Shared Identities: Our Interweaving Threads

Investigating multiple-choice discourse completion tasks on Japan's Center Exam

Eric Setoguchi
Kanda University of International Studies

Reference data:
Setoguchi, E. (2009). Investigating multiple-choice discourse completion tasks on Japan's Center Exam. In A. M. Stoke (Ed.), *JALT2008 conference proceedings*. Tokyo: JALT.

A new class of multiple-choice discourse completion tasks (MDCT) has appeared in the Japan English as a Foreign Language (EFL) assessment context, including the English listening section of the *National Center Examination for University Admissions*. A topic of interest given their newness is the validity of the MDCT item in this particular assessment use context, as they are generally thought of as pragmatic rather than listening skill assessments. In this study an exam comprised of MDCT items was administered to a sample of Japanese university students. Using classical test theory (CTT) and Rasch analysis it was investigated whether pragmatic factors in MDCT item design had a measurable effect on the examinees' test performance. The results suggest that MDCT item difficulty could be affected by the pragmatic abilities of the examinees, thereby bringing into question the validity of the item in English listening assessment and calling for the need for further research into this issue and tests that employ them for this use in the Japanese EFL assessment context.

日本での外国語としての英語(EFL)の評価の分野において、新しいタイプの多岐選択方式談話完成問題(MDCT)が登場した。これは、大学入試センター試験の英語リスニングテストにも採用されている。MDCTは、通常リスニング能力よりむしろ語用論的能力の評価に使用されているものなので、リスニングテストとしてのMDCTの妥当性の問題が懸念される。本研究では、実験的MDCT試験を日本人大学生に受けてもらい、古典的テスト理論(CTT)及びラッシュ分析を使用して、MDCTの項目における語用論的要素が被験者の成績に測定可能な効果を及ぼすのかどうかを調べた。分析の結果は、MDCTの項目難易度は被験者の語用論的能力によって影響されうること、従って英語リスニングテストとしての問題の妥当性が問われることを示唆している。分析の結果は、MDCTの項目難易度は被験者の語用論的能力によって影響されうること、従って英語リスニングテストとしての問題の妥当性が問われること、また日本のEFLの評価の分野において、MDCTを用いたテストに関するこの問題をより深く研究していく必要性があるを示唆している。

In Japan, interest is growing in the improvement of language assessment systems to incorporate communicative assessment. This growing interest is likely one result of a ripple effect of a larger reform movement to shift Japan's foreign language teaching style from an emphasis on a traditional, synthetic-based approach, heavy in grammar, vocabulary, and translation, to a communicative language

PAC7 at JALT2008: Shared Identities

teaching (CLT) approach. This ambitious reform project, and its success, has become a national preoccupation of sorts and was the central motivation behind a July 2002 mandate by the Ministry of Education, Science, and Technology known as "A Strategic Plan to Cultivate Japanese with English Abilities" (Ministry of Education, Culture, Sports, Science, and Technology, 2003). As schools and teachers at the primary, secondary, and post-secondary level move to make curriculum adjustments in response to the shift to communicative language learning, there has been an increasing trend away from language assessment as traditional grammar and vocabulary testing and towards the development of non-traditional communicative performance assessment. Investigations of group oral discussion task assessments (Bonk & Ockey, 2003), and video-based discourse completion task assessments (Tada, 2005), and others are recent examples of communicative performance assessments that are now being researched for future application in various Japanese English as a Foreign Language (JEFL) contexts.

In this atmosphere of growing receptiveness in the JEFL testing community for new forms of assessment, in 2006 the *National Center Examination for University Admissions* (*Daigaku Nyushi Sentaa Shiken*; hereafter the *Center Test*) a collection of nationwide university entrance exams, implemented a new English listening test, the *Center Test in English Listening*. Touted as a communicatively focused language assessment to reflect the new emphasis on communicative learning (Center Test, 2008), the test employs a modified version of the multiple-choice discourse completion task (MDCT), a non-traditional item type that is common in language and language testing research, but before that point in time had not been seen in large scale authentic language assessment.

Although they are new to real-world operational applications, MDCTs have been well studied for over a decade as experimental testing items (Hudson, Detmer, & Brown, 1995; Yamashita, 1996; Yoshitake, 1997; Jianda, 2007). Developing testing items from the experimental phase into real-world application should occur under strict consideration of several factors, including (but not limited to): assessment purpose, uses, users, and context. One area of concern is that MDCTs are being rapidly implemented into an operational assessment context before their potential has been well studied for specific intended uses. A critical aspect of test item quality is whether a given item accurately measures what it is intended to measure in a specific context of use, with as little error as possible. This study focuses on this aspect, formally known as construct validity, of the MDCT test item and its intended use in the context of the *Center Test in English Listening*. To situate this study in a larger context, given the tendency for testing items on university entrance examinations to be incorporated into curriculums and assessment at the JEFL educational level in an effect called "washback" (Brown, 1997), it is likely that EFL in Japan will see a significant increase in MDCT use beyond the university testing context. The contribution of this study to an understanding of how MDCTs function in the JEFL context represents a potential gain in our ability to design quality language assessments in the high-school classroom, university, and elsewhere.

PAC7 at JALT2008: Shared Identities

General information about the test and test usage

The *Center Test*, of which the *Center Test in English Listening* is a part, is a collection of standardized annual exams in different academic subjects, and is developed by Japan's National Center for University Admissions. A number of primary and secondary stakeholders (persons with invested interest in the test and what it provides) use the *Center Test* for a variety of different purposes. The stakeholders with the highest priority are probably universities, many of which utilize test scores as a part of their admissions process. Although the exact number of universities using the *Center Test* to make admissions decisions varies each year, a recent administration of the test (2007) was used by approximately 600 public and private universities, as well as junior colleges. Individual universities do not interpret students' *Center Test* scores in the same manner, but the *Center Test's* role in admissions processes can be divided into several categories: (a) use as the sole determiner of admission, (b) use in combination with additional assessment factors specific to each university to determine admission, and (c) use as a general qualifier to participate in a secondary university examination that will be used alone to determine admission.

The English Listening exam was first formally administered in 2006 and to date is the only listening exam in the Foreign Language subcategory of the *Center Test*. Based on statistics from 2006 and 2007, the English Listening exam was the second most-taken exam of the 34 exams comprising the Center Test, with 492,555 examinees in 2006 and 497,530 in 2007. These and other *Center Test* statistics are available publicly on the *Daigaku Nyushi Center* homepage.

As readers of this paper are likely already readily familiar with the *Center Test*, I'll conclude this section here by emphasizing two important points: (a) while the *Center Test* has a number of users and uses, the primary user and use of test scores is for a rather high-stakes decision (whether an individual gets admitted to a university or not), and (b) in terms of test-takers the *Center Test* is very high-volume. Particularly in high-stakes, high-volume testing contexts, the consequences of implications derived from test results highlight a critical need for accountability demonstrating that what a test result is intended to measure is what it actually does in practice. This issue in relation to the use of the MDCT item is essentially the motivation for addressing a need for a thorough construct validity study, which this study is intended as an early step of.

MDCTs in the test in English listening

The *Center Test in English Listening* (hereafter referred to as the *Center Test*) has 4 sections, with the second section comprised of multiple-choice discourse completion tasks. Examinees must complete 7 MDCT items, roughly 28% of their total exam score. The remaining sections of the exam are content-based listening items in which the examinees listen to short dialogues, lectures, or narrations and complete multiple-choice questions based on the listening passage. The 7 MDCT items are the only test items requiring examinees to make judgments of the appropriateness of dialogues in language situations.

The basic format of the MDCT items on the *Center Test* is shown below in Example 1. Examinees first listen to a short dialogue between two speakers and then read a short

PAC7 at JALT2008: Shared Identities

list of accompanying lines of dialogue on their test forms. To answer an item, examinees must select the line that most appropriately continues the dialogue.

Example 1

Examinees hear:

W: What did you do over the weekend?

M: Oh, I started reading a really good book.

Examinees see:

1. Really? What's it about?
2. Really? Why don't you like it?
3. Sure, I'll lend it to you when I'm done.
4. Sure, I'll return it to you later.

(from *Daigaku Nyushi Center Shiken*, 2007)

In this situation, a conversation about one of the speakers reading a good book, contextual information is encoded within the prompt dialogue itself. In other words, who is speaking, where the conversation is taking place, and the ultimate intent of either speaker is not explicitly made known to the examinee. In comparison, detailed situational descriptions are a common component of most MDCT models under investigation in current language assessment research (Hudson, Detmer, & Brown 1995; Jianda, 2007; Tada, 2005), as well as established language tests such as TOEFL that do not employ MDCTs but do have items based on conversational situations.

The MDCT items on the *Center Test* appear to be based on common conversational themes that appear in English communication textbooks used in Japanese high schools. The answer choices for each item are based on entirely different speech functions. In other words, as in Example 1, the four answer choices are not based on four different ways to ask what the book the woman read was about, but are four completely different statements, only one of which is considered correct. Distractors are identified based on how they conflict with information in the prompt or demonstrate a lack of context appropriateness. In contrast, MDCT items in the literature are frequently limited to the three major speech acts best understood by pragmaticians: apologies, requests, and refusals, and have answer choices based on variations of a single speech function that differ based on pragmatic variables.

From a test design perspective, these distinctions allow for some measure of freedom in item writing for MDCTs on the *Center Test,* as well as confidence in the content validity of the items for the target test population. *Center Test* MDCTs are also noticeably less pragmatically oriented compared to MDCTs being used as pragmatics assessments in the literature, at least in general test design. This distinction is actually a desirable one in terms of the construct validity of the *Center Test*, as will be discussed in depth below.

Construct validity of MDCTs

Validity theory, the dominant notion for the rating and evaluation of educational assessment (including that of language), has been greatly influenced by Messick's unified and comprehensive interpretation of test validity (Messick, 1996). Originally, Messick (1989) advocated that the primary component of validity is construct validity, the notion that any assessment should adequately measure the construct

PAC7 at JALT2008: Shared Identities

under investigation, but only that construct without influence by variance from undesirable effects (as cited in Norris, 2008, p. 44). It follows that the primary threat to validity is construct under-representation, when a test does not measure all of the intended construct, or construct irrelevant variance, when a test measures more than the intended construct.

The *Center Test in English Listening* is intended as a measure of Japanese high school students' English listening proficiency. As mentioned previously, the *Center Test in English Listening* was largely a response to an educational mandate from Ministry of Education in Japan. The mandate explicitly states that the test should meet the goal of improving the English oral communication abilities of Japanese learners, but makes no specific mention of pragmatic proficiency (Ministry of Education, Culture, Sports, Science, and Technology, 2003). Furthermore, there is strong evidence to suggest that instruction in English pragmatics is largely ignored, or at most improperly addressed in high school EFL education (Shimizu, Fukasawa, & Yonekura 2007). The concept of MDCT test construct in the JEFL assessment context is highly ambiguous, but there is no indication at this point that MDCT tests are being designed, deployed, or interpreted with pragmatics as a component of English listening.

Therefore based on Messick's definition, the MDCT item would have construct validity in this context if it could be demonstrated that the item adequately and exclusively assesses the listening proficiency of examinees, the singular construct of its intended use. It also follows that any variation in test performance due to pragmatic proficiencies of examinees represents undesired construct irrelevant variance. The obvious concern here is that all MDCT research currently focuses primarily on their potential as measures of pragmatic proficiency. Their appropriateness in the exclusive assessment of general language skills such as listening is unknown and unsubstantiated. Although the *Center Test* MDCT items were not implicitly designed for assessing knowledge of pragmatic strategies per se, and have been modified to make them less obviously pragmatic in orientation, strong empirical evidence is needed to rule out the possibility that MDCTs in the format that they appear on the *Center Test* covertly function to assess examinee pragmatic proficiency in addition to the intended construct, listening proficiency.

One potential mechanism for a pragmatic effect lies in the fact that examinees must still agree with the acceptability of the key to correctly answer an item. If a key is not written with pragmatic strategies that examinees agree with, regardless of its correctness compared to distractors, examinees may be less likely to select it. For example, past research indicates that Japanese EFL learners prefer indirect speech strategies when completing MDCTs (Rose, 1994, 1995). Such a finding suggests that Japanese EFL learners might not identify with language styles that are direct, and MDCT items using direct language in the keys could be more difficult for Japanese EFL learners to answer correctly. Furthermore, it is also possible that Japanese EFL learners are sensitive to other pragmatic variables in addition to the level of directness, and variations in any of these on the MDCT answer choices could play a role in test performance. This issue raises a critical point concerning the construct validity of the MDCT item for English listening assessment, and serves as the motivation and focus for this study.

PAC7 at JALT2008: Shared Identities

Research question and implications

As MDCTs in JEFL assessment are currently being used on a prominent English listening exam, the *Center Test*, under conditions of high-stakes, high-volume testing, it would be useful for an investigation of MDCT items to focus specifically on aspects of their construct validity in measuring L2 listening proficiency. A primary threat to construct validity is construct irrelevant variance, when a test measures more than the intended construct. To this end, the following research question is addressed in this study: Are JEFL examinees influenced by pragmatic variables in MDCT answer choices, and is there an observable effect caused by this influence on MDCT item performance?

At this point in the paper, it is also pertinent to address what kind of implications the results of this study and the research question it proposes can and cannot have on the use of MDCTs in JEFL assessment, and specific tests that employ them. The study conducted here is representative of an initial phase of a testing evaluation primarily concerned with item level investigation. In other words, at this point the study is interested in a general understanding of the MDCT item itself and how it functions, as opposed to directly evaluating any specific test which deploys it. Due to a need for this type of initial research before conducting more in depth test evaluation, as well as limitations in scope and materials available to the researcher at the time of this study, the intent of this study is not to interpret from the results a conclusive evaluative judgment about a specific test that employs MDCTs, such as the *Center Test*. Rather, I seek here to provide a study that serves as an information providing initial step, that, depending on its findings might suggest a need for a more in-depth evaluative study of tests like the *Center Test*, as well as to contextualize and provide data to aid in the understanding of any results that follow-up research might provide.

Because this study represents a preliminary rather than a conclusive phase in the evaluation of MDCTs and tests that employ them, and because the primary purpose here is to gather information about a specific function of an test item and not a test in operation itself, it is not necessary here to replicate authentic testing conditions. In fact, the conditions of this study deviate from conditions of the actual *Center Test* in several ways, as will be explained and justified below. To clarify this point further, if there is in fact an observable effect on MDCT item performance based on pragmatic proficiency of JEFL learners revealed in this study, it would suggest in general that MDCT items might not have construct validity in measurement of English listening proficiency of JEFL learners. Such a finding would not immediately provide conclusive evidence for evaluating the appropriateness or quality of tests like *Center Test* itself, and the researcher implores readers that to interpret it as such at this phase would be reckless and unsubstantiated. Such a finding would however strongly suggest a need for larger scope research in this area specific to operational conditions of actual tests like the *Center Test*, and other contexts where MDCTs are used in the measurement of JEFL listening proficiency. Such research in turn would take steps towards the generation of solid evidence for the quality and appropriateness of particular tests, and indeed should be of high priority.

PAC7 at JALT2008: Shared Identities

Method

Participants

Twenty-six Japanese university students participated in this study. The students were recruited from a private Japanese university, and all were actively pursuing EFL studies at the time of their recruitment.

Materials

The research instrument employed in this study is an English listening exam composed of 36 MDCT items similar to those appearing on Japan's *Center Test* in English Listening. Although authentic *Center Test* items are freely available, it was deemed necessary for the purposes of this study to modify the items slightly as well as supplement them. The motivation, specifics, and implications for these changes, as well as example items from the exam appears in the following section.

Measuring the effect of directness as a pragmatic variable on test performance

The MDCT items appearing on the exam used in this study needed to quantitatively assess the effect on examinee performance of pragmatics variables contained within MDCT answer choices. Therefore, a way to measure any such effect needed to be developed and operationalized within the exam. The use of authentic *Center Test* items as they are used in authentic conditions presented several problems.

Firstly, only 7 MDCTs appear on any given administration of the *Center Test*, an insufficient number of items for reliable statistical analysis. This issue was addressed by combining authentic item prompts appearing on the pilot, 2006, and 2007 versions of Japan's *Center Test*, and supplementing these with original prompts developed by the researcher to mimic their style, context, and difficulty level. This resulted in an exam containing 36 MDCT items, a sufficient number for reliable statistical analysis.

Second, answer choices on the Center Test closely mimic natural speech, which unfortunately contains a variety of speech styles across a diverse spectrum of pragmatic variables. Rather than attempt to deal with a multitude of variables, experimentally it would be better to begin by isolating a single pragmatic variable and investigating the potential of that variable to effect exam performance. Based on previous literature (Rose, 1994, 1995), it was decided that the pragmatic variable of directness had the highest potential to produce an effect. For many items, this involved writing a new answer key specifically designed to measure such an effect. The first step in this process was to divide the exam items into two categories based on the level of directness of the item's answer key, a constructed variable for this study called the indirectness factor. The indirectness factor was assigned either a (+) or (–) value. The values and their labels, indirectness factor (+) and indirectness factor (–), are defined here as follows: The answer keys of indirectness factor (+) items were designed to present a high level of acceptability to Japanese EFL students based on current literature on pragmatic behavior; additionally, for these, authentic answer keys that used strategies of indirectness,

PAC7 at JALT2008: Shared Identities

apology, excuse, and expressions of regret were selected, or authentic answer keys that did not employ such strategies were slightly modified to do so. The answer choices of indirectness factor (–) items were designed to present a low level of acceptability to Japanese EFL students based on current literature on pragmatic behavior. For these, entirely new answer keys that use strategies of directness and clearly lack the use of apology, excuse, and expressions of regret, even in situations where they are applicable, were written.

An example exam item from each category is shown in Table 1. The answer key for the *indirectness factor* (+) item, "I see. Sorry to bother you." employs an apology as an indirect and softening strategy. The answer key for the *indirectness factor* (–), "I'm taking a short rest first," lacks an indirect strategy, although it could be said that in some contexts one might be appropriate (an apology, excuse, etc.). Note that in both cases, both keys are the only possible correct answer to the item.

A significant difference in overall test performance between items with and without the *indirectness factor* was hypothesized. In other words, if pragmatic proficiency has an effect on performance, then *indirectness factor* (+) items would be significantly easier for examinees than *indirectness factor* (–) items. That is, Japanese EFL students were anticipated to more easily identify correct answers that use indirect and passive strategies than correct answers that use direct and aggressive strategies.

Table 1. Example exam items

indirectness factor (+) item	indirectness factor (–) item
Question #20 A: Excuse me. May I borrow that dictionary for a moment? B: Oh, I'm sorry. It's not mine. A: _____ Oh, it's yours? Thanks. I'll return it soon. Oh, it belongs to my professor. *I see. Sorry to bother you.	Question #17 A: Mr. Stevens, I finished making copies like you asked. B: Great job. Next, I need you to deliver a message to the office downstairs. A: _____ Thanks. I could use a rest. Yes, here is the message from downstairs. Great job. Thanks for all the help. *I'm taking a short rest first.

Note: In each item the answer choice marked with a (*) indicates the correct answer

Data collection

The test was administered to all participants as if it was an authentic testing condition. Although the exam itself is not meant to mimic or replicate the actual *Center Test*, the procedure for the MDCT item section itself as it appears on the actual *Center Test* was adopted. The participants first heard a set of instructions in Japanese, followed by recordings of the 36 dialogues. The voice actors for the dialogues were high proficiency Japanese speakers of English with native-like or near-native-like pronunciation. Each dialogue was read twice, and the participants had 12 second

PAC7 at JALT2008: Shared Identities

pauses between readings to record their answers on the test forms. The total time given for the test was approximately 30 minutes. At the end of the recording, the test forms were collected. The participants were not given additional time to review their answers after the recording finished.

Results

Classical testing theory analysis

Classical testing theory (CTT) and descriptive statistics serves as a useful starting point for investigating the differences in performance among items based on the *indirectness factor*. The mean score on *indirectness factor* (+) items, 14.54, is noticeably higher than for *indirectness factor* (–) items, 11.12 (Table 2). A t-test was performed on the data to determine if this difference in scores could be considered statistically significant in this study ($t = 23.36$, $df = 51$, $p < 0.00$). These results suggest that the level of directness of the answer choice had a measurable and statistically significant effect on MDCT item difficulty (α: $p = .05$).

Table 2. Descriptive statistics based on item categories

	indirectness factor		Overall
	(+)	(–)	
N	26.00	26.00	26.00
K	18.00	18.00	36.00
M	14.54	11.12	25.65
SD	2.34	3.57	5.43

FACETS analysis of performance

FACETS analysis, or multifaced Rasch, provides another method of investigating the score variation due to the *indirectness factor*. While classical testing theory can demonstrate one way of showing that participants scored significantly higher on *indirectness factor* (+) items, FACETS can provide supporting evidence for this with a more advanced simultaneous analysis of item difficulty, examinee ability, and subtest type, all on the same scale. It is beyond the scope of this paper for a detailed explanation of FACETS or Rasch analysis, however, it will be sufficient to mention that these are advanced analytical techniques for investigating tests and test performances. Sick (2008) provides an excellent concise discussion of Rasch measurement as it applies to language education.

The FACETS output comes in the form of a vertical ruler, as shown in Figure 1. The units of such a ruler are called "logits," with a logits visually represented in the leftmost column of the ruler. In this case, our ruler goes from -1 to 4 logits. The middle column represents the 26 examinees, and the rightmost column represents the *indirectness factor*. Paying attention to only the rightmost column, simply interpreted the higher a category falls on the logit scale the more difficult items in that category were for examinees. The FACETS analysis mirrored the results in the classical analysis, and indicated that the *indirectness factor* had a considerable effect on item difficulty. Items with the *indirectness factor* had an average logit score of -.55, while items without the factor had an average logit score of .55, a difference of a complete logit unit.

PAC7 at JALT2008: Shared Identities

Conclusion

Both classical testing theory and Rasch analysis methods indicate that the pragmatic variable of directness in MDCT items as used in this study can quite strongly influence item difficulty with JEFL examinees. If this is true in the case of directness, it could be reasoned that other pragmatic variables might also have similar effects. This finding brings into question the construct validity of the MDCT item in English listening assessment, due to the possibility of construct irrelevant variance of examinees being influenced by pragmatic cues imbedded in answer choices in their answering of items. If this is indeed occurring, it would be an undesirable source of error from a testing standpoint, as it would be possible in some cases for an examinee to correctly understand an MDCT prompt of a given item as they listen to it, the desired target proficiency, yet still incorrectly answer the item if the key for that item happened to be written using pragmatic cues that might be unfamiliar or unappealing for them. In such cases, the item and the test as a whole would not be providing an accurate assessment of examinees' English listening proficiency.

Future research

This study is a first attempt at what will likely need to be a prolonged and detailed investigation of the MDCT test item format in the JEFL context. In closing, it would be useful at this point to contextualize this study within that larger context and provide some perspective on what could be future research in this area.

First and foremost, it should be re-emphasized here that while this study contributes to the understanding of how the MDCT item functions with JEFL examinees, and suggests a need for further securitization of its use in listening proficiency assessment, it does not at this point directly substantiate any claims about the quality or appropriateness of actual operational tests currently employing them, including the *Center Test*. Although the *Center Test's* employment of MDCTs was a primary motivation for this study, and the MDCT format used by the *Center Test* a substantial inspiration for the items used in this study, further research is definitely needed, especially studies closely duplicating the actual conditions of the *Center Test.* Based on the findings of this study, it is the researcher's opinion that there is certainly a high need for such continued research, especially given the high-volume context of the *Center Test* and the possibility that even minor inaccuracies could lead to considerable consequences for examinees. Future studies should focus specifically on the *Center Test*, the high school JEFL learners that take it, as well as any other possible contexts of MDCT use in the JEFL context. Doing so would be a major step forwards in not only contributing to our understanding of such tests, but to improve our ability to interpret and employ them as responsibly as possible.

PAC7 at JALT2008: Shared Identities

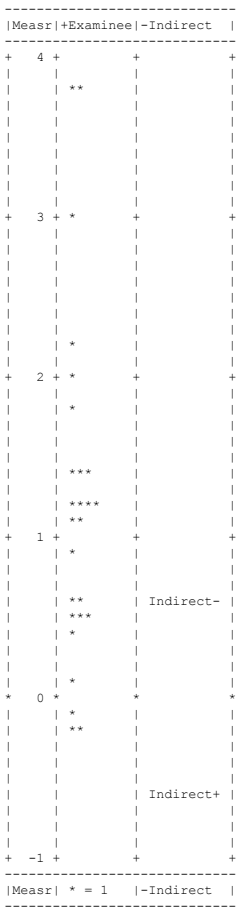

Figure 1. FACETS vertical ruler for the MDCT exam

Eric Setoguchi recently completed his Master's degree in Second Language Studies at the University of Hawaii at Manoa, and is currently a lecturer at the English Language Institute at Kanda University of International Studies. His research interests include language assessment in Japan, language test implementation, and task-based language teaching. <eric-s@kanda.kuis.ac.jp>

References

Bonk, W. J., & Ockey G. J. (2003). A many-facet Rasch analysis of the second language group oral production task. *Language Testing, 20*(89), 89-110.

Brown, J. D. (1997). English language entrance examinations in Japan: Myths and facts. *The Language Teacher, 19*(10), 21-26.

Center Test. (2006). Teacher committee evaluation of the Center Test in English listening. [Online] Available: <http://www.dnc.ac.jp/old_data/exam_repo/18/pdf/18hyouka63.pdf>

Hudson, T., Detmer, E., & Brown, J. D. (1995). *Developing prototypic measures of cross cultural pragmatics*. *Technical Report 7*. Honolulu, Hawaii: University of Hawaii, Second Language Teaching and Curriculum Center.

Jianda, L. (2007). Developing a pragmatics test for Chinese EFL learners. *Language Testing, 24*(3), 391–415.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241-256.

PAC7 at JALT2008: Shared Identities

Ministry of Education, Culture, Sports, Science, and Technology. (2003). Regarding the establishment of an action plan to cultivate Japanese with English abilities. [Online] Available: <http://www.mext.go.jp/english/topics/03072801.htm>

Norris, J. M. (2008). Validity evaluation in language assessment. New York: Peter Lang.

Rose, K. (1994). On the validity of discourse completion tests in non-Western contexts. *Applied Linguistics*, *15*(1), 1–14.

Rose, K. & Ono, R. (1995). Eliciting speech act data in Japanese: The effect of questionnaire type. *Language Learning, 45*(2), 191–223.

Shimizu, T., Fukasawa, E., & Yonekura, S. (2007, March). *Dearth of pragmatic information in Japanese high school English textbooks*. Paper presented at the 17th International Conference on Pragmatics & Language Learning, Honolulu, Hawaii.

Sick, J. (2008). Rasch measurement in language education: Part 1. *Shiken: JALT Testing & Evaluation SIG Newsletter, 12*(1). Retrieved March 17, 2009 from <http://jalt.org/test/sic_1.htm>

Tada, M. (2005). *Assessment of EFL pragmatic production and perception using video prompts*. Unpublished doctoral dissertation. Philadelphia: Temple University.

Yamashita, S. O. (1996). *Six measures of JSL pragmatics.* Honolulu, Hawaii: Second Language Teaching & Curriculum Center of University of Hawaii at Manoa.

Yoshitake, S. S. (1997). Measuring interlanguage pragmatic competence of Japanese students of English as a foreign language: A multi-test framework evaluation. Unpublished doctoral dissertation. Columbia Pacific University, Novata, CA.