MULTIPLE-CHOICE DISCOURSE COMPLETION TASKS IN JAPANESE ENGLISH LANGUAGE ASSESSMENT

ERIC SETOGUCHI

University of Hawai'i at Manoa

ABSTRACT

A new class of multiple-choice discourse completion tasks (MDCTs) is beginning to gain popularity in the Japan English as a Foreign Language (JEFL) assessment context. In this study, an experimental MDCT test was administered to a sample of Japanese university students. An item format analysis was conducted focusing on the construct validity and discrimination of MDCTs in measuring the English listening proficiency of JEFL speakers. Using a combination of classical test theory and Rasch analysis, test performance was analyzed in regard to two research questions: (a) whether a pragmatic proficiency construct is related to item difficulty, and (b) whether the use of different distractor types has an effect on item discrimination characteristics. The results suggest that a pragmatic proficiency construct plays a role in determining item difficulty on MDCTs, bringing into question the construct validity of MDCTs as a listening language proficiency measurement. Additionally, MDCT item discrimination might be affected by the type of distractors being used, hinting at possible ways to optimize discrimination of MDCTs in norm-referenced testing (NRT). Given the high probability of continued use of MDCTs in the JEFL context and the need for more investigation into these items, this study is hopefully an early step towards better and more informed MDCT test design and use.

INTRODUCTION

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 the countries' foreign language teaching style from an emphasis on a traditional, synthetic-based approach, heavy in grammar, vocabulary, and

Second Language Studies, 27(1), Fall 2008, pp. 41-101.

translation, to a communicative language 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 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 nationwide university entrance exam system, 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 a non-traditional item type never before used in large scale authentic language assessment, the multiple-choice discourse completion task (MDCT). MDCTs are a subclass of the discourse completion task (DCT), a pragmatic instrument that was first researched for its potential in English language assessment three decades ago (Levinston, 1975). One issue of concern is that MDCTs are being rapidly implemented into an operational assessment context before their potential has been well studied for the specific intended use. A lack in understanding of how MDCTs function in the JEFL context represents a potential threat to quality language assessment, given the extent to which they are already being used in operational evaluation and decision-making purposes. For this reason, the purpose of this study is to further investigate the function and quality of the MDCT item type in the JEFL assessment context.

This paper begins with a brief review of MDCT item research and identifies several issues with operational use of MDCT items as the motivation for the study. The main body of the paper outlines the creation, implementation, and findings of a preliminary study designed to lead to better understanding of these issues. It concludes with recommendations for areas of future research and some preliminary recommendations for improving the MDCT item format in current operational testing.

THE MDCT ITEM

Variation in MDCT Item Format and Use

Brown (2001) loosely defined a MDCT as a pragmatics instrument that requires students to read a written description of a situation and select what would be best to say in that situation from a set of choices (p. 301). While all MDCTs should share these general characteristics, a situational prompt and a set of answer choices, there is no singular definition that more explicitly lays out what an MDCT item should look like in order to be referred to as such. The potential of MDCTs in language assessment has been explored in a variety of settings and with examinees of more than one ethnicity, language, and proficiency level. A review of the literature reveals that MDCT item format differs across the context and purpose of the intended assessment in which they are being used, evolving and adapting to specific needs of various contexts of use (Yamashita, 1996, Yoshitake, 1997, Tada, 2005, Roever, 2006, Jianda, 2007).

MDCT Items on the Center Test in English Listening

This section of the paper presents and discusses the basic format of an MDCT item as it appears on the *Center Test* in English Listening, with some occasional referencing to new format elements introduced by this particular context of use that differ from most of the studies referred to above. A detailed discussion of the construct of this test (what the test is intended to measure) is reserved for a later section, but it is worth mention that based on available information about the test listening proficiency as distinct from pragmatic proficiency is the likely intended measurable target.

Examinees first listen to a prompt in the form of a short dialogue between two speakers, then read four accompanying lines of dialogue on their test form, as shown in Example 1. To answer the item correctly they select the line that most appropriately continues the dialogue. A conversation turn is assumed to take place. In other words, it is always assumed the next speaker is not the one that was heard last in the dialogue.

Example 1

Examinee hears:

W: What did you do over the weekend?

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

Examinee reads:

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.

(*Center Test in English Listening*, 2007)

Situational information in the prompt is delivered in the form of a conversation rather than a descriptive narrative. The impact of this alteration away from most other MDCT formats mentioned above is a decrease in the amount and detail of situational information about setting, situation, and roles provided to the examinee. Such a change is rather unusual considering that detailed situational descriptions are a common component of all MDCT variations under investigation in current language assessment research, e.g. those discussed thus far. Instead, in *Center Test* MDCTs, context-specific information relevant to each MDCT item is not provided to the examinee as a functional component of the item. In the case of the item shown in Example 1, a conversation about one of the speakers having read a really good book, the situation is only apparent as encoded information within the prompt dialogue itself. Who is speaking, where the conversation is taking place, and the ultimate intent of either speaker are typically not information that is made available to the examinee when the prompt occurs in this format.

MDCT items on the *Center Test* do not appear to be based on any of the three major speech acts: apologies, requests, and refusals, as do most other MDCT item formats appearing in the literature. One interpretation is that they could be based on common conversational topics appearing in English communication textbooks used in Japanese high schools. It could be said that a few of the more simple items to appear on the *Center Test* resemble the language routine based items from Roever (2005), but most items on the *Center Test* are more complex language tasks than what could be considered language routines.

Unlike all previous MDCT formats, distractors on the *Center Test* are not designed to represent different pragmatic strategies and formulas. Instead, they are designed to be truly incorrect answers that can be identified by non-pragmatic factors. Under close inspection, distractors on the *Center Test* fall into three major classes based on how they can be eliminated as possible answer choices by examinees: (1) fact explicit type, (2) fact implicit type, and (3) order type. A comprehensive discussion of each of the three classes follows accompanied by authentic examples.

Example 2 is a fact explicit type distractor. Factual information explicitly stated in the dialogue is contradicted by explicit information in the distractor. In this example, B's means of transportation is incorrectly referred to as "motorcycle". The distractor can be ruled out as a possible answer choice as long as the examinee was able to understand that B rides a "bicycle" to work.

Example 2: A fact explicit type distractor

A: So, how do you get to work?

B: Well, I live close by, so I just ride my bicycle.

distractor: I must get a motorcycle too.

Example 3 is a fact implicit type distractor. It does not contain an explicit factual conflict. Rather, the examinee must be perceptive to implicit information from the dialogue that is not explicitly or directly stated.

Example 3: A fact implicit type distractor

A: How about going to the Chinese restaurant for dinner?

B: Let's try a different restaurant tonight.

A: Why? I thought that was your favorite place.

distractor: Yes, but I don't like Chinese food.

In this case, it is only implicitly clear that it is unlikely that B does not like Chinese food, even though they state they do not want to go to the restaurant tonight. A's use of the particle 'the', and reference to the restaurant as B's 'favorite place', both implicities indicate that A and B frequent the Chinese restaurant. If these cues are understood, A's question "I thought that it was your favorite place" can be readily understood not to be questioning whether the Chinese restaurant is B's favorite place but an indirectly implied inquiry to why B suddenly wants to do somewhere else. In other words, the question does not preclude a direct answer but a divulging of the reason why B suddenly does not want to go to their favorite restaurant tonight. The correct answer for this particular item, “Yes, but their prices have gone up recently”, confirms that this was the intention. Examinees can only rule out fact implicit type distractors if they can successfully comprehend implicit cues from the prompt or distractor.

A number of existing studies suggest that comprehension of content that is not made explicit, such as that required in the example above, might pose a more difficult or at least cognitively different challenge for Japanese EFL learners than comprehension of explicit content (Takahashi & Roitblat, 1994; Taguchi, 2002, 2005). Another group of studies investigated the teachability of understanding of implied content to Japanese EFL learners. Kubota (1995) showed some success at teaching comprehension of English implicature to Japanese EFL students with explicit instructional methods. In general, instruction in implication and other pragmatic competencies are a neglected part of secondary level English curriculum in Japan (Kubota, 1995). Therefore, regardless of whether implication really is a higher cognitive challenge, if students who take the Center Exam are exposed to it through fact implicit type distractors, they would be expected to be more difficult simply because they measure a language proficiency that has not been learned in the classroom.

Example 4 Is an example of an order type distractor. Unlike with fact type distractors, there is nothing explicity or implicity stated in the distractor that contradicts the dialogue. Rather, the line is quite plausible in this situation, but is inappropriate in the particular order it occurs.

Example 4: An order type distractor

A: What did you do over the weekend?

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

distractor: Sure, I’ll lend it to you when I’m done.

The error in order can be related to timing, where the distractor appears too early or late to appropriately continue the dialogue, or role, where the distractor is not an appropriate line for the speaker whom the examinee is assuming the role of. The distractor in the example is a case of both. The distractor occurs too early in the conversation, and is a line that would be spoken by B, not A, the speaker who is designated to speak next. This fact can be demonstrated by logically continuing the conversation to the point where the distractor becomes appropriate, as shown below.

A: What did you do over the weekend?
B: Oh, I started reading a really good book.
A: Really? Could I borrow it?
B: Sure, I'll lend it to you when I'm done.

In order to rule out an order type distractor, examinees rely on their comprehension of how far the dialogue has progressed and which speaker is playing which role.

The Operational Testing Context of the Center Test

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 use the *Center Test* for a variety of different purposes. The stakeholders with the highest priority are universities, many of which utilize test scores as a part of their admissions process. 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 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 will not need further in depth knowledge about the *Center Test* for purposes of this paper, I'll conclude this section here by emphasizing two critical 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.

A summary of MDCT item format and context of use for the *Center Test in English Listening* is shown in Table 1. Note that the context of use of the *Center Test* is in a large scale gatekeeping assessment of listening proficiency, with real consequences for examinees.

Table 1
Test and Item Characteristics of the Center Test in English Listening

TEST FACTORS	*Center Test in English Listening* (Introduced 2006)
Language Context	EFL
Test Format	Aural
K	7 MDCTs, 28 total items
Intended use	Gatekeeping
EXAMINEE FACTORS	
N	~500,000
Participant nationality	Japanese
Language level	Various
MDCT FORMAT	
Situational prompt	Spoken dialogue
Content	Various, mostly taken from conversational topics in high-school textbooks
No. of distractors	4
Characteristics by which distractors are identified	Fact (implicit & explicit) and timing cues

MOTIVATION FOR THE STUDY

The *Center Test* is one context where MDCTs are already being used in high-stakes operational assessment context in the JEFL setting. There is a definite need to better understand how the MDCT item functions in the JEFL context, not only so that a substantiated argument can be made for or against their use on the *Center Test*, but to better inform further decisions about MDCT use in other contexts as well. Given time and scope limitations, this study focuses primarily on investigating two specific issues: (a) as MDCTs are traditionally measurements of pragmatic proficiency, what is their potential for use in other assessment purposes without introducing construct irrelevant variance, and (b) what is the relationship, if any, between distractor type and MDCT item discrimination.

Investigating MDCT Construct Irrelevant Variance

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 only measure all of and only the construct under investigation, and scores should not be influenced by variance from undesirable effects (as cited in Norris, 2008, p. 44). It follows that the primary threat to validity are 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. 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 only assesses the listening proficiency of examinees, the singular construct of its intended use. The obvious concern here is that all MDCTs 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. Despite modifications to answer choices to make them less obviously pragmatic in orientation, it cannot be ruled out that MDCTs on the *Center Test* covertly function to assess examinee pragmatic proficiency in addition to listening proficiency.

Some would argue that all language competencies, including listening proficiency, inherently include pragmatic competence. The model of language competence proposed by Bachman (1990) included pragmatic competence as an inseparable and necessary component. While the researcher would not argue the case that pragmatics plays a role in many if not all situation of language use, the point of concern here is how designers and users of MDCT tests in the JEFL context conceptualize what the MDCT item tests, as this will in turn shed light on what the construct of MDCT tests in Japan really is. 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 would 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 in high school EFL education (Shimizu et al., 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, in investigating the construct validity of MDCT items in the assessment of English listening proficiency in the JEFL context it will be assumed that any variation in test performance due to pragmatic proficiencies of examinees represents undesired construct irrelevant variance.

A Brief Review of Japanese and English Pragmatics

At this point, it will be useful to provide a summary of major studies that have investigated pragmatic differences in the Japanese and English language, and how they are related to potential causes for construct irrelevant variance when MDCT-based tests are given to JEFL examinees. It was pointed out earlier that the MDCT format requires examinees to judge the appropriateness of dialogue in the answering of items. Of concern is whether examinees would use linguistic cues alone in their judgments, or pragmatic cues as well. Rose (1994, 1995) demonstrated some evidence that JEFL learners were influenced by pragmatic cues of indirectness in their answering of MDCTs. A number of other studies in the JEFL context corroborate this hypothesis, and provide some context for it. Rose (1996) pointed out that a belief in the propensity of the Japanese language for indirectness has been a persistent fixture in the field of Japanese language and culture. Inspired largely by this characterization, Takahashi (1987) was the first to attempt to experimentally investigate the differences in directness in the language use of Japanese ESL and EFL speakers compared with that of native English speakers in their performance of speech acts. A similar experiment had been attempted earlier in the context of Israeli ESL learners, finding some evidence for the transfer of Hebrew speech patterns into English used by forty-four Israeli university students (Cohen & Olshtain, 1981). The motivation for these studies was the theory of "pragmatic transfer", defined in a previous study as ‘transfer of L1 sociocultural communicative competence in performing L2 speech acts’ (Takahashi & Beebe, 1989). The major task for Takahashi was demonstrating how Japanese sociocultural and communicative practices influence the L2 use of JEFL learners in speech act situations. By administering an open-ended DCT refusal task to sixty Japanese EFL and ESL learners and twenty native speakers of English, it was observed that higher proficiency Japanese English speakers in general used higher frequencies of indirect language softeners in their refusals, including intensifiers, excuses, and expressions of politeness. According to Takahashi, this finding could be interpreted as a transfer into the L2 of ‘the Japanese norm of avoiding direct expressions and sounding polite as possible’ (Takahashi T., 1987). It remains an ongoing question whether Japanese L2 English learners demonstrate a measurable preference for indirect English behavior as a result of their L1. Findings from Beebe, Takahashi, Ulitz-Weitz (1990) support this argument, while Fukushima (1990) and Rose (1992) provide evidence that Japanese EFL learners are in fact more direct then English native speakers in performing request speech acts. Several attempts to account for conflicting findings in the level of directness used by Japanese EFL learners have focused on individual variation in proficiency level as a factor. A surprising finding from Takahashi (1987) was that low proficiency Japanese ESL speakers and EFL Japanese speakers in general used higher frequencies of direct language in their refusals. This finding was hypothesized to be a reflection of the limitations in vocabulary of EFL and low proficiency speakers, which would not necessarily contradict observations of pragmatic transfer of indirectness from Japanese observed in higher proficiency behaviors. Studies attempting to confirm this trend found evidence to both support (Hill, 1997) and dispute (Maebashi et al., 1996; Takahashi S., 1996) that Japanese EFL learners would show increased use of indirect language in their L2 with increasing proficiency.

More recently, compelling evidence has suggested that the method of collection of speech act behavior has a significant impact of the nature of the behavior itself. Rose (1994) demonstrated that Japanese EFL speakers would use more direct language when given open-ended DCTs of request speech acts, but would favor indirect language when given MDCTs of request speech acts. This finding was further corroborated by a follow-up study in the same context (Rose & Ono, 1995).

The researcher would like to draw a brief distinction here between pragmatic behavior and pragmatic test behavior. We do not yet have a clear understanding of indirectness of L2 speech in Japanese EFL learners, nor can we say anything conclusive yet about the role of L1 transfer on this behavior. While further qualitative and quantitative behavioral studies like those above will be necessary to better explain the pragmatic behaviors of Japanese English learners, this study is concerned solely with test behavior and MDCT item quality, and would only provide a loose theoretical basis for making any conclusions about general pragmatic behavior. Therefore, this study does not claim to add to knowledge in the field of behavioral pragmatics, and the intent of the researcher is solely concerned with investigating the MDCT item itself and how JEFL learners interact with it.

What this is study is concerned with is the effect that JEFL learners' pragmatic behaviors will have on how they approach MDCTs as test tasks, and what contribution this will have on variation in test performance. As it has already been strongly established that the intended construct of MDCTs as they are currently being used in JEFL assessment is largely as listening proficiency measurements, any indication of performance variation due to pragmatic behaviors would be considered construct irrelevant variation, and would therefore be a threat to construct validity of MDCTs. In investigating the truth of this claim, this study hopes to contribute to making clearer what construct MDCTs test in the JEFL context, and how this should be incorporated into decisions of test use and interpretation.

Investigation of MDCT Item Discrimination

The *Center Test in English Listening* is a large scale norm referenced test (NRT) designed to produce a dispersion of scores over a very large population of examinees. None of the current research into MDCTs has investigated their potential in this context, and no evidence has been produced concerning the item discrimination behavior of MDCTs (the degree to which an item differentiates between examinees of different proficiency levels). In light of this fact, a second focus of this study is to evaluate the MDCT item for its potential in discriminating large populations of examinees. In multiple choice testing, distractor quality is one determining factor in item discrimination quality, in that distractors should be appropriately meaningful and plausible to examinees. Brown (2005) cautions for test designers to make sure all distractors in a multiple choice item are sufficiently plausible. As discussed above, MDCTs distractors on the *Center Test* fall into three different categories. In evaluating the item discrimination of the MDCT, this study empirically compares the discrimination behavior of distractor category types as a way of learning more about item performance and providing some evidence for more informed item design.

Addressing Limitations of Current Research

A secondary objective of this study is to address a lack in empirical, validation focused research on the *Center Test*. The *Center Test* and the *Daigaku Nyushi Center* have historically been subject to confidentiality requirements that have implications for the availability of information to researchers. Although the *Daigaku Nyushi Center* does track detailed statistics (e.g., item level statistics), these statistics are not available to the public or to researchers. As of yet, no studies have been published on the new *Center Test in English Listening*. Very little quality research in English has been done on the *Center Test*, thereby limiting the access of non-Japanese scholars (refer to Brown & Yamashita, 1995; Ingulsrud, 1994; Ito, 2005, for exceptions to this). Research into the new *Center Test in English Listening* is still lacking, and investigations into MDCT items as they appear on the test have been generally ignored. With the exception of Ito (2005), empirical, research on the *Center Test* in general has been lacking. The aim of this study is to provide item-specific data to reinforce the non-empirical observations in the literature, and provide a more concrete foundation for making practical improvements to the *Center Test* and MDCT testing in the future of JEFL assessment.

Purpose

As MDCTs in JEFL assessment are currently being used most prominently on an English listening exam under conditions of norm referenced testing (NRT), it would be useful for an investigation of MDCT items to focus on aspects of their validity in measuring L2 listening proficiency as well as their discrimination characteristics. To this end, the following two research questions are addressed in this study:

1. Is there an observable effect of examinee pragmatic proficiency on MDCT item performance?
2. What is the observed discrimination behavior of fact implicit type distractors when compared to between fact explicit type and order type distractors on *Center Test* MDCTs?

CREATING THE TEST INSTRUMENT

Initial Conceptualization

The instrument used in this study consisted of an English test composed of forty-two MDCT items of the same format as those appearing on Japan's *Center Test in English Listening*. The underlying purpose guiding the design of the test instrument was for participant response data to provide evidence related to two questions concerning the MDCT item type: (a) the role of a pragmatic construct in determining exam performance, and (b) the effect of fact implicit type distractors on item discrimination. In order to investigate these two research questions, this study utilized a unique research approach to item analysis. First, two MDCT item manipulation techniques were developed specifically for this study. The first, *indirectness factor*, refers to the level of directness of language in the answer key for a particular MDCT item. The second, *implicature factor*, indicates the presence or absence of fact implicit type distractors in a particular MDCT item. Manipulating MDCT items along the factors and observing the subsequent changes in examinee test performance could potentially provide valuable information.

For example, as directness is a pragmatic feature of language, manipulation of MDCT items along the *indirectness factor* and observing how this affects the relative difficulty of test items is one potential measure of variation in test performance due to pragmatic abilities of examinees. Comparison of examinee test behavior on MDCT items along the *implicature factor* might ascertain whether fact implicit type distractors possess any unique qualities in terms of MDCT item discrimination. A more in depth description of both variables and how they function in the test instrument can be found in a later section of this paper.

Both *indirectness factor* and *implicature factor* describe features of MDCT answer choices (distractors or answer keys), and are distinct from the prompts with which they combine to form a complete MDCT test item. Therefore, the first step in developing the test instrument was the process of developing MDCT item prompts, which would later be combined with answer choices to form complete MDCT test items for use on the test instrument. The following two sections of this paper are an overview of the development of item prompts.

Initial Review of MDCT Item Prompts

Listening prompts for MDCT items on the *Center Test* consist of two to four lines of actual dialogue between two speakers of English. This same format was adopted for use in the test instrument of this study. The listening prompts are a combination of those used in authentic items appearing on the prototype, 2006, and 2007 versions of the *Center Test* (available online from http://www.dnc.ac.jp/index.htm), supplemented with original prompts. Creating the test instrument entirely of authentic items was considered, but ultimately abandoned, since a number of items appearing on actual *Center Tests* were deemed inappropriate for use in the study. This included items thought to be obviously flawed or confusing, or not challenging enough for use with university EFL learners. Clearly, the exclusion of these items reduces the extent to which this study will be relatable to MDCTs on the actual *Center Test*, however the primary intent of this study is not to serve as an analysis of the *Center Test*, but an investigation of the MDCT item format itself and its' potential uses throughout the JEFL assessment context. This indirectly relates to the *Center Test* as the source of the MDCT format investigated and as one of many possible contexts of use, but is unconcerned with exclusively targeting MDCTs on the *Center Test* and therefore takes some liberties in the selecting of certain items and exclusion of others.

The exclusion of some items proved to interesting in and of itself, as poor items are evidence for the importance of careful item writing and reviewing before implementation in operational tests. For example, an item appearing on the 2007 *Center Test* that was deemed inappropriate for use in the study is shown in Example 5. The item contains a noticeable flaw in that there are multiple plausible answer choices.

Example 5

Examinee hears:

M: I'm worried about the dog.

W: Yeah, she hasn't eaten anything for two days.

M: Maybe we should take her to Dr. Thompson.

Examinee reads:

1. OK, I'll find something for her to do.
2. OK, I'll find something for her to eat.
3. OK, I'll take her for a walk tonight.
4. OK, I'll take her tomorrow evening.

Credit was only given if examinees marked (4) as their answer choice, which seems to be the best answer. However, upon close inspection answer choice (1) can also be correct. This is especially true if the pronoun “her” is interpreted to refer to “Dr. Thompson” instead of the dog, the latter being the likely intended reference, but not the only one. Items such as this are of poor quality and misleading, and perhaps indicate a need for better informed and more careful item design on the *Center Test*, in addition to the issues addressed in this study.

A total of thirteen authentic listening prompts from actual *Center Test* administrations were used in the test instrument (two from the prototype, five from the 2006, and six from the 2007 *Center Test*). An additional thirty prompts were developed by the researcher to best mimic the context and difficulty level of those appearing on the *Center Test*. In order to accomplish this, a careful review of prompts from the prototype, 2006, and 2007 versions of the *Center Test* was done. After review, it was apparent that much of the dialogue content of MDCT item prompts appeared to be based on material commonly used in English communication textbooks used in Japanese high school classrooms. As an example, the MDCT items appearing on the 2007 *Center Test* consisted of the seven language situations appearing in Table 2.

Table 2
Example Language Situations from Center Test MDCT Items

Item #	Language Situation
1	talking about weekend activities
2	talking about transportation to school
3	asking someone to deliver a message
4	asking about car repair costs
5	talking about a favorite restaurant
6	talking about the health of a pet
7	talking about vacation plans

With the exception of item 4, the language situations conform remarkably closely to set language topics and themes that are very common to classroom materials used by Japanese high school students. It was decided that textbooks used in high school English communication classes would be an appropriate reference for creating the additional items needed to complete the test instrument. Only textbooks approved for use in high-school classrooms by Japan's Ministry of Education were selected as suitable reference material for the developing and writing of prompts for the original thirty items of the test instrument.

Considering Situational Variables and Writing of Item Prompts

Situational variables have been a major component in the design of DCT item prompts in previous research studies, and this topic will be briefly addressed here. Roughly defined, situational variables are social properties associated with speech events, of which several have been classified. In their attempt to design a DCT section of a L2 pragmatic proficiency assessment, Hudson, Detmer, and Brown (1995) incorporated the three most dominantly studied situational variables: power, social distance, and imposition. Table 3 defines these variables in detail.

Table 3

Power and Distance Situational Variables in MDCT Item Prompts

Relative Power	The degree to which the speaker can impose his or her will on the hearer due to a higher rank within an organization, professional status, or the hearer’s need to have a particular duty or job performed. (+P) Speaker has a higher rank, title, or social position, or is in control of the assets in the situation. (=P) Speaker is of approximately the same rank, title, or social position (−P) Speaker has a lower/lesser rank, title or social position, or is not in control of the assets in the situation.
Social Distance	The distance between the speaker and the hearer. In effect, the degree of familiarity and solidarity they share as represented through in-group or out-group membership. (+D) Speaker and hearer do not know or identify with each other. They are strangers interacting due to social/life circumstances. (−D) Speaker and hearer know and or identify with each other. There is an affiliation between the speaker and hearer; they share solidarity in the sense that they could be described as working toward a common goal or interest.

adapted from Hudon, Detmer, & Brown (1995)

This framework has direct applications to the research of DCT items as pragmatic proficiency assessments. Bachman and Palmer (1996) defined sociolinguistic competence (a component of pragmatic proficiency) as the ability to employ language appropriate to a particular language use setting (as cited in Norris, 2001, p.248). Language use settings are defined in part by situational variables including interlocutor power, social distance, and level of imposition. By challenging examinees with language situations of varying power, distance, and imposition conditions, researchers and test designers can gather information about individual ability to deploy or identify appropriate pragmatic strategies for specific situations. Open-ended DCTs specifically target the ability to employ appropriate strategies in actual language use, while MDCTs target the ability to recognize these strategies among a series of choices.

The need to consider situational variation and situational variables in the design of MDCT item prompts for the test instrument used in this study is rather ambiguous. Complicating matters is the fact that the MDCT items appearing in the test instrument are designed to mimic those that appear on the actual *Center Test*, which as discussed earlier employ an entirely different system for framing answer choices that is unrelated to pragmatic strategy options. Therefore, in a testing situation where examinees are not presented with the challenge of having to recognize the appropriate pragmatic strategies that correspond to situational variables, the purpose of attending to such variables when designing MDCTs of this format is questionable. Nothing in the history of research into MDCT item design suggests a clear answer to this question. This issue presented a problematic dilemma in the design of the test instrument, as no justification could be given could be given for or against attention to situational variables in the design of the MDCT item prompts. Focusing solely on one combination of situational variables for the entire test, in other words forty-two prompts of −P/−D configuration for example, had been considered as a viable option. This was rejected however as it was felt this would result in a repetitive and unauthentic test to which examinees might respond negatively. In the end, it was decided that the best option was to balance the prompts on the test to include an equal proportion corresponding to each possible combination of situational variables. How this was accomplished for the forty-two items is summarized in Table 4.

Table 4
Item Distribution Across Situational Variables on the Test Instrument

		Power		
		+	=	−
Distance		7 items	7 items	7 items
		7 items	7 items	7 items

A number of important points regarding this approach require further explanation. First was the decision to exclude the imposition situational variable. One outcome of the review of prompts appearing on the *Center Test* referred to above was a finding that almost none included relevant imposition information. This is not altogether surprising, as the imposition variable is commonly omitted as irrelevant in studies using DCTs that do not contain request or apology speech acts (Rose, 1994). The *Center Test* and our test instrument contain no apology situations, and only request and refusal situations of low imposition (according to guidelines from Hudson, Detmer, and Brown, 1995). Therefore, the imposition variable was dropped in this study. A second point is the addition of a new category in the power variable, the equal power category (denoted as =P). Although rarely used in pragmatic studies, the *Center Test* and our test instrument contain several items with the speaker and hearer having approximately equal social status, including conversations between classmates, friends, and coworkers. To address such items this new category was created. Finally, it should be noted that although moderate attention was given to situational variables in terms of balancing and categorizing when designing the test instrument, they do not factor substantially in the final analysis section of the current study, which is primarily focused on issues of construct validity and item discrimination. The particular omission of situational variables is not expected to adversely effect the analysis conducted in this study, however a lingering questions remains whether situational variables in MDCT item design significantly effect examinee performance, which if true would suggest that specific attention to situational variables is something to be explored in future iterations of this research.

Using five government-approved high school English textbooks, thirty original MDCT item prompts and twelve authentic MDCT item prompts were combined into a series of forty-two prompts across six situational variable categories (Table 4). At this stage, four outsider raters were consulted to confirm the categorizations of the researcher. This step was seen as especially necessary given that MDCTs item prompts do not contain detailed descriptions of the speaking roles and setting of each language situation, as do most DCTs. Situational variables such as power and distance must be inferred from a few lines of dialogue. Four Japanese speakers of English were given a copy of the forty-two listening situations and asked to rate them in terms of power and distance variables according to modified guidelines from Hudson, et al. (1995). Each of the four raters was at an advanced level of proficiency and had at least three years of high-school and junior-high teaching experience in Japan in English communication classes. Table 5 shows the agreement percentages of each of the four raters for the variable categorizations. The table is sub-divided by categorization type, so that comparisons can be made between how the raters agreed with specific categorizations within each variable.

Table 5
Level of Agreement with Situational Variable Assignments

Situational	Percent Agreement			
Variable	Rater 1	Rater 2	Rater 3	Rater 4
Power				
+ (14)	71.4%	100.0%	100.0%	85.7%
= (14)	85.7%	92.9%	50.0%	78.6%
− (14)	71.4%	71.4%	57.1%	85.7%
Distance				
+ (21)	100.0%	100.0%	85.7%	90.5%
− (21)	100.0%	100.0%	95.2%	95.2%

The level of agreement about the variable categorizations was especially high in assignment of the distance variable. In other words, it was relatively clear from dialogues in the listening prompts what the power and distance relationship was between two speakers, even without detailed description of speakers and setting. The power variable proved to be more difficult for raters to perceive, especially with rater 4 who seemed to have trouble identifying situations of equal or negative power. Based on this data, individual prompts that showed 50% or more disagreement amongst raters (i.e., two or more of the raters disagreed with the researchers' categorization) were altered slightly to better emphasize the power relationship between speakers.

The Indirectness Factor

One of the main functions of the test instrument is to provide a means of quantitatively assessing whether examinee performance on MDCT items is affected by pragmatic proficiency. As discussed earlier, one of the primary reasons this might be happening on MDCT items as they appear in the *Center Test* is that examinees might be judging their selection of answer choices based on pragmatic appropriateness in addition to factual and chronological appropriateness. The challenge of this study was to come up with a way for this phenomenon to be empirically demonstrated within a sample of Japanese EFL examinees.

The first step in the process was to identify a single pragmatic feature to which Japanese EFL speakers would be the most likely to display high sensitivity. This feature could then be incorporated and experimentally manipulated in MDCT items on the test instrument with the hope of eliciting variation in overall test performance. Any observed variation could then be attributed, at least in part, to the pragmatic feature. Based on a number of pragmatic studies in the JEFL context, level of directness was chosen as the pragmatic feature for use in this study. A technique to incorporate level of directness into MDCT items was uniquely developed for this study, the *indirectness factor*. The *indirectness factor* refers specifically to the level of directness of an MDCT item answer key. It is assigned one of two values (+ or −). The values and their labels are described in detail below, followed by 2 examples used in the test instrument:

***Indirectness factor* (+).** The answer choices of positive *indirectness factor* items were designed to present a high level of acceptability to Japanese EFL students based on current literature on pragmatic behavior. These choices use strategies of indirectness, apology, excuse, and expressions of regret.

Examples:

Response to a student doing a favor for a professor (Item #3):

Thanks. I appreciate your help.

Response to a stranger not being able to fulfill a request (Item #14):

That's OK. Thanks anyway.

***Indirectness factor* (−).** The answer choices of negative *indirectness factor* items were designed to present a low level of acceptability to Japanese EFL students based on current literature on pragmatic behavior. These choices use strategies of directness and clearly lack the use of apology, excuse, and expressions of regret, even in situations where they are applicable.

Examples:

Response to a request for directions from a stranger (Item #10):

I don't know.

Response to a request by a student to a professor for a delay in an assignment (Item #11):

No, give it to me today.

If performance on MDCT items is influenced by examinee pragmatic proficiency in the area of directness, it was hypothesized that *indirectness factor* (+) items would be substantially easier for examinees than *indirectness factor* (−) items. That is, Japanese EFL students were anticipated to more easily identify correct answers in items that use indirect and passive strategies than correct answers on items that use direct and aggressive strategies.

The Implicature Factor

The second main function of the test instrument was to investigate the relation between MDCT item discrimination and distractor type. In multiple choice test items, the relative ease with which distractors can be dismissed by a group of examinees has implications for how the item will discriminate between examinees of different abilities.

There are at least three major types of distractors on MDCT items investigated in this study: fact explicit type, fact implicit type, and order. Nothing is really known about whether there are differences between the types in how easily they can be dismissed by JEFL learners. In particular, compelling evidence exists suggesting that rejection of fact implicit type type distractors might present a markedly higher challenge for JEFL speakers than either fact explicit type or order type distractors (Takahashi & Roitblat, 1994; Taguchi, 2002, 2005). In order to examine this issue, a second technique was uniquely developed for this study, the *implicature factor*. The *implicature factor* refers to the presence or absence of a fact implicit type distractor in an MDCT item. Similar to the *indirectness factor*, (+) and (−) values have been assigned based on whether the items have at least one fact implicat type distractor. The two categories are described in detail below, and Example 6 can be referred to as a typical example of a fact implicit type distractor:

***Implicature factor* (+).** Examinees must infer information from the dialogue on the basis of implicit meaning and apply this information to eliminate distractors and select the correct answer. Understanding implicit meaning may include inferring important information concerning the relationship of speakers, speaker opinion or stance, or the location or context of the dialogue, which are not directly stated in the dialogue.

***Implicature factor* (−).** Examinees do not have to infer information from the dialogue or perceive implicit meaning to eliminate distractors and select the correct answer. Examinees will be able to select the correct answer on the basis of their comprehension of the dialogue and the answer choices.

In this study, *implicature factor (+)* items only contain one fact implicit type distractor, which was done for two reasons: (a) to avoid creating items that might pose too difficult a challenge for students to complete given time constraints; and (b) given the difficulty of actually writing high quality fact implicit type distractors, creating more than one credible fact implicit type distractor per MDCT item proved to be nearly impossible. *Implicature Factor (−)* items replace the fact implicit type distractor with a fact explicit type distractor. All other distractors on the exam are order type distractors. In other words, *implicature factor (+)* items contain one fact type implicit distractor, one fact explicit type distractor, and one order type distractor. *Implicature Factor (−)* items contain two fact explicit type distractors and one order type distractor.

Realization of the Test Instrument

The next stage in development of the test instrument was incorporating *indirectness factor* and *implicature factor* into the writing of answer choices for each of the already existing forty-two item prompts. This process had do be done in a systematic way so that examinee performance on test items could be easily tracked, compared, and analyzed. Since each of the two factors had two possible configurations, resulting in four possible item configurations in total, it was decided to divide the test into four classes of items with equal numbers of items representing each of the four configuration types (Table 6).

Table 6
Schematic of Items on the Test Instrument

# of Items	Item Classification	Distractors (Total = 3)			Answer Key (Total = 1)
		#	Type	Implicature factor	Indirectness factor
9	Class A	1 2 3	fact explicit order fact implicit	+	+
9	Class B	1 2 3	fact explicit order fact explicit	−	+
9	Class C	1 2 3	fact explicit order fact implicit	+	−
9	Class D	1 2 3	fact explicit order fact explicit	−	−
6	linking items	unchanged from *Center Test*			unchanged from *Center Test*

Nine items of each type would appear on the test, each denoted with an alphabetic letter (A, B, C, or D). Type A items are *indirectness factor* (+) *implicature factor* (+) items. In other words, Type A items have answer keys written in indirect language style and contain one fact implicit type distractor. Type B items are *indirectness factor* (+) *implicature factor* (−) items, and have answer keys written in indirect language style and do not contain a fact implicit type distractor. Type C items are *indirectness factor* (−) *implicature factor* (+) items, and have answer keys written in direct language style and contain a fact implicit type distractor. Type D items are *indirectness factor* (−) *implicature factor* (−) items, and have answer keys written in direct language style and do not contain a fact implicit type distractor. Initially, a fifth class of six linking items was created for the test instrument, with items appearing exactly as they do on the actual *Center Exam*, in other words, with no modifications by the researcher to their prompts or answer choices. The original intention was to use these items as a check to ensure that items developed for use in the study were of approximately the same difficulty as those appearing on the authentic *Center Exam*, and relate findings of the study to critique of the *Center Exam*. Given the limited sample size of examinees participating in this study and differences between it and the intended target examinees of the *Center Exam*, the feasibility of making such a comparison was greatly reduced and the use of the items and intended was abandoned. While the linking items appear as artifacts in a few tables and figures in the findings section of the study, they otherwise are not considered as relevant in the remainder of this paper.

Using Table 6 as a guide, an appropriate set of answer choices for each item prompt were written resulting in nine items of each of the four classes. The items were randomized to appear in no particular order on the overall test, as were distractors and answer choices within each item. Appendix C is a summary table showing information for each item on the test instrument including a description of the prompt, situational variables, and class type. Appendix D is a master copy of the complete test instrument showing full prompt scripts and answer choices for each item.

In summary, the test instrument created for this study was a forty-two item MDCT exam (thirty-six items of which would ultimately be used). The items fall into four distinct sub-types, based on two item design factors that are designed to elicit variation in examinee performance that if observed would be relevant information to the research questions proposed by the study, mainly (a) whether examinees use pragmatic cues in the answering of MDCT items and (b) whether fact implicit type distractors pose a higher than average challenge to examinees relative to other types. In addition, MDCT item distractors of each of the three types classified in this study: fact explicit, fact implicit, and order type, are spread across the test items so that comparisons can be made between them.

METHOD

Participants

A total of thirty-seven participants were involved in this study. They were all Japanese university students pursuing English language education and currently enrolled in English language courses. Because the participants came from two different universities, at the outset of the study they were classified into two groups.

Group I consisted of twenty-six first and second year undergraduate students currently enrolled at a private university in central Japan. The students were members of a class participating in a one month study abroad program at an American university in the Spring of 2008 at the time of their participation, where they were enrolled in an English writing course. The majority of the students had never traveled abroad before. Group II consisted of eleven third and fourth year undergraduate students currently enrolled at a public university in southern Japan. The students were all enrolled in an intermediate level English communications course in Japan at the time of their participation. Some of the students had traveled abroad before for English study, and others had not.

As university level English learners, the participants in this study represent a population of learners that is distinct from the most likely target population of MDCT type exams in Japan, high school English learners. It is expected that the overall English proficiency of the participants will be higher than that of the typical high school learner given their additional language training. However, since there is a shared L1 and similar educational background between the participant group and the target population, it is anticipated that observable trends in score data might be expandable to the high school JEFL population as well. Of course, such a claim is only a reasonable hypothesis based on the context of this study, and future studies will have to be done on actual high school JEFL learners.

Materials

The material used in this study consisted of a carefully constructed test composed of MDCT type items like those appearing on Japan's *Center Test in English Listening*. The test contained forty-two MDCT items in total. This number incorporated concerns for adequate number of items for statistical analysis as well as an examinee fatigue factor.

Recall that the research questions of this study focus on an MDCT item quality analysis, not in the development of an actual MDCT test. On that point, the test developed and used in this study is not intended as an actual MDCT test, but an instrument created solely for use in this study to elicit information about JEFL examinee behavior on the MDCT item type, specifically with regard to questions of the relation between pragmatic proficiency and test performance, and discrimination behavior of different distractor types.

Procedure and Data Collection

The test was administered to the thirty-seven participants in two separate administrations. Group I was given the test in the last week of their one month study abroad program in the U.S. Group II was given the test during a special summer class session in Japan. In both administrations the test conditions reflected standardized testing procedure. The participants first heard a set of instructions in Japanese, followed by an audio recording of the test items. The voice actors for the dialogues used in the test were high proficiency Japanese speakers of English with native or near-native pronunciation. Each dialogue was read twice, and the participants had twelve second pauses between readings to record their answers on test forms. The total time for the test was approximately thirty 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.

Data Analysis

The test data of the thirty-seven examinees was analyzed using classical testing theory (CTT), Rasch analysis, and holistic item analysis. CTT was used to analyze overall examinee performance on the test, and investigate for differential performance on groups of items of the four different types. Item level CTT analysis (IF & ID) were calculated as well to estimate the difficulty and discrimination of individual test items. The major CTT analysis was conducted using the *Microsoft Office Excel*© statistics program. Additionally, *SPSS* was used for additional analysis purposes involving ANOVA (*SPSS* version 16.0), largely used to determine the significance level of differential performance observed through CTT. Rasch analysis was conducted using the FACETS many-facet Rasch Analysis Software© (Linacre, 1998). Finally, holistic item analysis was used to more closely investigate several test items that showed noticeably interesting characteristics, mainly unusual levels of difficulty or one or more distractors with marked effectiveness. Close inspection of individual items and their distractors condenses broad findings at the item type level into specific case-study exemplar items in an attempt to draw connections between item level characteristics to test level performance patterns.

RESULTS: PART I

CTT Analysis of Overall Performance on the Test

CTT test results for the examinees in Group I and Group II are shown in Table 10. The table contains both overall performance on the test as well as the groups' specific performance on the nine items of each subtype. The mean (M), standard deviation (SD), range, median, and minimum and maximum scores in each category were calculated to provide additional interpretations of examinee performance.

Table 7

Descriptive Statistics for Group I and Group II

Group I

	Item Type				Overall
	A	B	C	D	
N	26.00	26.00	26.00	26.00	26.00
K	9.00	9.00	9.00	9.00	36.00
M	7.15	7.39	5.77	5.35	25.65
SD	1.31	1.55	1.51	2.35	5.43
Range	6.00	7.00	7.00	9.00	20.00
Median	7.00	8.00	5.00	5.55	26.50
Minimum	4.00	3.00	3.00	1.00	16.00
Maximum	9.00	9.00	9.00	9.00	35.00

Group II

	Item Type				Overall
	A	B	C	D	
N	11.00	11.00	11.00	11.00	11.00
K	9.00	9.00	9.00	9.00	36.00
M	6.91	7.46	5.64	5.00	25.00
SD	1.14	1.51	1.50	1.79	4.07
Range	4.00	5.00	3.00	8.00	16.00
Median	6.00	8.00	6.00	5.00	25.00
Minimum	6.00	5.00	5.00	2.00	18.00
Maximum	9.00	9.00	7.00	9.00	33.00

Note that the six linking items have been omitted for this phase of the data reporting, as they were included as an indicator of the overall difficulty of the test instrument in comparison to the

Center Test and do not serve a purpose in answering the major research questions of the study. A perfect score on the exam was therefore set at 36 points as opposed to 42.

The mean score for overall performance on the test instrument was 25.65 points (71.3%) for Group I, and 25.00 points (69.5%) for Group II. Although it is difficult to make definitive judgments about the exam at this point in the analysis, it could be said that since the mean scores for both groups was above 50% the exam may have been slightly unchallenging for examinees. However, this finding is to be expected as the target users of the *Center Test* are high school seniors, yet all examinees participating in this study were university students who would be expected to have the advantage of additional language education. Interestingly, Group I scored slightly higher on the exam than Group II, despite being on average one to two years younger. While there were no instances of a perfect score, two examinees in Group I did come very close with scores of 35 (97.2%), and one in Group II achieved a 33 (91.7%). On the low end, two examinees in Group I received a score of 16 (44.4%), and one examinee in Group II only scored an 18 (50.0%). Score distribution for the overall exam is wide, with the range of scores in Group I at 20 points and Group II at 16 points. This suggests that there was some degree of variation in examinee ability that the test was able to discriminate.

At this point in the analysis, a t-test was conducted on the test data to determine if examinees in the two groups had scored statistically different from one another on the test. The analysis showed that statistical differences in test performance between Group I and Group II was not at the alpha level used for this study (α: $p = .05$), F (2,35) = 2.61, $p = .12$, so it was decided to consider both groups as a single thirty-seven examinee aggregate group for the remainder of the study. The separated data in Table 7 is combined into a single set in Table 8.

Table 8
Descriptive Statistics for all Examinees

	Item Type				Overall
	A	B	C	D	
N	37.00	37.00	37.00	37.00	37.00
K	9.00	9.00	9.00	9.00	36.00
M	7.08	7.41	5.73	5.24	25.46
SD	1.26	1.52	1.48	2.18	5.02
Range	6.00	8.00	6.00	9.00	20.00
Median	7.00	8.00	5.00	5.00	26.00
Minimum	4.00	3.00	3.00	1.00	16.00
Maximum	9.00	9.00	7.00	9.00	35.00

Analysis of Performance Between Item Types: CTT

A comparison across mean scores within items of each subtype indicates that there was noticeable variation in examinee scores between the item types. Examinees scored the highest on Type B items (average = 7.41), the lowest on Type D items (average = 5.24), with scores on Type A (average = 7.08) and Type C items (average = 5.73) falling in between. There seems to be a particular trend for examinees to score higher on Type A and Type B items than on Type C or Type D items, suggesting that MDCT items without the *indirectness factor* were markedly harder for JEFL students than those with the factor. Variation in examinee performance across the *implicature factor* does not appear to be happening based on comparison of mean scores. These findings suggest that the level of directness of an MDCT answer key has an effect on item difficulty, whereas the presence of a fact implicit type distractor does not. To confirm this observation, differences in means scores on each of the four item types were investigated with a one-way repeated measures ANOVA. The test confirmed that the difference in scores was statistically significant, F (3,144) = 14.84, $p < 0.00$. A post-hoc scheffé test was used to determine where the significant difference in scores was occurring, in other words, which of the four items types were contributing the most to score variation. The test showed that scores on Type A and Type B items as a pair differed from scores on Type C and Type D items as a pair ($p < 0.01$), but scores between these pairs, for example between Type A and Type B items, were not distinct enough to be considered statistically relevant. In other words, scores on Type A items were not significantly different from scores on Type B items, nor were there significant differences between scores on Type C items compared with Type D items. However, there was an overall difference between Type A and B items and Type C and D items.

Analysis of Performance Between Item Types: FACETS

By comparing mean scores on each item type using a CTT approach, it was able to be demonstrated that examinees scored significantly higher on Type A and Type B items than they did on Type C and D items. This finding suggests that there is a measurable effect of the *indirectness factor* on MDCT item performance for JEFL learners. In this study, FACETS analysis will be used as additional supporting evidence for the CTT findings concerning the effect of the *indirectness factor* and *implicature factor* on MDCT item performance. To meet this need, only a limited use of the many functions of FACETS analysis are required, therefore the presentation of findings in this section will be relatively basic in relation to those of more in depth FACETS studies, such as Bonk and Ockey (2003).

Analysis was carried out utilizing FACETS 3.63 (Linacre, 2007). The model was calculated using a three facet run, with items at the first facet, *indirectness factor* as the second, and *implicature factor* as the third. Both the *indirectness factor* and *implicature factor* facets each have two components, corresponding to the (+) and (−) values for each factor. The items facet was left intact representing all thirty-six items in order to give an overall impression of the difficulty of each item on the test relative to the examinees and the two factor facets. The FACETS computer program produces an output in the form of a vertical ruler, shown in Figure 1. The ruler is scaled in logits. A logit score is a direct representation of item difficulty; a higher logit score indicates high item difficulty and a lower logit score indicates a lower item difficulty.

Figure 1
FACETS Vertical Ruler

```
-------------------------------------------------------------------------
|Measr|+Examinee|+Item      |-Indirect  |-Implicature                   |
-------------------------------------------------------------------------
+   4 +         +           +           +                               +
|     |         |           |           |                               |
|     | **      |           |           |                               |
|     |         |           |           |                               |
|     |         |           |           |                               |
|     |         |           |           |                               |
|     |         |           |           |                               |
|     |         |           |           |                               |
|     |         |           |           |                               |
|     |         |           |           |                               |
+   3 + *       +           +           +                               +
|     |         |           |           |                               |
|     |         |           |           |                               |
|     |         |           |           |                               |
|     | *       |           |           |                               |
|     |         |           |           |                               |
|     |         |           |           |                               |
|     |         |           |           |                               |
|     | *       | 1         |           |                               |
|     |         |           |           |                               |
+   2 + *       +           +           +                               +
|     |         | 37        |           |                               |
|     |         |           |           |                               |
|     | *       |           |           |                               |
|     |         |           |           |                               |
|     | *       |           |           |                               |
|     | ****    | 17 18     |           |                               |
|     |         | 9         |           |                               |
|     | ****    | 19 29     |           |                               |
|     |         | 13 8      |           |                               |
+   1 + ***     + 38        +           +                               +
|     | ***     |           |           |                               |
|     | **      |           |           |                               |
|     |         |           |           |                               |
|     | **      | 30 31 35  |  Indirect-|                               |
|     | ****    | 2  38 33  |           |                               |
|     | **      |           |           |                               |
|     |         | 27 39     |           |                               |
|     |         |           |           |                               |
|     | *       |           |           |                               |
*   0 * *       * 36        *           * Implicature+     Implicature- *
|     | *       | 10 15 16  |           |                               |
|     | **      | 14        |           |                               |
|     |         |           |           |                               |
|     |         |           |           |                               |
|     |         |           |           |                               |
|     |         |           |  Indirect+|                               |
|     |         |           |           |                               |
|     |         |           |           |                               |
|     |         | 20 24 7   |           |                               |
+  -1 +         +           +           +                               +
|     |         | 21 26     |           |                               |
|     |         |           |           |                               |
|     |         | 3  40     |           |                               |
|     |         |           |           |                               |
|     |         |           |           |                               |
|     |         |           |           |                               |
|     |         |           |           |                               |
|     |         |           |           |                               |
|     |         |           |           |                               |
+  -2 +         + 11 41     +           +                               +
|     |         |           |           |                               |
|     |         |           |           |                               |
|     |         |           |           |                               |
|     |         |           |           |                               |
|     |         |           |           |                               |
|     |         |           |           |                               |
|     |         |           |           |                               |
|     |         | 22 32 5   |           |                               |
|     |         |           |           |                               |
+  -3 +         + 42        +           +                               +
-------------------------------------------------------------------------
|Measr| * = 1   |+Item      |-Indirect  |-Implicature                   |
-------------------------------------------------------------------------
```

It is relatively clear that the majority of examinees in this study outperformed the test, in other words, the average proficiency of examinees was too high for the average difficulty of the test. While the proficiency range of the examinees spans from just under 0.0 logits to just under

4.0 logits, the difficulty range of the items spans from -3.0 logits to just under 2.0 logits. In the ideal situation, the difficulty of items should precisely overlap with the proficiency range of examinees. This produces the best match between examinee proficiency and item difficulty, and will produce test scores reflecting the best separation of examinee proficiency. In this study, thirteen of thirty-six items fall entirely below the proficiency level of the examinees. This is not unexpected, as the university JEFL learner participants in this study should be of a higher language proficiency than the average high school JEFL learner for whom the MDCT item format was designed for, however it does represent a limitation in the study. Further examination of the examinee column reveals that the limited sample size used in this study approaches but does not quite produce a normal distribution, which should be noted as another limitation that is especially important to be addressed in future studies, in particular given that a condition of FACETS analysis is normally distributed data.

The statistics for the *indirectness factor* and *implicature factor* facet are shown in Table 9. Within the *indirectness factor* facet, *indirectness factor* (+) component had an average logit score of -.55 while *indirectness factor* (−) component had an average logit score of .55, a difference of one entire logit unit. In other words, as the *indirectness factor* (+) component has a higher logit score than the *indirectness factor* (−) component, MDCT items on the test instrument that were written using indirect language strategies were less difficult than items that were written using direct language strategies. This finding supports those from the CTT analysis, which found that the *indirectness factor* had a measurable effect on MDCT item difficulty for JEFL examinees. In contrast, component within the *implicature factor* facet did not seem to have an effect on item difficulty. The *implicature factor* (+) component had an average logit score of -.03, while the *implicature factor* (−) component had an average logit score of .03, a difference of a mere .06 logits. These findings are summarized in Table 9, along with the standard error and infit mean square values for each of the two factors.

Table 9
Indirectness Factor Facet and *Implicature Factor* Facet Summary of Statistics

	Measure	Measure *SE*	Infit mean square
Indirectness factor			
Indirectness factor (+)	-.55	.10	1.01
Indirectness factor (−)	.55	.09	.98
Implicature factor			
Implicature factor (+)	-.03	.09	1.05
Implicature factor (−)	.03	.09	.94

This finding also supports those from the CTT analysis, which found that unlike the *indirectness factor,* the *implicature factor* did not seem to have an effect on MDCT item difficulty, but as there is a noticeable discrepancy between the distribution of examinees and items along logits (the examinee group being higher than the items group), this finding might perhaps be relevant with low proficiency examinees. Repeating the study with a population of lower proficiency examinees that better match the items is recommended.

Analysis of Item Discrimination: Point-biserial Correlation Coefficient

The second of three research questions investigated by this study focused on the issue of MDCT item discrimination quality. The issue was raised whether a measureable effect on item discrimination could be attributed to distractor type in MDCT items. In particular, it was proposed that fact implicit type distractors would pose a higher than average challenge for JEFL learners and therefore might influence how MDCT items containing them discriminate within a population of JEFL examinees.

A straightforward approach to this issue is to first investigate item discrimination characteristics for all items on the test instrument and compare between the four item types. As Type B and Type D items are the items on the test that did not contain a fact implicit type distractor, any noticeable difference between their item discrimination and those of Type A and Type C items might be due to of the presence of such a distractor. In order to evaluate the item discrimination of items on the MDCT test, a statistical measurement called the point-biserial correlation coefficient (r_{pbi}) was used in this study. The r_{pbi} is essentially a measure of the degree to which individual items on a test are related to the total test scores, and is an appropriate measurement to apply in cases where test data is in the form of dichotomously coded items (correct or incorrect) on a continuous scale (Brown, 2005). A high r_{pbi} would be an indication that the item is discriminating well between the better students and the weaker students, while a low r_{pbi} would indicate that the item is not discriminating well between these two groups.

Coefficient measures for each item on the test are shown in Table 10. Note that r_{pbi} values for each item were obtained in relation to the total test score of particular item types, and not the overall total test score. In other words, the r_{pbi} for a Type A item was calculated using total test scores on Type A items only. If the assumption is that there are fundamental differences between item types on the test, it would not make sense to compare performances between items of different types as we would not necessarily expect performance on one item type to correlate with performance on another.

Table 10
Point-biserial Correlation Coefficients

Item Type							
A		B		C		D	
Item #	r_{pbi}	Item #	r_{pbi}	Item #	r_{pbi}	Item #	r_{pbi}
2	**0.60**	3	0.35	1	0.36	10	**0.55**
5	0.28	7	**0.44**	8	0.26	13	**0.66**
9	**0.60**	14	**0.56**	11	0.36	17	0.39
16	**0.46**	15	**0.69**	18	**0.60**	19	**0.42**
20	0.35	24	0.33	21	0.53	30	**0.58**
22	0.01	31	**0.53**	27	**0.41**	33	**0.50**
26	0.09	35	**0.68**	28	**0.44**	36	0.33
32	0.15	41	0.15	29	0.28	37	**0.65**
38	**0.50**	42	NA	40	0.35	39	**0.49**
M	0.34	M	0.47	M	0.40	M	0.51
SD	0.21	SD	0.17	SD	0.11	SD	0.12

Note: “NA” indicates an item for which there were no incorrect responses, therefore a r_{pbi} value cannot be calculated
Items in bold type are considered very good discriminating items

Values for r_{pbi} always fall between –1 and +1, but it is not an easy problem to definitively determine what cutoff value constitutes good item discrimination. Ebel (1979) proposed a set of guidelines for evaluating the quality discrimination of individual items when using point-biserial correlation coefficients (Table 11).

Table 11
Relating Point-biserial Correlation Coefficient and Item Discrimination Quality

r_{pbi}	item quality
> .40	very good items
.30 to .39	reasonably good items
.20 to .29	marginal items
< .19	poor items

from Ebel (1979)

According to Ebel, items with r_{pbi} greater than .40 are considered to be very good items. Looking at Table 10, Type D items seemed to perform the best with regard to item discrimination. Seven out of nine Type D items could be considered very good items, and the mean r_{pbi} for subtest D was the highest of the four types. Type B items performed nearly as well, with five out of nine considered very good items. In comparison, Type A and Type C items did not perform as well as either Type D or Type B items. This finding suggests that MDCT items that contain a fact implicit type distractor do not discriminate well when used in populations of JEFL examinees when compared to MDCT items that do not contain a fact implicit type distractor. The next section of this paper will investigate this issue in more detail.

Analysis of Test Errors Related to Distractor Type Selection

A finding of the previous section was that MDCT items that contained a fact implicit type distractor as one answer choice did not seem to discriminate among the sample population of JEFL examinees as well as those items that did not contain one. In order to investigate this issue further, a more detailed analysis of precisely how the three different distractor types behaved on the test was carried out.

Investigating the differences between fact explicit type, fact implicit type, and order type distractors is more complicated than our investigation has been thus far. Unlike other factors which were purposefully designed to occur in a pre-determined configuration across the test instrument, the three distractor types occur in irregular frequencies. In other words, the primary focus of the test instrument developed for this study was on measuring the specific effect of the indirectness and implicature factors, which are balanced across four item types to facilitate certain analytical approaches. The three distractor types occur in an irregular distribution across all the items, which does not facilitate similar analytical approaches. Despite such difficulties, this study will attempt to provide a preliminary investigation of this issue as possible with the less than ideal data available.

A central issue concerns why fact implicit type distractors seemingly lead to poor item discrimination. It will be helpful to examine how examinees of different overall proficiency levels are being distracted (or not distracted) by different distractor types. The first step in such an analysis is recognizing that behind every incorrect response on an MDCT item is a distractor that was mistakenly selected by an examinee. If one were first to classify the total incorrect responses for each examinee into groupings by the type of distractor mistakenly chosen, then order examinees based on overall test performance, the relative effectiveness of the different distractors across examinee proficiency could be compared. The analysis at this point returns to looking at the test instrument as a whole. In other words, the thirty-six items (one-hundred-eight distractors; three distractors per item x thirty-six items) will be considered, for the time being, as a single homogeneous test.

Table 12 shows the total number of incorrect responses by each examinee subdivided into the distractor type that was mistakenly selected. The table is organized with the highest performer at the top (#116, with one missed item) and the lowest performer at the bottom (#113, with twenty missed items).

Table 12
Total Errors Organized by Distractor Type

Examinee	Type			Total
	Fact-explicit	Order	Fact-implicit	
116	0	0	1	1
118	0	0	1	1
121	2	0	0	2
202	2	1	0	3
117	0	1	3	4
126	3	0	2	5
206	4	0	2	6
123	1	3	2	6
211	5	0	2	7
106	2	1	5	8
119	5	2	1	8
120	3	4	1	8
102	2	3	4	9
112	5	3	1	9
115	6	1	2	9
125	6	1	2	9
204	6	1	3	10
105	3	3	4	10
124	5	2	3	10
101	6	3	1	10
201	7	0	4	11
207	6	1	4	11
203	5	5	2	12
208	4	1	7	12
110	7	2	4	13
122	4	5	4	13
205	6	4	4	14
103	9	4	1	14
104	10	2	2	14
111	7	4	3	14
114	9	1	4	14
209	7	4	4	15
108	6	8	2	16
210	9	6	3	18
107	5	10	4	19
109	10	7	3	20
113	8	5	7	20
Total	185	98	102	385

Note:
-Participants are arranged by total exam score in descending order (100’s = Group I, 200’s = Group II)
-Fact explicit, order, and fact implicit distractor types appear on the exam in a (3:2:1) ratio

As expected, missed items in all types increase as we move down the list from the highest performing examinee to the lowest. Of the total 385 errors made by the twenty-six examinees, 185 were because of a selection of a fact explicit type distractor, 98 were because of a selection of an order type distractor, and 102 were because of a selection of a fact implicit type distractor.

Based on these numbers alone, it would appear that fact explicit type distractors were the most effective at distracting examinees and order type distractors were the least. Recall however that in the design of the exam (Table 9) there are three times as many fact explicit type distractors, and two times as many order type distractors as there are fact implicit type distractors. Therefore, we cannot consider distractor types simply as total errors, and we must adjust our data to reflect relative frequencies of the distractors as they appear on the exam. The following two statistical procedures incorporate this fact to produce adjusted findings regarding frequency of test errors by examinees in the study in relation to distractor types selected.

One statistical method that can be employed in the analysis of data organized into categories, such as test errors due to selection of certain distractor types, is the Chi-squared procedure (χ^2). Chi-squared can help us determine whether there is a relationship between distractor type and frequency of errors on the test. Table 13 shows the results of a Chi-squared procedure on the data in Table 12 carried out according to Hatch and Lazaraton (1991).

Table 13
One-way Chi-Squared Analysis of Errors Based on Distractor Type

Distractor Type	Observed frequency	Expected frequency	Obs f – Exp f	(Obs f – Exp f)^2	(Obs f – Exp f)^2 / Exp f
Fact Explicit	185	192.50	-7.50	56.25	.29
Order	98	128.33	-30.33	919.91	7.17
Fact Implicit	102	64.17	37.83	1431.11	22.30
					χ^2 = **29.76**

Note: The observed value for χ^2 is statistically significant ($p < 0.001$)

The calculated value for Chi-Squared of 29.76 is greater than the lowest alpha level for a Chi-squared procedure of this type of 13.82. This indicates that there is less than a 1 in 1000 chance that the distribution of distractor type selection observed in this study would occur by chance alone, and it can be reasonably stated that there is a relationship between distractor type and frequency of errors on the test. A closer examination of Table 13 reveals that a large contribution to the Chi-squared value comes from a larger than anticipated frequency of fact implicit type distractor selections, and a less than anticipated frequency of order type distractor selections. This would suggest that fact implicit type distractors are over-performing at causing examinee errors relative to other distractor types, and order type distractors are under-performing.

The data shown in Table 14 is another adjusted interpretation of the data in Table 15, with the data now adjusted to reflect relative frequencies of the distractor types so that we can make direct comparisons between them.

Table 14
Average "Hits" for Distractor Types Based on Participant Proficiency Level

Examinee Proficiency Level	Type			Mean
	Fact-explicit	Order	Fact-implicit	
High	4.12	2.78	9.26	5.39
Mid	9.41	5.56	17.13	10.70
Low	13.82	14.35	19.23	15.80
Mean	9.12	7.56	15.21	

The data have been compressed, with the thirty-seven examinees divided into three distinct proficiency groups: High (top 12 examinees), Middle (middle 12 examinees), and Low (bottom 13 examinees). The number in each cell is an estimation of the number of distractors of a particular type that are expected to distract an examinee of a given proficiency level, if they are exposed to one hundred distractors of that type. This variable is defined as "hits" in this study, short for "average hits per one hundred exposures." The formula used to achieve this estimation is as follows:

hits = Avg/Freq x 100

Avg = Average number of times a distractor type was selected by examinees in a proficiency level

Freq = Number of times a distractor type appeared on the test instrument

Using this formula, the hit value of 4.12 in the first cell of Table 14 means that if high proficiency examinees are exposed to one hundred fact explicit type distractors on an MDCT test, they would be expected to mistakenly select about 4 of them as their answer choice. Similarly, low proficiency examinees would be expected to mistakenly select about 14 fact explicit type distractors.

From the calculation of means in Table 14, it is now clear that fact implicit type distractors were the most effective at distracting examinees, with an average hit of 15.21 across all examinee proficiency levels. Fact explicit type distractors followed with an average hit of 9.12, with order type distractors last with an average hit of 7.56. These findings support the findings of the Chi-squared analysis in Table 13. Furthermore, with the additional consideration to examinee proficiency level, a number of new important observations can be made. It would appear from the data that the three distractor types are not discriminating equally across examinees of different proficiency levels. Only fact explicit type distractors seem to behave effectively in their increasing tendency to distract examinees based on proficiency level, and differentiate between all three levels. There is a nearly uniform 5 hit difference between each proficiency level with fact explicit type distractors. Order type distractors seem to only distract low level examinees, and do not seem to be very effective at distracting high and mid level examinees or differentiating between them. There is less than a 3 hit difference between high and mid level examinees with order type distractors, and nearly a 10 hit difference between mid and low level examinees. Fact implicit type distractors distract examinees of all levels with a relatively high rate, but are poor at differentiating between mid and low level examinees. There is a mere 2 hit difference between mid and low level examinees with fact implicit type distractors, but a 7 hit difference between mid and high level examinees.

These findings support the earlier finding that items with fact implicit type distractors had poor item discrimination with the sample population. It would now appear that while fact implicit type distractors are perhaps effective at discriminating between high and mid and high and low proficiency examinees, they are not effective at discriminating between mid and low examinees. Similarly, order type distractors are perhaps effective at discriminating between low and mid and low and high proficiency examinees, but not effective at discriminating between high and mid examinees. Only fact explicit type distractors seemed to discriminate well between all three proficiency levels. A graphical representation of these observations is shown in Figure 2. The x-axis of each graph depicts the examinees ordered from highest proficiency to lowest. The y-axis depicts the total number of mistakenly selected distractors resulting in a missed item. Each point on the graph represents the average number of distractors selected of three adjacent students. The fact explicit type distractor graph most clearly approaches a uniform line with a positive slope, a visual representation of discrimination over all proficiency levels. The order type distractor graph begins level then rises quickly, while the fact implicit type distractor graph rises somewhat then quickly levels off. Both leveled areas indicate areas of poor discrimination corresponding to the findings discussed above.

Additional Analysis Using IF and ID

Another common measurement of item difficulty is item facility (IF), a statistic equivalent to the percentage of students who correctly answer a given item. This statistic is related to a second common measurement for analyzing item discrimination, item discrimination (ID), equivalent to the difference between the IF of high performance examinees and low performance examinees. As these two statistics are very common to the field of test analysis, and serve as additional evidence for many of the findings mentioned above, Table 15 shows the ID and IF values for each item on the test instrument.

Figure 2
Total "Hits" Organized by Distractor Type Plotted in Decreasing Examinee Proficiency

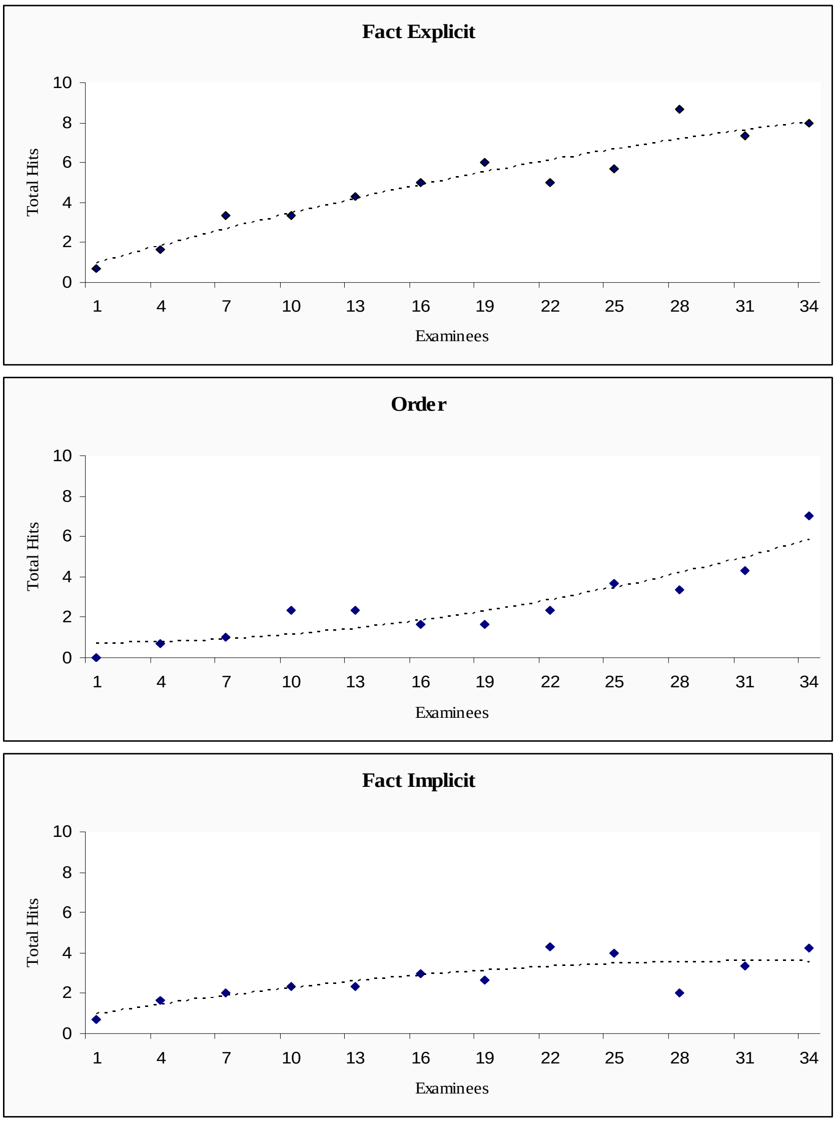

Table 15
IF *and* ID *Values*

Item #	Item Facility (*IF*)	Item Discrimination (*ID*)
1	.24	.19
2	.65	.46
3	.92	.23
5	.97	.08
7	.87	.31
8	.51	-.04
9	.43	.35
10	.78	.30
11	.95	.15
13	.51	.84
14	.81	.39
15	.78	.46
16	.78	.23
17	.41	.35
18	.41	.84
19	.49	.36
20	.87	.23
21	.89	.23
22	.97	.08
24	.87	.07
26	.89	-.01
27	.70	.22
28	.65	.46
29	.46	.02
30	.62	.61
31	.62	.69
32	.97	.08
33	.65	.38
35	.62	.53
36	.76	.46
37	.32	.51
38	.54	-.12
39	.70	.53
40	.92	.23
41	.95	.08
42	1.00	.00

RESULTS: PART II

Further Investigation of Fact Implicit Type Distractors

Analysis thus far has indicated that MDCT items with a fact implicit type distractor as one answer choice has a low item discrimination, a result of middle and low level examinees being equally distracted by them with regular frequency. An early hypothesis of this study was that fact implicit type distractors would present a higher than average challenge to JEFL learners, which seems to be supported by their generally high overall effectiveness at distracting examinees in the study. One problem with this conclusion is that the presence of a fact implicit type distractor did not seem to have an measureable effect on overall item difficulty based on the CTT and FACETS portion of the analysis. If fact implicit type distractors were truly more challenging, we would expect higher numbers of examinees to incorrectly select them, leading to lower overall performance and higher item difficulty.

Briefly returning to Figure 2 may shed some light on this issue. The "flatness" of the fact implicit type graph was an indicator of low item discrimination. However, turning attention to the maximum y-value of the graph it is noticeably lower in comparison to the other two graphs. An interpretation of this finding is that while examinees of all proficiency levels were consistently missing relatively similar numbers of items due to mistakenly selecting fact implicit type distractors, overall this was not happening with very high frequency. What may be happening is that a handful of fact implicit type type distractors turned out to be very effective at distracting examinees of all proficiency levels, but in general most were not. In order to investigate this claim, the next section of this paper will take a holistic analytical approach to a number of key items on the test.

Holistic Item Investigation

No item format analysis should be considered complete without dedicating some attention to looking at the items themselves in their complete form. This section of the analysis is devoted to a holistic investigation of several key items on the test instrument, with particular focus to those containing fact implicit type distractors. Four items are investigated in total, all of which contain fact implicit type distractors.

Item #1:
Examinee hears:
A: Hello! Are you Emiko?
B: Oh, yes I am.
A: Welcome to Canada! I'll be your host mother, my name is Beth.

Examinee reads:
(1) Is this your first visit to Canada?
(2) Hi! I'm so glad to be in Japan.
(3) Ok. Take me somewhere to eat.
(4) Hi Beth. I've missed you very much.

The analysis will begin with Item #1, the most difficult item on the test. The item is shown above as it appears on the test form used in the study. Item #1 is a Type C item, meaning it contains a fact implicit type distractor and the answer key is written using direct language strategies (answer choice (3)). The item was missed by 28 of the examinees, with only 9 answering the item correctly (IF = 0.24). Over 60% of examinees who missed the item mistakenly selected answer choice (4) as their answer, the fact implicit type distractor. The remainder choose answer choice (2) the fact explicit type distractor, or answer choice (1) the order type distractor with about equal frequency. One conclusion to be drawn from this item is that the JEFL learners participating in this study had difficulty recognizing that in a first encounter scenario it is not appropriate to express an emotion of missing the interlocutor. Furthermore, as the interlocutor is playing the role of a host parent, a higher social status position, the examinees may not have been comfortable selecting the direct answer choice (3) despite that it is the only correct answer in this situation.

Item #9:

Examinee hears:
A: Hello, are you Professor Hill? I'll be in your literature class this semester.
B: Oh, I see. I'm looking forwards to seeing you in class then.

Examinee reads:
(1) Oh. Are you a new student?
(2) Thank you. I'll do my best.
(3) I'm not good at math, but I'll do my best.
(4) Me too. I'm glad to be your student again.

Item #9 is a Type A item, meaning it contains a fact implicit type distractor and the answer key is written using indirect and softening language strategies (answer choice (2)). The item was missed by 21 of the examinees, with only 16 answering the item correctly (IF = 0.44). Over 60% of the examinees who missed the answer mistakenly selected answer choice (4), the fact implicit type distractor. A majority of the remainder selected answer choice (3) the fact explicit type distractor, and a very small percentage selected answer choice (1) the order type distractor. One conclusion that might be drawn from this item is that JEFL learners participating in this study had difficulties recognizing that as this is a first encounter scenario, answer choice (4) is impossible. Combined with examinee behavior on Item #1, this might be an indication that JEFL learners have difficulty with either identifying first encounter scenarios, or identifying appropriate language in such situations.

Item #20:
Examinee hears:
A: Excuse me. May I borrow that dictionary for a moment?
B: Oh, I'm sorry. It's not mine.

Examinee reads:
(1) Oh, it's yours?
(2) Thanks. I'll return it soon.
(3) Oh, it belongs to my professor.
(4) I see. Sorry to bother you.

Item #20 is a Type A item, meaning it contains a fact implicit type distractor and the answer key is written using indirect and softening language strategies (answer choice (4)). The item was missed by only 5 of the examinees, with 32 answering the item correctly (IF = 0.87). Of those examinees who missed the item, 2 selected answer choice (2) the fact implicit type distractor, and 3 selected answer choice (3) the order type distractor. Given the failure of the fact implicit type distractor to distract more than 2 of the examinees, the great majority of JEFL learners in this study were able to identify that apologizing and saying that an item does not belong to them implicitly communicates that they cannot loan it away.

Item #32:
Examinee hears:
A: Pardon me, how many stops is it until Higashi Station?
B: Oh, four stops I think. But I'm not so sure.

(1) Examinee reads:
(2) Sorry I can't help much.
(3) Ok, thanks anyway.
(4) But I'm sure it's only three stops.
(5) Five stops? Ok, thanks you.

Item #32 is a Type A item, meaning it contains a fact implicit type distractor and the answer key is written using indirect and softening language strategies (answer choice (2)). The item was missed by only 1 of the examinees, with 35 answering the item correctly (IF = 0.97). The examinee who missed the item selected answer choice (1) the order type distractor as an answer choice. No examinees selected answer choice (3) the fact implicit type distractor or answer choice (4), the fact explicit type distractor. Given the failure of the fact implicit type distractor to distract any of the examinees, the great majority of JEFL learners in this study were able to identify that as "A" initiates a fact finding inquiry in the dialogue, which implicitly means they are not in possession of the knowledge of interest (number of stops to Higashi station) and therefore cannot make the statement in answer choice (3).

In two of the items analyzed above, the fact implicit type distractor distracted a large number of examinees. On the other hand, in the other two items the fact implicit type distractor distracted very few or none of the examinees. In general, this behavior was the case throughout the test where about 4 of a total of 18 created fact implicit type distractors distracted large percentages of the examinees, and the remaining distractors were relatively ineffective. This finding explains both the poor item discrimination and lack of effect on item difficulty of fact implicit type distractors observed in this study. Both highly effective and largely ineffective fact implicit type distractors contributed to poor item discrimination, but there were simply not enough effective distractors to have a measurable effect on item difficulty in this particular study.

CONCLUSIONS

This study set out to investigate the MDCT item in the JEFL assessment context. From the onset, two research focuses were established: (a) the construct validity of MDCT items in listening proficiency assessment, and (b) MDCT item discrimination behavior as affected by distractor type. In the final section of the paper I will summarize each finding of the study, contextualize each finding within the relevant body of research in other studies, comment on the issues raised by each finding for MDCT testing, and finally make some recommendations for addressing those issues and areas of future research. The paper will end with a discussion of the limitations of the current study.

Findings

Statistical analysis demonstrated that MDCT items written without the *indirectness factor*, or items with answer keys written in a direct or aggressive style, were significantly more difficult for the sample of JEFL examinees to answer correctly. It was demonstrated that under a certain condition of marked directness in how an answer key was written, some JEFL learners could be induced to reject it due to pragmatic features, regardless of the fact that it represents the only correct answer in a given language context. This was not an unexpected finding, as a number of previous studies (Takahashi, 1987; Beebe et al, 1990; Rose, 1994; Rose et al, 1995) indicate that the JEFL learners might transfer a norm for marked indirectness in their L1 onto how they perceive the acceptability of language in their L2. If examines are indeed using pragmatic cues in addition to linguistic cues in the answering of MDCT items, the construct validity of MDCT tests for listening proficiency measurement is questionable. Furthermore, given issues of quality and sufficiency of English pragmatic language education observed in Japan (Shimizu et al, 2007), there is a question of whether MDCT items test skills or knowledge that examinees do not have access to, an implication for the ethical use of these items in the JEFL context.

Maximizing construct validity of MDCTs for listening assessment in the JEFL context is a matter of reducing the likelihood that examinees would use pragmatic cues to answer test items. To achieve this, it might be worthwhile to implement measures into the design and writing of MDCT items to align them more specifically to language styles used by the target examinee population, JEFL learners. Employing JEFL speakers as item writers over native English speakers combined with piloting of MDCT items prior to operational testing are two viable options.

The *implicature factor*, or inclusion of an implicit type distractor into an MDCT item, did not seem to significantly increase the difficulty of test items for JEFL learners. This finding was in contrast to an early hypothesis of this study, which predicted that implicit type distractors would present a higher than average challenge to JEFL learners to reject as possible answer choices, leading to more missed questions on MDCT items that contain them. As will be discussed below, a more in depth investigation of how different distractor types were functioning on the exam provided a possible explanation for this finding. Table 16 is a summary of the findings of this study concerning the *indirectness factor* and *implicature factor*.

Table 16
Summary of Findings for the Two Factors

Factor Type	Item Difficulty	Item Descrimination (r_{pbi})
Directness Factor	Makes items more difficult	Increases discrimination
Implicature Factor	No measurable effect	Decreases discrimination

An investigation of distractor type on the exam indicates that distractor types displayed unique discrimination behaviors when administered to JEFL examinees. Based on the sample of JEFL learners participating in this study, the optimal distractor type for discrimination of JEFL learners are fact explicit type distractors. The data suggests there are serious issues concerning the effectiveness of order type distractors and fact implicit type distractors. Fact explicit type distractors showed good discrimination between all proficiency levels. However, order type distractors only discriminated very low proficiency levels and might not pose a high enough challenge for university level JEFL learners or even high school learners. They seem to be rejected easily be learners of most ability levels, and only appear to effectively distract learners of very low ability level in the study. Fact implicit type distractors performed unpredictably, in some instances distracting high numbers of examines and in others distracting very few. Futhermore, fact implicit type distractors are the most difficult to write. A holistic investigation of individual items on the test provided evidence suggesting that while JEFL learners are poor at demonstrating understanding of implied information to identify first encounter language situations or rule out distractors in those situations, JEFL learners were able to demonstrate understanding of implied meaning in a request situation, as well as a knowledge sharing situation. This brings into question whether it is possible to write high quality fact implicit type distractors given limitation in our current understandings of how JEFL learners comprehend implied meaning in English, and how and where they acquire these skills.

Further research will have do be done on this distractor class in particular to better understand exactly what JEFL learners can do in terms of comprehension of implied meaning, and how this knowledge can be utilized in the design of reliable and high performing fact implicit type distractors. These are particularly interesting findings, as this is the first known study to indentify and empirically investigate MDCT item distractors as distinct types. Based on this study, a lack of consideration of distractor type on MDCT item design as it relates to target population might have negative effects on the reliability and quality of accurate discrimination of examinees. Table 17 is a summary of the findings of this study with regard to the three distractor types and their discrimination behaviors.

Table 17
Summary of Findings for Distractor Types

Type	Ease of writing	Discrimination
Fact Explicit Type	Very Easy	Overall good discrimination between students of low, mid, and high proficiency.
Order Type	Very Easy	Do not present a viable answer choice to examinees of most proficiency levels except very low ability. In general, only discriminate between the lowest level students and those of mid or high proficiency. Does not discriminate between students of mid or high proficiency.
Fact Implicit Type	Difficult	Depends highly on the particular information which is implicitly encoded. Will often present too high a challenge to all examinees, or too low a challenge to all examinees to provide useful discrimination information.

A more focused study centered on how to deploy MDCT distractor types to achieve maximum examinee discrimination is necessary before solid recommendations can be made about how and in what proportions each type should be utilized in operational testing. At this point however, several tentative recommendations could be attempted. Exclusive use of fact explicit type distractors might increase overall discrimination of MDCT tests, as this distractor type was the only one in this study to discriminate effectively. One complication to this is the indication that in some cases the adding of additional fact explicit type distractors to a single MDCT item results in markedly reduced effectiveness at distracting examinees. In other words, it may not be possible to design MDCT items of the format employed in this study with multiple viable fact explicit type distractors. More research will have do be done on this topic in particular, specifically how MDCT item format might be altered to allow for increased use of fact explicit type distractors and reduced dependency on the other distractor types.

Limitations of the Study

This study is the first attempt at what will likely need to be a prolonged and detailed investigation of the MDCT test item format in the JEFL context. As such, there a are number of limitations regarding the current study that will need to be addressed in further research into this topic.

First and foremost, the sample size of thirty-seven JEFL examinees participating in this study is too limited to reliably extend trends observed in the findings as applicable to JEFL learners as a population. The small sample size is a result of limitations in access to research participants at the institution at which this research was conducted. Due to the inclusion of FACETS analysis in the interpretation of data in this study, it will be especially critical that future studies employ significantly larger sample sizes that are more conductive to this type of statistical procedure.

A second limitation to the current study is that university JEFL learners were chosen as participants over high school JEFL learners, the most likely target population of MDCT type assessments given the specificity of the *Center Test* for high school learners. University learners were selected as participants in this particular study because of the relative ease at which they could be recruited compared to high school learners given the institution at which this research was conducted, and the researchers' limited status. Future studies should focus specifically on high school JEFL learners as well as other populations of learners as participants, in order to investigate whether the findings indicated in the results of this study are applicable to populations outside of JEFL university learners.

REFERENCES

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Beebe, L., Takahashi, T., & Uliss-Weltz, R. (1990). Pragmatic transfer in ESL refusals. In R. Scarcella, E. Andersen, & S. Krashen (Eds.), *Developing communicative competence in a second language* (pp. 55–73). New York: Newbury House.

Blum-Kulka, S., & Olshtain, E. (1984). Requests and apologies: A cross-cultural study of speech act realization patterns (CCSARP). *Applied Linguistics, 5(3),* 196–213.

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. (2005). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Brown, J.D. (2001). Pragmatics tests: Different purposes, different tests. In K. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 301–325). Cambridge, UK: Cambridge University Press.

Brown, J.D., & Yamashita, S.O. (1995). English language entrance examinations at Japanese universities: What do we know about them? *JALT Journal, 17(1),* 7–30.

Center Test. (2006). Teacher committee evaluation of the *Center Test* in English listening. Retrieved October 5, 2008 from http://www.dnc.ac.jp/old_data/exam_repo/18/pdf/18hyouka63.pdf

Cohen, A. & Olshtain, E. (1981). Developing a measure of sociocultural competence: the case of apology. *Language Learning*, 31(1), 113-134.

Fukushima, S. (1990). Offers and requests: Performance by Japanese learners of English. *World Englishes*, 9, 317-325.

Hatch, E. & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. New York: Newbury House Publishers.

Hill, T. (1997). Pragmatic development in Japanese learners: A study of requestive directness level. Unpublished doctoral dissertation, Temple University Japan, Tokyo.

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

Hudson, T. (2001). Indicators for cross-cultural pragmatic instruction: Some quantitative tools. In K. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 283–300). Cambridge, UK: Cambridge University Press.

Ingulsrud, J.E. (1994). An entrance test to Japanese universities: Social and historical context. In C. Hill & K. Parry (Eds.), *From testing to assessment: English as an international language (pp. 61–81).* New York: Longman.

Ito, A. (2005). A validation study on the English language test in a Japanese nationwide university entrance examination. *Asian EFL Journal, 7(2),* Article 6.

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

Kubota, M. (1995). Teachability of conversational implicature to Japanese EFL learners. *IRLT Bulletin, 9,* 39-65.

Levenston, E.A. (1975). Aspects of testing the oral proficiency of adult immigrants to Canada. In L. Palmer & B. Spolsky (Eds.), *Papers on language testing 1967–1974 (pp. 67–74).* Washington, DC: TESOL.

Maebashi, N., Yoshinaga, N., Kasper, G., & Ross, S. (1996). Transfer and proficiency in interlanguage apologizing. In S. Gass & J. Neu (Eds.), *Speech acts across cultures* (pp. 155–187). Berlin: Mouton de Gruyter.

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

Ministry of Education, Culture, Sports, Science, and Technology. (2003). Regarding the establishment of an action plan to cultivate “Japanese with English abilities.” Retrieved June 1, 2008, from http://www.mext.go.jp/english/topics/03072801.htm

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

Norris, J.M. (2001). Use of address terms on the German Speaking Test. In K. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 248–282). Cambridge, UK: Cambridge University Press.

Roever, C. (2006). Validation of a web-based test of ESL pragmalinguistics. *Language Testing, 23(2)*, 229–256.

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

Rose, K. (1992). Speech acts and questionnaires: The effect of hearer response. *Journal of Pragmatics,* 17, 49-62.

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, Hawai‘i

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

Taguchi, N. (2002). An application of relevance theory to the analysis of L2 interpretation processes: The comprehension of indirect replies. *International Review of Applied Linguistics*, 40, 151–176.

Taguchi, N. (2005). Comprehending implied meaning in English as a second language. *Modern Language Journal*, 89, 543–562.

Takahashi, S. (1996). Pragmatic transferability. *Studies in Second Language Acquisition, 18,* 189–223.

Takahashi, S., & Roitblat, H. L. (1994). Comprehension process of second language indirect requests. *Applied Psycholinguistics, 15,* 475–506.

Takahashi, T., & Beebe, L. (1987). The development of pragmatic competence by Japanese learners of English. *JALT Journal, 8,* 131–155.

Yamashita, S.O. (1996). Six measures of JSL pragmatics. Honolulu, HI: 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.

APPENDIX A: CATEGORIZATIONS OF TEST ITEMS

	Description of Item Prompt	Situational Variable		Class
		P	D	
	Responding to an Introduction from a host mother	-	+	C
	Responding to an indirect request from a friend/family member	=	-	A
	Responding to acceptance of request by a professor	+	-	B
	Responding to an inquiry from a stranger	+	+	LI
	Responding to decline of an inquiry from a stranger	=	+	A
	Responding to a request by a boss	-	-	LI
	Responding to an inquiry by a doctor	-	+	B
	Responding to a student saying they are tired	+	-	C
	Responding to encouraging comments from a professor	-	+	A
0	Responding to an inquiry for a good place to eat	=	+	D
1	Responding to a request from a student	+	-	C
2	Responding to news from a friend	=	-	LI
3	Responding to a request from a professor	-	-	D
4	Responding to a decline of an inquiry by a stranger	=	+	B
5	Responding to help from a store clerk	+	+	B
6	Responding to a complaint from a customer	-	+	A
7	Responding to a request from a boss	-	-	D
8	Responding to an offer from an airline clerk	+	+	C
9	Responding to a high price of a textbook	+	+	D
0	Responding to a decline of a request by a stranger	=	+	A
1	Responding to an inquiry from a professor	-	-	C
2	Responding to a request from a student	+	-	A
3	Responding to information about a movie from a friend	=	-	LI
4	Responding to a request by a custome	-	+	B
5	Responding to information about transportation to school by a friend	=	-	LI
6	Responding to a decline of a request by a student	+	-	A
7	Responding to information about a missed meeting by a friend	=	-	C
8	Responding to information from a store clerk	=	-	C
9	Responding to information from a student	=	+	C
0	Responding to an inquiry from a customer	-	+	D
1	Responding to an inquiry from a professor	-	-	B
2	Responding to information from a stranger	=	+	A
3	Responding to a compliment from a student	+	-	D
4	Responding to negative information from a student	=	-	LI
5	Responding to information from a host-mother	-	-	B
6	Responding to a compliment from a professor	-	-	D
7	Responding to negative information from a bus clerk	+	+	D
8	Responding to negative information from a store clerk	+	+	A
9	Responding to a compliment from a customer	-	+	D
0	Responding to an inquiry from a stranger	=	+	C
1	Responding to negative information from a student	+	-	B
2	Responding to an inquiry from a store clerk	+	+	B

Note: LI = Linking Item

APPENDIX B: TEST INSTRUMENT MASTER COPY

Notation:
F = Fact explicit distractor
I = Fact implicit distractor
O = Order distractor
* = Answer key
Note: Linking items have been omitted from this copy of the test instrument

Question #1
A: Hello! Are you Emiko?
B: Oh, yes I am.
A: Welcome to Canada! I'll be your host mother, my name is Beth.
B: ________________________________

F Hi! I'm so glad to be in Japan.
I Hi Beth. I've missed you very much.
O Is this your first visit to Canada?
* Ok. I'm very hungry.

Question #2
A: How about going to the Chinese restaurant for dinner?
B: Let's try a different place tonight.
A: Why? I thought it was your favorite place?

F Yes, but I really want to each Chinese food.
I Yes, but I don't like Chinese food
O That sounds good. Let's try it.
* Yes, but they've raised their prices.

Question #3
A: Jason, can you deliver this message to Professor Brown for me?
B: Sure Professor West, I'll do it right now.
A: ________________________________

F I see. I can do it myself then.
F Great. Thanks for the message.
O You're welcome. I'm happy to help.
* Thanks. I appreciate your help.

Question #5
A: Excuse me. Do you happen to know where the library is?
B: Sorry, I'm not from around here.
A: _______________________________

F Oh, you're from here?
I I see. Well, could you show me the way then?
O No problem. Sorry I couldn't help you.
* Ok, thanks anyway.

Question #7
A: Doctor, I've had a sore back since I woke up this morning.
B: OK. Tell me if that hurts.
A: Ouch! That's quite painful.
B: How about here?
A: ________________________________

F That was worse yesterday.
F It's been sore for a week.
O Ok, let me give you some medicine.
* Ouch! That hurts as well.

Question #8
A: Hello Mark. How are you?
B: Hi Professor Tom. I had 6 classes and baseball practice today, so I'm pretty tired.
A: __________________________________

F I see. Is soccer practice hard?
I I'm sorry. Better luck next time.
O I'll be fine, thanks for asking.
* I see. Goodbye.

Question #9
A: Hello, are you Professor Hill? I'll be in your literature class this semester.
B: Oh, I see. I'm looking forwards to seeing you in class then.
A: __________________________________

F Thanks. But, I'm not so good at math.
I Me too. I'm glad to be your student again.
O Oh. Are you a new student?
* Thank you. I'll do my best.

Question #10
A: I wonder if you could help me.
B: Yes?
A: Where's the best place to get something to eat around here?
B: __________________________________

F Oh, the bank is near the train station.
F Well, my favorite food is pasta.
O That sounds like a great place to eat. Thanks.
* I don't know.

Question #11
A: Brian, did you hand in your essay paper to me yet?
B: I'm sorry Professor James. I'm still working on it. Could I give it to you tomorrow?
A: __________________________________

F No, I'm sorry. I need it by next week.
I Oh, was the assignment too easy?
O Thanks. I won't forget.
* No, give it to me today.

Question #13
A: Excuse me Professor Brown. I wanted to ask you about our homework assignment.
B: Sure Beth. Can you come to my office tomorrow?
A: __________________________________

F Yes, but I'm not sure where the library is.
F Sorry. I'm a little busy the day after tomorrow.
O 3 o'clock is fine Beth. See you then.
* No. Let's meet the day after tommorow.

Question #14
A: Pardon me. Do you know what time it is?
B: I'm sorry, I don't know.
A: ________________________________

F Well, what about last night?
F Never mind. I can tell someone else.
O Sure, it's 11:30 right now.
* That’s Ok. Thanks anyway.

Question #15
A: Excuse me. Can I help you?
B: Yes, I’d like to buy a sweater, but I only have $40 to spend.
A: Well, this sweater is within your budget.
B: ________________________________

F Oh, but this sweater is too expensive for me.
F Well, I don’t like that shirt color though.
O Would you like to try it on?
* Oh, can I try it on?

Question #16
A: Excuse me. I’d like to buy this jacket. How much is it?
B: That jacket? The price is $250.
A: Oh my! That’s more expensive than I thought.
B: ________________________________

F Yes, that’s because it’s on sale.
I I'm glad you like the quality.
O Do you have one a little cheaper?
* Would you like to see a cheaper one?

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: ________________________________

F Thanks. I could use a rest.
F Yes, here is the message from downstairs.
O Great job. Thanks for all the help.
* I'm taking a short rest first.

Question #18
A: Hello, I'd like to check in. I'm flying to Osaka.
B: OK. Are you interested in paying an extra 200 dollars for a first class seat?
A: ________________________________

F No, first class is fine. Thank you.
I OK. I can't miss this flight so I have no choice.
O Here’s your ticket. Thank you.
* No I’m not.

Question #19
A: Hi. I’d like to buy the textbook for English 1A.
B: OK, here it is. The price is 100 dollars.

A: ________________________________

F I’m sorry. I only have $150.
F Oh! Why is the notebook so expensive?
O Yes, you can pay with credit card.
* Give me a student discount.

Question #20
A: Excuse me. May I borrow that dictionary for a moment?
B: Oh, I’m sorry. It’s not mine.
A: ________________________________

F Oh, it’s yours?
I Thanks. I’ll return it soon.
O Oh, it belongs to my professor.
* I see. Sorry to bother you.

Question #21
A: Sorry to keep you waiting Professor James. I’m ready for the meeting now.
B: We were worried about you. What happened?
A: ________________________________

F Did something happen to Professor James?
I I’m sorry. Am I too early?
O Oh my! Are you OK now?
* Nothing. Let’s start the meeting.

Question #22
A: Ann, what's wrong? You look pale.
B: Oh, Professor Brown. I have a cold today. Is it OK if I go home early?
A: ________________________________

F Yes, I should go home right now.
I OK, but try harder next time.
O OK, I will. Thank you.
* Sure, take care of yourself.

Question #24
A: Hi, can I take your order?
B: Yes. First I'd like a cup of tea.
A: ________________________________

F OK. Anything to drink?
F Anything else besides coffee?
O May I have cream for my tea?
* Ok. Would you like anything else?

Question #26
A: David, can you help me move the desks after class?
B: I'd be happy to Professor Thomas, but I’m afraid I have a dentist appointment.
A: ________________________________

F Ok, I hope you feel better soon.
I Thanks so much. Let’s finish moving before then.
O Sorry, I’ll help you next time.
* That’s Ok. I’ll ask someone else.

Question #27
A: I didn’t see you at the meeting last night.
B: Meeting? Oh, I totally forgot about it.
A: ______________________________

F Who cancelled the meeting?
I Oh, you never know about it?
O I’m sorry. I won't forget next time.
* You forgot! What’s wrong with you?

Question #28
A: Well Bob, how much is it going to cost to fix this old car of mine?
B: Hmm, since the engine needs major repairs, it’ll be at least 3000 dollars.
A: ______________________________

F Ok, I think I’ll buy the car.
I Great! That sounds very affordable.
O Ok, you can pay me next week.
* That’s too expensive. Goodbye.

Question #29
A: Hi, my name is Miki. Are you interested in joining our tennis club?
B: Well, I'm not very good at tennis.
A: ______________________________

F Great! We need more good tennis players.
I I see. Too bad you don't like tennis.
O Really? Can I join the club then?
* We don't need you then.

Question #30
A: Excuse me. Can I try this sweater on?
B: Yes, of course.
A: Oh no. This one is quite large. Do you have one in medium?
B: ______________________________

F Sorry, we don’t have a larger size.
F Would you like to see another color?
O Could I see another sweater?
* Uh, no.

Question #31
A: Enjoy your winter vacation Amy.
B: Thanks Professor Hudson. Actually, I’m going to go skiing at White mountain with some friends.
A: Really? Do you think there will be enough snow?
B: ______________________________

F Yes, it’s never crowded.
F Yes, my friends are all good skiers.
O Sounds fun. Have a good time.
* Yes, they got a lot a few days ago.

Question #32
A: Pardon me, how many stops is it until Higashi Station?
B: Oh, four stops I think. But I’m not so sure.

A: _________________________________

F Five stops? Ok, thank you.
I Are you sure? I'm positive it’s three stops.
O Sorry I can’t help much.
* Ok, thanks anyway.

Question #33
A: Oh, hello Professor Lynn. Did you get a haircut?
B: Does my hair look different?
A: Yes, it looks great.
B: _________________________________

F I don't like it very much either.
F Thank you. I bought it last week.
O Really? Was it expensive?
* I know. My hair is prettier than all the other teachers.

Question #35
A: Hi Mrs. Brown. My flight arrives in Hawaii tomorrow afternoon.
B: OK Takeshi. We’ll pick you up at the airport. Have a safe flight.
A: ________________________________

F Thanks! See you next week.
F Thanks! I had a really nice time in Hawaii.
O Can you tell me your flight number?
* Thank you. See you at the airport.

Question #36
A: Hello Professor Paul. Today’s class was very interesting.
B: Hi Ann. I’m glad you liked it. You're one of my best students.
A: _________________________________

F I’m sorry. I’ll try harder next class.
F Thanks. I promise to stay awake next time.
O It’s true. You really are a hardworking student.
* Yes. All my professors tell me that.

Question #37
A: I'd like two bus tickets please.
B: I'm very sorry. Our bus is full today and we can’t sell more tickets.
A: _________________________________

F OK. One ticket then please.
F These are the last two tickets? Wow, so lucky.
O Would you like to buy a ticket for tomorrow?
* Give me a break! I need two tickets.

Question #38
A: I really love this dress. Does it come in red?
B: Oh, I’m sorry miss. It only comes in green.
A: _________________________________

F Oh, I see. I'll just buy the red dress then.
I Oh, green is my favorite color.
O Would you like to try it on anyways?

* Oh, I see. What a pity.

Question #39
A: Here is your coffee. Can I get you anything else?
B: No, that’s OK. The meal was very delicious. Thank you.
A: _________________________________

F Would you like to order your meal now?
F Oh, what was the problem?
O I’ll be sure to come again soon.
* I know it was. Here's the bill.

Question #40
A: Excuse me. Where’s the nearest bank?
B: It’s next to the library. Do you know where that is?
A: _________________________________

F Yes, its 3:15.
I Oh, you don't know? Thanks anyways.
O Shall I show you the way?
* No. Take me there.

Question #41
A: Thank you Mark. Your class presentation was very interesting today.
B: But Professor Adams, I was so nervous. I hope I didn't make too many mistakes.
A: _________________________________

F Oh, I’m glad you weren’t nervous.
F Oh, was the exam too hard?
O Thanks. I’ll try harder next time.
* Don’t worry. You did very well.

Question #42
A: Good afternoon. Can I help you?
B: I hope so. I’m looking for a nice present for my mother.
A: How much did you want to spend?
B: _________________________________

F The price is $50.
F I’d like to buy at least two gifts.
O I see. Is that all you can spend?
* About $20 at most.