How Authentic Is the English Listening Section of the NCT

for the EFL Context in Japan?

Development based on the TOEFL® iBT English listening section

TANAKA Nozomi, SAGE Kristie

センター試験のリスニングは日本の EFL 環境でどれだけ使えるか

—— TOEFL® iBT のリスニングセクションに基づく考察 ——

田　中　望　美*・セージ・クリスティー

要旨

本研究は Sage と Tanaka（2006b）による TOEFL® iBT とセンター試験の聴解を比較・対照した前研究に引き続き、テスティングを真正性の面からさらに深く考察してゆく。センター試験のための学生に対して不利になりうる教育を懸念し、センター試験の問いと基準（若しくは、対象言語領域）がこれまでの研究や理論及び TOEFL® iBT の特徴に基づいて、厳密に調査されている。本研究の構成は、1）過去の研究における真正性と相互性に関する概念、2）テストにおける伝達能力の側面、3）センター試験と TOEFL® iBT のテストの問いと対象言語領域の比較及び分析、4）聴解の真正性の今後の方向性と提案、となっている。要するに、TOEFL® iBT が対象言語領域で必要とされる能力を測定している点、また統合テスト形式を採用しているという 2 点から、本研究はセンター試験の聴解は TOEFL® iBT の特徴に倣ってさらに改定していくことができると示唆する。

キーワード：センター試験、真正性、テストタスク、基準、対象言語領域、相互性、伝達能力、統合テスト

1. Introduction

In previous research Sage & Tanaka (2006b) discussed the National Centre Test's (NCT) English listening section (pioneered in 2006) in terms of construct validity, based on Bachman and Palmer's test usefulness framework (1996, p.18) (see Figure 1). Due to the pivotal role that the NCT plays in the secondary and tertiary environments, hence the larger educational system in Japan, it is considered beneficial to conduct further investigation into the NCT's test usefulness for the development and quality control of this standardized test (see Figure 2). For this paper, the element of test usefulness that will be focused on is authenticity via using qualitative analysis. Authenticity is a controversial issue in the Second Language Acquisition (SLA) field. It is often defined in a variety of ways dependent on the specific research area of SLA, such as classroom activities and materials (Thomas, 1995); teaching styles; and, testing or assessment (Barlow,

2000). As Thomas (1995, p.3) admits, "I am left with proposing a definition that will doubtless satisfy no one completely".

This article will endeavour to reduce some ambiguity by addressing authenticity from the perspective of testing. One key to achieving authenticity in testing, according to McNamara (1996, p. 43) is, "the degree of correspondence between the characteristics of a given test task and of a particular language use task." As Carroll (cited in Doyle, 1991) emphasizes, an important facet in language testing should always be authenticity, since from tests, we make inferences (McNamara, 2000). As McNamara (2000) highlights, the difficulty involved in making tests authentic is well-known. Nevertheless, this paper proposes that the English listening section of the TOEFL® iBT test tasks have higher authenticity than the NCT's and are more reflective of real life language use. Additionally, since the link between content validity and authenticity is considered significant it will be explored. Also of import is the interplay between authenticity and interactiveness, which refers to a test task's language quality and hence, is closely connected to construct validity (Bachman & Palmer, 1996).

Usefulness=Reliability+Validity (Construct & Content) + Authenticity+Interactiveness+Impact+Practicality

Figure 1 : A graphic representation of test usefulness adopted from Bachman and Palmer (1996, p.18)

2. Contemporary tests: national and international

To address the aforementioned theoretical aspects in more practical terms, this paper attempts to briefly cover how content validity, authenticity, and interactiveness contribute to making the TOEFL® iBT English listening section more integrative. In saying this, it is important to note that the TOEFL® test has undergone continuous revision since it started in 1976 (ETS, 2005). Conversely, the English listening section of the NCT was inaugurated in 2006 (It was piloted and revised prior to its first administration) (NCUEE, 2006b). Due to decades of development, the TOEFL® iBT would be expected to inevitably provide a better test construct. Moreover, it is recognised worldwide as an international test of English language communication. By conducting a parallel comparison of the NCT and TOEFL® iBT English listening sections, the areas where this test could be further developed and made more authentic for communicative purposes will be highlighted. Drawing attention to the areas which can be improved for future versions would simultaneously encourage it to become more in line with the Action Plan of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) to cultivate "Japanese with English language abilities'".

3. MEXT's Action Plan

The MEXT Action Plan emphasizes listening and speaking abilities for the primary purpose of "English as a means for communication" (MEXT, 2003). Furthermore, MEXT's Action Plan states that Japanese need to be well versed in the international language of English to be able to interact in the world arena, and as a result, for the Japanese to be better understood by the international community in professional and/or research fields. MEXT mentions, in the evaluation system section of the Action Plan, that while "… entrance examinations tend to serve as final goals," TOEFL, TOEIC and STEP are important benchmark tests (MEXT, 2003, p. 15).

This is supported by McNamara (2000, p.74) who states, "the existence of tests such as TOEFL [which are] used as gate-keeping mechanisms for international education, and administered to huge numbers of candidates all over the world, has effects beyond the classroom, in terms of educational policy and the allocation of resources to education." Brown (1995ab) also discusses the broad applications of the use of TOEFL, which is often referred to as a model example of a Norm-Referenced Test (NRT), administered worldwide, and taken by millions people annually. As a matter of fact, MEXT also has adopted the TOEFL test as an indicator of

students' academic achievements (Toyama, 2003). Thus, the ubiquity of the TOEFL test in Japan has also been established.
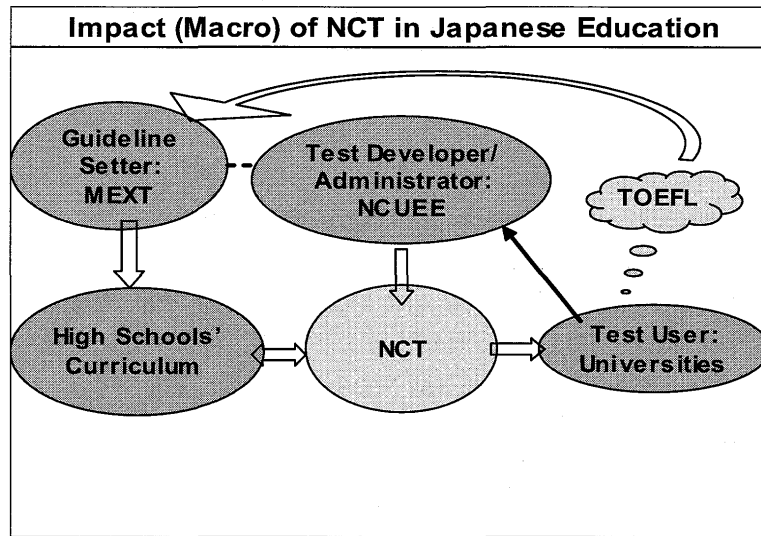

Impact (Macro) of NCT in Japanese Education

Figure 2 : NCT's Impact on Japanese Education
(Sage & Tanaka, 2006a)

4. Literature review

4.1. Test validity

Validity is a fundamental notion to consider during the processes of test development and test evaluation (Chalhoub-Deville & Turner, 2000) (see Figure 1). McNamara (2000) defines content validity as "the extent to which the test appropriately samples from the domain of knowledge and skills relevant to performance in the criterion" (p. 132). In addition, Fulcher (1999) cites Messick (1989) to highlight the importance of consensus among professionals' judgments for: a) an item's content relevance; and b) the level to which the appropriate domain is covered or not by the knowledge and skills being tested. Commitment to ensuring that a test generalizes to the target domain and follows test specifications (which provide the details for doing this), in the case of an achievement test, is vital. To establish any one test's content involves careful sampling from the domain which the test is ultimately examining (McNamara, 2001). Thus, when selecting a test's content, it must be synchronous with the domain, and the need for empirical confirmation is pertinent (Fulcher, 1999). If in the case that the test is a proficiency test, it ought to be testing future and "real life" language known as criterion. The criterion, however, is unobservable in a test. Instead, the test provides a medium for measuring performances from which inferences about performance in the criterion can be made (McNamara, 2000) (see Figure 3).
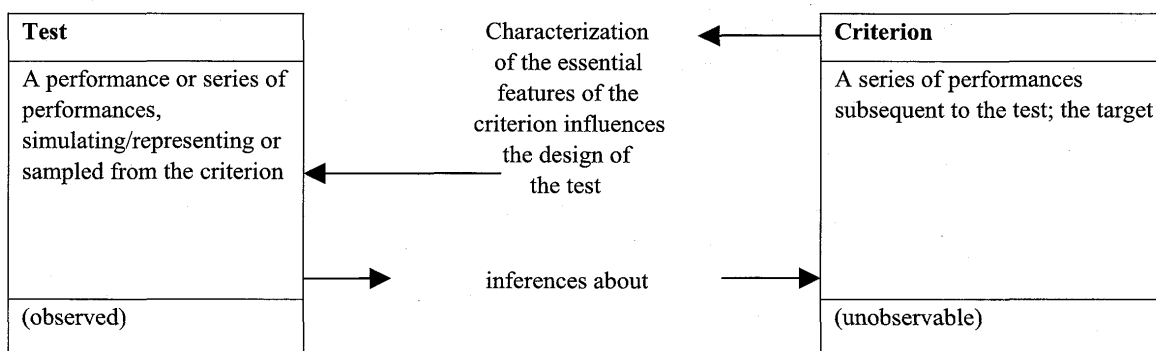
Test	Characterization of the essential features of the criterion influences the design of the test	Criterion
A performance or series of performances, simulating/representing or sampled from the criterion		A series of performances subsequent to the test; the target
	inferences about	
(observed)		(unobservable)

Figure 3 : Test and criterion (McNamara, 2000, p.8)

4.2. Test authenticity and communicative testing

From Bachman & Palmer's (1996) test usefulness framework, it can be seen that authenticity is an important element (see Figure 1). Bachman and Palmer (1996) state that performance demanded from language tests needs to correspond with the language use required in the target language use (TLU) domain (see Figure 4). These researchers further emphasize that authenticity is the quality which interprets the test taker's performance for generalization to its criterion. However, McNamara (2000) points out that there are limitations to authenticity in testing, including: the difficulty of making the test items replicate the TL domain; observer's paradox; and, classroom oriented lessons, which are not realistic.

Regarding tests which examine communicative ability, Weir (1990) emphasises that the test construct must facilitate communication tasks which closely resemble the situations a test-taker would face in the TL domain. In other words, they are more authentic. Additionally, Buck (2001, p.83) indicates authenticity to be a core element of communicative testing and points out that "the important thing is not what a person knows about the language, nor how grammatically correct they are, but whether they can actually use it to communicate in the TLU situation, [that is] the real-world situation in which the language will be used." This can be exemplified by the following. When a student listens to a lecture during it, the lecturer may invariably ask students to provide a real-life example for what they have just heard. In this way, a test with high authenticity might ask test takers to watch a lecture via a computer-simulated video and provide an example. On the other hand, a test with lower authenticity might adopt a construct where students would be expected to listen from an audio recording to a lecturer's script and complete a gap-fill exercise on his talk.
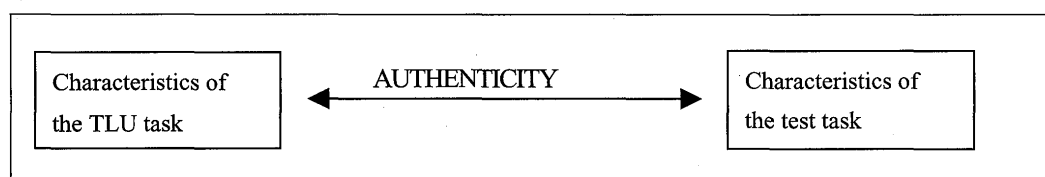

Figure 4 : Authenticity (Bachman and Palmer, 1996, p.23)

4.3. Interactiveness and tests

Another important element of Bachman and Palmer's test usefulness framework is interactiveness, which is defined as "the extent and type of involvement of the test taker's individual characteristics in accomplishing a test task" (Bachman & Palmer, 1996, p.25) (see Figure 1 & 5). Specifically, a test taker's language ability, including language knowledge, strategic competence or metacognitive strategies, topical knowledge, and affective schemata, are counted as such characteristics (Bachman & Palmer, 1996) (see Figure 5). According to Bachman and Palmer (1996), for a test task to show a high level of interactiveness depends on its degree of correspondence with construct validity. Thus, the importance of well-defined test taker characteristics and the construct is clear (Bachman & Palmer, 1996). Since if this is not the case, it is difficult to infer language ability based on an examinee's test performance when the test task does not demand that their language knowledge is used, despite a high level of interaction (Bachman & Palmer, 1996).
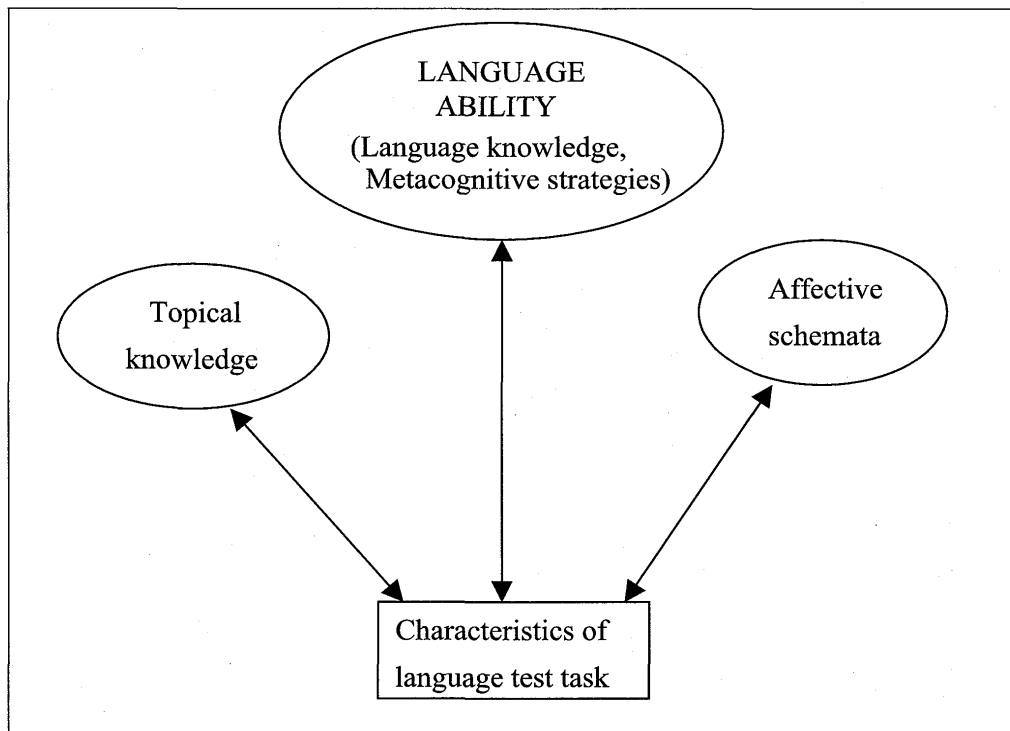
Figure 5 : Interactiveness: Main facets of a language test task's characteristics (Bachman and Palmer, 1996, p.26)

4.4. Distinction between authenticity and interactiveness

Since authenticity and interactiveness are both inextricably linked to construct, validity is required first to clearly establish how they differ. For interactiveness, it is the degree of the test taker's involvement when they are solving questions which assess: communicative competence; background knowledge; and, affective schemata (Bachman & Palmer, 1996). However, in the case of authenticity, it is highly dependent upon the level to which the test materials and conditions are able to replicate the TLU situation (McNamara, 2000; Bachman & Palmer, 1996).

4.5. Communicative competence and testing

As previously discussed, the MEXT Action Plan advocates the improvement of English language ability to equip Japanese with skills to communicate in the language, internationally. With regard to communicative testing, Fulcher's (2000) research is relevant, as he articulates the manner in which the communicative approach is affecting modern day language testing. He claims that language assessment has begun to be looked at in terms of communicative tests involving: a) performance; b) authenticity; and, c) scores based on real-life outcomes (Fulcher, 2000). Since the significant teaching methodology of communication is Communicative Language Teaching (CLT), prominent researchers make the consensual assumption that this pedagogy is informed by communicative competence models (Celce-Murcia, et. al, 1995).

Initially, models of language knowledge paid little attention to the role of language performance (McNamara, 2000). However, in the early 1980s Canale and Swain's (1980) work on communicative competence included related aspects (McNamara, 2000). By 1995, this model had been further built upon with Bachman and Palmer (1995) proposing communicative language abilities to be in a "hierarchical, multi-level form" but with components sharing the competencies of: discourse; linguistic; actional; sociocultural; and, strategic (Celce-Murcia et.al., 1995). Weir (1990) states that models of language competence provide a potentially useful framework for the design of language tests. Many researchers believe that not only the knowledge of these competencies, but also the learner's actual ability to use these skills or knowledge in meaningful communicative situations are important in CLT (Morrow,1979; Canale and Swain, 1980). Thus, if it is envisioned to test

authentic communication in Japan, it is vital that some sort of communicative competence model be incorporated. In this way when developing tests it can provide for a solid guidance more effectively promoted through test specifications.
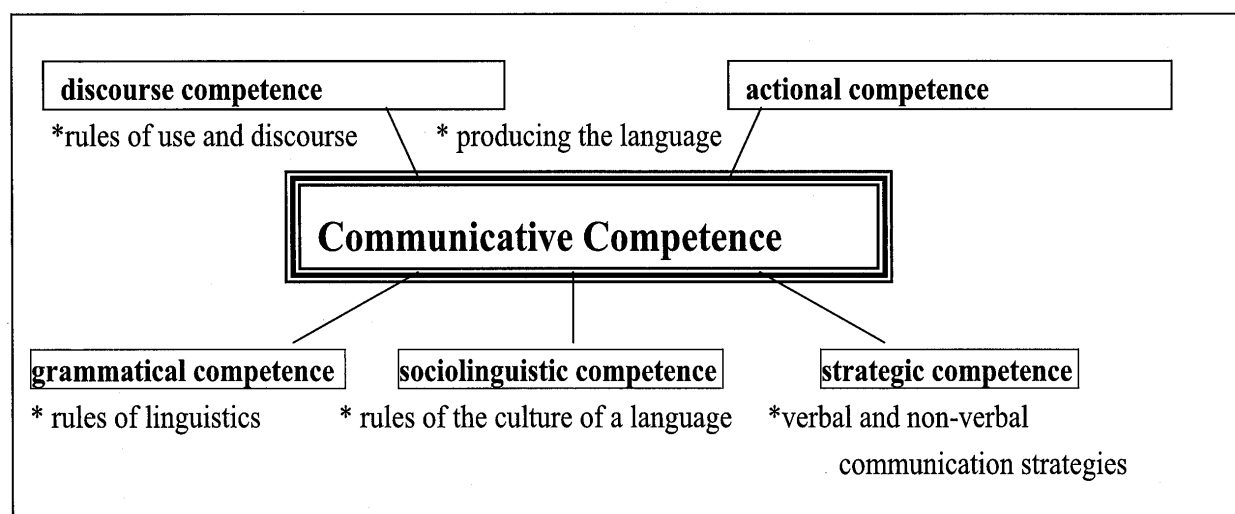

Figure 6 : An adapted communicative competence model (Adapted from Canale and Swain, 1980; and Celce-Murcia et. al. 1995)

5. Research Method

5.1. Approach to the research

The research methodology that has been chosen is qualitative, therefore, it is more ethnographic rather than hypotheses-based. Furthermore, the authenticity focus of this research can be delved into as findings are presented in a more unconstrained fashion (Larsen-Freeman & Long, 1991. p.11). Hence, the following analysis will compare the English listening sections of the NCT and TOEFL® iBT by centring on the relationship between the characteristics of their respective test tasks and will present the effectiveness of their respective domains of generalization. While an expansion of the larger connections in which authenticity plays an integral part has been presented in Figure 7, it is the aim of this qualitative research to analyse the tests in line with the parts of this diagram which directly affect authenticity.
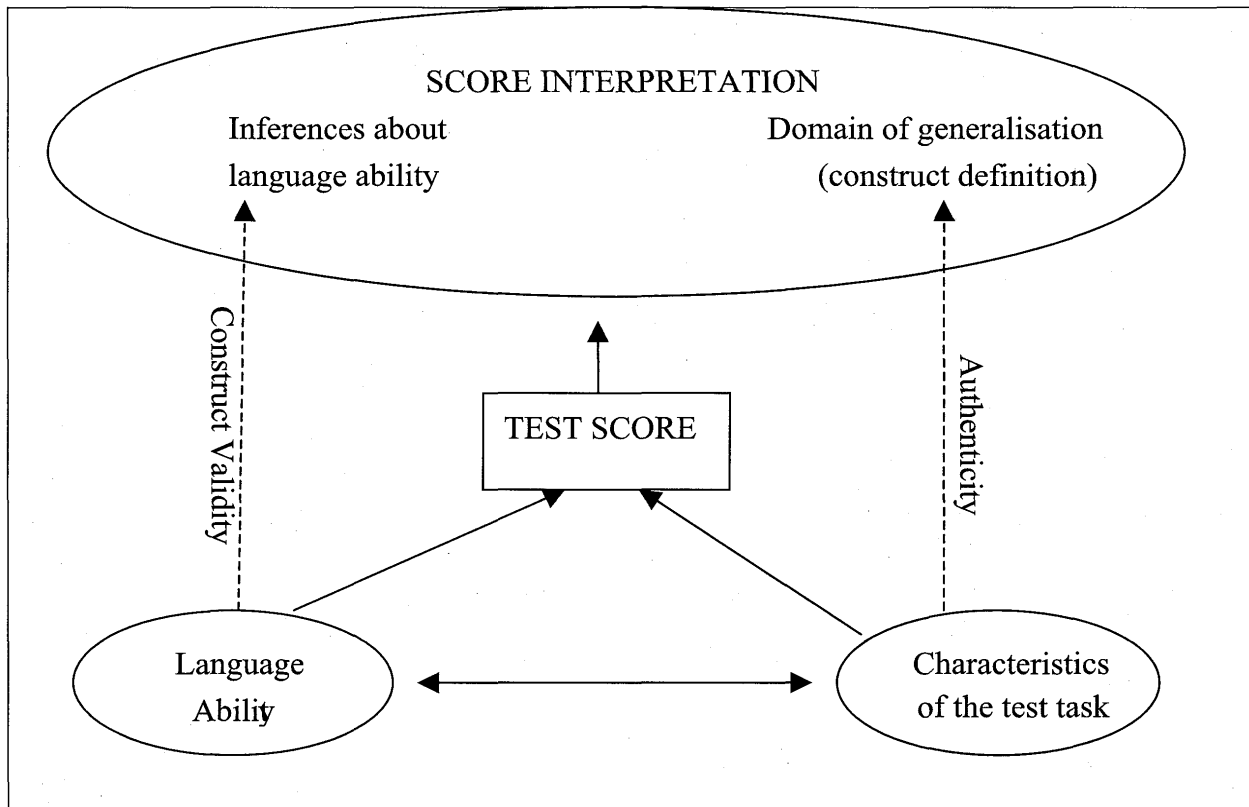
Figure 7 : Construct validity of score interpretations (Bachman and Palmer, 1996, p.22)

5.1.1. Selection of two tests

The newest version of TOEFL, the TOEFL® iBT, is used as a model examination because of its worldwide market use and emphasis on interactive, authentic, and communicative testing approaches (ETS, cited in Sage & Tanaka, 2006b). It also impacts vastly on English education in Japan (see Sage & Tanaka, 2006b for details). Since TOEFL® iBT commenced in 2005, there are no past papers yet available for public viewing. Therefore, the 2006 ETS Testing Guidelines have provided the basis to compare its authenticity with that of the NCT. Thus, the actual 2006 English listening section test of the NCT was compared to sample questions from the TOEFL® iBT.

5.1.2. Presentation of findings

First and foremost, the respective features of the English listening sections of the TOEFL® iBT and NCT will be presented adjacently in a table format (see Table 3). Then, sample questions for both tests will be looked at by analysing their criterion, purpose, format (method and construct), and content (situations), after which the test task's levels of authenticity will be critiqued and discussed. Finally, based on the aforementioned, suggestions are offered for ways that the English listening section of the NCT can become more theoretically accountable for achieving communicative competence goals. Addressed also will be the TLU domain of the TOEFL® iBT, which covers both academic English and conversational English, considered a fitting combination for the university domain, irrespective of EFL or ESL environments.

5.2. Authenticity and interactiveness of the TOEFL® iBT and NCT
5.2.1. TOEFL® iBT and the EFL environment

Although it is acknowledged that the primary purpose of the TOEFL® iBT is to examine the appropriateness of prospective students who wish to enter a university in the United States (U.S.), that is, an

ESL environment. As discussed previously in this article, MEXT recommends, and universities in Japan do use TOEFL scores for a variety of different purposes and MEXT recommends this as well. Thus it is a relevant test for the Japanese EFL environment. Unlike the NCT English listening section, the TOEFL® iBT requires that students are able to perform both academic and non academic listening related language functions for the following examples:

> 1) Operating in an academic environment: read textbooks, listen to lectures, take notes, ask questions, write reports, take exams, consult with professors, discuss class with classmates, deal with administrative matters, and so on.
>
> 2) Operating in a non-academic environment: to carry out daily life and functional conversations (buy things, order food, take a bus, et cetera.)

In Japanese university contexts, students do use English in academic situations such as, to listen to lectures by native speakers, to visit an English-speaking professor's office, ask questions, write reports, take tests, and read English textbooks. Therefore, even though it is a placement test test for an EFL environment, the TOEFL® iBT is applicable for Japanese universities, provided there are content-based courses in English offered. In fact, such courses are becoming more widespread in Japan (Furmanovsky, 1996). For example, at the University of Aizu, 50% of faculty members are non-Japanese (Sawyer, 1998). Courses in economics, literature, politics, law, and other subjects, are taught in English in many major universities[2].

5.2.2. Analysis of the two tests: authenticity and interactiveness

Table 3 shows whether the NCT and TOEFL® iBT include elements of authenticity and interactiveness or not. Test questions of the NCT and TOEFL® iBT have been categorized and analyzed. When comparing the NCT column with the TOEFL® iBT column, common features for both can be clearly seen. That is, for the NCT crosses (X) dominate its respective TLU column. This indicates that the passages are purely conversational and include no academic content. Juxtaposed to this, the TOEFL® iBT has many circles (O) under its TLU column. This is primarily attributed to the ESL environment context, when English is used as the common medium for interaction in university settings such as, lectures, conversations with professors and among classmates, and in administrative offices. The manner in which the analysis has been conducted for Table 3 gives rise to this paper's argument that the TOEFL® iBT's English listening section has higher authenticity than the NCT's, as it contains more academic listening problems.

Table 3 : Authenticity and interactiveness of the TOEFL® iBT and NCT

Categories	TOEFL® iBT		NCT	
	test	TLU	test	TLU
Situations	Academic:		Short passages:	
	− lecture by professor	○	− talking about his/her	△
	− class discussion	○	experience	
	Conversational:		− leaving a message on an	△
	− on campus between		answering machine	
	・　　students	○	− orientation announcement	△
	・　　student &	○	− story telling	○
	administration officer	○	Conversational:	
	・　　student & professor		− at a restaurant between	
			・　　customer & waiter	×
			− on a bus between	
			・　　passengers	×
			− airport between	
			・　　friends	×
			− in the office between	
			・　　office workers	×
			− at a department store between	
			・　　customer & clerk	×
			− in other places between	
			・　　friends	×
			・　　strangers	×
			− on campus between	×
			・　　workers	×
			・　strangers	×
			・　student & secretary	△
			・　friends or classmates	
Format	Independent skills	×	Multiple choice questions	×
	・Multiple choice questions (reading of choices involved)		(reading of choices involved)	
	Integrated skills			
	・writing, speaking, &			
	reading involved	○		
Strategic competence	Note taking	○		
Discourse competence	Grasp overview & main point	○	Understanding cues (voice tone & hesitation)	○
	Understanding non-verbal cues	○		
	Understanding the function behind what is said (=MI) (see eg. T-4)	○		
	Understanding the speaker's attitude (=DM) (see eg. T-5)	○		
	Understanding discourse organization (=EGI) (see eg. T-6)	○		
	Connecting content (=EGI) (see eg. T-7)	○		
	Making inferences (=MI) (see eg. T-8)			

Table 3: Authenticity and Interactiveness of the TOEFL® iBT and NCT (Continued...)

Categories	TOEFL® iBT		NCT		
	test	TLU	test	TLU
Grammatical competence	Gist-Content (see eg.T-1)	○	Understand special expressions	○
	Gist-Purpose (see eg. T-2)	○		
	Detail (see eg. T-3)	○		
Background knowledge	Read & interpret visuals (such as tables, graphs, etc.)	○		

NOTE: X = rarely seen in the TLU domain; △ = unlikely to happen in the TLU domain; ○ = of high occurrence in the TLU domain

6. The TOEFL® iBT

6.1. The TOEFL® iBT's authenticity

The TOEFL and IELTS tests measure the English ability of non-native speaking students who wish to study in English speaking nations (McNamara, 2000). The scores obtained from these two major English tests are used to ascertain whether a student could operate competently in a university environment (domain of generalisation) where English is the medium of instruction. Therefore, the test content and tasks of these tests correspond to the TLU domain, and if they are comprised of university lectures and associated teaching and learning activities, that is, whether they are authentic or not (Bachman & Palmer, 1996).

It has been demonstrated by Table 3 that the test method for the TOEFL® iBT and NCT is different and each test has its own unique characteristics. This table also highlights the TOEFL® iBT as being more complex. The major difference is the integrated section of the TOEFL® iBT, which as previously stated, is considered to have relatively high authenticity, primarily attributed to the nature of the test questions and response formats closely simulating actual situations in which students participate in university environments abroad or content-based academic lectures conducted by native speakers, in an EFL environment. This is due to the fact that the TOEFL® iBT includes integrated tasks such as reading a passage, followed by listening to a lecture, then speaking or writing a summary of their opinion on what they have just read and listened to. Supportive of this is Low (1986), who states high authenticity and context adherence as major advantages of the integrated testing method.

6.2. The TOEFL® iBT's interactiveness

As for its interactiveness, the TOEFL® iBT incorporates graphs and visual material into its rubric to support academic lectures, and candidates are required to manipulate this supplemental information to understand the content. Furthermore, after closer investigation of the integrated tasks, it can be seen that the test takers need to use their topical knowledge and various linguistic and cognitive strategies to state their opinions or write summaries (see Table 3). Most of all, the integrated skills tasks reflect the aforementioned realistic situations, and ask for language performance and productivity from test takers, which resemble the demands of American universities or content-based English courses in EFL environments. Therefore, the TOEFL® iBT is considered to also be high in interactiveness (as well as authenticity).

7. NCT

7.1. Authenticity

From Buck's (2001) viewpoint, the NCT English listening section could be seen to incorporate aspects of communicative language testing, such as dialogues between students on campus and functional language at

a restaurant. However, for students who expect to enter universities in Japan, that is, an EFL environment, the TLU required is that of understanding lectures in English and communicating with English speaking professors, not ordering food at a restaurant, as in Figure 8 (see Example C-1). However, in some lower level English conversation classes native speakers require students to practice these types of activities. In order to justify the se claims, it would be necessary to conduct extensive research as to what conversational textbooks are being used, which although beyond the scope of this present article, is recommended for future research. The research would need to find out for each university, how many classes are: 1) content-based and conducted in English, 2) conversational and conducted in English and which textbooks are used, and 3) content-based courses such as literature, but which are not conducted in English.

Example C-1

Question No. 2

W: What would you like to order?
M: I'd like eggs with toast and, uh, sausage…no, sorry, I'd rather have bacon.
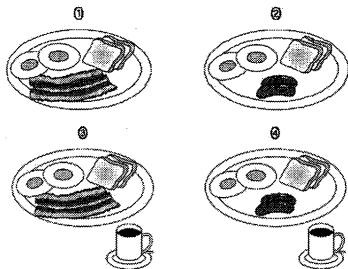W: OK. And coffee?
M: Sure.

Figure 8 : NCT³ question number 2, Section 1

When checking the correspondence between the test tasks of the TOEFL iBT and NCT, for the most part, the questions of the NCT are conversational. Depending on the ratio of conversational to content-based courses conducted in English, inferences made from these tests scores cannot be perceived as resembling accurately the English skills required for students to effectively operate in the TLU domain. This is attributed to the responses of these questions, which are in multiple-choice format. These only promote discrete linguistic comprehension, not broader contextual understanding required for a class conducted in English by a native speaker, regardless of whether it is content-based or conversational. Therefore, the multiple-choice construct is low in terms of authenticity, as the scores obtained are not able to effectively reflect the domain of generalisation.

Questions such as example C-2 (see Figure 9) are more authentic, as some interaction is required by making test takers involved in the monologue. Therefore selecting an appropriate sentence from four choices is not a feature of being able to operate in a native speaker's class (the TLU domain), again irrespective of whether it is content-based or conversational. To generalise to the university domain, chief responses would include productive responses such as writing a report or spoken replies, including simply replying or giving a presentation. Hence, interpreting and processing content for the individual is an important and authentic activity for the TLU domain. Yet some questions, as in example C-3 (see Figure 9), do not require test takers to understand its content. Example C-2 does not even require the test taker's comprehension of the passage, as C-2 can be easily answered by selecting a virtually identical sentence to that which is heard from the listening task ("get the teacher's permission" and "get permission from teachers"). Furthermore, the ability to easily "match" the question with the response makes the distractors redundant (Popham, 1990).
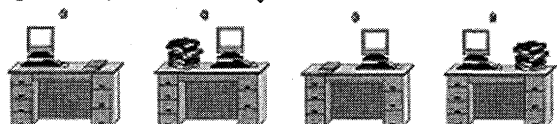
Example C-2	Example C-3
Script: This afternoon, all new students are to go to the assembly hall to sign up for classes. .... Go to the assembly hall and get the teacher's permission for the classes you want to take. Then write your name down on the class lists. Q: What must new students do at the assembly hall to sign up for their classes? ① Get permission from the teachers. ② Hand in the course descriptions ③ Pick up a registration form. ④ Write the teachers' names on the class lists.	selecting the drawings or numbers which are described [Listen to the conversation] W: Nice office, George. Which desk is yours? M: It's over there — the one with a computer on the left and some books on the right. [Question] Which is George's desk?

Figure 9 : NCT question numbers 2 and 3, Section 2

7.2. NCT interactiveness

Moreover, it could be said that the multiple-choice format employed tends to obstruct achieving high levels of test-task interactiveness. Interactiveness can also be confounded by the relative ease for a test taker to guess the correct answer. In some questions, the most appropriate response can be easily inferred by only listening to the last sentence. Question 11 is a good example (see Figure 10). For obvious linguistic reasons, choices ① and ④ are unnatural responses to the question "Do you have an appointment?" Also, response ③ is probable only if the question is along the lines of "Would you like to make an appointment?", or "Would you like to see him?", that is, using the modal verb "would". Therefore, it is clear that the most appropriate response for question 11 can be obtained by a basic process of elimination by merely listening to the last sentence. In other words, it is difficult to identify how much of the question that the test taker has actually comprehended since the distractors are inadequate (Popham, 1990) (For further information on test item discrimination and difficulty, see Sage and Tanaka, 2006b).

> [From the conversation script]
>
> A: May I help you?
>
> B: Yes, I' d like to see Professor Baker.
>
> A: He' s not in now. Do you have an appointment?
>
> [From the question sheet]
>
> ① Uh, no. Could I cancel the appointment?
>
> ② Uh, no. Could I make one?
>
> ③ Yes, I'd like to make one.
>
> ④ Yes, I'll be happy to help you.

Figure 10 : NCT question number 11

8. Conclusion

Listening assessment can be integrative, as shown by the TOEFL® iBT test. Furthermore, as Buck (2001) states, it ought to reflect the competencies of the person combined with the listening task that they must perform in the "real world." In essence, this is authenticity (Alderson and Banerjee, 2002). From the qualitative analysis conducted in this paper, it has been concluded that the NCT listening test tasks are neither

as integrative nor communicative as they could or perhaps should be if the MEXT guidelines were followed and they reflected the future language use required of the TLU domain. Through the comparison of the TOEFL® iBT and NCT English listening sections, it is proposed that the TOEFL® iBT is higher in the notions of communicative competence, and elements from Bachman & Palmer's (1996) test usefulness framework, namely authenticity, validity and interactiveness. Hence it is a more authentic gauge of the TLU use domain. This article attributes this to the multiple-choice construct that the NCT listening section employs, which itself constrains the aforementioned, authenticity. The best way to maximize its authenticity would be to adopt integrated test tasks to improve interactiveness, as the methodology of the TOEFL® iBT does.

However, it is not the purpose of this article to entirely discredit the NCT as a test and put the TOEFL® iBT on a pedestal in response. As a matter of fact, the ubiquity of TOEFL and its degree of influence, not only as a test of English language proficiency, but also as a test which has been adopted by many institutions as a placement test, an achievement test, and even as a progress test, is a matter of concern (Brown, 1995a, pp.14-15; 1995b, p.40). Nonetheless, these applications are supported by MEXT, and furthermore the CIEE Nihon Daihyobu (2004) found, in Japan, that 312 universities are currently using the TOEFL score, in a variety of different ways, including using it as a substitute for the English section of their entrance examination. Since TOEFL is designed to measure the proficiency of students who expect to enter universities in English speaking countries, as a whole, it is an inauthentic test to assess students' English ability to operate in EFL situations at Japanese universities. Thus, the NCT, which is designed for the Japanese context, could be effectively embraced, and in doing so, woud further develop the NCT English listening section for Japan's EFL university environment. The only way to make the NCT listening section a more authentic test is to include test tasks which examine listening as an interactive, integrative communicative skill in terms of academic and conversational situations, both of which occur in English language classes at Japanese universities.

9. Future Directions

The following proposals are put forward in order to provide some ways in which specific areas of the NCT can be developed in the future.

Proposal one:

Since the purpose of the NCT is to assess students' English ability to operate in English classes in Japan, it is essential to include not only conversation but also academic listening test tasks which reflect the Japanese university environment. This could be achieved by including test items which are more academically hued, and include academic genre and contextual applications. By doing this, discourse competence (see communicative competence theory) can be addressed. Further, actional competencies and task interactiveness could be enhanced by accompanying test items with visuals such as graphs and tables which support the content of the listening sections more appropriately and resemble those that occur in universities in Japan (TLU domain). As the TOEFL® iBT already includes academic listening tasks, it would contribute significantly to the NCT if test developers were to examine the features of CLT methodologies and communicative competencies used in this test.

Proposal two:

It is fundamental to develop a detailed set of test specifications[4] and make them a transparent document available for public reference. This would clarify the goals of the NCT, notably if CLT and communicative competence methodologies were written into the test specifications. It would also provide a framework, when writing tests, for test developers to follow consistently and on an annual basis. Yet, to set reasonable goals

for the developers of the NCT's listening section, it is necessary to investigate what English listening ability is actually required in universities in Japan and try to improve authenticity based on this research and data collection.

Proposal three:

Harmful washback can be caused by teachers "teaching to the test" or when students partake in extra classes administered exclusively for passing the test (McNamara, 2000; Bachman & Palmer, 1996). It is commonly known that many high schools in Japan offer special preparation courses for the NCT, and preparing for the listening section is now prominent. To combat this, if the focus of the NCT were to change to that of an achievement test, rather than a proficiency test, then test specifications would need to be developed. Furthermore, emphasis would move away from learning English to pass the test which is an inauthentic process in itself. The TOEFL test is informed by detailed test specifications, addressing methodically the test's aims and goals. To follow suit may mean that the NCT English listening section would reduce its negative washback.

Notes:

1   For MEXT's Action Plan, visit http://www.mext.go.jp/b_menu/houdou/15/03/03033101/001.pdf
    For MEXT guidelines, visit http://www.mext.go.jp/english/shotou/030301.htm
2   University A: English language/conversation: 336 classes, content courses: 26 classes
3   2006 NCT's English Listening Section is available at NCUEE homepage (NCUEE, 2006a)
4   The design decisions in the test, including: the length and structure of each part of the test, the type of materials with which candidates will have to engage, with the source of such materials if authentic, the extent to which authentic materials may be altered, the response format, the test rubric, and scoring (McNamara, 2000, p.31-32)

References

Alderson, J. C., & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching*, 35, 79-113. (SS&H Library High Use)

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

Brown, J.D. (1995a). Differences between Norm-Referenced and Criterion-Referenced Tests. In J.D. Brown and S.O. Yamashita, Eds. *Language Testing In Japan*. pp. 12-19. Tokyo: Japan Association for Language Teaching

Brown, J.D. (1995b). Developing Norm-Referenced language tests for program-level decision making. In J.D. Brown and S.O. Yamashita, Eds. *Language Testing In Japan*. pp. 40-47. Tokyo: Japan Association for Language Teaching.

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

Chalhoub-Deville, M. & Turner, C. (2000). What to look for in ESL admission tests: Cambridge certificate exams, IELTS, and TOEFL. *System*, 28, 523-39.

Canale, M. & M. Swain (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1,1: 1-47.

Celce-Murcia, M., Dornyei, Z. & Thurrell, S. "Communicative Competence: A Pedagogically Motivated Model with Content Specifications." *Issues in Applied Linguistics* 6.2 (1995) : 5-35.

Coombe, C., Kinney, J., Canning, C. (1998). Issues in foreign and second language academic listening assessment. In *Current Trends in English Language Testig*, Conference Proceedings for CTELT 1997 and 1998, Vol. 1 Edited by Coombe, C. A. Arabia: TESOL Arabia. 53

Doye, P. (1991). *Authenticity in Foreign Language Testing*. Report. Access date August 1, 2006.
    <http://www.eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/0000000b/80/23/2b/24.pdf>

ETS. (2005). *ETS TOEFL, Test and Score Data Summary: 2004-2005 Test Year Data*. Princeton, NJ: Educational Testing Service.

ETS. (2006a). *The Official Guide to the NEW TOEFL® iBT.* The McGraw-Hill Companies, Inc.

Fischer, R. (1981). *Towards a Practical Model of a Communicative Competence Testing Program.*

Flowerdew, J. (Ed.). (1994). *Academic listening: Research perspectives.* Cambridge: Cambridge University Press.

Fulcher, G. (1999). Ethics in Language testing. *TAE SIG Newsletter* – Special Conference Issue, Volume 1, No. 1. INDICATE PAGES OF ARTICLE.

Fulcher, G. (2000). Computers in Language Testing In Brett, P. and G. Motteram (Eds) 2000. *A Special Interest in Computers.* Manchester: IATEFL Publications, 93-107.

Furmanovsky, M. (1996). *Content Video in the EFL Classroom.*
  <http://eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/0000000b/80/21/f4/3f.pdf>

Hughes, A. (1989). *Testing for language teachers.* Cambridge: Cambridge University Press.

Katz, R., and Ralston, M. (1990). From the Classroom to the Real World: Listening Attack Strategies. Paper presented at the Annual Meeting of the Teachers of English to Speakers of Other Languages (24th, San Francisco, CA, March 6-10, 1990).

Larsen-Freeman, D., & Long, M.H. (1991). Chapter 2. Second languageacquisition research methodology. In *An Introduction to Second Language Acquisition Research.* London: Longman, pp.11-51.

Low, G.D. (1986). 'Storylines and other developing contexts in use-of-language test design,' *Indian Journal of Applied Linguistics* (Special Issue on Language Testing).

McNamara, T.F. (1996). *Measuring second language performance.* London and New York: Addison Wesley Longman.

McNamara, T. (2000). *Language testing.* Oxford: Oxford University Press

Morrow, K. (1979). Communicative language testing: revolution or evolution? In C.J. Brumfit and K. Johnson, Eds. *The communicative approach to language teaching.* Oxford: Oxford University Press, 143-157.

NCUEE. (2006a). "平成18年度センター試験（本試験）問題：外国語－リスニング＆音声問題＆スクリプト."
NCUEE homepage. Access date April 20, 2006. <http://www.dnc.ac.jp/center_exam/18exam/18hon_mondai.html>

NCUEE. (2006b). "英語リスニング試行テスト実施結果の概要" *NCUEE homepage.* Access date October 15, 2006.
  <http://www.dnc.ac.jp/center_exam/18exam/listening.html>

Popham, W.J. (1990). *Modern educational measurement: a practitioner's perspective.* Englewood Cliffs, N.J.: Prentice Hall.

Sage, K. and Tanaka, N. (2006a). "Future English language testing direction: focus on listening." Presentation at *the 5th Annual JALT Pan-SIG Conference, Theme: Authentic Communication.* Shizuoka, Japan: Tokai University, May 13-14, 2006.

Sage, K. and Tanaka, N. (2006b). "So what are we listening for? A comparison of the English listening constructs in the Japanese National Centre Test and TOEFL® iBT." *Authentic Communication: Proceedings of the 5th Annual JALT Pan-SIG Conference.* (May 13-14, 2006. Shizuoka, Japan: Tokai University College of Marine Science.) pp 74-98. < http://www.jalt.org/pansig/2006/HTML/SageTanaka.htm>

Sauvignon, S. J. (1991). Communicative language teaching (CLT) : Practical understandings. *Modern Language Journal,* 83 (4), 494-517.

Sawyer, D. (1998). *From Non-communicative Exercises to Technical Writing: Profile of a Two-Semester Preparatory Sequence.*
  <http://eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/0000000b/80/11/29/27.pdf>

Toyama, A. Minister of Education, Culture, Sports, Science and Technology. "Regarding the Establishment of an Action Plan to Cultivate Japanese with English Abilities". *MEXT.* 31 March 2003. Access Date April 2, 2006
  <http://www.mext.go.jp/english/topics/03072801.htm>

Weir, C. (1990). *Communicative language testing.* New York: Prentice Hall.

Appendix A: Analytical Comparisons between NCT and TOEFL® iBT

1) Comparing goals and contents of questions
2) Considering question types and required ability for target situations

Category	TOEFL® iBT	The National Center Test
Authenticity	(Positive) Independent skills = long lectures with graphs and visuals, realistic situations (such as discussion, lecture, conversation between teacher and student, etc.) Integrated skills = very realistic (read the textbook → go to the classes and listen to professor's lecture → make notes and write summary, state your opinion, etc.) (Negative) Independent skills = multiple-choice can only assess students' receptive skills and cannot determine whether the students can communicate (or use English correctly in the TL domain) or not.	(Positive) · Questions like example C-3 are somewhat authentic because it is interactive by involving test takers in the conversation, yet it still can measure only their receptive skills. · The lecture part is realistic by adopting a mini version of actual classes in universities. (negative) · Multiple-choice can only assess students' receptive skills and we cannot determine whether the students can communicate (or use English correctly in the TL domain) or not. · Conversations are not very useful in universities in Japan (need to include conversation between professor and student) · Some questions like example C-1 do not require test takers to understand the content.
Validity: Face Validity	(Positive) · TOEFL® iBT was developed and introduced in order to: 1) measure the ability to communicate successfully in academic settings 2) reflect how language is really used 3) keep up with the best practices in language learning and teaching · More than 5,000 colleges, universities, and licensing agencies in 90 countries accept TOEFL scores.	(Positive) · The NCT was developed in order: 1) to measure the students' current ability to enter college 2) for universities and colleges to assess students' aptitude and ability · NCUEE conduct a pilot of the listening test before the actual NCT and revise/improve its adoptability/quality · As a matter of fact, all of the National and Prefectural Universities and 439 private universities and colleges combine the NCT with their entrance examinations.

Continued...
Appendix A: Analytical Comparisons between NCT and TOEFL® iBT(Cont...)

Validity: Test Validity		
	Test Content (Positive) The TOEFL® iBT reflects what the examinees need to perform in actual classroom settings and professors' offices in English speaking countries. Also it adopts many different situations including: 1) a lecture where the professor is the only speaker; 2) a lecture where the professor and students speak (discussion type) ; and, 3) conversations during the professors' office hours, with teaching assistants, librarians, bookstore clerks, etc. (for example, test format's description of integrated tasks)	Test Content (Positive) A conversation between the professor and student in the office can happen in the correspondent criterion in Japanese universities. (Negative) A conversation (such as ordering food, etc.) may not happen in EFL situations. (example is C-1 in Figure 8)
	Test Method and Test Construct Very close to criterion by adopting integrated assessment Each passage is quite a long · Conversation (3 minutes long, 12-25 exchanges) · Lectures (3-5 minutes long each, 500-800 words)	Test Method and Test Construct Multiple-choice situations will never occur in the targeted criterion · Conversation (10-30 seconds long, 2-10 exchanges) · Speeches (30 seconds, 90-100 words) · Monologue (60 seconds, 184 words)