Reconsidering the Measurement of Pragmatic Knowledge Using a Reciprocal Written Task Format

Kirby Cook Grabowski[1]
Teachers College, Columbia University

ABSTRACT

Current influential models of communicative language ability (Bachman & Palmer, 1996; Purpura, 2004) represent language knowledge as a set of separate yet related components, including grammatical and pragmatic dimensions, which interact with background knowledge, metacognitive strategies, and contextual features (Chapelle, 1998) in language use. Although some researchers have attempted to measure aspects of pragmatic knowledge, the vast majority have not incorporated a clearly articulated pragmatic component into the test construct. The purpose of the current study is to investigate the extent to which scores from a test designed for the current study can be interpreted as indicators of test-takers' grammatical and pragmatic ability. This study attempts to address some of the limitations of prior pragmatics assessment research, namely, the issues of construct underrepresentation, the lack of highly contextually constrained reciprocal tasks, and the use of less sophisticated statistical tools to support claims of validity. This paper examines the construct validity of a test based on Purpura's (2004) theoretical model, which specifies the grammatical and pragmatic (sociolinguistic, sociocultural, and psychological) components of language knowledge. This model accounts for the way in which grammatical resources are used to convey a range of pragmatic meanings in language use. The analysis explored a broader range of features of pragmatic knowledge than had previously been investigated.

INTRODUCTION

Since language tests can be door-openers or gate-keepers (Bachman & Purpura, in press), the constructs underlying language tests should reflect theoretical notions of what it means to know a language. Drawing on the work of Canale and Swain (1980), Bachman and Palmer's (1996) and Purpura's (2004) models of communicative language ability (CLA) represent language knowledge as a set of separate yet related components, including grammatical and pragmatic dimensions, that interact with background knowledge, strategic competence, and contextual features (Chapelle, 1998) in language use. As such, many applied linguists have argued that pragmatic knowledge is a critical dimension of CLA, one that is integral to a learner's underlying

[1] Kirby Cook Grabowski is a doctoral student in Applied Linguistics at Teachers College, Columbia University. Her research interests include the assessment of pragmatic knowledge, language program evaluation, and language study abroad. Correspondence should be sent to Kirby Cook Grabowski, 525 W 120th Street, Box 66, New York City, NY 10027. E-mail: kjc33@columbia.edu.

ability. Still, the vast majority of language testers have opted to operationalize other aspects of CLA in their assessments, often to the exclusion of pragmatic knowledge.

Even the well-regarded, high-stakes English language assessments have limitations with respect to the representation of pragmatic knowledge in their constructs, even though pragmatic meanings are often elicited in the test tasks. For instance, the Internet-based Test of English as a Foreign Language (TOEFL iBT) and the International English Language Testing System (IELTS) do not explicitly represent pragmatic dimensions in their speaking rubrics. By contrast, the Test of Spoken English (TSE) does include pragmatic dimensions (e.g., sociolinguistic competence and functional competence) as performance indicators in addition to linguistic and discourse dimensions in the scoring rubric; however, like the speaking sections for the TOEFL iBT and the IELTS, it uses a holistic rating scale. Thus, while reporting a global score may be the most practical approach for large-scale assessments, it is nearly impossible to determine the extent to which pragmatic knowledge contributes to a test-taker's overall score. Therefore, the inferences made about the test-takers cannot be extended to the role of pragmatic knowledge in a broader model of CLA from these tests. While it is true that the question of whether or not to test pragmatic ability is widely debated, usually based on issues of test fairness and the host of difficulties associated with assessing it (McNamara & Roever, 2006), these issues do not negate the importance of pragmatic knowledge in CLA: If the claim is that the purpose of the test is to measure a learner's overall language proficiency, a clearly articulated pragmatic knowledge component should be part of the test construct.

With that said, since the early 90s, a growing number of promising empirical studies have investigated various pragmatic dimensions in assessment contexts, albeit on a much smaller scale than in the tests mentioned above. The most influential research thus far has involved a battery of language tests created by Hudson, Detmer, and Brown (1992, 1995) and those following their framework (Enochs & Yoshitake-Strain, 1996, 1999; Liu, 2006; Yamashita, 1996; Yoshitake-Strain, 1997). Hudson et al. (1992, 1995) outlined their test development process and identified the variables to be measured. They used three types of test tasks (e.g., indirect, semi-direct, and self-assessment measures) that involved situations in which power, social distance, and absolute ranking of imposition were systematically varied. To assess appropriateness, they elicited the production of three commonly researched speech acts (e.g., requests, refusals, and apologies) as an indication of cross-cultural pragmatic ability in Japanese non-native English speakers. Although Hudson et al.'s (1992, 1995) research is compelling in that it described a rigorous test development process, the model of pragmatic knowledge upon which their test tasks were based was limited. More specifically, pragmatic knowledge was operationalized only in terms of power, distance, and absolute ranking of imposition, generally incorporated under the rubric of sociolinguistic knowledge (Purpura, 2004). In addition, the tests scored only for the test-taker's productive capacity with respect to the three speech acts mentioned above, tapping into conveyance of implied meanings while disregarding the comprehension or interpretation of implied meanings – an essential aspect in pragmatic knowledge.

Even though other researchers have attempted, with varying degrees of success, to measure other aspects of pragmatic knowledge, namely, the production of sociocultural aspects of speaking (Cohen & Olshtain, 1981) and the comprehension of implied meaning (Bouton, 1988, 1994; Roever, 2006; Taguchi, 2005), the vast majority of research in this field has followed Hudson et al.'s (1995) framework (Enochs & Yoshitake-Strain, 1996, 1999; Liu, 2006; Yamashita, 1996; Yoshitake-Strain, 1997), and therefore, carries similar limitations. As such, research in pragmatics testing has left language testers unclear as to how second language (L2) learners can use language to communicate a number of psychological stances (e.g., sarcasm, positive and negative affect; Purpura, 2004) or to convey sociocultural meanings (Cohen, 1995; Cohen & Olshtain, 1981). Language testing is also unclear as to how different elicitation methods affect pragmatic performance or how pragmatic meanings conveyed by L2 speakers might interact with varying degrees of context. In fact, most of the research in this area has defined context very narrowly, usually only in terms of what are deemed to be fixed, sociolinguistic features (i.e., power, distance, and absolute ranking of imposition), which according to Brown and Levinson (1987) and others (Coulter, 1989; Jefferson, 1989; Schegloff, 1987), may actually be more fluid and co-constructed in real-life contexts. As such, pragmatics assessment research has not yet addressed a widely held belief that language and social and/or linguistic context have a mutually reciprocal and compounding effect (for a complete discussion, see Duranti & Goodwin, 1992). In this view, language can shape context in the same way that context can shape language during an interaction (Heritage, 1984; Johnson, 2001). Therefore, in order for pragmatic meanings to be properly conveyed and interpreted by test-takers, and systematically scored by raters, constraints in the form of rich contextual features must be given (Purpura, 2004). These constraints will help generate responses that are not only sociolinguistically, socioculturally, psychologically, and rhetorically appropriate, but also expected and assumed within the broader, systematic, and sociocultural context that is real-life communication.

Although prior empirical research in assessing pragmatics has laid the groundwork for future inquiry, there have been some limitations. Most notably, as discussed above, construct underrepresentation has been a consistent concern in tests that purport to measure pragmatic knowledge while operationalizing only certain dimensions (e.g., sociolinguistic knowledge) to the exclusion of others (e.g., psychological knowledge). Since a clear and well-defined construct should be at the heart of every language test (Chapelle, 1998), Purpura's (2004) model of language ability distinguishes the grammatical and semantic meanings (i.e., literal and intended meanings) of an utterance from the pragmatic, or implied, layers of meaning that are contextually driven, and often not derived solely from the arrangement of the words in syntax. These pragmatic meanings, involving appropriateness, conventionality, naturalness, and acceptability, can only be determined in high context situations; Purpura argues that the more indirect an utterance, the richer the contextual features need to be in order for meanings to be decoded. Purpura's pragmatic component also includes a more comprehensive treatment of meaning than prior representations of pragmatic features (Bachman & Palmer, 1996) in that his model accounts not only for the contextual and sociolinguistic meanings in language use, but also for the sociocultural, psychological, and rhetorical meanings as conveyed and interpreted in language use.

In Purpura's (2004) pragmatic component, knowledge of sociolinguistic meaning is defined in terms of social identity markers (e.g., gender, age, status), social meanings (e.g., power and politeness), register variation and modality, social norms, preferences, expectations, and genres (e.g., academic, English for Specific Purposes). Knowledge of sociocultural meaning is defined in terms of cultural meanings (e.g., cultural references, metaphor), cultural norms, preferences, and expectations (e.g., naturalness, frequency and use of apologies, formulaic expressions), and modality differences (e.g., speaking, writing). Knowledge of psychological meaning is defined in terms of affective stance (e.g., sarcasm, deference, importance, anger, irony). Knowledge of contextual meaning is defined in term of interpersonal meanings, in-group references, and metaphor. Knowledge of rhetorical meaning is defined in terms of coherence and genres. As such, pragmatic meaning "embodies a host of implied meanings that derive from context relating to the interpersonal relationship of the interlocutors, their emotional or attitudinal stance, their presuppositions about what is known, the sociocultural setting of the interaction and participation of an interlocutor during talk-in-interaction" (Purpura, 2004, p. 262). This model also takes into account how pragmatic (i.e., implied) meanings are superimposed, or layered, onto the grammatical structures and semantic meanings in language use. Therefore, since grammatical knowledge is inextricably linked to pragmatic meanings in this model, Purpura believes that a test designed to measure pragmatic knowledge should also include the measurement of a learner's grammatical knowledge. (For a complete representation of the model, see Purpura, 2004.)

A second limitation in the assessment of pragmatics has been the lack of validity evidence for widely-used, nonreciprocal tasks, such as multiple-choice (MC) discourse completion tasks (DCTs) and limited production DCTs, which inauthentically reduce the response to a single turn; this is potentially problematic given that researchers know that many speech acts generally occur over several turns (Korsko, 2004; Levinson, 1983), and "their exact shape takes into account interlocutor reactions" (McNamara & Roever, 2006, p. 63). In addition, response patterns in DCTs have been shown not to resemble real-life communication (Beebe & Cummings, 1985, 1996; Golato, 2003; Wolfson, Marmor, & Jones, 1989). In other words, though the purpose of DCTs is to measure language ability in a communicative context, they do not adequately represent the interactive process between interlocutors. This limitation can be addressed by incorporating reciprocal tasks, such as role plays, which have been argued to elicit production data most closely resembling naturally occurring negotiated speech (Clark, 1992; Edmondson, House, Kasper, & Stemmer, 1984; Scarcella, 1979; Trosborg, 1987). Role plays have the potential to reflect both the production and perception of appropriateness in discourse (Purpura, 2004). Reciprocal tasks have a degree of reactivity (Bachman & Palmer, 1996) not represented in MC or limited production DCTs. Under this definition, each interlocutor's response has an effect on the subsequent responses of his interlocutor and vice versa. Unlike an MC DCT or limited production DCT in which the test-taker does not interact with an interlocutor, reciprocal tasks allow the test-taker to "receive feedback on the relevance and correctness of the response, and the response in turn affects the input that is subsequently provided by the interlocutor" (Bachman & Palmer, 1996, p. 55). As a result, reciprocal tasks are considered interactive, since they represent the interaction between receptive and productive skills in language use (Brown, Hudson, Norris, & Bonk, 2002). This notion of reciprocity is analogous to the conceptualization of social and linguistic context as negotiated and dynamic in interaction.

One type of reciprocal task, the interactive DCT has been used in discourse analysis to measure aspects of pragmatic knowledge. For instance, Korsko's (2004) discourse analytic study analyzed the complaint patterns between women in several-turn conversations using an interactive discourse completion test (IDCT), made up of four reciprocal tasks. These tasks differ from more traditional DCTs in that they involve interactive negotiation on the part of *two* participants instead of one. The IDCT used in her study was in a written format using role plays, and was designed to elicit a sequence of spontaneous conversational exchanges between two interlocutors working in pairs (Kettering, 1975). Results from her study demonstrated that the speech patterns elicited through the IDCT were relatively long and negotiated, and unlike a traditional, single-response DCT, resembled natural turn-taking behavior and showed a more authentic progression of meaning over the course of several turns.

A third limitation in prior research in the assessment of pragmatics has been in the statistical procedures used to analyze the test data. Most researchers in pragmatics assessment have primarily used statistical procedures (e.g., correlations and factor analysis) as a basis for validity evidence that are less powerful than some of the more sophisticated analyses available, including many-facet Rasch measurement. A rigorous statistical investigation of the underlying test construct should be employed, including work on test bias and differential item functioning (DIF; McNamara & Roever, 2006). Statistical analyses can also be supported by qualitative methods, such as discourse analysis, which can be used to investigate response content and patterns in the data (for a complete discussion, see Lazaraton, 2002). Data triangulation of this type can lend further support to claims of validity by linking evidence in the responses to the meanings conveyed. Addressing the three limitations outlined above would incorporate theory about grammatical and pragmatic knowledge in a comprehensive underlying test construct, further explore the practicality and validity of role plays in measuring pragmatic meanings, and provide statistical evidence about the hypothesized model on which the test is based.

Purpose of the Study

The primary purpose of the current study was to investigate the extent to which scores from the test designed for this study could be interpreted as indicators of test-takers' grammatical and pragmatic knowledge. In other words, this study investigated the main effects of the components of language ability as operationalized in this test (i.e., grammatical control and sociolinguistic, sociocultural, and psychological appropriateness). The purpose of the study was also to examine how grammatical and pragmatic knowledge can be assessed through reciprocal tasks. Therefore, language ability was operationalized in the context of contextually rich, written interactive DCTs using role plays, in which test-takers were required to both convey and interpret implied meanings in situated, language use. Although the tasks did not directly assess these features in the context of speaking, the IDCT was meant to simulate a conversational context. Many-facet Rasch measurement was used to support claims of validity of the instruments in the study. In addition, interactional sociolinguistic (IS) methods (Gumperz, 1982; Schiffrin, 1994; Tannen, 2004) were used to help show how the range of contextually situated pragmatic meanings could be identified and interpreted qualitatively in the task responses. This analysis added substantive and complementary support for the ratings and statistical procedures used. This type of qualitative procedure can also be used in further rater training. This study represents a multidimensional, interdisciplinary approach to an extremely important yet generally underappreciated research area that may have implications not only for language testing, but also for other areas of applied linguistics.

Research Questions

This study addressed the following research questions:

1. To what extent are the test tasks effective in eliciting the sociolinguistic, sociocultural, and psychological meanings of pragmatic knowledge in the responses?
2. To what degree can the components of grammatical and pragmatic knowledge be rated consistently?
3. What is the relationship between grammatical and pragmatic knowledge in this test?

4. What are the main effects of test-taker ability, rater severity, language knowledge component difficulty, and scale functionality in this test?
5. Is there evidence of bias in the test?

METHOD

In order to address the research questions outlined above, this study used a mixed design (Grotjahn, 1987), which allowed for data triangulation (Brown & Rodgers, 2002). Specifically, the first part of the study used an exploratory-interpretive paradigm (Grotjahn, 1987) involving an interactional sociolinguistic (IS) analysis of a typical native speaker response. This analysis was used to investigate the efficacy of the test tasks in eliciting a range of pragmatic meanings. The second part of the study used an exploratory-quantitative-statistical paradigm (Grotjahn, 1987) in order to investigate the main effects and interactions effects of test facets on test-taker performance. The following subsections include a description of the participants and how they were selected, a description of the instrument and procedures used to elicit and score the data, and a description of the analytic and statistical procedures used to analyze the data.

Participants

Ninety adult English language learners from the Teachers College, Columbia University Community English Program (CEP) participated in this study. The CEP students were heterogeneous with respect to gender, age, native language, educational background, years studying English, and years living in the United States. Considering the nature of the language used in the task descriptions and the assumed language ability needed to complete the tasks, all students in the current study were enrolled in advanced class levels.

In addition, 10 adult native speakers (NSs) of English participated in this study. These participants were heterogeneous with respect to gender, age, and educational background, and were purposefully sampled. Since the *native speaker norm* for language ability is generally recognized to be a theoretical rather than an empirically proven concept (Kasper, 1995; McNamara & Roever, 2006; Roever, 2000), these test-takers served to help establish a baseline understanding of pragmatic use as it relates to the test tasks. Two trained raters, doctoral students in Applied Linguistics at Teachers College, Columbia University scored the performance samples.

Instruments

The Test

In the current study, the construct of second language (L2) knowledge was operationalized in terms of Purpura's (2004) definition of language ability, including both grammatical and pragmatic knowledge dimensions. The first component of L2 knowledge, grammatical knowledge, was defined in terms of knowledge of both grammatical form and semantic meaning. The second component, pragmatic knowledge, related to knowledge of implied meanings in language use, and included the ability to understand and convey sociolinguistic, sociocultural, and psychological meanings. Since the test used in the current study aimed to assess the test-takers' ability to both convey and interpret implied meanings, tasks with rich situational features were employed in an attempt to control for variability in the test-takers' interpretation of the task context. Therefore, knowledge of contextual meaning was not measured in the scoring rubric. Knowledge of rhetorical meaning, which involves the use and understanding of coherence and genre (e.g., the discourse structure of a business meeting), was also not explored in this study, under the assumption that all test-takers were familiar with the structure of the genre used in this test (i.e., informal conversation).

The six reciprocal test tasks were designed to elicit conventional conversational and turn-taking behavior in pairs, and required the test-takers to both convey and interpret implied meanings in online performance. The reciprocal feature is especially important in light of current language testing theory, which views context as dynamic, negotiated, and co-constructed by interlocutors, turn by turn (He & Young, 1998; McNamara, 1996). In this view, context is a social/psychological accomplishment (Douglas & Selinker, 1985), comprised of both linguistic context (e.g., the language addressed to, or used in the presence of, the language user, and co-text) and situational context (e.g., the social, physical, psychological, and temporal situation the language activity is taking place in; Celce-Murcia & Olshtain, 2000; Douglas, 2000). Therefore, since pragmatic meanings are intrinsically linked to how an utterance relates to the context of the situation, in the current study, IDCTs using role plays were used in order to operationalize the reciprocal affect of context and meaning in the context of conversation. Following Korsko (2004), the role play responses are in written format.

In terms of the input, the tasks were highly contextually constrained with respect to the role and situation information given, and the communicative goal of each task was specified, so that the content and purpose were clear (Douglas & Selinker, 1989, 1993). Although the importance of speech acts in pragmatics is clear, given that most speech acts are realized indirectly (Leech, 1983), speech acts do not constitute pragmatic knowledge in its entirety. With respect to tests of pragmatics that only consider speech acts, McNamara and Roever (2006) caution that "conclusions drawn from [them] and decisions based on scores would have to be fairly limited and restricted" (p. 63). As such, the test tasks in the current study did not specify that test-takers use a particular speech act to complete the task (although some may be used more frequently than others in the responses). Instead, a communicative goal was given under the theory that it is possible to go about achieving that goal in pragmatically different ways (i.e., with different speech acts), while still being appropriate to the situation. Therefore, the focus was on the knowledge of appropriateness in understanding and conveying a range of implied meanings, not specifically on the use of certain speech acts. Nonetheless, it is important to note that the appropriate understanding and use of speech acts is part of sociocultural knowledge in Purpura's (2004) model, and is therefore measured in this test. The current approach also attempts to reconcile the dilemma of operationalizing affective and volitional factors, which are also involved in test performance (McNamara, 1997). In addition, in order to incorporate a certain amount of interactive complexity (Edmondson et al., 1984), negotiation (Billmyer & Varghese, 2000), and emotion (Dougill, 1987, as cited in Hudson et al., 1995; Korsko, 2004; Rintell, 1989) into the situations, conflict was operationalized in the task descriptions by making both interlocutors feel that their position or perspective was the right one; in other words, they were given equal power. Social distance was also varied along a continuum, going from a parent-child relationship to strangers, and the situations all represented a moderate degree of imposition.

In terms of the expected response, the role and situation information given to the participants for all six tasks was designed to have interlocutors convey and understand a range of literal and implied meanings. Therefore, each task was designed to elicit sociolinguistic, sociocultural, and psychological meanings on the part of the test-takers. These cues were embedded in the situations and/or explicitly stated in the task directions.

Task 1 involved a discussion between two interlocutors in relation to excessive noise; Task 2 aimed for an exchange between two friends about damaged property; Task 3 was designed to foster a negotiation between a parent and child about proper behavior; Task 4 involved a discussion between friends, one of whom is chronically tardy; Task 5 aimed for a discussion between strangers about an expensive watch; Task 6 was designed to elicit a confrontation between strangers in a grocery store. These tasks can be further categorized into two groups. The first group (task 1: excessive noise; task 2: damaged property; and, task 4: chronic tardiness) was taken from prior research (Abé, 1982; Giddens, 1981; Holmes, 1988, 1989; Olshtain & Weinbach, 1987, 1993; Schaefer, 1982, as cited in Korsko, 2004) and then modified to reflect an equal power differential between the interlocutors; the second group (task 3: parent-child negotiation; task 5: expensive watch; and task 6: grocery store confrontation) was based on real-life occurrences or observations.

The Rubric

An analytic scoring rubric was used to score the test data. The rubric included a total of four scales, each variable rated on a five-point scale, with scoring ranging from 0 (no effective use) to 4 (effective use). In addition to sociolinguistic, sociocultural, and psychological appropriateness, grammatical control was also scored since language users must incorporate grammatical resources in order to realize pragmatic meanings in language use (Purpura, 2004). Also, since researchers are aware that a highly developed knowledge of grammar does not necessarily guarantee a comparable level of pragmatic knowledge (Bardovi-Harlig, 1999), findings from this study were used to explore the extent to which this hypothesis obtains in this test. Each variable had corresponding descriptors for each level of effectiveness of the response. The rubric can be seen in Appendix A.

Equipment and Software

Microsoft Excel for PC Version 5.0 was used to enter and organize the test data. Data were then exported to other statistical programs. The item-level response data were exported to SPSS Version 14 for PC. Frequencies, descriptive statistics, and bivariate correlations were calculated using SPSS. The test data were converted and exported to FACETS Version 3.4 (Linacre & Wright, 1992) for many-facet Rasch analyses, to examine test-taker ability, rater severity, component difficulty, and scale functionality.

Procedures

Test Administration

A task packet labeled either *Person A* or *Person B*, including a set of directions, was assigned randomly to each test-taker by the test administrator (see Appendix B). One person in the pair was Person A (see Appendix C for the tasks for Person A) and the other was Person B (see Appendix D for the tasks for Person B), and would remain so for the duration of the test.

Next, the administrator read aloud the directions to the participants, while they followed along on their own directions sheet. An example of a task and conversation response example was also provided. The administrator gave clarifying information and answered any questions the participants had about the procedure. Participants were then instructed to open their task packets and read the first role and situation information.

Beginning with Task 1, Person A was asked to begin writing what s/he would say in the first turn. The conversation unfolded from there with no other specifications about conversational principles (e.g., turn-taking rules, interruptions), and continued until it came to its natural conclusion. The first person to write alternated from task to task, with the participants completing Tasks 1 through 6 on their own. Neither person saw their partner's role and situation information. The conversations were written on a task worksheet. An example of the task worksheet can be seen in Appendix E.

Scoring

Test performance data for each participant were analyzed and scored by two raters, based on an analytic scoring rubric, with the variables of grammatical control, and sociolinguistic, sociocultural, and psychological appropriateness. All variables were rated on a five-point scale. Raters were normed using a norming packet (adapted from Hudson et al., 1995). In the case that there was more than one level discrepancy between the judges, the raters were normed again and rescored the tests. Ratings were initially entered by hand by each rater onto a scoring matrix after which they were entered into an Excel spreadsheet.

Analyses

Interactional Sociolinguistic Analysis

The most commonly used technique in oral language testing research, conversation analysis (CA), has been employed to describe and analyze patterns and sequencing in speaking data (for a complete discussion, see Lazaraton, 2002). Interactional sociolinguistics (Gumperz, 1982; Tannen, 2004) is similar to CA in that both are concerned with social order and the reciprocal relationship of context and meaning. However, while researchers using CA avoid making claims based on speaker intention and do not assume fixed conceptualizations of social identity (Lazaraton, 2002), IS analysis allows researchers to incorporate broader social context, including speakers' and hearers' background knowledge, into the interpretation of meaning. In the IS view, the contextualization process "is achieved through links between language and participants' knowledge of situation" (Schiffrin, 1994, p. 9), and limits the possibilities with respect to what a speaker may mean or a hearer may interpret in a given context. Since interpreting the implied meanings conveyed in the task responses was found to require the integration of the contextually rich background task information in addition to the linguistic context into the analysis, IS, as opposed to CA, was employed in the current study.

An IS analysis (Gumperz, 1982; Schiffrin, 1994; Tannen, 2004) of native speaker responses from all six tasks was first done in order to determine the potential of the tasks to elicit contextually situated pragmatic meanings. More specifically, the IS analysis was done in order to uncover evidence in the data (e.g., contextualization cues) of sociolinguistic, sociocultural, and psychological meanings as conveyed and interpreted by the interlocutors. Contextualization cues are empirical evidence in the data that help to uncover not only singular instances of implied meaning, but also the progression of meaning, and how meanings shape the context, both in terms of the language used and the shifting pragmatic features of the interaction. In addition, since a certain measure of practicality is desirable in test scoring, it was also important to show how the meanings, both conveyed and interpreted, could readily and systematically be identified by raters in the performance samples. The purpose of this analysis is to highlight a bigger picture of meaning, specifically the sociolinguistic, sociocultural, and psychological meanings in the responses. Obviously, to a certain extent, pragmatic meanings are present throughout the conversation; however, this analysis attempts to isolate the most salient and meaningful examples of pragmatic overlay within the responses. Although IS methodology is generally used on spoken, rather than written data, the written IDCT format in the current study was designed to simulate a conversation; test-takers were encouraged to write what they would actually say in the situations. A brief analysis of one of these conversations will be presented in the Findings section. The conversation was typed exactly as it was written on the task worksheet.

Statistical Procedures

All statistical analyses were performed using data only from the 90 non-native speaker test-takers. Descriptive statistics for the item-level data were calculated in order to examine the data for central tendency and dispersion. The mean, standard deviation, skewness, and kurtosis were examined. In addition, descriptive statistics for each dimension of language ability were also calculated in order to examine their relative difficulty. Inter-rater reliability was also calculated using a Pearson product-moment correlation, based on average ratings on scores for the four variables for each test-taker. Internal consistency reliability estimates using Chronbach’s alpha were calculated to examine the homogeneity of the items. A bivariate correlation using a Pearson product-moment correlation was calculated and examined for the extent to which grammatical and pragmatic knowledge were related in the test.

Since performance assessment necessarily involves human judgment (McNamara, 1996), many-facet Rasch measurement (Linacre, 1989; Linacre & Wright, 1993; McNamara, 1996) was used in order to ensure the trustworthiness of the non-native speaker pragmatic test data and to support claims of validity of the underlying test construct by identifying potential sources of variance in test-taker ability. A Rasch measurement model takes into account different test facets that can affect test-taker performance, and can adjust test-taker ability to represent a fairer estimate by comparing test facets on the same scale. A Partial Credit model (Wright & Masters, 1982) was used in order to compare how the raters interpreted the rating scales in the rubric. This was done based on the assumption that the analytic scale categories were not necessarily interpreted in the same way by the raters across the four language knowledge dimensions. Four facets were considered in this study, namely, test-taker ability, rater severity, language component difficulty, and scale functionality. All test facets must be examined before any generalizations are made about test-taker performance (Lunz & Wright, 1997). The Partial Credit form of the Rasch model used in this study is as follows:

$$\log (P_{nijk} / P_{nijk}\text{-}1) = B_n - C_j - D_i - F_{ik}$$

P_{nijk} = probability of test-taker *n* being awarded a rating of *k* when rated by rater *j* on component *i*

P_{nijk}-1 = probability of test-taker *n* being awarded a rating of *k-1* when rated by rater *j* on component *i*
B_n = ability of test-taker *n*
C_j = severity of rater *j*
D_i = difficulty of component *i*
F_{ik} = difficulty of achieving a score within a particular score category (*k*) on a particular component (*i*)

The first facet, test-taker ability, reflects the test-takers' knowledge and skill with regard to the underlying test construct. This facet should demonstrate a certain degree of variability as the purpose of the test is to measure test-takers' relative ability. The second facet is the rater which shows the relative severity of raters' scoring behavior. As this was a performance test, rater severity is especially important since test-takers' expected scores may differ from those assigned by the raters. If the difference is significant, this could signal the need for more rater training, or more seriously, affect the validity of the inferences made about test-takers' abilities (Tsai, 2004). The third facet, item difficulty, is defined in terms of language knowledge components. The scores used were averages of the two raters' scores for each component, across all six tasks, and refer to the construct being measured in this test (i.e., grammatical control and sociolinguistic, sociocultural, psychological appropriateness). Scale functionality, the fourth facet, was also examined by language knowledge component. This facet helps to show how the raters interpreted the rating scale differentially for each observed variable. A bias analysis was also conducted to uncover any (systematic) interaction effects in the data.

RESULTS

Interactional Sociolinguistic Analysis

A content analysis, using Interactional Sociolinguistic methods, of a prototypical example of native speakers' (NS) responses was performed in order to show how pragmatic meanings could be uncovered in the performance samples. More specifically, this was done to demonstrate how sociolinguistic, sociocultural, and psychological meanings, as conveyed and interpreted, were elicited by the test tasks. Obviously, to a certain extent, pragmatic meanings are present throughout the conversation; however, this analysis will attempt to isolate a few examples of pragmatic overlay within the response.

The example presented in Figure 1 is from Task 2, which simulates a telephone conversation between friends, both students, about a very important book that Person B let Person A borrow. In the task description, Person B is very upset because, after the book was returned, s/he noticed it was damaged. However, in the task description, Person A was told that nothing was wrong with the book when s/he returned it. Person B calls Person A, and the task begins with Person B taking the first turn (1).

FIGURE 1
Sample 1/Task 2/Person 39A and Person 39B/NS

1	Person B	Hey, how're you doing?
2	Person A	Hey, not too bad.
3	Person B	Thanks for returning my book. Looks like you took it for quite a ride…
4	Person A	What do you mean?
5	Person B	Well, there are some coffee stains, and it seems like there are some pages missing.
6	Person A	Really? I went out of my way to make sure nothing happened to it. I even made a special cover for it. I'm not sure how that happened.
7	Person B	But when I gave it to you, it was like new… I don't even drink coffee.
8	Person A	I don't know what else to tell you. It was in perfect condition when I gave it back. Has someone else been using it without your knowledge?
9	Person B	No, you're the only one I lent it to. It's a really important book, so I try not to lend it out.
10	Person A	Someone must have taken it without your knowing because I remember its mint condition clearly when I gave it back.
11	Person B	OK, well thanks. I'll see you later.

In this example, sociolinguistic meaning being conveyed can first be seen in the opening turns ("Hey, how're you doing?," line 1; and "Hey, not too bad," line 2). The casual nature of the interaction is the first indication that the two interlocutors have a close relationship. This type of greeting between friends is to be expected. An ease of communication is also seen between the interlocutors throughout the response, showing a level of comfort that is common between friends. For instance, in lines 3 ("Looks like you took it for quite a ride…"), 5 ("…it seems like there are some pages missing"), 7 ("But when I gave it to you, it was like new") and 9 ("No, you're the only one I lent it to"), Person B shows a reluctance to directly accuse Person A of causing damage to the book. This shows how, sociolinguistically, a friend might approach a delicate subject with prudence.

In this same sequence, sociocultural meaning being conveyed can be seen in Person B's determination and unwillingness to accept the repeated denials by Person A. In other words, since the book is important (i.e., the stakes are high), socioculturally Person B would be expected to be persistent in an attempt to uncover the guilty party; an immediate acceptance of Person A's accounts of innocence by Person B would not illustrate the gravity of the (supposed) offense. Comparably, in turns 4 ("What do you mean?"), 6 ("Really?…I'm not sure how that happened"), 8 ("It was in perfect condition when I gave it back"), and 10 ("Someone must have taken it without your knowing”), Person A responds to each of Person B’s indirect accusations with an indirect denial. Socioculturally, this progression is very natural in terms of how this type of inquiry would unfold in this particular situation.

In line 3, Person B conveys sociocultural meaning by using figurative speech (“Looks like you took it for quite a ride”) to convey an indirect accusation of wrongdoing. Person A’s subsequent turn (“What do you mean?”; line 4) shows an appropriate interpretation of Person B’s implied meaning as being some sort of request to ratify the accusation, rather than a literal reference to some mode of transportation. In turn 4, Person A is implying that s/he misunderstands why Person B might say something like that. This is an appropriate and expected response considering that Person B is told s/he returned the book in the condition in which it was received. These turns show a high level of sociocultural knowledge on the part of both interlocutors. The conveyance and interpretation of meaning in line 10 (“mint condition”) also shows a similar situation with respect to sociocultural meaning through figurative speech being appropriately conveyed and interpreted.

The psychological meanings being conveyed in this example become more and more overt as the conversation unfolds. For example, in line 3, the first indirect accusation by Person B is presented it in a sarcastic and seemingly light-hearted way (“Looks like you took it for quite a ride…”); Person B is showing little emotional distress at first. By line 9, however, Person B is much more overt with his/her affect and uses a direct and contradictory “No” to convey this. Interestingly, an affective shift in Person B from persistence to concession can clearly be seen from lines 9 to 11. Line 9, “No, you’re the only one I leant it to,” shows the most overt indication that Person B is upset and determined to uncover the culprit. However, in line 11 (“OK, well thanks”), Person B shows resignation, essentially gives up trying to get an apology or explanation from Person A, believing him or not, and ends the conversation. Psychologically, Person A shows a comparable progression in affective stance being conveyed in the conversation early on, in line 4, by saying, “What do you mean?” Person A shows little emotion with regard to Person B’s prior comment; s/he just appears confused. However, by line 8 (“I don’t know what else to tell you”), Person A is implying that s/he has offered up a number of plausible alternative explanations, and refuses to be pressed any further. This shows evidence that Person A feels no responsibility for what happened to the book. This is expected since Person A was told in the task directions that s/he did not damage the book.

In this NS example, both interlocutors were given the highest score for each of the five variables. It is clear from the performance sample that both interlocutors showed effective use of pragmatic knowledge in this context through appropriate sociolinguistic, sociocultural, and psychological meanings. Although high scores were generally consistent across most NS test-takers, it is interesting to note, however, that not all NS test-takers received the highest scores across all variables. This shows evidence that the native speaker norm for pragmatic use may be better represented in terms of a range*,* rather than a prescriptive guideline or cut-off score.

This type of analysis was performed on all of the native speaker response samples and some of the non-native speaker samples, for all of the tasks. On the part of native speakers, all tasks were found to elicit the three types of pragmatic meanings operationalized in the test. An analysis of several non-native speaker examples from all of the tasks indicated that the tasks have the potential to elicit these three dimensions of pragmatic knowledge on the part of non-native speakers as well, though in varying degrees based on their level of pragmatic knowledge.

Descriptive Statistics

The means for the tasks ranged from 2.62 for Task 1 (excessive noise) to 2.91 for Task 3 (parent-child negotiation) out of a total of 4. The standard deviations ranged from .46 for Task 1 (excessive noise) to .66 for Task 5 (expensive watch). All values for skewness, ranging from -.25 to .20, centered around zero and were within the acceptable range of ±3.0. All values for kurtosis, ranging from -.84 to .94 were also within the acceptable range. These figures indicate that the data seem to be normally distributed. The descriptive statistics for the item-level data are presented in Table 1.

TABLE 1
Descriptive Statistics for the Six Tasks (N = 90)

	Task	Mean	SD	Skewness	Kurtosis
1	Excessive noise	2.62	.46	.01	.94
2	Damaged property	2.77	.57	.20	-.19
3	Parent-child	2.91	.61	.05	-.84
4	Chronic tardiness	2.70	.53	-.19	-.53
5	Expensive watch	2.69	.66	-.25	-.60
6	Grocery store	2.80	.55	.02	.34

As seen in Table 1, the means for all six tasks were in close proximity to one another, demonstrating that the tasks were very similar in terms of their difficulty for the test-takers. Likewise, a fairly limited range in standard deviations indicates a similar distribution in test scores across tasks.

Four variables (i.e., grammatical control and sociolinguistic, sociocultural, and psychological appropriateness) were rated on a scale of 0 to 4, and the final scores were the averages of the raters for each variable, across all six tasks. The means for the four language knowledge components ranged from 2.71 for sociocultural appropriateness to 2.85 for grammatical knowledge. Standard deviations ranged from a low of .49 for psychological appropriateness to a high of .57 for sociolinguistic knowledge. Values of skewness and kurtosis were all within the acceptable range. These findings are presented below in Table 2.

TABLE 2
Descriptive Statistics for the Four Language Knowledge Components (N = 90)

Component	Mean	SD	Skewness	Kurtosis
Grammatical control	2.85	.57	.28	-.01
Sociolinguistic appropriateness	2.74	.57	.13	-.71
Sociocultural appropriateness	2.71	.52	.18	.45
Psychological appropriateness	2.74	.49	.19	.13

As seen in Table 2, test-takers' scores were consistently, albeit slightly, higher for grammatical knowledge than for any of the other pragmatic dimensions. This is not unexpected given the fact that grammar is very often much more emphasized in instruction—even at the advanced level. Although the means indicate that the test-takers received relatively high scores across all components, these findings highlight the relative difficulty of pragmatic strategies when compared with grammatical knowledge, even at the advanced level (though this may or may not be the case for learners at lower levels).

Inter-rater Reliability

In order to examine the consistency in the raters' assignment of scores, inter-rater reliability was calculated for the entire test as well as each individual language knowledge component using a Pearson product-moment correlation. Inter-rater reliability for the test was relatively high at .82, showing evidence that the raters were consistent in their assignment of scores across the variables. Inter-rater reliability for the average scores on each variable across all six tasks ranged from .68 for sociocultural control to .78 for grammatical control, indicating a moderate to moderately high degree of agreement between raters across all variables. It is not surprising that grammatical control showed the highest degree of rater agreement, since the raters probably had more prior experience rating grammatical features than any other variable. However, moderately high correlations for all the other variables indicate that the raters were relatively consistent, even though these variables may have been less familiar to them. All of these correlations were statistically significant at the .01 level. These figures can be seen in Table 3.

TABLE 3
Inter-rater Reliability for the Four Language Knowledge Components (N = 90)

	Inter-rater reliability
Grammatical control	.777**
Sociolinguistic appropriateness	.757**
Sociocultural appropriateness	.697**
Psychological appropriateness	.759**
No. of variables = 4	.815**

***p* < .01

Internal Consistency Reliability

The internal consistency reliability of the test, estimated using Chronbach's alpha, was .85, indicating a relatively high degree of homogeneity of the test tasks. In other words, there is evidence that the test tasks were all measuring the same underlying factor (i.e., language ability in this test). This high reliability indicates that the test-takers were performing consistently across tasks. These estimates are presented in Table 4.

Following reliability analyses, corrected item-total correlations for each task on the test were obtained. The corrected item-total correlation is a measure of the relationship of the item in the scale to the entire scale, when that given item is not included in the correlation calculation. The individual item is not considered in this calculation since it may inflate the coefficient. All items in the scale had item-total correlations above .3, showing evidence that they were reliably measuring the same thing other items in the scale were also measuring.

TABLE 4
Reliability Analysis of the Six Tasks (N = 90)

	Corrected item-total correlation	Alpha if item deleted
1	.633	.829
2	.528	.845
3	.715	.809
4	.688	.816
5	.743	.804
6	.531	.844
	No. of items = 6	.850

The reliability for each of the language knowledge components from the hypothesized model was then calculated in order to examine the degree to which the scores across these dimensions were consistent. The high degree of consistency across the language knowledge components provides evidence that language knowledge components were quite homogeneous (i.e., they were measuring the same underlying factor) and that test-takers performed consistently across these four dimensions. These findings are presented in Table 5.

TABLE 5
Reliability Analysis of the Four Language Knowledge Components (N = 90)

	Corrected item-total correlation	Alpha if item deleted
Grammatical control	.784	.904
Sociolinguistic appropriateness	.859	.876
Sociocultural appropriateness	.892	.866
Psychological appropriateness	.722	.922
	No. of items = 4	.918

Only one variable, psychological appropriateness, if deleted, would result in a slight increase in the reliability of the scale. Considering that it would be a minimal increase, and in the interest of retaining as many variables as possible, psychological appropriateness was included in the remaining analyses.

Correlation Analyses

The relationship between the language knowledge dimensions in the test was calculated using Pearson product-moment correlations. Grammatical and pragmatic knowledge dimensions are hypothesized to be related, but separate components of language knowledge (Bachman & Palmer, 1996; Purpura, 2004). Since language users must incorporate grammatical resources in order to realize pragmatic meanings in language use, there is substantive reasoning to believe that they would also be related in this test. The correlations ranged from a low of .57 for grammatical control and psychological appropriateness, to a high of .83 for sociolinguistic and sociocultural appropriateness, providing evidence that these language knowledge components do exhibit some unique variance. This indicates that all dimensions are related but separate in the operationalized definition of language knowledge in this test. All correlations were statistically significant at .01 and are presented in Table 6.

TABLE 6
Relationship between Language Knowledge Components (N = 90)

	GRAM	SL	SC	PSY
Grammatical control	1			
Sociolinguistic appropriateness	.802**	1		
Sociocultural appropriateness	.770**	.828**	1	
Psychological appropriateness	.569**	.675**	.784**	1

**$p < .01$

Many-facet Rasch Analysis

Main Effects

A four-facet model was used to examine the main effects of test-taker ability, rater severity, language knowledge component difficulty, and scale functionality. FACETS produces a vertical map on a logit scale, detailing information about each of the facets, as seen in Figure 2.

The logit scale, as seen in the far left column, provides an equal interval representation of all the test facet measures so that they can be interpreted with respect to one another on the same scale. The second column shows test-taker ability. In Rasch analysis, average test-taker ability is set at zero on the logit scale, so a high value indicates a correspondingly high level of ability, and a low value indicates a lower ability level. There was a relatively large spread of test-taker ability. The logit measures of the test-takers ranged from a low of -6.37 to a high of 7.29, suggesting that the test-takers exhibited a wide range of ability level. Infit statistics of the test-takers were also examined. These statistics provide information about the extent to which the ability estimates fit the model, in this case, by variable. This is expressed in terms of the degree of fit between the expected and observed data for each test-taker. In this analysis, the standardized infit, rather than infit mean-square, was used since it is more accurate when $n < 400$ (McNamara, 1996). According to McNamara (1996), standardized infit values between -2 and 2 are within an acceptable range. Values outside of this range indicate greater or lesser variability than is predicted by the model. This could compromise the validity of the inferences made about a test-taker's ability. All test-takers showed good fit in terms of their ability level in this test. Although there was nearly a 14-logit spread, a separation index of 1.87 with a reliability of .57 was found for the ability measures. Separation index is an indication of how many real levels of test-taker ability are represented by the data. The reliability of separation in this case indicates the extent to which there are real differences between ability levels. In this case, the data represent just under two levels of ability. Although the relatively large logit spread of test-taker ability suggests a wide range of levels, further examination shows that the majority of test-takers are clustered between -2 and 3 logits. As such, even though a wide range of ability levels may be represented, the relatively low separation index may indicate that there were minimal differences in the ability levels of the test-takers. This is not unexpected in the current study given that all non-native speaker test-takers were at the advanced level.

FIGURE 2
FACETS Summary (test taker ability, rater severity, component difficulty, scale functionality)

```
                                     Vertical = (1A,2A,3A) Yardstick (columns,lines,low,high)= 0,4,-5,6
--------------------------------------------------------------------------------------------------------------------------------------------
|Measr|+test taker                                                                              |-Raters    |-Items|SL   |SC   |PSY  |GRAM |
--------------------------------------------------------------------------------------------------------------------------------------------
+   6 + A4   A44  B12  B2                                                                       +           +      +(4)  +(4)  +(4)  +(4)  +
|     |                                                                                         |           |      |     |     |     |     |
|     | B6                                                                                      |           |      |     |     |     |     |
|     |                                                                                         |           |      |     |     |     |     |
+   5 +                                                                                         +           +      +     +     +     +     +
|     |                                                                                         |           |      |     |     |     |     |
|     |                                                                                         |           |      |     |     |     |     |
|     |                                                                                         |           |      |     |     |     |     |
+   4 + A2                                                                                      +           +      +     +     +     +     +
|     |                                                                                         |           |      |     |     |     |     |
|     |                                                                                         |           |      |     | --- |     |     |
|     |                                                                                         |           |      |     |     |     |     |
+   3 + A31                                                                                     +           +      +     +     + --- +     +
|     |                                                                                         |           |      |     |     |     |     |
|     | A30  B23  B30  B35  B8                                                                  |           |      | --- |     |     |     |
|     |                                                                                         |           |      |     |     |     | --- |
+   2 + A35                                                                                     +           + PSY  +     +     +     +     +
|     |                                                                                         |           |      |     |     |     |     |
|     | B19  B27                                                                                |           |      |     |     |     |     |
|     | A37                                                                                     |           |      |     |     |     |     |
+   1 + A27  A6   B18  B33                                                                      +           +      +     +     +     +     +
|     | A23                                                                                     |           | SC   |     |     |     |     |
|     |                                                                                         | Rater 2   |      |     |     |     |     |
|     | A13  A17  A5   B14  B21  B31                                                            |           |      |     |     |     |     |
*   0 *                                                                                         *           *      * 3   * 3   * 3   * 3   *
|     | A15  A18  A7   B42                                                                      |           |      |     |     |     |     |
|     | A1   A3                                                                                 | Rater 1   |      |     |     |     |     |
|     | B22                                                                                     |           | SL   |     |     |     |     |
+  -1 + A38  A47  B20  B29  B37                                                                 +           +      +     +     +     +     +
|     |                                                                                         |           |      |     |     |     |     |
|     | A12  A19  A28  A32                                                                      |           |      |     |     |     |     |
|     |                                                                                         |           |      |     |     |     |     |
+  -2 + B47                                                                                     +           + GRAM +     +     +     +     +
|     | B43                                                                                     |           |      |     |     |     | --- |
|     | B24                                                                                     |           |      | --- |     |     |     |
|     | A29                                                                                     |           |      |     |     |     |     |
+  -3 +                                                                                         +           +      +     +     + --- +     +
|     | A10                                                                                     |           |      |     |     |     |     |
|     | B26                                                                                     |           |      |     | --- |     |     |
|     |                                                                                         |           |      |     |     |     |     |
+  -4 + A22  A49                                                                                +           +      +     +     +     +     +
|     | B25  B48                                                                                |           |      |     |     |     |     |
|     |                                                                                         |           |      |     |     |     |     |
|     |                                                                                         |           |      |     |     |     |     |
+  -5 + A11  A33  A36  A41  A42  A46  A48  A50  A8   A9   B10  B11  B13  B36  B38  B44  B45  B50 +           +      +(2)  +(2)  +(2)  +(2)  +
--------------------------------------------------------------------------------------------------------------------------------------------
|Measr|+test taker                                                                              |-Raters    |-Items|SL   |SC   |PSY |GRAM  |
--------------------------------------------------------------------------------------------------------------------------------------------
 GRAM = Grammatical control   SL = Sociolinguistic appropriateness  SC = Sociocultural appropriateness   PSY = Psychological appropriateness
```

The third column is rater severity, with the most severe rater toward the top of the logit scale and the more lenient rater toward the bottom. Rater differences can account for a large proportion of the variability in a set of scores (Linacre, 1989); therefore, it is critical to examine the model to determine if rater severity is compromising the validity of the inferences to be made from test-takers' scores. As seen in Figure 2, Rater 2, is more lenient than Rater 1, with logit measures of -.48 for Rater 2 and .48 for Rater 1. In order to determine the extent to which the expectations of rater behavior fit the model, it is important to examine the degree of fit for each rater. In this analysis, the standardized infit was again used since $n < 400$. Both raters were within the acceptable range, providing evidence that they were self-consistent. In addition, although neither rater showed greater variation than expected, a spread of .96 logits indicates that the raters were somewhat different in the way they assigned scores across components. This is reinforced by a moderate separation index of 1.53 with a reliability index of .45, indicating about one-and-a-half real levels of rater severity. The larger the separation index and the higher the reliability, the more the differences between raters are systematic (McNamara, 1996). Although the separation index is not large, values greater than zero may indicate that the raters are in need of further training with regard to scores on components. As such, a bias analysis, discussed in the following section, was conducted to explore the data for systematic interaction effects.

The fourth column is component difficulty, with the most difficult toward the top of the logit scale, and the least difficult toward the bottom. In this test, four variables were scored by the raters, including grammatical control and sociolinguistic, sociocultural, and psychological appropriateness. Since these language knowledge components are hypothesized to be related yet separate in this test, a test-taker may perform differentially depending on the difficulty of the variable in question. The data show a moderate logit spread of nearly three logits. Grammatical control, at -2.12 logits, was the easiest and psychological control, at .73 logits, was the most difficult. This result lends evidence in support of the distinctness of grammatical knowledge and pragmatic knowledge in this test. It is also an indication that a highly developed knowledge of grammar does not necessarily guarantee a comparable level of pragmatic knowledge (Bardovi-Harlig, 1999). This spread also indicates that the variables are measuring a moderate range of ability levels. In addition, standardized infit statistics indicate that the data fit the model. For the most part, the observed variables are measuring the majority of test-taker ability levels represented, clustering between -2 and 3 logits. A relatively high separation index of 4.49, with a reliability of .89 lends further support to the notion that there are real differences in the difficulty of the variables. With four variables and nearly four-and-a-half levels of difficulty, the data provide evidence that the variables represent different language knowledge components in the underlying construct. The correlation analysis of the test variables in the prior section also lends support to the notion that the underlying language knowledge components are separate but related in this test.

Columns 5 through 8 represent the raters' interpretation of the four analytic rating scales in this test. Although these scales may appear to be equal-interval on a rubric, in practice, raters may interpret the levels of the scales differentially for each variable. For instance, raters may be especially harsh on a particular variable, using only the lower score spectrum of the scale when scoring the variables; in contrast, they may be lenient for another, primarily using the higher score spectrum. Consequently, this can make a particular score more easy or difficult to receive than the model predicts. Or, a rater may apply a certain score more or less frequently than the model predicts. Either situation may affect the test-takers' scores, and therefore the validity of any inferences made from them. In addition, in order for the scale to be functioning properly, the rating scale categories must be in ascending order (i.e., with the theoretically easiest rating to receive at the bottom of the logit scale, and the most difficult at the top), corresponding to increasing thresholds of difficulty as represented in the rubric. In other words, the horizontal lines in each column graphically represent the point at which test-takers are more likely to receive one scale category (e.g., 4) over another (e.g., 3).

The scoring rubric used in the current study consisted of four scales with each variable being rated on a 5-point scale, with scores ranging from 0 (“no effective use”) to 4 (“effective use”). As can be inferred from Figure 2, the scores of 0 and 1 appeared least frequently in the average variable scores, with no test-takers’ average scores in those categories. There is substantive support for this result due to the fact that the test-takers were advanced level ESL students and would be less likely to have “generally ineffective use” or “no effective use” of these language knowledge components. However, this finding may mean that the current rubric is not ideal for this population. In other words, the categories at the upper end of the rating scale may need to be further stratified into two or more levels in order to provide finer distinctions in ability level. The limited categories used by the raters means that the rubric may be better suited to a broader sample of test-takers (i.e., including those at lower ability levels). Nonetheless, a score of 3 was seen most often across all four variables (57% for grammatical [GRAM] control; 58% for sociolinguistic [SL] appropriateness; 73% for sociocultural [SC] appropriateness; and 59% for psychological [PSY] appropriateness). Interestingly, a score of 4 on grammatical control was given 32% of the time; however, a score of 4 was less frequently given on the three pragmatic variables (26% for SL, 8% for SC, and 9% for PSY), indicating that although test-takers may have the grammatical resources necessary to realize pragmatic meanings in discourse, they may not have the pragmatic knowledge necessary to be considered effective in performance. In other words, pragmatic appropriateness first assumes a certain level of underlying grammatical knowledge. By contrast, had the average scores for grammar been lower, it would imply that test-takers could convey a range of pragmatic meanings at a higher level than they could their grammatical knowledge. This was not the case in this test.

FACETS produces a scale measurement report showing the calibrated difficulty measures of the scales, or the raters’ interpretation of difficulty of each level on the rubric. If the step calibration measures are ordered, it provides evidence that the scale is functioning properly. As evidenced by the average measures, all four language knowledge component scales showed a monotonic increase in difficulty level as the scores went up; thus, the scales for these variables appear to have been functioning properly. In other words, the average test-taker observed scores increase as the variable difficulty level also increases, in this case from 2 to 3, and then from 3 to 4. These measures can be seen in Table 7.

TABLE 7
Average Test-taker Ability Measure and Outfit Mean-square Indices (N = 90)

	GRAM		SL		SC		PSY	
Category label	*Average measures*	*Outfit MnSq*	*Average measures*	*Outfit MnSq*	*Average measures*	*Outfit MnSq*	*Average measures*	*Outfit MnSq*
2	-1.30	1.2	-2.71	1.4	-3.65	1.4	-3.55	1.1
3	.31	.6	.29	3.4	-.78	.6	-.99	1.3
4	3.76	.7	3.68	.8	4.45	.2	3.37	.3

Note. GRAM = Grammatical control; SL = Sociolinguistic appropriateness; SC = Sociocultural appropriateness; PSY = Psychological appropriateness

Another statistic provided in the table, the outfit mean-square index, offers additional evidence about the data model fit, namely, the discrepancy between the observed and expected scores for test-takers within specific scale levels. Ideally, this index would be close to 1.0 for each rating scale category, indicating that the expected average test-taker score closely mirrors the observed score. However, if this measure is greater than 2.0, the scale category may not be functioning as intended, and raters may be assigning unexpected scores at a given scale level. For the score category of 3 for sociolinguistic appropriateness, the outfit mean-square index was 3.4, indicating that the raters may have been assigning unexpected scores at this particular scale level. As such, this may signal the need for better rubric descriptors, and/or further rater training with respect to sociolinguistic appropriateness. All other outfit mean-square indices were within the acceptable range.

Bias Analysis

A bias analysis was conducted to reveal any differences between the expected and the observed values, in order to uncover any systematic interaction effects in the data. In general, it is likely (and desirable) that raters will differ slightly in their rating behavior (McNamara, 1996); therefore, as long as raters show at least a relatively high degree of consistency, rater differences do not significantly affect the scores, and can often be diminished with more rater training. However, discrepancies can become especially problematic when a rater is exhibiting a systematic pattern in their scoring behavior that is not accounted for by the model. As seen earlier, the inter-rater reliability was moderately high and the raters showed good data model fit; however, the Rasch analysis also indicated more than one level of rater severity in the test. This finding supported the need for a bias analysis. By contrast, test-taker ability and component difficulty fit statistics showed that the difference between the expected and observed scores was not significant, and the data fit the model.

The bias model *z*-score for each test facet was examined for acceptability within the range of +2 to -2 (McNamara, 1996). If there is systematic difference between the expected and observed measures for rater behavior, it may mean that test-takers were assigned a significantly higher (or lower) score than was warranted by their ability level. A *z*-score with a value greater than 0 shows that a rater is systematically more severe in their scoring behavior than is predicted by the model for that particular item. A *z*-score with a value less than 0 suggests that a rater's scoring behavior is systematically more lenient than is predicted by the model for that item. Values outside the range of +2 to -2 indicate that the variability in that item may be due to systematic bias on the part of the rater. *Z*-scores for both raters showed no significant interaction effects with respect to test-takers or language knowledge component difficulty. This lends evidence in support of the validity of the test data.

DISCUSSION AND CONCLUSION

The current study investigated the extent to which scores from the test designed for this study can be interpreted as indicators of test takers' grammatical and pragmatic knowledge as defined in Purpura's (2004) model of language ability. This study also examined how grammatical and pragmatic knowledge can be assessed in the context of contextually rich reciprocal tasks. In an attempt to better measure both the conveyance and interpretation of a range of implied meanings in situated language use, the assessment tool used in the current study was comprised of six paired, written IDCT role-play tasks. Many-facet Rasch measurement was employed to analyze the main effects of the components of language ability operationalized in this test (i.e., grammatical control and sociolinguistic, sociocultural, and psychological appropriateness). This analysis was used to support claims of validity of the instruments used in this study, reflecting current trends in research in educational measurement and psychometrics. The findings, as they pertain to each of the seven research questions, will be discussed in turn.

Research Question 1: To what extent were the test tasks effective in eliciting the sociolinguistic, sociocultural, and psychological meanings of pragmatic knowledge in the responses?

An interactional sociolinguistic analysis was performed in order to uncover pragmatic meanings encoded in the responses through the identification of contextualization cues. The purpose of this analysis was to uncover the extent to which the test tasks were effective in eliciting a range of pragmatic meanings. In the current study, a native speaker example was used to exemplify a prototypical sample of a high level of performance. The analysis provided examples of sociolinguistic, sociocultural, and psychological meanings as conveyed and interpreted by both interlocutors in the responses. Not only did the test tasks elicit the pragmatic meanings hypothesized in Purpura's (2004) model of language ability, but also this analysis showed how the appropriateness of these meanings can be systematically identified by raters. This type of analysis was critical to show substantive evidence for the scoring procedures as well as to support claims of validity of the underlying test construct.

Research Question 2: To what degree can the components of grammatical and pragmatic knowledge be rated consistently?

Inter-rater reliability for all four language knowledge components indicated that there was a moderate to moderately high degree of agreement between the two raters. In addition, Rasch analysis revealed that rater fit was within an acceptable range with respect to these components. Although there was some question about the relative severity of the raters, a subsequent bias analysis revealed that there were no interaction effects between raters and components or test-takers.

Even though the findings indicated that the scores obtained for the current study were trustworthy, all variables in the rubric, with the exception of grammatical control and, to some extent, sociolinguistic knowledge, were relatively unfamiliar to the raters. Considering the limited number of studies that have investigated pragmatic knowledge in the context of performance assessment, the relatively high inter-rater reliability shows evidence that it may be possible to reliably and systematically score aspects of pragmatic knowledge, even on reciprocal tasks similar to the ones used in the current study. Nonetheless, measures of scale functionality revealed that more rater training may prove beneficial in achieving an even greater degree of consistency. In addition, substantive reasoning would also give cause for providing more rater training to increase the raters' familiarity with the pragmatic features in question.

Research Question 3: What is the relationship between grammatical knowledge and pragmatic knowledge in this test?

The relationship between grammatical knowledge and pragmatic knowledge (defined in the current study as sociolinguistic, sociocultural, and psychological knowledge components) was first explored using correlation analysis. This analysis revealed a moderately high correlation between grammatical knowledge and pragmatic knowledge, showing evidence that although the components are related in this test, they still exhibit some unique variance. This result is expected since grammatical knowledge and pragmatic knowledge were hypothesized to be related in this underlying construct.

In terms of degree of difficulty, Rasch analysis revealed that achieving a high score for pragmatic appropriateness was more difficult than for grammatical control in this test. This provides evidence that a high level of grammatical ability does not necessarily predict a comparably high level of pragmatic use in this test. This relationship lends further support to the distinctness of grammatical and pragmatic knowledge in Purpura's (2004) model of language ability.

Research Question 4: What are the main effects of test-taker ability, rater severity, language knowledge component difficulty, and scale functionality in this test?

In terms of test-taker ability, the Rasch analysis provided evidence that the test successfully separated out the test-takers into a wide range of ability levels. Although a relatively low separation index and reliability indicated minimal differences in the candidate ability measures, this finding was expected considering that the test-takers were all advanced learners.

In terms of ratings, there was evidence that the raters were self-consistent and did not exhibit greater than expected variation in their assignment of scores. Although Rater 2 was more severe than Rater 1, the difference was not significant. Therefore, the raters were relatively interchangeable. Given the significant impact that raters can have on the reliability of scores and the inferences made from them, this finding lends further support to the validity of the scores.

With respect to language knowledge components, there was evidence that the test variables not only were reliably measuring a wide range of ability levels, but also were significantly measuring reliably different underlying variables. Since the focus of English language teaching is often grammar, the finding that grammatical control was the least difficult component was not surprising. The hierarchical difficulty relationship of grammatical knowledge and pragmatic knowledge components lends support to Purpura's (2004) model in which grammatical and pragmatic knowledge are distinct traits. In addition, psychological appropriateness, the most difficult component in this test, may perhaps be the least concrete of all of the underlying variables considering its affective nature. It would conceivably be more difficult for English language learners to master the appropriate use and interpretation of emotion and psychological stance than it would be for more overt meanings such as those associated with sociolinguistic features (e.g., social distance). Sociocultural knowledge, second in difficulty, may arguably require more exposure to the target culture in order to be incorporated into a learner's language knowledge system (Kasper & Rose, 2001).

The analysis of scale functionality of the four variables indicated that the rating scales seemed to be functioning properly. However, the findings also revealed that one rating scale category for sociolinguistic appropriateness may need to be examined further and revised.

Research Question 5: Is there evidence of bias in the test?

A bias analysis was carried out with respect to raters. The findings revealed no systematic interaction effects in the data for either rater. The absence of significant bias in the test scores lends further evidence in support of the trustworthiness of the data.

Limitations

There are a number of limitations in this study that may affect the validity of the inferences made from the test-takers' scores as well as the generalizability of the results. Two of the most salient issues will be discussed below.

First, since practicality is a main concern in most assessment contexts, written, as opposed to spoken, role-play tasks were employed in this study. DCTs can be a practical and systematic tool for capturing conversational behavior (Beebe & Cummings, 1985, 1996; Blum-Kulka & House, 1989; Korsko, 2004). Written IDCTs may have an advantage over traditional, single-response DCTs in their reciprocal representation of context and language, and the progression of meaning over several turns. However, efforts to maximize the practicality of a test do not always justify the resulting reduction in the generalizability or validity of the inferences made from test-takers' scores when there are other more authentic task formats available. Clearly, the written format used in this study has face validity issues with respect to any claim that it is measuring pragmatic aspects in the context of speaking, not the least of which is that written response data cannot easily reflect aspects of spoken data that are not transcribable (Cohen & Olshtain, 1981; Scarcella, 1979). As McNamara and Roever (2006) note:

> There are many differences between written and spoken language with regard to hesitation phenomena, tone of voice, facial expression, gesture, and a number of other nonverbal cues that interlocutors use to contextualize their utterance and convey meaning. None of these are available in DCTs, or any written instrument for that matter. (p. 66)

In the current study, this limitation can be seen as a result of the IS analysis. Although the IS analysis showed evidence that the tasks had the potential to elicit a range of pragmatic meanings, many contextualization cues that could be used to show evidence of these meanings if the data were in a spoken format (e.g., overlap, paralinguistic cues, opting out) were not available to the raters in scoring the written responses. Therefore, there is no way to know if or how these features would affect the scores. In addition, the written format introduced the potential for test-takers' writing ability to be confounded with the abilities being measured in the test (i.e., grammatical and pragmatic knowledge in the context of speaking). There is clearly a need for more research in the use of written and oral role play formats and the meanings they elicit. As a result, an oral role play version of the test is currently being developed in an attempt to assess grammatical and pragmatic knowledge in the context of a direct speaking test.

Second, it has been argued that tasks involving two interlocutors have an advantage over indirect or semi-direct (monologic) measures in that they allow for a theoretically more authentic representation of interaction (Stansfield & Kenyon, 1992). As such, there is a growing body of research in paired and group task formats as evidenced by the 2007 International Language Testing Association / American Association for Applied Linguistics joint symposium on pair work in L2 learning and assessment. While some research has shown the paired format to be an acceptable method of eliciting language ability in the context of speaking, research has been inconclusive about the extent of positive or negative effects of a test partner on performance (Berry, 1997; Foot, 1999; Galaczi-Dimitrova, 2003; Iwashita, 1999; O'Sullivan, 2002; Saville & Hargreaves, 1999; Taylor, 2001), due, perhaps in part, to the great number of variables that can have a combined effect on performance (Luoma, 2004). Consequently, some researchers question the fairness of rating discourse that is jointly constructed in any testing situation (Brown, 2003; McNamara, 1997).

In the current study, the test-takers were paired with a partner for the test; therefore, there is the possibility that one test-taker's ability level may have unfairly influenced the other's scores. Although attempts were made to minimize this threat by only involving advanced learners in the test, an investigation into the effect of test partner on scores was beyond the scope of this study. However, it is important to note that the current study is very exploratory, and therefore, the findings are not meant to be generalized to any other contexts than this particular test represents; in other words, this test is not intended for high-stakes use.

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. James Purpura for his support and confidence in this project. I would also like to extend my gratitude to Christine Jacknick for giving up hours of her free time during the rating process. Thank you also to the anonymous reviewers at the Web Journal for their patience and insight during the review process.

REFERENCES

Abé, J. (1982). *An analysis of the discourse and syntax of oral complaints in Japanese.* Unpublished master's thesis, University of California, Los Angeles.

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

Bachman, L., & Purpura, J. (in press). Language assessments: Gate-keepers or door openers? In B. M. Spolsky & F. M. Hult (Eds.), *Blackwell handbook of educational linguistics*. Malden, MA: Blackwell.

Bardovi-Harlig, K. (1999). Researching method. In L. Bouton (Ed.), *Pragmatics and language learning* (Vol. 9, pp. 236-264). Urbana-Champaign: University of Illinois, Division of English as an International Language.

Beebe, L., & Cummings, M. (1985, April). *Speech act performance: A function of the data collection procedure?* Paper presented at the TESOL Convention, New York, NY.

Beebe, L., & Cummings, M. (1996). Natural speech act data versus written questionnaire data: How data collection method affects speech act performance. In S. M. Gass & J. Neu (Eds.), *Speech acts across cultures: Challenges to communication in a second language* (pp. 65-86). Berlin: Walter de Gruyter.

Berry, V. (1997, March). Gender and personality as factors of interlocutor variability in oral performance tests. Paper presented at the Language Testing Research Colloquium in Orlando, FL.

Billmyer, K., & Varghese, M. (2000). Investigating instrument-based pragmatic variability: Effects of enhancing discourse completion tests. *Applied Linguistics*, *21*, 517-552.

Blum-Kulka, S., & House, J. (1989). Cross-cultural and situational variation in requesting behavior. In S. Blum-Kulka, J. House, & G. Kasper (Eds.), *Cross-cultural pragmatics: Requests and apologies* (pp. 123-154). Norwood, NJ: Ablex.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, *20*, 1-25.

Brown, J. D., Hudson, T., Norris, J., & Bonk, W. (2002). *An investigation of second language task-based performance assessments*. Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Brown, J. D., & Rodgers, T. S. (2002). *Doing second language research*. Oxford, UK: Oxford University Press.

Brown, P., & Levinson, S. (1987). *Politeness: Some universals in language usage.* Cambridge, UK: Cambridge University Press.

Bouton, L. (1988). A cross-cultural study of ability to interpret implicatures in English. *World Englishes*, *17*, 183-196.

Bouton, L. (1994). Conversational implicature in the second language: Learned slowly when not deliberately taught. *Journal of Pragmatics*, *22*, 157-167.

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

Celce-Murcia, M., & Olshtain, E. (2000). *Discourse and context in language teaching.* Cambridge, UK: Cambridge University Press.

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. Bachman & A. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp.32-70). Cambridge, UK: Cambridge University Press.

Clark, H. (1992). *Arenas of language use*. Chicago: The University of Chicago Press.

Cohen, A. (1995). Investigating the production of speech act sets. In S. M. Gass & J. Neu (Eds.), *Speech acts across cultures: Challenges to communication in a second language* (pp. 21-44). Berlin: Mouton de Gruyter.

Cohen, A., & Olshtain, E. (1981). Developing a measure of sociocultural competence: The case of apology. *Language Learning*, *31*, 113-34.

Coulter, J. (1989). *Mind in action*. Atlantic Highlands, NJ: Humanities Press.

Douglas, D. (2000). Assessing languages for specific purposes. Cambridge, UK: Cambridge University Press.

Douglas, D., & Selinker, L. (1985). Principles for language tests within "discourse domains" theory of interlanguage. *Language Testing*, *2*, 205-226.

Douglas, D., & Selinker, L. (1989). Markedness in discourse domains: Native and non-native teaching assistants. *Papers in Applied Linguistics*, *1*, 69-82.

Douglas, D., & Selinker, L. (1993). Performance on general versus field-specific tests of speaking proficiency. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 235-256). Alexandria, VA: TESOL Publications.

Duranti, A., & Goodwin, C. (1992). *Rethinking context*. Cambridge, UK: Cambridge University Press.

Edmondson, W., House, J., Kasper, G., & Stemmer, B. (1984). Learning the pragmatics of discourse: A project report. *Applied Linguistics*, *5*, 113-127.

Enochs, K., & Yoshitake-Strain, S. (1996). Self-assessment and role plays for evaluating appropriateness in speech act realizations. *ICU Language Research Bulletin*, *11*, 57-76.

Enochs, K., & Yoshitake-Strain, S. (1999). Evaluating six measures of EFL learners' pragmatic competence. *JALT Journal, 21*, 29-50.

Foot, M. (1999). Relaxing in pairs. *ELT Journal, 53*, 36-41.

Galaczi-Dimitrova, E. (2003). Interaction in a paired speaking test: The case of the First Certificate in English. *Research Notes*, *14*, 19-23. Cambridge, UK: UCLES.

Giddens, D. (1981). *An analysis of the discourse and syntax of oral complaints in Spanish*. Unpublished master's thesis, University of California, Los Angeles.

Golato, A. (2003). Studying compliment responses: A comparison of DCTs and naturally occurring talk. *Applied Linguistics*, *24*, 90-121.

Grotjahn, R. (1987). On the methodological basis of introspective methods. In C. Faerch & G. Kasper (Eds.), *Linguistic perspectives on second language acquisition* (pp. 54-81). Clevedon, UK: Multilingual Matters.

Gumperz, J. (1982). *Discourse strategies*. Cambridge, UK: Cambridge University Press.

He, A., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young & A. He (Eds.), *Talking and testing: The discourse approaches to the assessment of oral proficiency* (pp. 1-24). Amsterdam: John Benjamins.

Heritage, J. (1984). *Garfinkel and ethnomethodology*. Oxford, UK: Blackwell.

Holmes, J. (1988). Paying compliments: A sex-preferential politeness strategy. *Journal of Pragmatics*, *12*, 445-465.

Holmes, J. (1989). Sex differences and apologies: One aspect of communicative competence. *Applied Linguistics*, *10*, 194-213.

Hudson, T., Detmer, E., & Brown, J. D. (1992). *A framework for testing cross-cultural pragmatics.* Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Hudson, T., Detmer, E., & Brown, J. D. (1995). *Developing prototypic measures of cross-cultural pragmatics* (Tech. Rep. No 7). Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Iwashita, N. (1999). The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing*, *8*, 51-66.

Jefferson, G. (1989). Preliminary notes on a possible metric which provides for a "standard maximum" silence of approximately one second in conversation. In D. Roger & P. Bull (Eds.), *Conversation: An interdisciplinary perspective*. Clevedon, UK: Multilingual Matters, 166-196.

Johnson, M. (2001). *The art of non-conversation*. New Haven, CT: Yale University Press.

Kettering, J. (1975). *Developing communicative competence: Interaction activities in English as a second language*. Pittsburgh: University of Pittsburgh, Center for International Studies.

Kasper, G. (1995). Routine and indirection in interlanguage pragmatics. In L. Bouton (Ed.), *Pragmatics and language learning* (Vol. 6, pp. 59-78). Urbana-Champaign: University of Illinois at Urbana-Champaign.

Kasper, G., & Rose, K. (2001). Pragmatics in language teaching. In K. Rose & G. Kasper (Eds.), *Pragmatics and language teaching* (pp. 1-9). Cambridge, UK: Cambridge University Press.

Korsko, P. (2004). *The narrative shape of two-party complaints in Portuguese: A discourse analytic study.* Unpublished doctoral dissertation, Teachers College, Columbia University, New York City.

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests*. Cambridge, UK: Cambridge University Press.

Leech, G. (1983). *Principles of pragmatics*. London: Longman.

Levinson, S. (1983). *Pragmatics*. Cambridge, UK: Cambridge University Press.

Linacre, J. (1989). *Many-faceted Rasch measurement*. Chicago: MESA Press.

Linacre, J., & Wright, B. (1992). *FACETS: Rasch measurement computer program* (Version 2.6). Chicago: MESA Press.

Linacre, J., & Wright, B. (1993). *FACETS: Many-facet Rasch analysis* (Version 2.68). Chicago: MESA Press.

Liu, J. (2006). *Measuring interlanguage pragmatic knowledge of EFL learners*. Frankfurt am Main, Germany: Peter Lang.

Lunz, M., & Wright, B. (1997). Latent trait models for performance examinations. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 80-88). Munster, Germany: Waxmann.

Luoma, S. (2004). *Assessing speaking*. Cambridge, UK: Cambridge University Press.

McNamara, T. (1996). *Measuring second language performance: A new era in language testing*. New York: Longman.

McNamara, T. (1997). "Interaction" in second language performance assessment: Whose performance? *Applied Linguistics*, *18*, 44-466.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension.* Malden, MA: Blackwell.

O'Sullivan, B. (2002). Learner acquaintanceship and OPT pair-task performance. *Language Testing*, *19*, 277-295.

Olshtain, E., & Weinbach, L. (1987). Complaints: A study of speech act behavior among native and nonnative speakers of Hebrew. In J. Vershueren & M. Bertuccelli-Papi (Eds.), *The pragmatic perspective: Selected papers from the 1985 International Pragmatics Conference* (pp. 196-208). Amsterdam: John Benjamins.

Olshtain, E., & Weinbach, L. (1993). Interlanguage features of the speech act of complaining. In G. Kasper & S. Blum-Kulka (Eds.), *Interlanguage pragmatics* (pp. 108-122). New York: Oxford University Press.

Purpura, J. (2004). *Assessing grammar*. Cambridge, UK: Cambridge University Press.

Rintell, E. (1989). That reminds me of a story: The use of language to express emotion by second-language learners and native speakers. In M. Eisenstein (Ed.), *The dynamic interlanguage: Empirical studies in second language variation* (pp. 237-257). New York: Plenum Press.

Roever, C. (2000). *Web-based test of interlanguage pragmatics*. Retrieved on January 27, 2006, from http://pages.prodigy.net/crhnl/start.htm

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

Saville, N., & Hargreaves, P. (1999). Asessing speaking in the revised FCE. *English Language Teaching Journal*, *53*, 42-51.

Scarcella, R. (1979). On speaking politely in a second language. In C. Yorio, K. Perkins, & J. Schachter (Eds.), *On TESOL '79* (pp. 275-287). Washington, DC: TESOL.

Schegloff, E. (1987). Between micro and macro: Contexts and other connections. In J. Alexander, B. Giesen, R. Munch, & N. Smelser (Eds.), *The micro-macro link* (pp. 207-234). Berkeley: University of California Press.

Stansfield, C., & Kenyon, D. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System*, *20*, 347-364.

Tannen, D. (2004). Interactional sociolinguistics. In U. Ammon, N. Dittmar, K. Mattheier, & P. Trudgill (Eds.), *Sociolinguistics: An international handbook of the science of language in society*. Berlin: Walter de Gruyter.

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

Taylor, L. (2001). The paired speaking test format: Recent studies. *Research Notes*, *2*, 14-15. Cambridge, UK: UCLES.

Trosborg, A. (1987). Apology strategies in natives/non-natives. *Journal of Pragmatics*, *11*, 147-167.

Tsai, C. (2004). *Investigating the relationship between ESL writers' strategy use and their second language writing ability*. Unpublished doctoral dissertation, Teachers College, Columbia University, New York City.

Wolfson, N., Marmor, T., & Jones, S. (1989). Problems in the comparison of speech acts across cultures. In S. Blum-Kulka, J. House, & G. Kasper (Eds.), *Cross-cultural pragmatics: Requests and apologies* (pp. 174-196). Norwood, NJ: Ablex.

Wright, B., & Masters, G. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA Press.

Yamashita, S. (1996). *Six measures of JSL Pragmatics* (Tech. Rep. No. 14). Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

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

APPENDIX A
Analytic Scoring Rubric

	Grammatical control ▪ *phonological/graphological forms and meanings; lexical forms and meanings; morphosyntactic forms and meaning; cohesive forms and meaning; information management forms and meanings; interactional forms and meanings;* ▪ *meaningfulness in terms of the conveyance of literal meaning;* ▪ *ability to get their point across.* ***Response demonstrates:***	***Sociolinguistic appropriateness*** ▪ *social identity markers of age, gender, status, and group;* ▪ *cultural identity markers (dialect, nativeness);* ▪ *social meanings(power, politeness);* ▪ *social norms, preferences, and expectations;* ▪ *register variation (modalities and genres).* ***Response demonstrates:***	***Sociocultural appropriateness*** ▪ *cultural meanings (references, metaphor, figurative meanings);* ▪ *cultural norms, preferences, and expectations (naturalness, frequency and use of apologies, formulaic expressions, collocations);* ▪ *modality differences (speaking, writing).* ***Response demonstrates*:**	**Psychological appropriateness** ▪ *attitude (sarcasm, irony, understatement, humor, deference, criticism,);* ▪ *affect (anger, impatience)* ***Response demonstrates:***
4	effective use.	effective use.	effective use.	effective use.
3	generally effective use.	generally effective use.	generally effective use.	generally effective use.
2	somewhat effective use.	somewhat effective use.	somewhat effective use.	somewhat effective use.
1	generally ineffective use.	generally ineffective use.	generally ineffective use.	generally ineffective use.
0	no effective use.	no effective use.	no effective use.	no effective use.

APPENDIX B
Test Directions

Directions

In your task packet, you will find six everyday situations that you will act out with a partner. For each, you will be given a description of the situation and the role that you will play. Keep in mind that you should play the role that you have been given, even if the role is unfamiliar to you. Do the best you can.

After you have read the description of the situation and your role, you will complete a conversation with your partner by writing it on the task worksheet, and not by speaking. Write what you think you would *actually* say to the other person in this situation.

When finished, pass the task worksheet to your partner to respond. Take turns passing the worksheet back and forth until the conversation is finished. You will do this same thing for each of the six tasks. You should think of this task as "speaking on paper".

Here is an example of a situation like the ones you will see in your packet. Each partner will be given a different role for the same situation. This is an example of what one partner will see.

Example Task:

Roles

You: You work at a large company

Your partner: A co-worker and friend

Situation

Your partner has recently received a promotion. You worked together on several successful projects in the last year. You did more work on those projects and feel that you deserved the promotion instead. You run into your partner at the office. You have not seen him/her since the promotion.

During your conversation, make sure your partner knows you are disappointed.

Use the Task Worksheet *to write what you would say. You will write first.*

On the next page, there is an example of the worksheet that you will use to complete the conversation with your partner. Read the sample conversation provided:

Task Worksheet

PERSON A will write first

Person A	Person B
Hey. Congratulations! I heard about the promotion.	
	Yeah, thanks. Honestly, I really thought they were going to give it to you. You've been working so hard!
Well, I really was happy for you that you got it, but you know, our boss has passed me over so many times in the past four years, it just makes me not want to stay. You know what I mean?	
	Yeah. I'm really sorry. I'd be disappointed too. But don't leave now. Your promotion might be right around the corner.
Yeah, maybe you're right, I guess. I just feel like I do more than just about anyone around here.	
	(Continue the conversation…)

The conversation that you create with your partner should continue until its natural conclusion, that is, until you feel the conversation is finished. You can use the back of the pages, if necessary.

The person who will write first will alternate from task to task. The directions inside your task packet will tell you who will write first.

In your folder you will also find a task worksheet packet. Use the worksheets provided for your conversations.

To clarify:

1. Do not describe what you would *do* in the situation, but write what you would actually *say* to your partner in the situation described.

2. Please do not talk to your partner or show him/her the information in your task packet.

3. Take as long as you need, but keep in mind that there are six situations and an hour has been allotted for the entire exercise. So, take about 3 to 10 minutes to do each situation. Be as spontaneous and as quick as you like.

4. Write clearly so that your partner can read your handwriting.

5. Dictionaries are not allowed to be used on this test.

APPENDIX C
Partner A Test

Tasks for Person A

Do not turn the page until
the administrator tells you to do so.

A

Task 1

Roles

You: a tired person

Your partner: your 25-year-old neighbor

Situation

It is 1:30 am on a Wednesday night. You have been trying to fall asleep for some time now, but you can't because of all the noise coming from your neighbor's apartment. This has been an on-going problem with your neighbor. You get out of bed, go over to your neighbor's apartment, and knock on the door. Your neighbor opens the door.

During the conversation, make sure your partner knows that you are frustrated by the situation.

Use the *Task 1 Worksheet* to write what you would say. You will write first.

A

Task 2

Roles

You: a student

Your partner: a student and friend

Situation

You borrowed an important book from your partner. You were very careful with it while you had it. You gave it back to him/her last week. You are at home relaxing when the phone rings. You answer it.

During the conversation, make sure your partner knows you don't think you did anything wrong.

Your partner will write first. Wait until your partner hands you the *Task 2 Worksheet* to respond.

A

Task 3

Roles

You: a parent

Your partner: your 10-year-old child

Situation

You are at home waiting for your partner to come home from school. When your partner comes in the door, s/he calls the teacher "a jerk", and runs upstairs to his/her bedroom. You follow upstairs and knock on the door to find out what's going on.

During the conversation, make sure your partner knows how you feel about his/her behavior.

Use the *Task 3 Worksheet* to write what you would say. You will write first.

A

Task 4

Roles

You: a student

Your partner: a student and friend

Situation

You had arranged to meet your partner 35 minutes ago at a certain café to have a cup of tea. All of your friends know that you are always late, so you are not really worried. No one has ever said anything to you about it. You arrive just as your partner is in the process of leaving.

During the conversation, make sure your partner knows that you don't think s/he should be mad.

Your partner will write first. Wait until your partner hands you the *Task 4 Worksheet* to respond.

A

Task 5

Roles

You: a teacher

Your partner: a stranger at a party

Situation

You are at a cocktail party. Your partner is standing across the room, wearing a beautiful watch. You desperately want to know how much it cost, because you want one just like it, but aren't sure you can afford it. You walk over to introduce yourself.

During the conversation, try to get your partner to tell you how much the watch cost.

Use the *Task 5 Worksheet* to write what you would say. You will write first.

A

Task 6

Roles

You: a person in a hurry

Your partner: a stranger

Situation

You have been waiting in line at the grocery store for 15 minutes and are very late for an important meeting. The person in front of you (your partner) drops all of his/her groceries. It will take him/her a while to pick things up, so you step around and go up to the cashier. You don't think there is anything wrong with this because you don't want to waste any more time.

During the conversation, try to make your partner let you go to the cashier first.

Your partner will write first. Wait until your partner hands you the *Task 6 Worksheet* to respond.

APPENDIX D
Partner B Test

Tasks for Person B

Do not turn the page until
the administrator tells you to do so.

B

Task 1

Roles

You: a 25-year-old

Your partner: your annoying neighbor

Situation

It is 1:30 am on a Wednesday night. You and some of your friends are having a good time in your apartment. You are all listening to music and laughing. There is a knock on the door, so you go to the door and answer it. It's one of your neighbors. You find this neighbor very annoying because s/he is always telling you to turn down your music.

During the conversation, make sure your partner knows you don't think you are too loud.

Your partner will write first. Wait until your partner hands you the *Task 1 Worksheet* to respond.

B

Task 2

Roles

You: a student

Your partner: a student and friend

Situation

Your partner borrowed a book of yours that was very important to you. S/he returned it with coffee stains and a few pages missing. You decide to give him/her a call. S/he answers the phone.

During the conversation, make sure your partner understands that you are upset.

Use the *Task 2 Worksheet* to write what you would say. You will write first.

B

Task 3

Roles

You: a 10-year-old child

Your partner: your parent

Situation

You are very mad because your teacher failed you for cheating on a test today. You **did not** cheat. You come home from school, call your teacher "a jerk", and run upstairs. Your partner follows you upstairs and knocks on your door.

During the conversation, make sure your partner knows that you don't think you did anything wrong.

Your partner will write first. Wait until your partner hands you the *Task 3 Worksheet* to respond.

B

Task 4

Roles

You: a student

Your partner: a student and friend

Situation

You arranged to meet your partner at a certain café to have a cup of tea. You have been waiting at the café for him/her for more than 30 minutes. Every time you arrange to meet, s/he is at least 20 to 30 minutes late. You are in the process of leaving when your partner arrives.

During the conversation, make sure your partner knows that you don't like his/her behavior.

Use the *Task 4 Worksheet* to write what you would say. You will write first.

B

Task 5

Roles

You: unemployed

Your partner: a stranger at a party

Situation

You are at a cocktail party. You are wearing your best outfit including a new watch. You bought it recently and it was very expensive. Since you don't have a job, you are embarrassed to say how much you spent. Your partner comes over to you and starts a conversation.

During the conversation, don't tell your partner how much you spent.

Your partner will write first. Wait until your partner hands you the *Task 5 Worksheet* to respond.

B

Task 6

Roles

You: a person in a hurry

Your partner: a rude person

Situation

You have been waiting in line at the grocery store for 15 minutes. You are next in line and the clerk calls "Next!" As you go to the counter, you drop the groceries you were holding. No one helps you. As you are picking things up, your partner, who is in line behind you, starts to step around you toward the cashier. You were next and don't want to have to wait any longer.

During the conversation, try to make your partner let you go to the cashier first.

Use the *Task 6 Worksheet* to write what you would say. You will write first.

APPENDIX E
Sample Task Worksheet

Task 1 Worksheet

PERSON A will write first

Person A	Person B