 162-192). New Jersey: Ablex Publishing corporation.
- Freedle, R., & Kostin, I. (1996). The prediction of TOEFL listening comprehension item difficulty for minitalk passages: Implications for construct validity (*Research Reports* No. 56). Princeton, NJ: Educational Testing Service.
- Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing*, 16, 2-32.
- Geranpayeh, A., & Taylor, L. (2008). Examining listening: developments and issues in assessing second language listening. *Cambridge ESOL Research Notes*, 32, 2-5.
- Gimson, A. C., & Cruttenden, A. (1994). *Gimson's pronunciation of English* (5th ed.). London: Arnold.
- Goh, S. T. (1990). The effects of rhetorical organisation on expository prose on ESL readers in Singapore. *RELC Journal*, 21, 1-13.
- Goh, C. C. M. (2000). A cognitive perspective on language learners' listening comprehension problems. *System*, 28, 55-75.
- Goh, C. C. M. (2008). Metacognitive instruction for second language listening development: Theory, practice, and research implications. *RELC Journal*, 39 (2), 188-213.
- Goh, C., & Taib, Y. (2006). Metacognitive instruction in listening for young

- learners. *ELT Journal*, 60 (3), 222-232.
- Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education* (Studies in Language Testing, 25). Cambridge: Cambridge University Press.
- Griffiths, R. (1990). Facilitating listening comprehension through rate-control. *RELC Journal*, 21 (1), 55-65.
- Griffiths, R. (1991). The paradox of comprehensible input: Hesitation phenomena in L2 Teacher-Talk. *JALT Journal*, 13 (1), 23-38.
- Halliday, M. A. K., & Hassan, R. (1976). *Cohesion in English*. London: Longman.
- Hasan, A. S. (2000). Learners' perceptions of listening comprehension problems. *Language, Culture, and Curriculum*, 13 (2), 137-153.
- Henning, G., Hudson, T., & Turner, J. (1985). Item response theory and the assumption of unidimensionality for language tests. *Language Testing*, 2 (2), 141-154.
- Henrichsen, L. E. (1984). Sandhi-variation: A filter of input for learners of ESL. *Language Learning*, 34 (3), 103-126.
- Hughes, R. (2003). Testing the visible: Literate biases in oral language testing. *Journal of Applied Linguistics*, 1 (3), 295-309.
- IIBC (International Institution of Business Communication) (2010). *TOEIC Test Data Analysis 2009*.
- Imura, H. (2011). The influence of test format on performance: Focusing on the presentation of questions and answer options in multiple-choice listening tests. *ARELE (Annual Review of English Language Education in Japan)*, 22, 361-376.
- Ito, Y. (2006). The significance of reduced forms in L2 pedagogy. In J.D. Brown & K. Brown (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 17-25). Honolulu: National Foreign

- Language Resource Centre, University of Hawai'i at Manoa.
- Ito, H., Kawamura, A., Shimada, Y., Nishihara, M., & Funato, S. (2007). Daigaku shingaku yoteisha-wo taishou-to shita eigo nouryokusiken-no kokusai-hikaku [A comparison of high-stakes English tests used for gate-keeping purposes for the universities or colleges: *The Centre Test* in Japan and *Matriculation Examination* in Finland]. *Shikoku-eigo kyouiku gakkai kiyo* [Journal of Shikoku English Education Society], 27, 11-26.
- JACET (2003). *JACET List of 8000 Basis Words*. Tokyo: JACET (Japan Association of College English Teachers).
- Jacewicz, E., Fox, R.A., O'Neill, C., & Salmons, J. (2009). Articulation rate across dialect, age, and gender. *Language Variation and Change*, 21, 233-256.
- Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000 framework: A working paper* (TOEFL Monograph Series Report No. 16). Princeton, NJ: Educational Testing Service.
- Jensen, C., & Hansen, C. (1995). The effect of prior knowledge on EAP listening-test performance. *Language Testing*, 12 (1), 99-119.
- Joyce, P. (2008). *Linguistic knowledge and psycholinguistic processing skills as components of L2 listening comprehension* (Unpublished doctoral dissertation). Roehampton University/ University of Surrey, Roehampton/ Surrey.
- Kachru, B. B. (1992). Teaching world Englishes. In Kachru, B. B. (Ed.), *The other tongue: English across cultures* (2nd ed., pp. 355-365). Urbana, IL: University of Illinois Press.
- Kasahara, K. (2009). The effects of lexical richness on the item difficulty of the 2008 Center Listening Test. *ARELE (Annual Review of English Language Education in Japan)*, 20, 191-200.

- Khalifa, H., & Weir, C.J. (2009). *Examining reading*. Cambridge: Cambridge University Press.
- Kim, H. (2006). World Englishes in language testing: A call for research. *English Today* 88, 22 (4), 32-39.
- Koreman, J. (2006). Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech. *Journal of the Acoustical Society of America*, 119 (1), 582-596.
- Kostin, F. (2004). Exploring items characteristic that are related to the difficulty of TOEFL dialogue items (*Research Reports* No. 79). Princeton, NJ: Educational Testing Service.
- Kougo, H. (2006). Eigo lisuninngu testo ga motarashita mono to kongo no kadai [What the introduction of the listening comprehension component of the Centre Test has brought about and the relevant issues to be solved]. *Daigaku Nyushi Forum*, 29, 3-7. Tokyo: Daigaku Nyushi Centa [National Centre for University Entrance Examination].
- Kroll, B. (1977). Combining ideas in written and spoken English: A look at subordination and coordination. In E.O. Keenan & T. L. Bennett (Eds.), *Discourse across time and space* (pp. 69-108). Los Angeles: Department of Linguistics, University of Southern California.
- Kuroda, W. (2012). Nihon-no eigo kyouiku-kara 'jinbun-kei baiasu' wo torinozoke [Remove 'humanity bias' from English education in Japan]. *Rikou-kei eigo-kyouiku wo kangaeru* [Reconsidering English education for the students majoring in science and engineering], 11-27. Tokyo: Institute for Digital Enhancement of Cognitive Development, Waseda University.
- Lado, R. (1961). *Language testing: The construction and use foreign language tests*. London: Longman.
- Leeser, M. J. (2004). The effects of topic familiarity, mode, and pausing, on

- second language learners' comprehension and focus on form. *Studies of Second Language Acquisition*, 26, 587-615.
- Linacre, J. M. (2011). Winsteps (Version 3.70.0) [Computer software]. Retrieved from <http://www.winsteps.com>. [2011, 5 July]
- Linacre, J. M. (2007). *A user's guide to WINSTEPS MINISTEP: Rasch-model computer programs*. Retrieved from <http://www.winsteps.com>. [2011, 5 July]
- Lund, R. J. (1991). A comparison of second language listening and reading comprehension. *The Modern Language Journal*, 75 (2), 196-204.
- Lynch, T. (1988). *Grading foreign language listening comprehension materials: The use of naturally modified interaction* (Unpublished doctoral dissertation). University of Edinburgh, Edinburgh.
- Maclay, H., & Osgood, C.E. (1959). Hesitation Phenomena in Spontaneous English Speech. *Word*, 15, 19-44.
- Markham, P. L. (1988). Gender and the perceived expertness of the speaker as factors in ESL listening recall. *TESOL Quarterly*, 22 (3), 397-406.
- Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, 8, 1-71.
- Matsusaka, H. (1986). *Eigo onsei gaku nyumon* [Introduction to English phonetics]. Tokyo: Kenkyu-sha Shuppan.
- Matsusaka, H. (1995). Listening comprehension: The phonetic factor. *Gakujutsu Kenkyu*, 43, 51-61. Tokyo: The Graduate School of Education of Waseda University.
- McKay, S. L. (2002). *Teaching English as an international language: Rethinking goals and approaches*. Oxford: Oxford University Press.
- McNamara, T. (1996). *Measuring second language performance*. Essex: Addison Wesley Longman Limited.

- Mecartty, F. H. (2000). Lexical and grammatical knowledge in reading and listening comprehension by foreign language learners of Spanish. *Applied Language Learning*, 11 (2), 323-348.
- Messick, S. (1989). Validity. In R. L. Lion (Ed.), *Educational measurement* (3rd ed., pp. 3-103). New York: National Council on Measurement in Education/American Council on Education.
- Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13 (3), 241-256.
- MEXT (2007). Retrieved from <http://www.mext.go.jp/english/shotou/030301.htm> [2007, 10 March]
- MEXT (2008). Retrieved from <http://www.mext.go.jp/bmenu/toukei/001/08030520/004.htm> [2011, 15 July]
- MEXT (2010). *Koto gakko gakushu sidou youryo kaisetsu: Gaikokugo hen, Eigo hen* [The Course of Study for foreign languages for upper secondary school: English]. Tokyo: Kairyu-do.
- MEXT (2011a). Retrieved from http://www.mext.go.jp/component/b_menu/other/_icsFiles/afieldfile/2011/08/11/1309705_3_1.pdf [2011, 24, Aug.]
- MEXT (2011b). Retrieved from http://www.mext.go.jp/component/b_menu/other/_icsFiles/afieldfile/2011/08/04/1308729_3.pdf [2011, 11, Sep.]
- Miller, G. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2), 81-97.
- Mueller, G. A. (1980). Visual contextual cues and listening comprehension: An experiment. *The Modern Language Journal*, 64, 335-340.
- Murdock, B. B. Jr. (1962). Serial position effect of free recall. *Journal of*

Experimental Psychology, 64, 482-488.

Murphey, T. (2006, February 7). Practical reasons for praising entrance exams. *The Japan Times*. Retrieved from

<http://search.japantimes.co.jp/cgi.-bin/ek20060207al.html>

NCUEE (2006). *Daigaku nyusi centa shiken no mondai to seito* [Questions and answers of *the Centre Test*]. Tokyo: Daigaku Nyusi Centa [National Centre for University Entrance Examination].

NCUEE (2007). *Daigaku nyusi centa shiken no mondai to seito* [Questions and answers of *the Centre Test*]. Tokyo: Daigaku Nyusi Centa [National Centre for University Entrance Examination].

NCUEE (2008). *Daigaku nyusi centa shiken no mondai to seito* [Questions and answers of *the Centre Test*]. Tokyo: Daigaku Nyusi Centa [National Centre for University Entrance Examination].

NCUEE (2009a). *Daigaku nyusi centa shiken no mondai to seito* [Questions and answers of *the Centre Test*]. Tokyo: Daigaku Nyusi Centa [National Centre for University Entrance Examination].

NCUEE (2009b). Retrieved from

http://www.dnc.ac.jp/modules/center_exam [2009, 3, June]

NCUEE (2010) *Dagaku nyusi centa shiken-no mondai to seito* [Questions and answers of *the Centre Test*]. Tokyo: Daigaku Nyusi Centa [National Centre for University Entrance Examination].

NCUEE (2011a). Retrieved from

<http://www.dnc.ac.jp/modules/news/content0433.html> [2011, 24, Aug.]

NCUEE (2011b). Retrieved from

http://www.dnc.ac.jp/modules/center_exam/content0408.html [2011,24, Aug.]

Negishi, M., Matsuzawa, S., Sato, T., Toyoda, Y., & Nakano, T. (2010).

Daigaku-nyushi-ga kawareba eigo-kyouiku-mo kawaruka [Will changes

- to entrance examinations to universities or colleges change formal English education at secondary schools in Japan?] *Eigokyouiku* [The English Teachers' Magazine], 59, (5), 10-19. Tokyo: Taishu-kan.
- Nishigori, D., & Kuramoto, N. (2007). Nihon-no daigaku nyusio-wo meguru shakai shinrigakuteki kousei kenkyu-no kokoromi: eiou nyushi-ni kansuru bunseki [Proposing a fairness research in social psychology over university admissions in Japan: Analysis of the 'Admission Office Examination']. *Nihon Tesuto Gakkai-shi* [Japanese Journal for Research on Testing], 3, 148-160.
- Nissan, S., DeVincenzi, F., & Tang, K. L. (1996). An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension (*Research Reports* No. 51). Princeton, NJ: Educational Testing Service.
- O'Malley, J., Chamot, A., & Küpper, H. L. (1989). Listening comprehension strategies in second language acquisition. *Applied Linguistics*, 10, 418-437.
- Otomo, K. (1996). *Komoku otou riron nyumon* [Introduction to Item Response Theory]. Tokyo: Taishukan-Shoten.
- Peterson, P. W. (1991). A synthesis of methods for interactive listening. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (pp. 106-122). Boston: Heinle & Heinle Publishers.
- Richards, J. C. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly*, 17 (2), 219-240.
- Richards, J. C. (1990). *The language teaching matrix*. Cambridge: Cambridge University Press.
- Richards, J. C., Platt, J., & Platt, H. (1985). *Dictionary of language teaching & applied linguistics*. London: Longman.
- Rose, D. (2008). Vocabulary use in the FCE Listening test. *Cambridge*

ESOL Research Notes, 32, 9-16.

- Rost, M. (1990). *Listening in language learning*. New York: Longman.
- Rost, M. (2002). *Teaching and researching in listening*. Harlow: Pearson Education.
- Rost, M. (2005). L2 listening. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 503-527). Mahwah, NJ: Lawrence Erlbaum.
- Rubin, J. (1994). A review of second language listening comprehension research. *The Modern Language Journal*, 78 (2), 199-221.
- Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception and psychophysics*, 2, 437-442.
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50 (4), 696-735.
- Sage, K., & Tanaka, N. (2006). So what are we listening for? A comparison of the English listening constructs in the Japanese National Centre Test and TOEFL iBT. *Authentic Communication: Proceeding of the 5th Annual JALT Pan-SIG Conference* (pp. 74-98).
- Saida, C., & Yanagawa, K. (2011). Development of Kanagawa prefecture high school English Tests with a common-item design: Test evaluation by IRT equating. *Japanese Journal for Research on Testing*, 7, 122-132.
- Sasaki, M. (2000). Effects of cultural schemata on students' test-taking processes for cloze tests: a multiple data source approach. *Language Testing*, 17 (1), 85-114.
- Schmidt-Rinehart, B. C. (1994). The effects of topic familiarity on second language listening comprehension. *The Modern Language Journal*, 78

(2), 179-189.

- Shaw, S. D., & Weir, C. J. (2007). *Examining writing*. Cambridge: Cambridge University Press.
- Shea, D.P. (2009). The university English entrance exam and its implied effect on EFL pedagogy in Japan. *Keio gijyuku-daigaku hiyosi kiyo, eigo eibei bungaku* [Bulletin of English language and Literature of Hiyoshi Campus of Keio University], 55, 97-132.
- Shiki, O., Mori, Y., Kadota, S., & Yoshida, S. (2010). Exploring differences between shadowing and repeating practices: An analysis of reproduction rate and types of reproduced words. *ARELE (Annual review of English language education in Japan)*, 21, 81-90.
- Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: The effect of text and question type. *Language Testing*, 8, 23-40.
- Shizuka, T. (2007). *Kisokara fukaku rikaisuru Rasch modeling* [Introduction to Rasch Modeling]. Osaka: Kansai-Daigaku Shuppan-bu.
- Siddharthan, A . (2006). Syntactic simplification and text cohesion. *Language and Computation*, 4 (1),77-109.
- Skehan, P. (1996). A framework for the implementation of task-based Instruction. *Applied Linguistics*, 17 (1), 38-62.
- Smith, Jr. E. V. (2005). Effect of item redundancy on Rasch item and person estimates. *Journal of Applied Measurement*, 6, 147-163.
- Speidel, G. E., Tharp, R.G., & Kobayashi, L. (1985). Is there a comprehension problem for children who speak nonstandard English? A study of children with Hawaiian-English backgrounds. *Applied Psycholinguistics*, 6, 83-96.
- Suzuki, K. (2008, November). *Focusing on sandhi-variation for improving unsuccessful EFL learners' listening*. Paper presented at the annual conference of Society of English Language and Literature of Waseda

- University, Waseda University, Tokyo.
- Suzuki, S., Takashima, H., Matsuzawa, S., Kadonaga, N., Oh, S., Yoshino, T., & Namai, K. (2009). Global-ka jidai-ni okeru daigaku kyouiku-to nyushi seido [College educations and entrance examinations in the era of globalisation], *Daigaku-nyusi kenkyu-no doukou* [The research trend of college entrance examinations], 26, 7-64.
- Tauroza, S., & Allison, D. (1990). Speech rates in British English. *Applied Linguistics*, 11 (1), 90-105.
- Taylor, L., & Geranpayeh, A. (2011). Assessing listening for academic purposes : Defining and operationalising the test construct. *Journal of English for Academic Purposes*, 10, 89-101.
- Thompson, I., & Rubin, J. (1996). Can strategy instruction improve listening comprehension? *Foreign Language Annals*, 29 (3), 331-342.
- Tsui, A. B. M., & Fullilove, J. (1998). Bottom-up or top-down processing as a discriminator of L2 listening performance. *Applied Linguistics*, 19 (4), 432-451.
- Uchida, T., Kikuchi, K., Nakaune, N., Maekawa, S., & Ishizuka, T. (2002). Eigo lisuningu tesuto-ni okeru onsei-no jikankouzou-to teijijyouhou-no youshiki-ga koumoku tokusei-ni ataeru eikyou [Effects of temporal structure and mode of presentation on item characteristics on an English listening comprehension test]. *Kyouiku Shinri-gaku Kenkyu* [Japanese Journal of Educational Psychology], 50, 1-11.
- Uchida, T., Sugisawa, T., & Ito, K. (2010). Daigaku nyushi centa siken-no lisuningu tesuto-ga sokutei-suru gakuryoku-no fuchi [Structure of academic skills measured by the listening comprehension test in National Centre Test]. *Nihon Tesuto Gakkai-shi* [Japanese Journal for Research on Testing], 6, 103-111.
- Underwood, M. (1989). *Teaching listening*. London: Longman.

- Van Dijk, T., & Kintsch, W. (1983). *Strategies of discourse comprehension*. London: Academic Press.
- Vandergrift, L. (2006). Second language listening: Listening ability or language proficiency? *The Modern Language Journal*, 90 (1), 6-18.
- Vanderplank, R. (1993). 'Pacing' and 'spacing' as predictors of difficulty in speaking and understanding English. *ELT Journal*, 47 (2), 117-125.
- Voss, B. (1979). Hesitation phenomena as sources of perceptual errors for non-native speakers. *Language and Speech*, 22 (2), 129-144.
- Voss, B. (1984). Perception of first-language and second-language texts: Comparative study. *Bielefelder Beitrage zur Sprachehroforchung*, 13, 131-153.
- Wagner, E. (2010). Test-takers' interaction with an L2 video listening test. *System*, 38, 280-291.
- Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing*, 13 (3), 318-333.
- Watanabe, Y. (2004). Teacher factors mediating washback. In L. Cheng & Y. Watanabe with A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 129-146). Mahwah, NJ: Lawrence Erlbaum.
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.
- Widdowson, H. (1978). *Teaching language as communication*. Oxford: Oxford University Press.
- Wierzbicka, A. (2003). *Cross-cultural pragmatics: The semantics of human interaction*. Berlin: Mouton de gruyter.
- Wilson, M. (2002). Discovery listening—improving perceptual processing. *ELT Journal*, 57 (4), 335-343.

- Winsteps (Version 3.72.0) [Computer software]. Beaverton, Oregon: Kagi.
- Wolff, D. (1987). Some assumptions about second language text comprehension. *Studies in Second Language Acquisition*, 9, 307-326.
- Wu, Y. (1998). What do tests of listening comprehension test?—A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing*, 15, 21-44.
- Yamauchi, Y. (2002, August). *Onsei jyoho shorimoderu no shiten kara no eigo lisuningu no Konnanten no bunseki: kurikeashi kikukoto no genkai to 'Howa ten' wo kokufuku surutame no CALL shitemu no kaihatu* [An analysis of listening difficulty from the perspective of a listening processing model: Limitations of repeated input texts and the development of a CALL system to overcome 'saturation' point]. Paper presented at the 42nd annual conference of LET (The Japan Association for Language Education and Technology), Otsuma Women University, Tokyo.
- Yanagawa, k., & Green, A. (2008). To show or not to show: The effects of item stems and answer options on performance on a multiple choice listening comprehension test. *System*, 36 (1), 107-122.
- Yoshida, K. (2010). *Academikku eigo nouryoku tesuto (TEAP) kaihatu-no imi* [Meaning of the development of an English proficiency test, TEAP]. *Eigokyouiku* [The English Teachers' Magazine], 59, (5), 26. Tokyo: Taishu-kan.
- Zen-eiren (*Zenkoku Eigo Kyouiku Kenkyuu Dantai Rengou-kai* [Association of High School English Teachers in Japan]) (2007). *Kyouiku kenkyuu dantai no hyouka* [Evaluation by educational institutions]. In Daigaku Nyusi Centa [National Centre for University Entrance Examination] (Ed.), *Daigaku nyusi centa shiken no mondai to seito* [Questions and answers of the Centre Test] (pp. 378-380). Tokyo: Daigaku Nyusi Centa

[National Centre for University Entrance Examination].

Zhao, Y. (1997). The effects of listeners' control of speech rate on second language comprehension. *Applied Linguistics*, 18 (1), 49-68.