Tabaran Institute of Higher Education
Iranian Journal of Language Testing
Vol. 3, No. 2, October 2013
ISSN 2251-7324

Received: January 8, 2013 *Accepted: September 10, 2013*

The Development and Validation of a Second Language Listening Reduced Forms Test

Paul Joyce[1]

Abstract

This paper reports upon the development of a test of second language (L2) connected speech comprehension. Despite the importance of connected speech to L2 listening comprehension, there remains the absence of a theoretically and empirically sound means of measuring learners' understanding of it. Thus, the development of the reduced forms test was undertaken to address this need. The assessed material was delivered through a dictation that contained a wide range of frequently occurring reduced forms. To ensure the trait purity of the instrument, the dictation consisted of a series of short decontextualized sentences that were of great lexical and syntactic simplicity. The test underwent piloting with Japanese university students who were from a false beginner to upper intermediate proficiency level. During the test development, both the Classical Testing Theory and Item Response Theory approaches to test evaluation and item selection were utilized. The second version of the connected speech dictation test was administered to 548 participants. The findings showed that all of the items fitted the Rasch model and, therefore, the test is considered a valid measure of reduced forms in English as a second language listening comprehension. Furthermore, the results indicated that Japanese L2 learners have difficulty recognizing even the most frequently used English words when they are spoken in fluent native speaker discourse. It was concluded that the teaching of reduced forms should constitute a more important part of the L2 listening curriculum.

Keywords: *Second language, listening, connected speech, reduced forms, test development, testing*

1. Introduction

It is an all too familiar occurrence. Despite years of developing their English language skills through classroom-based materials, second language (L2) learners of English frequently find themselves wholly unequipped to understand natural native speaker conversation. While there are numerous potential reasons for this unpreparedness (see Rubin, 1994), one of the

1 Department of Law, Kinki University, Japan, pauljoyce@hotmail.com

primary obstacles is likely to be students' unfamiliarity with connected speech (see Brown & Kondo-Brown, 2006a). In the classroom, the dialogues and practice conversations used are typically, "full of clearly pronounced and articulated speech" (Rogerson, 2006, p. 85) and in order to facilitate comprehension teachers simplify and exaggerate the language they use (Ellis, 1985). On the other hand, outside of the classroom, rather than seeking to maximise clarity, native speakers usually prefer to "draw[n] sounds together with the purpose of saving time and energy" (Clarey & Dixson, 1963, p. 12). As a result, when listening to typical native speaker discourse, the L2 listener faces two major obstacles for which they are often inadequately prepared. Firstly, in contrast to written communication, there are rarely gaps between lexical items. Consequently, the listener has to detect the location of word boundaries without the aid of explicit markers. Secondly, and perhaps more importantly, words in natural speech do not have a stable auditory form. Instead, the immediate phonetic environment of a word greatly influences its acoustic realization (Pickett & Pollack, 1963). As a result, there is a clear difference between fluent speech and a series of highly distinct individual words (Bond & Garnes, 1980). Such phonological modification in fluent spoken communication has long been assumed to increase L2 comprehension difficulty (Belasco, 1965; Bonk, 2000; Cauldwell, 1996; Goh, 2000; O'Malley, Chamot & Kupper, 1989; Rubin, 1994). This supposition has been confirmed by L2 componential studies that have associated an increased understanding of phonological modification with improved aural proficiency (Joyce, 2011; Matsuzawa, 2006).

2. Review of the Related Literature

The phonologically modified or reduced form of spoken language refers to the collective processes which "reduce[s] the overt markedness, or perceptual saliency, of morphemes" (Henrichsen, 1984, p. 103). These reduced form processes occur in many different languages and are governed by a complex set of language specific phonotactic rules (Dirven & Oakenshott-Taylor, 1984; Grosjean & Gee, 1987; Roach, 2001). Buck (2001) asserts that the three most important reduced speech phonological changes in English are assimilation, elision, and intrusion. Assimilation refers to the blending of words at their boundaries. For instance, the pronunciation of *won't you* as *wonchoo*. Elision involves the omission of an individual phoneme to simplify pronunciation. For example, *where he* is frequently pronounced without the second /h/ consonant. Lastly, intrusion is the introduction of a phoneme between words. This often occurs when two vowel sounds meet, such as the

introduction of a /w/ between *too* and *easy*. Phonological modification is also apparent in the pronunciation of grammatical words, such as articles, prepositions, pronouns, conjunctions, and relative pronouns (Avery & Ehrlich, 1992). Grammatical words can be articulated in both strong and weak forms, depending on whether the word receives sentence stress. Since such lexical items are infrequently emphasized, these words are usually uttered in their weak unstressed form. For instance, the conjunction *and* is usually unstressed, as in *fish 'n chips*. Reduced forms are sometimes assumed to only be common in fast or informal speech. However, as Bowen notes, they are "an ever-present phenomenon in oral English" (1975, p. 226).

Despite the importance of phonological modification to L2 listening comprehension, not only is there considered to be a paucity of research in the area (Brown, 2006), the teaching of reduced forms has also been hampered by the limited attention that connected speech has received in pedagogical materials (Rogerson, 2006). While the reasons for this neglect are unclear, a contributory factor is the absence of a theoretically and empirically sound means of measuring learners' understanding of connected speech. After all, without an appreciation of students' ability to comprehend reduced forms, it is difficult to research and evaluate the importance of focusing on connected speech or the benefit derived from such study. However, although a widely recognized test of phonological modification knowledge has yet to emerge, there have been a number of studies that have measured students' comprehension of connected speech.

The first major reduced forms study was conducted by Henrichsen (1984). To assess the influence of connected speech on listening comprehension, two dictation tests were employed. The two 15-item instruments were drawn from a well-established dictation test, the Integrative Grammar Test (IGT) (Bowen, 1976). After hearing each of the sentences, the 47 high-level students and the 18 low-level learners were required to transcribe the full form of the words that they heard. The learners were evaluated on their ability to accurately record the second word of each sentence. While both of the dictation tests contained identical sentences, half of the sentences in each were assigned to the presence of reduced forms condition, and half to the absence of reduced forms condition. Henrichsen's (1984) study clearly showed the influence of connected speech upon comprehension. However, the dictation tests that were used both contained a number of relatively complex grammatical points and the length of the sentences employed was also likely to have strained short-term memory. For instance, one of the sentences used was: "Who would have thought he would ever remember her?" Consequently, the research results are likely to have confounded the

comprehension of reduced forms with syntactic knowledge and working memory capacity. Furthermore, the reliability of the test results varied between .31 and .68 (K-R21) (see Brown & Kondo-Brown, 2006b). Therefore, there was a large degree of random error contained in the scores.

In a further reduced forms study, Ito (2001) partially replicated Henrichsen's (1984) work. Nine advanced and nine intermediate participants completed two 20 sentence dictation tests. After hearing each of the sentences, the participants were required to transcribe the full form of the words that they heard. The learners were evaluated on their ability to accurately reproduce two words that contained reduced forms from each of the sentences. To control the grammatical complexity of the sentences, the researcher only selected sentences that contained structural forms from a beginner level grammar text (Azar, 1996). On the basis of the results, it was concluded that the presence of reduced forms, and the nature of the phonological modification, influenced the participants' listening comprehension. The reliability of the participants' scores suggested that the dictation test was reasonably suited to the purposes of this study (Cronbach alpha = .78). However, since the participants' mean scores were high (Advanced: 87%; Intermediate: 73%), a ceiling effect is likely to limit the reliability of the test for highly proficient learners. Furthermore, since Ito (2001) found some of the participants had to be discarded from the study because of their "extremely low proficiency" (p. 104), the wider application of the test could also be restricted by its difficulty for lower ability learners.

Brown and Hilferty (1986) investigated the effectiveness of teaching reduced forms. In their study, 32 participants were randomly divided into two groups. The two sets of students were administered a grammar test (the IGT) (Bowen, 1976), a reduced form dictation, and a norm-referenced multiple choice test. The tests were given both at the beginning and the end of the month-long study. During the intervening period, the treatment group received a daily 10-minute session on reduced forms and the control group spent a similar amount of time studying minimal pairs. The participants' understanding of reduced forms was evaluated through their ability to transcribe a short conversation that they heard three times. The conversation contained 45 words that were subjected to phonological modification, and the learner received a point for each one of these words that they correctly transcribed. Although the actual tests used have not been provided, Brown and Hilferty (1989; 2006b) and Brown (2006) have supplied example reduced forms dictations that are similar to the original. In an example conversation, after hearing "Whenerya goin' ta Peking?" (Brown & Hilferty, 1989, p. 28), the participants would be awarded a point for correctly writing the full form of each of

the underlined words: "When are you going to Peking?" (ibid.). While there were not found to be any statistically significant differences between the scores from the two groups on the pre-tests, the treatment group scored statistically significantly higher on the post-test grammar and reduced form dictation measures. The results suggest that students can improve their understanding of reduced forms through focusing on the phonotactic rules of the language and the dictation test provided an efficient means of evaluating the students' understanding of a large number of reduced forms. However, as the assessed material was presented as part of a conversation, there is likely to be some item interdependence across sentences. For example, after understanding the example test question above about the trip to Peking, much of the assessed content on the following line, "I am going to go on Sunday." (ibid.) could be predicted. Furthermore, owing to the interconnected conversational nature of the dictation, it would be difficult to remove material and replace it with alternative items that might be preferable for psychometric or other reasons. Lastly, although Brown and Hilferty (1986) did not present the reliability of their test results, from the descriptive data that they provided, it was possible to calculate the reliability (K-R21) of the test results for their control (pre: .65; post: .53) and treatment (pre: .53; post: .86) groups. The internal consistency results for the test scores were mediocre. However, it is important to note that by combining the results for all of their students, a higher reliability estimate is likely to have emerged.

In a second study to explore the effectiveness of teaching reduced forms, Matsuzawa (2006) gave 20 students four hours of training in the understanding of natural native speaker speech, over a month long period. In order to assess the participants' ability to grasp reduced speech, the learners were required to transcribe 30 sentences in pre and post tests. To help facilitate this process, the researcher informed the examinees how many words were in each of the sentences by providing a pair of parentheses for each vocabulary item. Only those words that were subjected to phonological modification were included in the test scoring. In the following example question, the assessed test items have been italicised: (What) (are) (you) (up) (to) ?. After comparing the pre and post test results, despite the moderate internal consistency of the test (K-R21 reliability = .63) (see Brown & Kondo-Brown, 2006b), there was found to be a statistically significant improvement in the participants' scores. Furthermore, there was discovered to be a high correlation between English proficiency and reduced-forms comprehension (r = .72), though the statistical significance of this finding was unreported. The study both reinforced the value of teaching reduced forms and the importance of connected speech to listening comprehension. However, by supplying

the participants with information about how many words were contained in each of the assessed sentences, the learners are likely to have been more able to rely on declarative L2 syntactic knowledge to help them answer the test questions. Furthermore, despite the relatively low proficiency of some of the students in the study (TOEIC scores ranged from 380 to 850), the learners were assessed on their ability to recognized some comparatively low frequency vocabulary items such as attic, fattening and mutton. Since such words are highly likely to have been unknown to some of the participants, to some degree, the study conflated knowledge of reduced forms with L2 vocabulary breadth. Thus, through the influence of L2 syntactic and vocabulary breadth knowledge, the trait purity of the connected speech test is likely to have been compromised.

Finally, as part of a study that investigated the relationship between L2 listening proficiency and various linguistic and psycholinguistic components, the importance of reduced forms to aural proficiency was explored (for more details, see Joyce, 2011). In total, there were 443 learners who completed the research instruments, which were all undertaken within a two week span. The reduced forms construct was operationalized through the connected speech test that is described in this paper. There was found to be a high Pearson product moment correlation between the participants' reduced forms scores and a composite listening proficiency measure ($r = .64$, $p < 001$) that combined the results from the listening section of the TOEFL and an in-house university listening test. Thus, as previously observed by Matsuzawa (2006), there was discovered to be a strong relationship between knowledge of reduced forms and L2 listening proficiency. The correlation results also provided evidence to support the concurrent validity of the reduced forms test that is the subject of this paper. A strong correlation coefficient was also recorded between connected speech recognition and aural syntactic knowledge ($r = .60$, $p < .001$). The results confirmed Brown and Hilferty's (1986) findings that grammatical knowledge can help compensate for a loss of saliency in the acoustic signal. Nevertheless, after subjecting the data to multiple regression analysis, it was found that knowledge of reduced forms made a substantial unique contribution to the prediction of L2 listening performance ($\beta = .29$, $p = < .001$). Thus, phonological modification knowledge and syntactic knowledge were found to tap overlapping but distinct capabilities.

From reviewing the research in this area, it is clear that a limited knowledge of reduced forms is likely to be reflected in diminished listening comprehension. However, there remains scope to improve on the measurement of L2 reduced forms knowledge. Therefore, the purpose of this paper is to introduce some of the issues concerned with testing phonological knowledge and to describe the process of developing an L2 reduced forms measure. I will

first introduce the test development methodology used. This will be followed by an account of the results from the two stages of test piloting, and a discussion of their implications.

3. Methodology

3.1 The Principles of Test Development

Test design. As discussed in the Introduction, there is currently no generally accepted approach to measuring L2 learners' understanding of phonologically modified speech. Nevertheless, it is notable that the primary studies on reduced forms have all included dictation tests (e.g. Brown & Hilferty, 1986; Henrichsen, 1984; Ito, 2001; Matsuzawa, 2006). As a format, there are a number of advantages to dictations. Notably, since each orthographic word within a text can be classed as a separate item, dictation tests tend to contain a large number of questions. As a result, they are recognised to yield highly reliable scores (Fountain & Nation, 2000). In addition, through the manipulation of the content, length, and delivery of the text, there is tremendous flexibility in the construct that they can target. Since it was decided that a dictation test was going to be used to assess L2 phonological modification knowledge, three measures were undertaken to safeguard the trait purity of the test.

Firstly, the aural material contained in the test was decontextualized. That is, rather than heard as a unified text; examinees were presented with single sentences that were divorced from the wider communicative context. Secondly, the length of each decontextualized sentence was carefully restricted. This was done to ensure that the participants' short-term memory (STM) did not become strained. Once STM capacity becomes stretched, those listeners' with a greater short-term memory store would be advantaged. Since STM was not part of the target construct, it was essential that the assessed material was sufficiently short to avoid this issue. There is evidence to suggest that STM is restricted to about seven units of information (Miller, 1956). Therefore, for the purposes of this study, the number of lexical items incorporated in the assessed sentences was limited to seven. Thirdly, the linguistic difficulty of the dictated material needed to be carefully controlled. When the perceptual saliency of the input is reduced, there is potential for listeners to employ their grammatical knowledge of the language to compensate for shortcomings in the acoustic signal. To make sure that the trait purity of the reduced forms measure did not decline, it was important that all of the participants were well acquainted with the structural forms used in the test. To ensure that this was the case, it was decided that the only grammatical structures used in the dictation would be those that were also contained in Essential Grammar in Use (Murphy, 2003), an elementary level self-study book. The structures contained in the textbook are

primarily focused on the present simple, present continuous, present perfect, simple negation, and common modal verbs. Regarding the lexical difficulty of the test, it was ensured that all of the test items were included in the list of core junior high school vocabulary items (The course of Japanese lower secondary school English, n.d.) that had been taught to all of the participants. This word list contains 508 of the most frequent words in the English language such as the months of the year, colours, and personal pronouns. In summary, as noted by Buck (2001), "…when segments are very short, and they do not challenge the test-taker, writing down a few words of spoken text is little more than a simple transcription exercise. The listening skills involved are probably just word recognition" (p. 77).

Test content. The content validity of the test was hampered by a lack of quantitative data on the frequency of common reduced forms. To compensate for this knowledge gap, researchers have suggested a number of examples of connected speech that they consider to be of importance. In a text intended to improve students' listening skills, Weinstein (1982) suggests twenty high-frequency relaxed speech patterns. Ur has also contributed thirty-four "fast colloquial" (1984, p. 46) reduced forms that she believes are either difficult to identify or pronounce in native-fashion. In addition, as part of a teaching programme and research study, Brown and Hilferty (1989) identified seventy-four examples of connected speech that they deemed salient. Lastly, Bond (2001) has recommended forty items that she considered to be essential for aural comprehension. In totality, the reduced forms cover a wide range of language, including examples of assimilation, elision, and intrusion. In addition, although each of the lists displays a differing emphasis, the researchers exhibit a great deal of agreement. Most notably, there is concurrence on the reduced perceptual saliency of certain grammatical words, especially modal and auxiliary verbs, and personal pronouns. An example of this agreement is: "you shouldn't have" (Ur, 1984, p. 46), "shoulda (should have)" (Brown & Hilferty, 1989, p. 27), and "should have + consonant (shoulda)" (Bond, 2001). In the absence of a more systematically compiled list of reduced forms, the four inventories formed the basis of the test material. Since most of the reduced forms that were specified in the lists only contained two words, in devising the test items, the reduced language segments were combined. For example, the separately listed, do you and want to could be united to form part of the single sentence, Do you want to do it?

Scoring procedure. In principle, each of the vocabulary items used within the test was employed as an individual item and scored using Rasch's (1960) one-parameter model (see Item Response Theory section for more details). However, as has been carried out in previous studies (e.g. Brown & Hilferty, 1986), to construct a complete question or statement that

could be used in the test, it was sometimes necessary to add a word that was not reduced. For instance, in the sentence Where are you going to play?, since the last word was not subject to phonological modification, it was excluded from the scoring. In the test transcription (see Appendix 2), the unassessed vocabulary items are not underlined. To receive a point for the correct identification of a word, the item had to be spelt correctly, in its full unabbreviated form.

3.2 Statistical Analyses

In order to help construct and validate the reduced forms test, two measurement models were employed: Classical Testing Theory (CTT) and Item Response Theory (IRT).

3.2.1 Classical Testing Theory

CTT analysis focuses on the difficulty, discrimination, and reliability of both tests and their individual test items. The CTT and IRT approaches to the use of item facility (IF) data differ. CTT advises the use of items that are of mid-range difficulty. For reasons that will be discussed in the Item Response Theory section below, in accordance with a widely recommended guideline (see Henning, 1987; Tuckman, 1972), the number of items with an IF of between .33 and .67 was monitored. Regarding item discrimination (ID), although there is no universally agreed minimum point at which an ID figure is considered sufficiently high, a point biserial correlation of .25 or above is widely regarded as acceptable (Henning, 1987). Lastly, while an internal reliability estimate in excess of .70 is commonly cited as acceptable (e.g. Kline, 1999; Nunnally, 1978), an internal consistency in excess of .80 was sought.

Although classical data analysis provides a useful guide to the psychometric characteristics of a test, it also has some serious limitations. Since classical analysis is sample-dependent, classical statistics are limited to the particular group of participants who are administered a test. Likewise, when using classical analysis, an individual's score is limited to the particular set of items contained in the test. A statistical model that overcomes these shortcomings is IRT.

3.2.2 Item Response Theory

As mentioned above, IRT statistics are independent of the participants or items used. Consequently, it is possible to compare individuals across different tests and items across different groups of test takers. For the purposes of this research, the one-parameter or Rasch model was used. The IRT analysis was performed using Quest (Adams & Khoo, 1993).

Person and item difficulty estimates refer to a single dimension along which test takers can be placed according to their ability, and items can be located in terms of their difficulty. As mentioned in the Classical Testing Theory section, CTT recommends the use of items that are of mid-difficulty. In contrast, in order to most effectively discriminate between test takers, IRT item difficulty values should mirror the person ability estimates. Since it was envisioned that the final version of the reduced forms test would be a balanced test of mid-difficulty level, it was expected that most of the test takers would average between 33 and 67 percent. Therefore, both from the CTT and IRT perspectives, it was important that a large proportion of the test items also fell within this range of difficulty. For this reason, as mentioned in the CTT section above, the number of mid-difficulty test items was monitored. However, since it was expected that there would be test takers who fell outside of the mid-difficulty range, it was also important that there were test items to match their proficiency levels. Therefore, while a preponderance of mid-difficulty items was sought, items throughout the difficulty scale were desired. Fit statistics relate to how well the data fits the statistical model. In accordance with recommended practice (McNamara, 1996), the items with an infit mean square of less than .75 or greater than 1.3 were excluded from the test.

3.3 General Procedure

As discussed in the Test Design section, the concept of phonological modification knowledge was operationalized through a dictation test that consisted of a series of short sentences. Each of the test sentences was only played once. The assessed material was delivered as naturally as possible, in terms of both speed and reduced forms. To forewarn the participants of the onset of the assessed material, the question number immediately preceded each sentence. Between each of the sentences, there was a fifteen second gap to allow the learners to transcribe the full form of the words that they had heard. One point was awarded for each correctly identified assessed word. The students recorded their answers on prepared test forms. Prior to beginning the test, the learners both read and heard the test instructions and an example sentence. The listening material was produced and delivered through high quality audio equipment. The test audio recording is available upon request.

3.4 Participants

The data was collected in Japan at a university specializing in foreign languages. All of the participants were native Japanese L1 speakers, who were enrolled as full-time English language major undergraduates. The learners could broadly be described as being at a false

beginner to an upper intermediate level. When expressed in terms of performance on the paper and pencil TOEFL, the participants' scores ranged from approximately 357 to 513 (see Bonk, 2001), which converts to scores of between 70 and 180 on the TOEFL Computer-Based Test. Since the selection of the participants was governed by the cooperation of their EFL instructors, a convenience sample was used. All of the students who had the opportunity to participate in the study chose to do so.

4. Results

4.1 Test Administration 1

Procedure. In the first administration, it was important to pilot a large number of items in order to develop a sizeable bank of psychometrically high quality items that could be used in the second test administration. However, the use of a very long test would have risked test fatigue and a decline in the reliability of the scores. Therefore, through the utilization of common anchor items, it was possible to trial three versions of the test and place all of the questions on the same scale. Each of the three research instruments contained 85 items, which were spread over 20 short sentences. The test forms were anchored together using 44 items that were common to all three of the tests. A total of 131 students participated in the pilot study. The participants were given 15 seconds to transcribe each of the sentences and the tests took around eight minutes to administer.

Results and discussion. As can be seen in Table 1, the CTT results from the three versions of the test were found to be fairly similar. The mean average scores were 61 percent (Form One), 59 percent (Form Two), and 58 percent (Form Three). Thus, despite the extremely high frequency of the vocabulary employed, the learners were collectively unable to identify a large proportion of the material that was presented to them.

Table 1. Descriptive Statistics for Test Administration 1

	n	**k**	**mean**	**SD**	**skew.**	**kurt.**	**min.**	**max.**	**rel. (α)**
Form One	50	85	51.74 (61%)	7.08	.47	-.16	39	69	.77
Form Two	38	85	49.74 (59%)	8.01	-.23	2.33	27	73	.82
Form Three	43	85	49.56 (58%)	9.31	.18	-.25	33	72	.86

However, only a relatively small proportion of the items fell in the IF range of 33 to 67 percent. In the case of Form One, there were 23 (27%) such items, and for Forms Two and

Three, there were 25 (29%) and 29 (34%), respectively. The relative paucity of items of a mid-difficulty level is partially explained by the number of test questions that were answered correctly by all or none of the participants. In total, there were twelve such items in Form One, five in Form Two, and five in Form Three.

In terms of the internal consistency of the test scores, the three dictations fared well. As displayed in Table 1, the Cronbach reliability coefficients were .77 (Form One), .82 (Form Two), and .86 (Form Three). The high reliability was in part due to the relatively wide dispersion of the scores. As can be seen in Table 1, the SD of the three tests was consistently high and ranged from 7.08 for Form One to 9.31 for Form Three. The high internal consistency was also a product of the large number of items meeting the .25 ID target. There were 31 (36%) items exceeding this mark in Version One of the test, 42 (49%) in Version Two, and 44 (52%) in Version Three. Yet, since there were 44 anchor items that were contained in all three test versions, there were actually only 87 different test items that met the ID criterion. Lastly, the scores from each of the three tests formed normal distributions in most respects. However, the kurtosis value for Form Two (2.33) was indicative of a non-normal score distribution. If this kurtosis result was repeated in the second administration of the test, there would be cause for concern. However, since the results from the other tests were normally distributed and only a small proportion of the items from Form Two were to be re-used, the errant kurtosis value was not considered to be a serious issue.

Turning to the IRT findings, the results show that the test remained slightly less difficult than the participants were able. That is, as can be seen in Table 2, the item estimates fell between .00 (Form One) and .07 (Forms Two and Three), whereas the person estimates ranged from .40 (Form Three) to .57 (Form One).

Table 2. Inferential Statistics for Test Administration 1

	Person Estimates			**Item Estimates**			
	Mean	**SD**	**Rel.**	**Mean**	**SD**	**Rel.**	**Misfitting Items (<.75, >1.3)**
Form One	.57	.67	.79	.00	1.90	.94	0
Form Two	.51	.74	.84	.07	1.99	.95	13
Form Three	.40	.84	.87	.07	2.02	.96	10

Finally, as can be seen in Table 2, there were a total of 23 items that did not fit the IRT statistical model. Of these, there were 13 overfitting items, which added very little unique information. Conversely, there were 10 items that had too little fit with the model. Both the overfitting and underfitting test questions were omitted from the second test administration.

In preparation for the second pilot test, the most effective items from the first administration were selected. However, the revision of the test was complicated by the nature of the test format. As previously discussed, the research instrument took the form of a dictation test with the items corresponding to individual orthographic words from a series of short sentences. Owing to the structural relationship between the test questions, the inclusion of each particular item was contingent upon the inclusion of other words with which it was presented. Therefore, when selecting test questions for the main study, rather than simply comparing the individual items, the statistical values for sentence groups of items were examined. When evaluating the relative merit of the various sentences, the ID, IF, and fit statistics were compared. Furthermore, as well as validity and reliability, test efficiency was also an important concern. Since testing time was limited, it was essential that as much information as possible could be collected in the shortest possible time. Of the test questions used in Test Administration One, 69 items from 13 sentences were selected. To provide encouragement to the lower proficiency participants, the sentences were ordered from the easiest to the most difficult.

4.2 Test Administration 2

Procedure. A revised version of the reduced forms test was administered to ensure that the items were functioning appropriately. A total of 548 students participated in the pilot study. The participants were given 15 seconds to transcribe each of the sentences and the tests took around six minutes to administer. The students' test paper and a transcript of the test are available in Appendices 1 and 2.

Results and discussion. The descriptive findings showed that the internal consistency of the test scores was high (Cronbach alpha = .88). This was primarily a product of the large number of items (74%) that reached or exceeded the ID target of .25. This finding represents a substantial gain in test score consistency over the first test administration. The improvement was due to the increase in the size of the sample population, and the selection of the most psychometrically sound items from the three pilot tests.

Table 3. Descriptive Statistics for Test Administration 2

k	mean	SD	skew.	kurt.	min.	max.	rel. (α)
69	42.09 (61.09%)	8.44	.01	-.21	14.00	68.00	.88

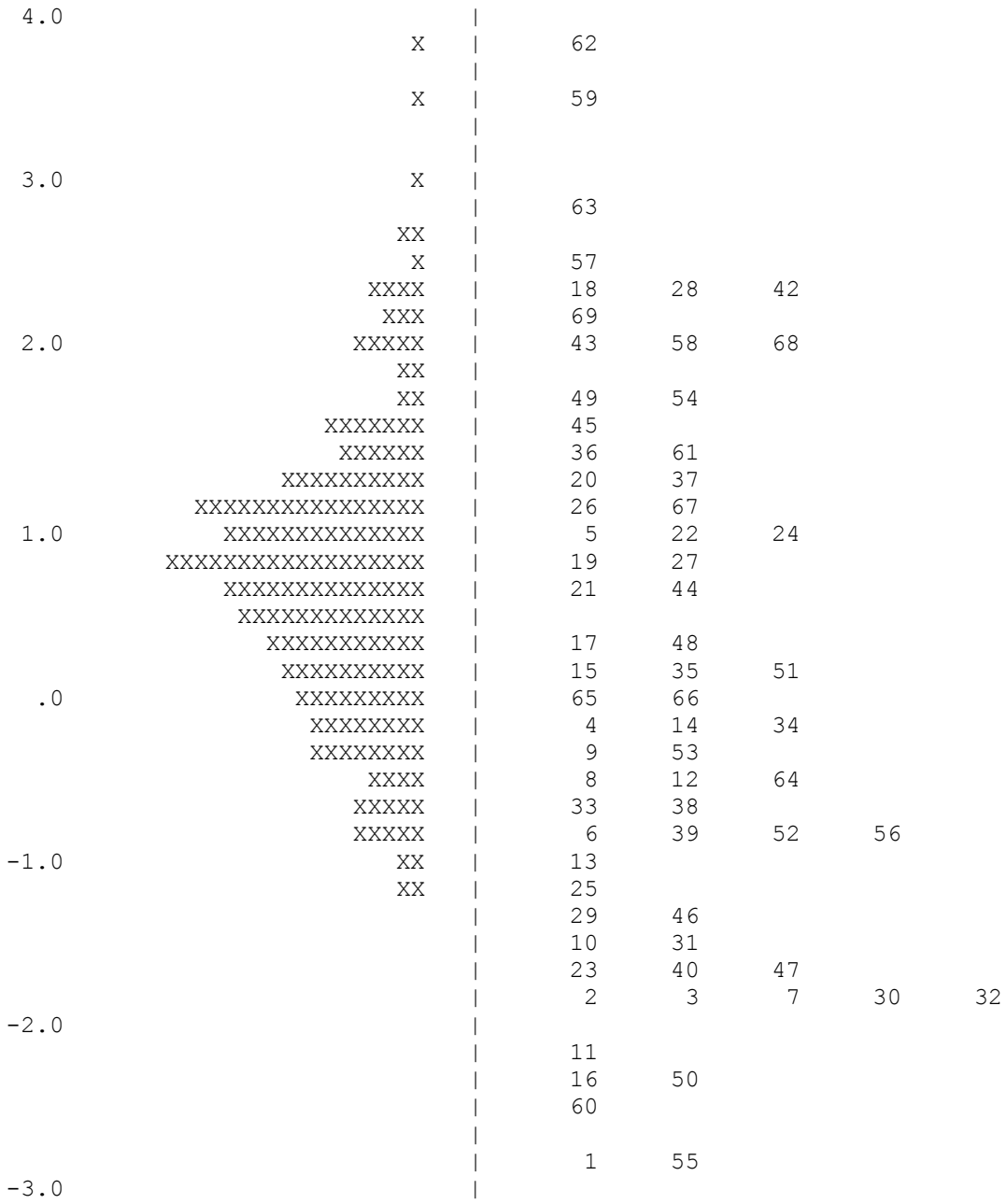
The test average (61%) suggests that the participants had an incomplete acquisition of the reduced form of even very high frequency vocabulary. Despite the overall test difficulty, only 23% of the test items fell within the target IF range. While this was less than hoped, the descriptive results suggest that the test performed satisfactorily.

Table 4. Inferential Statistics for Test Administration 2

Person Estimates			Item Estimates			
Mean	SD	Rel.	Mean	SD	Rel.	Misfitting Items (<.75, >1.3)
.70	.99	.89	.00	1.64	.99	0

The IRT results were consistent with the descriptive results. The inferential findings are displayed through a Wright map (see Figure 1). A Wright map displays the ability of the participants and the difficulty of the test items through two vertical histograms. The left side refers to the participants and the right side to the items. The logit scale in Figure 1 ranges from -3 to 4 logits. The most able candidates and the most difficult questions are positioned towards the top, and the least able learners and least difficult items towards the bottom. Since each X on the map represents a group of three participants, the case estimates at the ends of the score distribution are not displayed.

Figure 1. Wright Map for Test Administration Two



Each X represents three participants.

As can be seen through the Wright map, the participants were found to be more able than the items were difficult. While the mean person ability was .70 (SD = .79), the person ability estimates varied between -2.02 and 5.45. On the other hand, the average item difficulty was .00 (SD = 1.64) and the test question difficulty ranged between -2.91 and 3.88. Although the difficulty of the items should ideally have mirrored the person ability estimates, there were test questions spread almost throughout the logit scale. One important area for

improvement concerns the gap in the item distribution between Items 17 and 21. Since there were a large number of test-takers in this range, finding items to fill the gap is an avenue for future test development. Nevertheless, there were test items at enough difficulty levels to accurately evaluate the ability of the participants. From examining the Wright map, it is also noticeable that there were a large number of test questions that were more than one logit less difficult than most of the least proficient persons were able. Although these items provided little useful information about the ability level of the learners, they performed an important function. Due to the structural relationship between the test questions, the inclusion of the easiest test questions often enabled the use of more effective test items. Furthermore, by providing some attainable items for the least able participants, such students are likely to have been encouraged to try harder, which improves the reliability of the test. Lastly, the IRT results also showed that all of the Item Infit Mean Square values (.81 to 1.21) fell within an acceptable range. For the fit statistics for each of the items, see Appendix 3. On the basis of the statistical results, it was concluded that the reduced forms test constituted a reliable means of measuring the target construct.

5. Conclusion

This paper has summarized the development and initial validation of an L2 reduced forms test. As has been discussed, the first phase of test construction focused upon the design and content of the measure. This was followed by two stages of piloting and the analysis of the results through both Classical Testing Theory and Item Response Theory to ensure that the measure was reliable, valid and administratively efficient. Given the surprising lack of focus on reduced forms (Brown, 2006), it is hoped that the development of the reduced forms test will contribute towards a greater interest in the teaching, research, and acquisition of connected speech.

In terms of the results themselves, it was notable how difficult the participants found the tests. As previously discussed, the research instrument consisted of a series of short sentences of low syntactic and lexical difficulty. Since the participants were majoring in English language and had at least seven years of English language education, they might have been expected to score more highly. However, the study has confirmed that even L2 majors have difficulty recognising the most frequently used English words in native speaker discourse, when they are uttered naturally. Nevertheless, the performance of the participants in this study compares favourably with those from other phonological modification research studies. For instance, the control group in Brown and Hilferty's (1986) study recognized a

mere 35 percent of the highly frequent words that they heard. Furthermore, at the first attempt, Pemberton (2003) found that his participants were only capable of transcribing 50 percent of words from the most frequent 1000 words of the language.

This study was not designed with the purpose of discussing L2 listening pedagogy. However, the test difficulty findings have brought us to this point. The teaching of listening is becoming more responsive to learner needs. Nevertheless, a large part of most listening classes consists of students processing short aural texts and answering comprehension questions. It is unclear how the overwhelming emphasis on testing listening comprehension, rather than teaching it arose (Sheerin, 1987). Yet, given that the foundations of L2 listening proficiency have been far from well understood, it is understandable that textbook writers have primarily sought to promote L2 listening acquisition through a question and answer approach. Nevertheless, rather than employing naturally occurring language, it has been found that, "textbook dialogues do not reflect the ways in which real talk is produced in actual interactions" (Jones & Ono, 2000, p. 12). In order to improve L2 listening pedagogy, it is proposed that there should be a greater controlled focus on reduced forms, especially for students at lower proficiency levels. Through highlighting and practicing phonologically modified language forms, the learners' awareness of these critical skills can be heightened and improved. Furthermore, by focusing on selected skills, students can make perceptible strides in their learning and maintain their motivation. And, as learners become more aware of connected speech, they can practice more with authentic materials to automate their understanding. There are a wide range of activities that can be used to teach and practice reduced forms. Amongst many others, these include various types of cloze exercises, dictations and pair/group work activities (see Hough, 1995; Kobayashi & Linde, 1984; Rost & Stratton, 1980). As part of this process, the use of short tests, such as the one reported in this study, can help quantify improvement. The idea of focusing on specific aural sub-skills in order to improve overall listening performance is not new (see Field, 1998). However, due to the growing weight of evidence on the importance of reduced forms (e.g. Joyce, 2011; Matsuzawa, 2006), a focused pedagogic approach towards connected speech can be pursued with greater confidence.

The findings from this study need to be interpreted in light of its limitations. The research was conducted with eighteen to twenty-two year old Japanese university students, who ranged in proficiency from a false beginner to upper intermediate level. The homogeneity of the sample population had its benefits. For instance, since the participants were of a similar age, shared a common L1, and had experienced a comparable secondary

education; many important background variables were controlled. However, the uniformity of the sample also limits the range of participants with which the test can reliably be used. Before drawing any conclusions about the appropriateness of the test for a more diverse population, the piloting of the materials with participants from different L1 and educational backgrounds is especially important. A further limitation pertains to the scoring method. The award of a mark for correctly reproduced items in the same sentence is a violation of the local item independence assumption of the Rasch model. Instead of using Rasch's (1960) one parameter model scoring method, it would have been more appropriate to apply Master's (1982) partial credit model. Through this approach, items that are interdependent are considered as one polytomous item.

This study also raises a number of areas for future enquiry. A particularly pressing issue is the need for an empirically based account of the frequency of reduced forms. As previously discussed, given the lack of a more methodologically sound inventory, we are currently reliant upon the intuitions of researchers for our understanding of which aspects of connected speech are the most important. This lack of knowledge has contributed to the neglect of connected speech, and has thereby constrained the teaching, testing, and study of L2 listening. Furthermore, there is ample opportunity to explore the most effective means of teaching reduced forms. Through further research, it is anticipated that these issues will be better understood in the future. In addition, an analysis of the relationship between item difficulty and item content would shed light upon the relative comprehensibility of different kinds of reduced forms, provide further evidence for the validity of the test, and enable a comparison between the actual hierarchy of difficulty and what theory would predict.

References

Adams, R.J., & Khoo, S.T. (1993). Quest: The interactive test analysis system (Version 2.1.) [computer program]. Camberwell, Australia: Australian Council for Educational Research.

Avery, P., & Ehrlich, S. (1992). *Teaching American English pronunciation*. Oxford: Oxford University Press.

Azar, B. S. (1996). *Basic English grammar* (2nd ed.). Upper Saddle River, NJ: Prentice Hall Regents.

Belasco, S. (1965). Nucleation and the audio-lingual approach. *Modern Language Journal*, *49*, 482-489.

Bond, K. (2001). Reduced forms. Retrieved January 8, 2013 from http://www3.telus.net/linguisticsissues/ReducedForms.html

Bond, Z.S., & Garnes, S. (1980). Misperceptions in fluent speech. In R.A. Cole (Ed.), *Perception and production of fluent speech.* (pp. 115-132). Hillsdale, NJ: Erlbaum.

Bonk, W.J. (2000). Second language lexical knowledge and listening comprehension. *International Journal of Listening*, *14*, 14-31.

Bonk, W.J. (2001). Predicting paper-and-pencil TOEFL scores from KEPT data. *Research Institute of Language Studies and Language Education, Kanda University of International Studies, 12,* 65-86.

Bowen, J.D. (1976). Current research on an integrative test of English grammar. *RELC Journal, 7*, 30-37.

Brown, J. D. (2006). Authentic communication: Whyzit importan' ta teach reduced forms? In T. Newfields, et al. *Authentic Communication: Proceedings of the 5th Annual JALT Pan-SIG Conference.* (pp. 13-24). Retrieved January 8, 2013 from http://jalt.org/pansig/2006/HTML/Brown.htm

Brown, J.D., & Kondo-Brown, K. (Eds.). (2006a). *Perspectives on teaching connected speech: To second language speakers*. Honolulu, HL: Second Language Teaching & Curriculum Center, University of Hawaii Press.

Brown, J.D. & Kondo-Brown, K. (2006b). Testing reduced forms. In J.D. Brown & K. Kondo-Brown (Eds.), *Perspectives on teaching connected speech: To second language speakers* (pp. 247-264). Honolulu, HL: Second Language Teaching & Curriculum Center, University of Hawaii Press.

Brown, J., & Hilferty, A. (1986). Listening for reduced forms. *TESOL Quarterly*, *20*(4), 759-763.

Brown, J., & Hilferty, A. (1989). Teaching reduced forms. *Gendai Eigo Kyoiku*, January, 26-28.

Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.

Cauldwell, R.T. (1996). Direct encounters with fast speech on CD-audio to teach listening. *System*, *24*(4), 521-528.

Clarey, M. E., & Dixson, R. J. (1963). *The second pattern of English*. New York: Harper & Row.

Dirven, R., & Oakeshott-Taylor, J. (1984). Listening comprehension (Part 1). *Language Teaching*, *17*, 326-342.

Ellis, R. (1985). *Understanding second language acquisition*. Oxford, England: Oxford University Press.

Field, J. (1997). Notes on listening: Variability and assimilation. *Modern English Teacher*, *6*(1), 51-52.

Field, J. (1998). Skills and strategies: towards a new methodology for listening. ELT *Journal, 52*(2), 110-118.

Fountain, R.L., & Nation, I.S.P. (2000). A vocabulary-based graded dictation test. *RELC Journal, 31*(2), 29-44.

Goh, C.C.M. (2000). A cognitive perspective on language learners' listening comprehension problems. *System, 28*, 55-75.

Grosjean, F., & Gee, J.P. (1987). Prosodic structure and spoken word recognition. *Cognition, 25*, 135-155.

Henning, G. (1987). *A guide to language testing: Development, evaluation, research.* Cambridge, MA: Newbury House.

Henrichsen, L.E. (1984). Sandhi-variation: A filter of input for learners of ESL. *Language Learning*, *34*(3), 103-126.

Hough, D. (1995). *Before HearSay: Basic listening for the classroom*. New York: Addison-Wesley.

Ito, Y. (2001). Effect of reduced forms on ESL learners' input-intake process. *Second Language Studies*, *20*(1), 99-124.

Jones, K., & Ono, T. (2000). Reconciling textbook dialogues and naturally occurring talk: What we think we do is not what we do. *Arizona Working Papers in SLAT*.

Joyce, P. (2011). Componentiality in L2 listening. In B. O'Sullivan (Ed.), *Language testing: Theory and practice*. (pp. 71-93). Palgrave MacMillan, London.

Kline, R.B. (2005). *Principles and practice of structural equation modelling* (2nd ed.). New York: Guilford.

Kobayashi E., & Linde, R. (Eds.). (1984). *Practice in English reduced forms*. Tokyo: Sansyusya.

Matsuzawa, T. (2006). Comprehension of English reduced forms by Japanese business people and the effectiveness of instruction. In J. D. Brown, & K. Kondo-Brown, (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 59-66). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.

McNamara, T.F. (1996). *Measuring second language performance*. London: Longman.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*(2), 81–97.

Murphy, R. (2003). *Essential grammar in use with answers: A self-study reference and practice book for elementary students of English.* Cambridge: Cambridge University Press.

Nunnally, J.C. (1978). *Psychometric theory*. New York: McGraw-Hill.

O'Malley, J.M., Chamot, A., & Kupper, L. (1989). Listening comprehension strategies in second language acquisition. *English for Specific Purposes, 9*, 33-47.

Pemberton, R. (2003). Spoken word recognition and L2 listening performance; An investigation of the ability of Hong Kong learners to recognise the most frequent words of English when listening to news broadcasts. Unpublished Ph.D. thesis. University of Wales, Swansea.

Pickett, J.M., & Pollack, I. (1963). Intelligibility of excerpts from fluent speech: Effect of rate of utterance and duration of excerpt. *Language and Speech, 6*, 151-164.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. (Expanded edition, Chicago: The University of Chicago Press, 1980)

Roach, P. (2001). *English phonetics and phonology* (3rd ed.). Cambridge: Cambridge University Press.

Rogerson, M. (2006). Don'cha know? A survey of ESL teachers' perspectives on reduced forms instruction. In J.D. Brown & K. Kondo-Brown (Eds.), *Perspectives on teaching connected speech: To second language speakers* (pp. 85-98). Honolulu, HL: Second Language Teaching & Curriculum Center, University of Hawaii Press.

Rosa, M. (2002). Don'cha know? A survey of ESL Teachers' Perspectives on Reduced forms instruction. Retrieved January 8, 2013 from http://www.hawaii.edu/sls/uhwpesl/21(1)/Rosa.pdf

Rost, M. A., & Stratton, R. K. (1980). *Listening in the real world: Clues to English conversation*. Tucson, AZ: Lingual House.

Rubin, J. (1994). A review of second language listening comprehension research. *Modern Language Journal, 78* (2), 199-221.

Sheerin, S. (1987). Listening comprehension: Teaching or testing? *ELT Journal, 41*(2), 126-131.

The course of Japanese lower secondary school English. (n.d.). Retrieved January 8, 2013 from http://www.ne.jp/asahi/efl/2ndsc/CourseStudyLower.html

Tuckman, B.W. (1972). *Conducting Educational Research* (2nd ed.). NY: Harcourt.

Ur, P. (1984). ***Teaching listening comprehension***. Cambridge: Cambridge University Press.

Weinstein, N. (1982). *Whaddaya Say?*. Culver City. California: ESL Publications.

Appendix 1: Test Paper

Connected Speech Exercise: Instructions

Name: _____

You will hear 13 sentences. There will be a pause after each sentence. During the pause, write down the sentence that you have just heard on the line provided. Here is an example, do not write this time, just listen.

The correct answer was, *Excuse me you are too earl*y. When you write the sentence down, please use regular words only. So, *Scuse, yor tooery* would be incorrect. Also, please do not use contractions such as *you're*. Please spell the words that you write correctly! This is important. All the sentences you hear will be grammatically correct. Each sentence will be spoken only once. If you have any questions, raise your hand and ask your teacher now.

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____
7. _____
8. _____
9. _____
10. _____
11. _____
12. _____
13. _____

Appendix 2: Test Transcription

1. You would not tell him.
2. Do you want to do it?
3. I have got an idea.
4. He has got to call him.
5. Where are you going to play?
6. He would not see it.
7. There is not a lot of them.
8. I think he likes it.
9. He does not want her.
10. Sorry, I do not know him.
11. It is because of her.
12. Get your bag we are going.
13. There is another of them.

Appendix 3: Item Fit Statistics

Item	Infit Mean Square	Outfit Mean Square	Infit *t*	Outfit *t*	Item	Infit Mean Square	Outfit Mean Square	Infit *t*	Outfit *t*
1	1.00	.99	.1	.1	36	1.11	1.12	2.2	1.4
2	1.04	1.68	.3	2.2	37	1.09	1.08	2.0	1.0
3	1.04	1.18	.3	.7	38	.98	.94	-.3	-.4
4	.99	.94	-.2	-.6	39	.96	.93	-.6	-.4
5	.94	.90	-1.8	-1.3	40	.88	.62	-1.1	-1.9
6	1.03	1.37	.5	2.3	41	.94	.82	-.5	-.7
7	.99	.92	.0	-.2	42	.89	.79	-1.4	-1.7
8	.99	.89	-.1	-.8	43	.94	.88	-1.0	-1.1
9	.96	.86	-.7	-1.2	44	1.01	1.00	.4	.0
10	.99	.92	-.1	-.3	45	.93	.95	-1.4	-.6
11	.95	.69	-.2	-1.0	46	.99	.90	-.1	-.5
12	.94	.88	-1.1	-.9	47	.98	.85	-.1	-.6
13	.95	.90	-.7	-.6	48	1.00	.97	-.1	-.3
14	.98	.96	-.4	-.3	49	1.03	1.02	.6	.2
15	.90	.85	-2.5	-1.7	50	.99	1.39	.0	1.2
16	1.00	1.02	.1	.2	51	1.21	1.33	4.9	3.3
17	.99	.95	-.2	-.5	52	1.07	1.18	1.0	1.2
18	1.05	1.01	.7	.1	53	1.06	1.07	1.2	.7
19	.85	.81	-4.3	-2.6	54	1.04	1.01	.8	.1
20	.97	.97	-.7	-.4	55	1.05	1.51	.3	1.1
21	.86	.80	-4.2	-2.8	56	1.13	1.18	1.8	1.2
22	1.03	1.07	.6	.9	57	.94	.80	-.6	-1.3
23	1.03	1.02	.3	.1	58	.99	1.05	-.1	.5

24	.92	.91	-2.3	-1.2	59	1.06	1.15	.5	.7
25	1.10	1.14	1.2	.8	60	.98	.86	.0	-.3
26	.83	.79	-4.7	-2.9	61	1.04	1.13	.8	1.5
27	.88	.83	-3.6	-2.3	62	.81	.58	-1.2	-1.7
28	1.15	1.21	1.8	1.5	63	.95	.98	-.4	-.1
29	.92	.95	-.8	-.2	64	1.20	1.40	3.4	3.0
30	.96	1.07	-.3	.4	65	.97	.96	-.7	-.4
31	1.03	1.11	.3	.6	66	.97	.91	-.7	-.9
32	1.03	1.26	.3	1.0	67	1.02	1.00	.5	.1
33	.96	.94	-.7	-.4	68	1.07	1.13	1.1	1.2
34	1.08	1.28	1.8	2.6	69	1.06	1.14	.9	1.1
35	1.17	1.29	4.1	3.0					