JACET Journal 60 (2016) 97–115

Examining the Authenticity of the Center Listening Test: Speech Rate, Reduced Forms, Hesitation and Fillers, and Processing Levels

YANAGAWA, Kozo
Hosei University

Abstract

This study examined the extent to which a listening comprehension component of the Center Test in Japan attempts to be representative of real-life listening in relation to the nature of the input texts and the level of cognitive processing the test items require the test takers to be engaged in. The construct of the listening test was operationalized through the review of L2 listening theory, Course of Study, and local English teachers' views. Three components of the construct were elicited in terms of the input texts and one component was elicited in terms of the test items. These components were speech rate, reduced forms, hesitations and fillers, and processing level. Quantitative analysis and rater judgments were conducted using the Center Listening Test administered from 2007 to 2009. The results revealed that overall, the test items represent a range of cognitive processing levels, but the input texts do not necessarily represent the nature of real-life spoken texts. Recommendations are made to improve the Center Listening Test, and discussions consider how high-stakes achievement listening tests can be authentic.

Keywords: high-stakes achievement listening test, authenticity, input texts, processing levels

Introduction

Language testing cannot *per se* be authentic or real-life in so far as it is a test (Spolsky, 1985), and no listening test is an exception to this (Alderson, 2005). The use of multiple-choice questions (henceforth, MCQs), for example, is a commonly accepted compromise to the authenticity of language tests (Alderson, 2005). While this is a dilemma for the majority of language test developers, the relevant issue to be addressed concerns how and to what extent a particular listening test can be as "authentic" as possible in particular contexts. The notion of *interactional authenticity* proposed by Bachman (1991) and later renamed *interactiveness* (Bachman & Palmer, 1996) provides a theoretical framework for test developers to handle the dilemma. *Interactional authenticity* is concerned with the extent of the correspondence in knowledge, ability, and skills to be engaged between the test tasks and the target language use (TLU) tasks that the test takers are expected to encounter outside non-test situations. The higher the level of correspondence, the more interactionally authentic the test task is, and consequently the more valid the inferences that can be drawn from the test performances are. The notion of interactional authenticity is thus different from the conventional idea of authenticity, which exclusively focuses on the characteristics of the input texts.

In a similar fashion to Bachman (1991) and Bachman and Palmer (1996), Weir (2005) stresses the importance of eliciting real-world cognitive processing by establishing a symbiotic relationship between test tasks (*context validity*) and test takers' language abilities (*cognitive validity*). The notion of symbiotic relationship holds that the more representative the choice of tasks is, the more closely the processes elicited by the test tasks can replicate those processes that language users would employ in the real world (Weir, 2005).

Following Weir (2005), Field (2013) explicates cognitive validity by establishing two aspects that listening tests in particular are supposed to achieve: the similarity and comprehensiveness of cognitive processing. Similarity of processing refers to the extent to which the actual processes adopted during a listening test are similar to those that would be employed in TLU situations, whereas comprehensiveness of processing refers to the extent to which the items tap into a broad range of cognitive processes. If similarity is not achieved, Field (2013) equates the divergence to *construct-irrelevant variance* (Messick, 1989), whereas if comprehensiveness is not achieved, he equates the incompleteness to *construct under-representation* (Messick, 1989). That is, cognitive validity is considered an essential part of construct validity.

The cognitive validity of listening tests is most likely to be compromised by those three test components that include input texts (i.e., recording), test format (e.g., MCQ or summary writing), and test items (Field, 2013). The use of scripted texts, for example, should be a commonly accepted compromise that constitutes construct under-representation, especially when it comes to high-stakes tests. Furthermore, no matter what type of format listening tests employ, they cannot be free from the construct-irrelevant variance associated with a particular test format, such as guessing in MCQ or writing ability in summary writing. Furthermore, the presence and preview of test items prior to listening inevitably invites test takers to engage in strategy use (Yanagawa & Green, 2008), which makes their cognitive processing during the test different from that engaged in by language users in TLU situations (Field, 2013; Wu, 1996). Reducing construct under-representation and construct-irrelevant variance to a minimum is necessary so that listening tests can enhance cognitive validity or interactional authenticity.

Field (2013) investigated the cognitive validity of Cambridge English listening tests, and found that the higher the test level is, the wider the range of cognitive processing levels that are covered. Note that Cambridge English is not a national achievement test that needs to accord with a particular national syllabus. Exploring the cognitive validity is all the more necessary if the tests are high-stakes achievement tests, because they are more likely to cause profound and larger effects on ELT than those proficiency tests that do not comply with a national syllabus. Nevertheless, little research has been conducted into the cognitive validity of the Center Listening Test (henceforth, CLT), which is the main focus of this study.

In addition to cognitive validity or interactional authenticity, a national high-stakes achievement test is supposed to achieve consequential validity (Kane, 2002; Messick, 1996). Consequential validity includes washback effects on teaching and learning as testing programs purport to promote certain outcomes as the "engines of reform" (Kane, 2002, p. 33). In fact, the CLT was introduced to enhance students' practical communication abilities in

English in 2006 (MEXT, 2002), because this skill has been identified as an immediate national priority in Japan (Watanabe, 2013). Research shows, however, that the CLT has limited washback effects on ELT in Japan, particularly on the students (e.g., Yanagawa, 2014), suggesting that consequential validity has not been achieved. Although there is in essence some distance between achievements at schools and proficiency in real-world communicative events, school learning, which is carefully structured step by step, and the appropriate assessment which follows is indispensable if students are to reach proficiency (Shohamy, 1996). In this sense, the quality of achievement tests should be examined. Thus, this study aims to reveal the extent to which the CLT achieves interactional authenticity or cognitive validity and to help the CLT to achieve more consequential validity.

Context of the Study: The CLT in Japan

The Center Test is a unified national achievement test for all high school subjects including English, in Japan. Designed and produced by the National Center for University Entrance Examinations (henceforth, the NCUEE), the Center Test purports to measure the students' achievement level at the point of finishing the final year of upper-secondary education, and the content must be aligned with the Course of Study for secondary schools, as prescribed by the Ministry of Education, Culture, Sports, Science and Technology (henceforth, MEXT).

Used for gate-keeping purposes by all national and local public universities and more than 90% of private universities across Japan, the Center Test is also a high-stakes test intended to discriminate between candidates according to their academic proficiency level (NCUEE, 2007a). Thus, the Center Test has attracted more than 500,000 test takers every year over the last decade (NCUEE, 2015b) and is very influential not only for high school students who want to progress to tertiary institutions but also for formal English education conducted at high schools (Watanabe, 2013).

The CLT is composed of four major parts, each of which addresses different listening skills as Table 1 shows. The number of items is 25, all of which employ MCQs. The CLT is administered by an audio device with a headset, which is distributed individually to each test taker, suggesting how high-stakes the CLT is. The voice actors are native speakers of English with North American accents.

Table 1
The CLT (Based on NCUEE, 2015a)

Part		Targeted abilities and skills/situations or topics	# items	# texts
D	1	The ability to understand conversations (approx. 30 words)/shopping, travelling, computer, etc.	6 (24%)	6
	2	The ability to understand the purpose and situation of the conversation and the function of language by listening to conversations (20 to 30 words)/classroom activity, watching movies, restaurant, etc.	7 (28%)	7
	3A	The ability to understand the information, the situation, and the speakers' intention by listening to medium-length conversations (approx. 50 words)/DVD rental, lost things, a problem with a cup of coffee, etc.	3 (12%)	3
	3B	The ability to understand the information, situation, and speakers' intention by listening to a longer conversation (approx. 150 words)/London Paralympics.	3 (12%)	1
M	4A	The ability to grasp the main points of medium-length monologues (approx. 100 words)/wedding anniversaries in the U.K., a hotel advertisement, and the national flag of Palau.	3 (12%)	3
	4B	The ability to grasp the main points of a longer monologue (approx. 200 words)/ Helen Keller and a dog.	3 (12%)	1
		Total	25	21

Note. D: dialogues, M: monologues

Tanaka and Sage (2007) and Ito, Kawamura, Shimada, Nishihara, and Funato (2007) investigated the construct validity of the CLT and found a low validity of the inferences drawn from CLT scores. They attribute this primarily to the lack of interactive and integrated tasks in the CLT, where the test takers are given the role of "passive overhearers" (Ito et al., 2007) rather than interlocutors in non-test situations, which require a combination of speaking and listening abilities. Note that, however, the current practice of the CLT measures the ability to respond appropriately in TLU situations in Part 2 (see Table 1), and that a larger coverage of interactional listening could invite a construct-irrelevant variance to listening ability (i.e., speaking ability) to the extent that the score interpretation can be contaminated.

More importantly, these earlier studies did not explore the interactional authenticity or cognitive validity of the CLT, a key notion for language test validation (Bachman, 1991; Field, 2013; Weir, 2005). Although the TLU situations are specified in the Course of Study and the domain is operationalized by the NCUEE as different listening skills as seen in Table 1 (NCUEE, 2007a, 2008a, 2009a), the question is still open to what extent the characteristics of the domain of the test tasks found in the CLT correspond to those of real-life domain of the TLU situations. That is, we don' know to what extent interactional authenticity or cognitive validity is achieved in terms of input texts and test items, which are most likely to be under-representative or construct irrelevant.

Meanwhile, there is an immediate need to validate the current CLT to provide useful insights into the development of a new test. A government panel recently proposed that the current Center Test will be reformed (MEXT, 2015). While the discussion of English tests by the government panel focuses on the introduction of speaking and writing sections in order to have a four-skills English test, a discussion on how the quality of the existing listening and reading sections can be improved is lacking. This study will inform the development of the new test.

Operationalizing the Listening Construct

The interactional authenticity and cognitive validity of listening tests resides on the symbiotic relationship between the test tasks, including the input texts and test items, and test takers' cognitive processing (language abilities) elicited through the test tasks (Weir, 2005). The listening construct was therefore primarily operationalised by the input texts (i.e., recorded material) and the test items. A review of the Course of Study and local English teachers' perceptions of the CLT was also conducted to operationalize the construct, since the NCUEE admits that it acknowledges the Course of Study in the development of the CLT (NCUEE, 2007a, 2008a, 2009a). The NCUEE also explicitly states in Uchida and Otsu (2013) that the CLT should follow the actual teaching and learning at schools, which is most accessible by referring to English teachers (Winke, 2011). For this reason, the perceptions of English teachers, as representatives of the stakeholders of the CLT, were referred to for the construct definition. The NCUEE collects English teachers' views of the CLT from two different sources and reports them every year (NCUEE, 2007a, 2008a, 2009a), to both of which this study refers. One is from the Association of High School English Teachers in Japan (*Zen-eirenn*), whose membership reaches approximately 60,000, and the other comprises anonymous English teachers, although the NCUEE does not make clear how and how many of the latter were recruited. The perceptions of another stakeholder, high school students, were not referred to because their perceptions were not considered necessary for the purpose of theoretical construct definition.

Among the myriad characteristics of the input texts (Buck, 2001; Révész & Brunfaut, 2013), three acoustic features that are unique to listening (but not common between reading and listening) and considerably affect listening processing were chosen as the key components of the construct. They are a) speech rate (e.g., Buck & Tatsuoka, 1998), b) reduced forms (e.g., Brown & Brown, 2006), and c) hesitations and fillers (e.g., Blau, 1991). Research (e.g., Goh, 2000; Field, 2003, 2008) shows that the extent of the correspondence in listening skills or processing to be engaged between the test tasks and TLU tasks is considerably dependent on these three features, because they considerably affect the first phase of L2 listening processing, including perception, word recognition, and lexical segmentation. If speech rate is too fast or "natural" for L2 listeners, then the first phase of processing will not work. The spoken texts delivered with "natural" speed and associated with many reduced forms make word boundaries in continuous sound streams blurry, and word recognition and segmentation more difficult for the listeners (Goh, 2000). Hesitations and fillers may allow L2 listeners more time to process the input text, but they may require readjustment on the part of the listeners, by constantly demanding that they abandon successful attempts to decode the input text. Thus, speech rate, reduced forms, and hesitations and fillers were operationalized key components of the construct.

Among the characteristics of the listening test items (Buck, 2001), the processing level in which each item attempts to engage the test takers (e.g., Field, 2013) was operationalized by the component of the construct for the purpose of this study. Each component is explained and described below.

Speech rate. "Natural" speech rate presents L2 listening difficulties (Buck & Tatsuoka, 1998). While what constitutes "natural" speech rate is a complex issue, Tauroza and Allison (1990) provide the baselines for typical speech rates relative to different text types: approximately 263 syllables per minute (spm) for conversations and 249 spm for radio monologues. Note that, however, the proficiency level of test takers should not be neglected in considering the appropriate natural speech rate of an achievement listening test in particular (Green, 2014; Wagner, 2014a). This is where a necessary compromise should be made in the development of achievement listening tests, especially when they are targeting the students with lower proficiency level. Field (2013) revealed that the listening comprehension component of the Preliminary test, which forms part of Cambridge English and is considered to test the ability levels similar to the CLT, employs an approximate speech rate of 236 to 247 spm for dialogue items and 256 spm for monologue items. Both Cambridge English: Preliminary and the CLT seem to target test takers of approximately A2 or B1 Level on the CEFR (Cambridge English, 2015). Thus, it was considered reasonable to assume that the appropriate natural speech rate for the CLT be operationalized at approximately 240 spm for dialogue items and 250 spm for monologue items.

Local English teachers (i.e., *Zen-eirenn* and anonymous high school English teachers) are very positive about the current speech rate of the CLT, saying, "The speech rate is appropriate and should be acceptable, taking into account the current practice of English teaching at upper-secondary schools in Japan" (NCUEE, 2009a). They even add that the CLT provides a "standard model of speech rate for ELT in Japan" (NCUEE, 2009a). The Course of Study, on the other hand, does not mention this component.

Reduced forms. Another key component of the input texts is reduced forms. They include, for example, assimilation and weak forms. Reduced forms are unique to listening as they never occur in reading where writing systems are established and word boundaries are clearly marked by white spaces (Alderson, 2005). In listening, reduced forms make word boundaries in connected speech ambiguous, and word recognition and segmentation more difficult for listeners than readers (Goh, 2000). This is more problematic for Japanese learners of EFL in particular, who are predominantly exposed to written text rather than to spoken text because of the English entrance examinations to higher education institutions, which prioritize reading comprehension or receptive vocabulary knowledge. Consequently, their phonological expectations might be biased largely by the written text (Field, 2003), and their perceptions of the spoken text are weak.

Local English teachers seem to support the current practice of omitting reduced forms from the CLT, saying that the NCUEE should maintain the current "clarity of articulation" (NCUEE, 2008a) with which the test takers can cope. The Course of Study does not mention reduced forms at all.

Hesitations and fillers. Hesitations and fillers are another key component unique to listening to on-line human speech. Research shows that two-person face-to-face conversations contain 5.5 to 6 hesitation and filler words per 100 (Oviatt, 1995; Bortfeld, Leon, Bloom, Schober, & Brennan, 2001) while monologues contain 3.6 words (Oviatt, 1995). Meanwhile, the effect of hesitations and fillers on listening comprehension for L2 listeners is inconclusive (see Blau, 1991 versus Voss, 1979). They may allow L2 listeners more time to process the input text or they may require readjustment on the part of the listeners. Either way, the ability to cope with hesitations and fillers is clearly important in the real world (e.g., Wagner, 2014b), and accordingly, it should be tested.

Local English teachers seem to support the current practice of hesitations and fillers in the CLT (NCUEE, 2008a, 2009a), and the Course of Study does not mention hesitations or fillers at all.

A range of processing levels. In order for listening tests to achieve interactional authenticity or cognitive validity, listening processing during the listening test should be representative of that which language users would employ during real-life listening (e.g., Weir, 2005). Sometimes language users only have to recognize words or comprehend an utterance to meet their communicative needs. At other times, they need to interpret or infer beyond the textual information. That is, listeners in real-world listening events are engaged in a wide range of processing levels. In test situations, where the items chiefly determine the levels of processing (Field, 2013, p. 137), the test should therefore include those items that elicit processes that cover as many as possible of a range of processing levels if it claims to achieve cognitive validity. The CLT is no exception to this.

The Course of Study specifies "grasping the main points of the input texts" as an important listening component of the construct (MEXT, 2007), and this component is claimed by the NCUEE to be a chief target of Part 4 of the CLT as presented in Table 1. Also, local English teachers highly value those items that require the test takers to assemble the information scattered around a text into a coherent mental representation (NCUEE, 2009a), even though this argument requires empirical support. That is, both the Course of Study and local English teachers seem to be positive that the CLT should include the items requiring a broad range of processing levels.

Little research, however, has been conducted to provide empirical evidence about the extent to which the CLT reflects the three acoustic features and a wide range of processing levels while remaining in accordance with the Course of Study and local English teachers' views. Neither has been conducted to verify the NCUEE's claim that Part 4 of the CLT measures the ability of test takers to grasp the main points of the spoken texts (NCUEE, 2007a, 2008a, 2009a) as specified in the Course of Study (MEXT, 2007). That is, we do not know whether or to what extent the CLT achieves interactional authenticity or cognitive validity and consistency with the Course of Study. This question remains open and there is an imminent need for it to be answered, since the CLT causes potentially profound backwash effects on ELT in Japan, and the answers may also inform the development of the new test (MEXT, 2015) by providing a viable proposal for revising and improving the quality of the current CLT. Thus, the following two research questions were posed in this study.

RQ 1: To what extent is the CLT similar to real-life listening in relation to the nature of the input texts, namely speech rate, reduced forms, and hesitations and fillers?

RQ 2: To what extent is the CLT comprehensive of real-life listening in relation to cognitive processing levels? In particular, does Part 4 of the CLT measure the listening ability to catch the main points of the input texts?

Method

The 2007, 2008, and 2009 forms of the CLT were used for investigation into the *speech rate, hesitations and fillers*, and *processing level*, whereas the 2008 and 2009 forms were used for *reduced forms* as shown in Table 2. The use of the earlier forms of the CLT was considered not to lower the validity of the investigation to a significant degree, because both the earlier forms and the latest one share the same syllabus, namely the Course of Study (MEXT, 2007), and it was assumed that no significant change has been made to the test tasks (NCUEE, 2007a, 2008a, 2009a, 2015a). The materials were obtained from the NCUEE website (2007b, 2008b, 2009b) where all of the test items, scripts, and listening sound files were available for the public.

Speech rate was calculated, and hesitations and fillers were counted. Reduced forms and processing level were subjected to expert judgment, following Révész and Brunfaut (2013), where two raters rated the degree to which listening texts are explicit on a five-point Likert scale. In the present study, two raters (raters A and B) judged reduced forms on a seven-point Likert scale (7 = *very much reflective of real-life listening* to 1 = *not at all reflective of real-life listening*) while three raters (raters A, C, and D) judged the processing level. Quantitative analysis followed the rater judgment of processing level as seen in Table 2.

Rater A was an expert in English phonetics and phonology while rater B was an expert in language testing. Both raters, whose first language was Japanese, were so regularly exposed to natural English in their everyday lives and had such an expertise regarding English phonetics that the unavailability of native speakers of English in the judgment was considered as a compromise and rater training was not conducted. Rater C had wide theoretical knowledge and practical experience in test construction with a PhD in the field of language testing, and rater D was an experienced high school teacher with an MA degree in the field of L2 listening. Rater training was conducted to enhance the validity of the judgment of processing level.

Table 2
Method for the Components of the Listening Construct

RQ	Components	Forms	Quantitative analysis	Rater judgment
	speech rate	'07–'09	✓	
RQ1	reduced forms	'08–'09		✓
	hesitations & fillers	'07–'09	✓	
RQ2	processing level	'07–'09	(✓)	✓

Note. For the analysis of processing level, rater judgment was followed by quantitative analysis.

Research Question 1

Speech rate. *Speech rate* was calculated by separating dialogue items in Parts 1 to 3 from monologue items in Part 4, since the speech rate baseline varies relative to text type (Field, 2013; Tauroza & Allison, 1990). The researcher calculated the speech rate for each item text (excluding instructions and test questions) and computed the mean number of spm for each part for each form, given that each part targets different listening skills (see Table 1). In the present study, following the L2 listening literature (Field, 2013; Tauroza & Allison, 1990), the natural speech rate was operationalized at approximately 240 spm for the dialogue items and 250 spm for the monologue items.

Reduced forms. Reduced forms (Brown & Brown, 2006; Field, 2008) involved assimilation, weak form, elision, and formulaic expressions. Formulaic expressions were operationalized in the present study as a sequence stored in the mental lexicon of native speakers of English in the form of language chunks. They include syntactic relationships such as [wɑnə] for *want to* and lexical phrases such as [mɔːmɔː] for *more and more* (see Appendix A, Field, 2008, p. 155).

Four different excerpts from four different parts for each of the forms of the CLT administered in 2008 and 2009 were selected (see Table 2), so that the materials the raters were asked to listen to and rate could be representative of the CLT. Eight texts were used as the material to be rated. A rating sheet was developed (see Appendix A), addressing each component of reduced forms, and a holistic judgment of reduced forms. A holistic judgment (overall impression) of naturalness was included as an item since the addition of a holistic rating could provide a more appropriate indication of naturalness than an aggregate of judgments of each distinct feature of reduced forms.

The rating sheet employed a seven-point Likert scale. Five (indicating "to some extent") on the scale was taken as the minimum requirement for acceptable representativeness. Since four indicated only "uncertain," it was considered insufficient to reflect naturalness, while six indicating "fairly" was considered strictly as a minimum level of adequacy. The rating sheet was presented to rater A, an expert in English phonetics, prior to the coding to ensure that the sheet was valid for judging reduced forms. In response to his comments, a few minor changes were made to the examples of each feature to make it clearer for the raters. The actual question (prompt) provided to the raters was "To what extent do you think the CLT represents real-life listening in terms of the features (assimilation, weak form, elision, formulaic expressions)?" The raters were asked to participate in the rating in January 2011. In cases where two raters differed in their ratings, a final judgment was made by the researcher.

Hesitations and fillers. *Hesitations and fillers* referred to filled pauses (non-lexical sounds such as *uh, uhm,* and *er*), fillers (i.e., *you know, well, I mean,* and *let me see*), repetitions for repairs, and false starts and subsequent repairs. False starts were defined as utterances that are started and then either abandoned or reformulated in some way (Foster, Tonkyn, & Wigglesworth, 2000). The researcher attempted to identify all of the hesitations and fillers across the 2007–2009 forms by checking the scripts available on the website of the NCUEE (2007b, 2008b, 2009b), while listening to the CD recordings of the three forms.

Research Question 2

A range of processing levels. Henning (1991) and Alderson (2005) informed the definition of *a range of processing levels*. Henning (1991) classified three-sentence dialogue items of the listening component of the Test of English as a Foreign Language Paper-based

Test (TOEFL PBT) by processing hierarchy in accordance with the breadth of the textual information (one, two, and three sentences) needing to be processed. This approach would facilitate rater judgment so that the consensus among the raters would be easier to reach (Henning, 1991). Alderson (2005) classified listening test items into three types according to what sort of listening abilities are measured: items to measure the ability to make inferences on the basis of what was heard, items to measure the ability to identify the main point or idea, and items to measure the ability to listen intensively for specific details. Following Henning's (1991) and Alderson's (2005) approaches, the present study established three hierarchical processing levels, namely situation-model level, discourse-model level, and propositional- or word-processing level. Situation-model level refers to the level of understanding of the whole text, including inferencing or interpreting. In real-world listening events, listeners are often supposed to interpret or infer beyond the textual information (Alderson, 2005). Discourse-model level refers to the level of understanding of two or more utterances by finding the coherence between them or identifying the main idea. Propositional- or word-processing level is only concerned with comprehending an utterance or recognizing particular words or specific details. The actual question provided to the three raters (raters A, C, and D) was, "What processing level is required to arrive at a correct answer?"

A rater training session was held to improve the validity of the ratings. The listening comprehension component of Cambridge English: Preliminary was used for the practice as it was considered to be at a similar proficiency level to the CLT. The three raters were asked to rate the seven items from Part 1 of the Cambridge English: Preliminary according to the three nominal scales, and to convene in Tokyo in 2008 to discuss any difficulties they had experienced in their coding, and the discrepancies between the coding. This training session and the discussion that followed were assumed to facilitate their ratings and raise the validity of the rater judgment.

The three raters judged the processing level in August 2009, when a packet of coding materials was sent to them, including a coding sheet, a set of CD recordings, the test brochures for each form of the CLT, the answer keys, and the tape scripts. Their rate of agreement will be presented in the results section.

Results

Research Question 1

Speech rate. The results for the speech rate are presented in Table 3 in relation to the text types. Interestingly, the mean speech rate did not significantly vary across the text types and forms, converging around 210 to 215 spm, except for the relatively faster speech rate of 229 spm for the monologue items in the 2009 form. Since approximately 240 and 250 spm were the operationalized natural speech rates for dialogue and monologue items, respectively, the result of a range of 210 to 215 spm for dialogue items and a range of 210 to 229 spm for monologue items was found to be slower than the baseline, suggesting that the CLT under-represents authentic speech rate.

Table 3

Speech Rate by Syllables per Minute

	1	2	3A	3B	*Mean (SD)*	4A	4B	*Mean (SD)*
2007	201	228	209	223	215 (10.8)	209	217	213 (4)
2008	184	217	217	221	210 (15.1)	215	204	210 (5.6)
2009	229	222	201	204	214 (11.7)	223	234	229 (5.4)

Reduced forms. Table 4 shows the raters' coding results for each component of reduced forms and the holistic impression. The two raters (raters A and B) agreed on the ratings for weak form, elision, and holistic impression, where both "A" and "B" were indicated in the same cell in Table 4, and their ratings were taken as final (thus, Table 4 does not show the researchers' codings for those forms). When the two raters disagreed, the researcher's rating was referenced. If the researcher's rating was identical with either rater A or B, then that rating was taken as final. This occurred for assimilation. When 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. It is important to note that the difference in judgment between raters A and B was within one scale point for four items among the five, suggesting that this judgment was overall reliable.

Five on the seven-point Likert scale was taken as the cut-off point for acceptable "naturalness." Although formulaic expression was rated as "uncertain" (4), the other three components, namely weak forms (6), assimilation (5), and elision (5) were judged acceptable. Above all, holistic judgment was rated as reflecting real-life listening "to some extent" (5), suggesting that overall the CLT reflects a realistic level of reduced forms.

Table 4

Ratings of Naturalness of Reduced Forms

	(7)	(6)	(5)	(4)	(3)	(2)	(1)
Assimilation		B	**A, R**				
Weak form		**A, B**					
Elision			**A, B**				
Formulaic expressions			B	**A**	R		
Holistic impression			**A, B**				

Note. A and B: raters, R: the researcher. Bold indicates the final decision for each component. R is only referred to when the codings of raters A and B differed. (7):very much (6):fairly (5):to some extent (4):uncertain (3):not sufficiently (2): little (1): not at all

Hesitations and fillers. This analysis was conducted according to text type. The total numbers of words for dialogue items for the 2007, 2008, and 2009 forms were 602, 601, and 490 words, respectively, while it was 408, 486, and 620 words for monologue items. Few *hesitations* and *fillers* were found across the text types and across the forms. In dialogue items as shown in Table 5, five fillers (w*ell* (4) and *you know* (1)) and one filled pause (*uhm*) (1.2%) occurred for the 2007 form, four fillers (*well* (4)) and five filled pauses (*uhm*) (1.5%) for the

2008 form, and four fillers (*well* (3) and *let me see* (1)) and one filled pause (*uhm*) (1.5%) for the 2009 form. These figures are far below the established baseline (i.e., 5.6 to 6% by Oviatt, 1995) for conversations. In the monologue items, only one false start (*The number is, oh*) (0.2%) occurred, only for the 2009 form. This figure is again much lower than the baseline for monologue items (3.6% by Oviatt, 1995). No repetition for repairs occurred across the three forms. These results suggest that the CLT clearly under-represents a realistic level of hesitations and fillers.

Table 5

Frequency of Hesitations and Fillers for Dialogue Items

	Filled pauses (*uh)*	Fillers (*you know, well)*	Repetition for repair	False start	# of words
2007	1	5	0	0	602
2008	5	4	0	0	601
2009	1	4	0	0	490

Research Question 2

A range of processing levels. The three raters' (raters A, C, and D) coding diverged, as shown by the rates of exact agreement, namely 52%, 44%, and 16% for the 2007, 2008, and 2009 forms, respectively. Consequently, the final decision about each coding was made according to the agreement of the codings: where two or three raters agreed, this decision was taken as final. The researcher's coding was referred to only when the three raters' codings differed, which occurred for five codings out of the total number of items (5/75, 7%) across the three forms (four items for the 2007 form, one for the 2008 form, and none for the 2009 form). That is, at least two raters reached an agreement for 70 items (93%), which could have compensated for the low rate of exact agreement among the three. Examples of test items for each processing level are illustrated in Appendix B.

As Table 6 shows, in the results for the required processing level for each form, three hierarchical processing levels are required across the forms. This result suggests that the CLT represents the comprehensiveness of the processing level required of those in non-test situations. In addition, relatively more items at the discourse-model level were found across the forms than the items that required propositional- or word-processing levels; 14 items (56%), 8 items (32%), and 12 items (48%) were identified at the discourse-model level for the 2007, 2008, and 2009 forms respectively while 8 items (32%), 6 items (24%), and 8 items (32%) were identified at the propositional- or word-processing levels. This is consistent with the perceptions of local English teachers who acknowledge those items that require a higher level of processing (NCUEE, 2009a).

Table 6
Processing Level

	Propositional- or word-processing	Discourse-model	Situation-model
2007	8 (32%)	14 (56%)	3 (12%)
2008	6 (24%)	8 (32%)	11 (44%)
2009	8 (32%)	12 (48%)	5 (20%)

A sub-component question of RQ 2 was whether Part 4 of the CLT measures the test takers' ability to identify main points of the input texts as specified in the test specification (NCUEE, 2007a, 2008a, 2009a). The results showed that five (28%) and seven items (39%) out of the total number of eighteen items in Part 4 across the three forms (six items for each form, see Table 1) were found to be situation-model and discourse-model levels, respectively, whereas six items (33%; two items for the 2008 form and four items for the 2009 forms) were found to be propositional- or word-processing level. The result, that twelve items (28% + 39% = 67%) cover either the situational- or discourse-model level, partly verifies the NCUEE's claim that Part 4 measures the ability to identify the main points of the input texts. Meanwhile, another result, that no item in Part 4B for the 2009 form was found to be either situation-model or discourse-model level, shows that Part 4 does not necessarily address the main points of the input texts but also specific details of the input texts, suggesting that the NCUEE's claim is not fully supported.

Discussion and Recommendations

Research Question 1

RQ 1 asked to what extent the CLT is similar to real-life listening in relation to the acoustic features of the input texts. The results showed that overall, the CLT includes a realistic level of reduced forms, but it is not representative of real-life listening to the extent that it reflects "natural" speech rate and a realistic level of hesitations and fillers. The current speech rate, approximately 210 to 215 spm, was found to be relatively slower than the operationalized natural speech rate, approximately 240 to 250 spm. Fewer hesitations and fillers (1.2 to 1.5%), false starts (0.2%), and repetitions (0%) in particular, were found than the established baseline, 3.6 to 6 per cent. These findings are not consistent with the perceptions of the local English teachers, who accept the current practice as "natural" (NCUEE, 2009a).

The lack of "natural" speech rate and hesitations and fillers can partly be explained by the NCUEE's attitude that they should keep mirroring the current practice in high school classrooms in Japan (Uchida & Otsu, 2013), where the speech rate of recorded materials can be slowed down and include few hesitations and fillers. Another explanation is the respect for tradition regarding enunciation used in high-stakes listening tests, where "polished" (Wagner, 2014b) and professional recordings are considered fair to students, because they are assumed to display "good" pronunciation.

Here are specific and feasible recommendations for the improvement of the nature of the input texts of the CLT. First, the speech rate should be increased to 240 spm for dialogue items on average and 250 spm for monologue items. The proposed speech rate, 240 spm for dialogue items in particular, is still so much slower than natural speech rate (263spm), by 7.6%, that this should be acceptable even to targeted EFL learners with a lower proficiency level. Note that L2 listening difficulties for EFL learners are largely attributed to a lack of exposure to spoken texts delivered at a natural speed. The introduction of "natural" speech rate into the CLT is expected to encourage the introduction of spoken texts delivered at a "natural" speed into high school classrooms whereby students can become more familiar with real-life listening tasks. Second, a more realistic level of hesitations and fillers, which make up approximately 6 per cent of any conversation and 3.6 per cent of any monologue, is recommended. In addition to making the input text more real-life, a more realistic level of hesitations and fillers may facilitate L2 listening processing by providing increased processing time for the lower proficiency test takers (Blau, 1991). These two small but significant changes to the CLT, it is hoped, will change teachers' attitudes toward the choice of listening materials to be used in the classrooms (Wagner, 2014b), and thereby students' attitudes toward learning to listen.

Some may argue that the CLT should be as it is and keep following the current teaching practice (Uchida & Otsu, 2013), given that it is an achievement test targeted at EFL learners with a lower proficiency level and that English teachers seem satisfied with the quality of the input texts of the current CLT (NCUEE, 2009a). However, since the CLT has not achieved fully interactional authenticity, it is our collective responsibility as test developers and language experts to make necessary and feasible changes to the CLT.

Research Question 2

RQ 2 asked to what extent the CLT is comprehensive of real-life listening in relation to the cognitive processing tapped into by the test items. The results showed that the CLT is comprehensive of real-life listening to the extent that it attempts to require the test takers to engage in a broad range of cognitive processing levels. This confirmed the consistency between the local English teachers' perceptions and the CLT as to the importance accorded to the cognitive processing of the listeners. Although it is difficult to say what proportion of each processing level is representative of real-life listening, this result should be considered good practice and maintained.

RQ 2 also asked whether Part 4 of the CLT taps into the ability to grasp the main points of the input texts. The result, that nearly two-thirds (67%) of the items in Part 4 measure this ability, partly verifies the NCUEE's claim, confirming the consistency between the Course of Study, the test specification, and the CLT. The result, on the other hand, suggests that the test specification for Part 4 should be modified to state that it assesses both main and non-main points of the input texts (see Table 1). This modification would enhance the consistency between actual test contents and test specifications, and lead to higher validity of the score-based interpretations and use.

Conclusion

This study examined the extent to which the CLT attempts to be representative of TLU tasks. The results showed that the CLT attempts to be representative in terms of the cognitive processing that test items require the test takers to engage in, but it is not necessarily representative in terms of the acoustic nature of the input tests. This result can provide two major implications for the development of achievement listening tests. First, even for tests targeted at lower proficiency EFL students, it is achievable to represent a range of processing levels, while according with the syllabus and English teachers' views. This is noteworthy in that the coordination between the syllabus, the test, and the teachers' support may produce the greatest effects on teaching and learning, and that processing cannot be a compromise to the authenticity of achievement listening tests where it is very often almost impossible to achieve full authenticity in terms of the input texts and test format. The remaining issue concerns how the representation of the test items in relation to processing levels should represent the cognitive processing language users would employ in real-world listening events, given the gap between test developers' intention (i.e., items) and test takers' behavior (Wu, 1998). Second, it is necessary for test developers and stakeholders, including English teachers in particular, to be aware of and reconsider their underlying assumption that "polished" or "good" pronunciation is only appropriate and fair for a high-stakes test. If EFL learners are predominantly exposed to slowed down and "polished" enunciation, without hesitations and fillers, they are most likely to be deprived of the opportunity not only to attune to "natural" spoken text but also to enhance those listening skills, abilities, and strategies that would be useful in real-world listening events. The CLT should play a leading role in overcoming the tradition.

While the present study has implications for the improvement of the CLT, it also shows that high-stakes achievement listening tests in general can only be possible if they strike a balance between the syllabus, candidates' proficiency levels, stakeholders' views, necessary compromises relating to the administration, and authenticity, confirming that the issue of authenticity in high-stakes achievement listening tests is complex and will continue to be a major challenge that language testers have to tackle.

There are many limitations to this study, only a few of which are named here due to space limitation. First, the use of an impressionistic measure to investigate *a range of processing levels* and the low rate of exact agreement among the raters that followed, is a major limitation of this study. Neither did this study explore what processing level the test takers are actually engaged in during the test. The test takers may not necessarily be engaged in the processes that the raters in the present study have identified for each item (e.g., Wu, 1998). Introspective studies, such as the use of verbal protocols, should potentially allow for more accurate and in-depth analysis regarding their actual processing during L2 listening tests. Second, the number of phonological features (or the operationalization of interactional authenticity) investigated in the study was very limited. Further study should explore interactional authenticity more extensively by investigating other phonological features such as rhythm, pitch, and articulation rate. Third, this study did not explore the latest versions of the CLT. Further study should confirm the findings of this research by examining the latest forms of the CLT. It is for these reasons that the results of this study should be interpreted with reservations.

In conclusion, this study has revealed that the CLT, on one hand, attempts to be representative of real-life listening in terms of the processing levels the language users are engaged in, suggesting that the item quality of the CLT is potentially high and that the CLT is in tandem with the Course of Study and local English teachers. On the other hand, this study has shown that the CLT under-represents the acoustic features of real-life spoken texts, suggesting that the text characteristics require improvements to achieve interactional authenticity. It is hoped that further improvements, which may include the introduction of integrated tasks, will help the CLT and the new replacement test (MEXT, 2015) to serve as “engines of reform” and furnish tangible effects on ELT in Japan, which are yet to be seen.

References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment.* London: Continuum.

Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly, 25,* 671–704. doi:10.2307/3587082

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Blau, E. K. (1991). More on comprehensible input: The effect of pauses and hesitation markers on listening comprehension. ERIC DOC No. ED 340 234.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech, 44,* 123–147. doi:10.1177/00238309010440020101

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: University of Hawai'i.

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

Buck, G., & Tatsuoka, K. (1998). Application of the rule-spaced procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15,* 119–157. doi:10.1177/026553229801500201

Cambridge English. (2015). *The Cambridge English Scale.* Retrieved May 7, 2015, from http://www.cambridgeenglish.org/exams/cambridge-english-scale/

Field, J. (2003). Promoting perception: Lexical segmentation in L2 listening. *ELT Journal, 57,* 325–333. doi:10.1093/elt/57.4.325

Field, J. (2008). *Listening in the language classroom.* Cambridge: Cambridge University Press.

Field, J. (2013). Cognitive validity. In A. Geranpayeh, & L. Taylor (Eds.), *Examining listening* (pp. 77–151). Cambridge: Cambridge University Press.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21,* 354–375. doi:10.1093/applin/21.3.354

Goh, C. C. M. (2000). A cognitive perspective on language learners' listening comprehension problems. *System, 28,* 55–75. doi:10.1016/S0346-251X(99)00060-3

Green, A. (2014). *Exploring language assessment and testing: Language in action.* London: Routledge.

Henning, G. (1991). A study of the effects of variation of short-term memory load, reading response length, and processing hierarchy on TOEFL listening comprehension item performance (*TOEFL Research Reports 33*). Princeton, NJ: Educational Testing Service.

Ito, H., Kawamura, A., Shimada, Y., Nishihara, M., & Funato, S. (2007). *Daigaku shingaku* *yoteisha-wo taishô-to shita eigo nouryokusiken-no kokusai-hikaku* [A comparison of high-stakes English tests used for gate-keeping purposes for universities or colleges: *The Center Test* in Japan and *Matriculation Examination* in Finland]. *Shikoku-eigo kyôiku gakkai kiyo* [Journal of Shikoku English Education Society], *27*, 11–26.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*, 31–41.

Messick, S. (1989). Validity. In R. L. Linn (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*, 241–256. doi:10.1177/026553229601300302

MEXT (2002). *Eigoga tsukaeru nihon-jin-no ikusei-no tameno sennryaku kôsô-no sakutei-ni tsuite* [Strategic plan to cultivate Japanese with English abilities]. Retrieved March 10, 2007, from http://www.mext.go.jp/b_menu/shingi/chousa/shotou/020/sesaku/020702.htm

MEXT (2007). The course of study for foreign languages. Retrieved March 10, 2007, from http://www.mext.go.jp/english/shotou/030301.htm

MEXT (2015). *Kôdai setsuzoku-kaikaku jikkô purann* [Action plan for reforming the linkage between high schools and higher education]. Retrieved May 7, 2015, from http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo12/sonota/1354545.htm

NCUEE (2007a). *Daigaku Nyūshi Sentā shiken-no mondai-to seitô* [Questions and answers of the Center Test]. Tokyo: Author.

NCUEE (2007b). 19 *risningu* [Listening for the 2007 form]. Retrieved June 3, 2009, from http://www.dnc.ac.jp/modules/center_exam

NCUEE (2008a). *Daigaku Nyūshi Sentā shiken-no mondai-to seitô* [Questions and answers of the Center Test]. Tokyo: Author.

NCUEE (2008b). 20 *risningu* [Listening for the 2008 form]. Retrieved June 3, 2009, from http://www.dnc.ac.jp/modules/center_exam

NCUEE (2009a). *Daigaku Nyūshi Sentā shiken-no mondai-to seitô* [Questions and answers of the Center Test]. Tokyo: Author.

NCUEE (2009b). 21 *risningu* [Listening for the 2009 form]. Retrieved June 3, 2009, from http://www.dnc.ac.jp/modules/center_exam

NCUEE (2015a). *Daigaku Nyūshi Sentā shiken-no mondai-to seitô* [Questions and answers of the Center Test]. Tokyo: Author.

NCUEE (2015b). *Shigansha-sū jyukensha-sū tô-no suii* [Transition about the number of applicants and actual test takers]. Retrieved May 11, 2015 from http://www.dnc.ac.jp/data/suii/suii.html

Oviatt, S. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language, 9*, 19–35. doi:10.1006/csla.1995.0002

Révész, A., & Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition, 35*, 31–65. doi:10.1017/S0272263112000678

Shomahy, E. (1996). Language testing: Matching assessment procedures with language knowledge. In Birenbaum, M., & Dochy, F., (Eds.), *Alternatives in assessment of achievements, learning processes, and prior knowledge* (pp. 142-160). Boston, MA: Kluwer Academic Publishing.

Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing, 2,* 31-40. doi: 10.1177/026553228500200104

Tanaka, N., & Sage, K. (2007). How authentic is the English listening section of the NCT for the EFL context in Japan? *Ningen Bunka Ronsou, 9,* 113-129.

Tauroza, S., & Allison, D. (1990). Speech rates in British English. *Applied Linguistics, 11,* 90-105. doi:10.1093/applin/11.1.90

Uchida, T., & Otsu, T. (2013). *Daigaku Nyūshi Sentā shiken-eno lisuningu tesuto-no dônyūni itaru rekishitekikeii-to sono hyôka* [The historical background of English listening comprehension tests in the National Center Test and their evaluation]. *Nihon Testo Gakkaishi* [Japanese Journal for Research on Testing], *9,* 77-84.

Voss, B. (1979). Hesitation phenomena as sources of perceptual errors for non-native speakers. *Language and Speech, 22,* 129-144. doi:10.1177/002383097902200203

Wagner, E. (2014a). Assessing listening. In A. J. Kunnan (Ed.), *The companion to language assessment,* vol. 1 (pp. 47-63). Malden, MA: Wiley Blackwell.

Wagner, E. (2014b). Using unscripted spoken texts in the teaching of second language listening. *TESOL Journal, 5,* 288-311. doi:10.1002/tesj.120

Watanabe, Y. (2013). The National Center Test for University Admissions. *Language Testing, 30,* 565-573. doi:10.1177/0265532213483095

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* Basingstoke: Palgrave Macmillan.

Winke, P. (2011). Evaluating the validity of a high-stakes ESL test: Why teachers' perceptions matter. *TESOL Quarterly, 45,* 628-660. doi:10.5054/tq.2011.268063

Wu, Y. (1998). What do tests of listening comprehension test?—A retrospective study of ESL test-takers performing a multiple-choice task. *Language Testing, 15,* 21-44. doi:10.1177/026553229801500102

Yanagawa, K. (2014). Introduction of the Center Listening Test: Perceptions of English teachers and students. *Eigo-Tembô* [ELEC Bulletin], *121,* 66-73.

Yanagawa, K., & Green, A. (2008). To show or not to show: The effect of item stems and answer options on performance on a multiple-choice listening comprehension tests. *System, 36,* 107-122.

Acknowledgements

The author is grateful to Miyoko Kobayashi and three anonymous reviewers for their constructive comments on an earlier version of this paper.

Appendix A

Ratings of Naturalness of Reduced Forms

Please check ✓ the appropriate box.	(7)	(6)	(5)	(4)	(3)	(2)	(1)
Assimilation (e.g., *As you know*)							
Weak form (e.g., *John has seen it.*)							
Elision (e.g., *tell her, next station*)							
Formulaic expressions (e.g., *kina, wanna, gonna, more and more*)							
Holistic impression							

Note. (7): very much (6): fairly (5): to some extent (4): uncertain (3): not sufficiently (2): little (1): not at all

Appendix B (key is underlined)

Situation-model level: Question 14 in Part 3 of the 2008 form

Woman: 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.

Man: So, what'll you do?

Woman: Well, clean clothes are the most important thing.

[Question: What will the woman buy first?]

1) A suitcase. 2) A TV. 3) A washing machine. 4) Some clothes.

Discourse-model level: Question 21 in Part 4 of the 2009 form

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, uhm, I hardly stay in my hotel room, so I'll contact you again tomorrow. Thanks. Talk to you later.

[Question: 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.

Propositional- or word-processing level: Question 4 in Part 1 of the 2008 form

Woman: Hi, I'd like five ten-cent stamps and two one-dollar stamps, please.

Man: All right. Anything else?

Woman: No, but I only have a twenty-dollar bill.

Man: No problem.

[Question: How much are the stamps?] 1) $2.50 2) $5.10 3) $17.50 4) $20.00