Tabaran Institute of Higher Education ISSN 2251-7324
Iranian Journal of Language Testing
Vol. 4, No. 1, March 2014

Expanding Traditional Testing Measures with Tasks from L2 Pragmatics Research

Kathleen Bardovi-Harlig[1], *Sun-Young Shin*[2]

Abstract

This article argues that testing in pragmatics has for too long relied on the same six measures of pragmatics assessment introduced by Hudson, Detmer, and Brown (1992, 1995). We demonstrate that there is a wealth of potential test formats in the L2 pragmatics acquisition literature that are as yet untapped resources for pragmatics testing. The article first reviews definitions of pragmatics that are useful in guiding the design and development of pragmatic measures and subsequent scoring. It then discusses the principles of language assessment as they have been applied to tests of pragmatics. Next it assesses and reports on current interest in pragmatics testing in language programs through informal interviews conducted with researcher-teachers on current practices in pragmatics testing. We then introduce tasks that are used in pragmatic research which are innovative in the context of assessment, and address the potential of each task to enhance task authenticity, their practicality for testing, and their potential for broadening our construct representation.

Keywords: *pragmatics assessment; reliability; validity; authenticity; practicality*

1. Introduction

This paper explores the contribution that tasks used in pragmatics research can make to assessment of pragmatics. So far, the six measures of pragmatics assessment introduced by Hudson, Detmer, and Brown (1992, 1995) (namely oral, written, and multiple-choice DCTs, role plays, and two types of self-assessment) have been the most widely used means of measurement. Reliability, validity, and practicality of such tools have been investigated by subsequent studies (Brown, 2001; Hudson, 2001; Liu, 2006, 2007; Roever, 2006, 2007; Rose, 1994), but the studies have reported conflicting findings regarding these measures, calling into question their viability as the only means of assessing pragmatic knowledge. Without the development of new types of test items, the field is at an impasse. This paper makes several recommendations of item types based on empirical pragmatics research.

We first review definitions of pragmatics, then discuss principles of testing as they have been applied to tests of pragmatics. The next section briefly considers the interest in pragmatics assessment in foreign language programs. We then consider in some detail six types of

1Indiana University, USA. E-mail:bardovi@indiana.edu
2Indiana University, USA. E-mail: shin36@indiana.edu

pragmatics assessment formats that have been developed for pragmatics research that have not yet been considered for testing. We conclude with a proposal for integrating new items into systematic tests of pragmatics.

2. Definitions of Pragmatics

The definition of pragmatics that is adopted for a testing project will influence the orientation of the test. Levinson (1983) observed that the study of pragmatics has traditionally encompassed at least five main areas: deixis, conversational implicature, presupposition, speech acts, and conversational structure. Within applied linguistics, pragmatics research has focused on the investigation of speech acts and to a lesser extent conversational structure and conversational implicature, and has included address terms as well.

Kasper (1996, p. 146) offered the following inventory of topics which had been covered in interlanguage pragmatics up to that time, any of which could be considered for assessment: "nonnative speakers' perception and comprehension of illocutionary force and politeness; their production of linguistic action; the impact of context variables on choices of conventions of means (semantic formulae or realization strategies) and forms (linguistic means of implementing strategic options); discourse sequencing and conversational management; pragmatic success and failure; and the joint negotiation of illocutionary, referential, and relational goals in personal encounters and institutional settings." With emphasis on language users, Crystal (1997, p. 301) defines pragmatics as "the study of language from the point of view of users, especially of the choices they make, the constraints they encounter in using language in social interaction and the effects their use of language has on other participants in the act of communication."

An even broader definition is found in the *Handbooks of Pragmatics*: *Pragmatics across Languages and Cultures* (Trosborg, 2010, p. v), "Pragmatics is understood in a broad sense as the scientific study of all aspects of linguistic behavior. These aspects include patterns of linguistic action, language functions, types of inferences, principles of communication, frames of knowledge, attitude, and belief, as well as organizational principles of text and discourse. Pragmatics deals with meaning-in-context, which for analytical purposes can be viewed from different perspectives (the speaker's, recipient's, analyst's, etc.). It bridges the gap between the system side of language and the use side, and relates both of them at the same time."

A nontechnical definition of pragmatics is offered to test-takers in the *TOEFL Planner* by ETS. In a section called "Listening for pragmatic understanding" (p. 29), students are told how to practice for the pragmatics listening items: in nontechnical language, they are told to pay attention to illocutionary force (what the speaker hopes to accomplish) with examples of speech acts, formality (formal or casual), degree of certainty, prosodic information (stress and intonation, and the way it conveys meaning) and speaker point of view.

All of these definitions are relevant to assessment specialists in the development of pragmatics tests, as each reveals a different view of the main construct. The explicit naming of the processes (perception, comprehension, and production) and the components of pragmatic knowledge in Kasper's definition provides a checklist which may be especially helpful in achieving an appropriate level of construct representation.

If the first question in testing pragmatics is "what is pragmatics?" (Grabowski, 2008; Roever, 2011), then the second question is "what is a pragmatics test?" Recent discussions have questioned the lack of breadth in pragmatics testing, focusing on the dominance of speech acts. It

should be noted that speech act dominance in testing mirrors that in interlanguage pragmatics research more generally. For example, Grabowski (2008, p. 158) questions whether Liu's (2006) test of pragmatics—which tested either apologies or requests—represents pragmatics in a general sense or more accurately represents specific knowledge of one of two speech acts. Such a test is a case of construct underrepresentation. In contrast to current practice, a pragmatics test can have many different types of items, just as a grammar test does, targeting different areas of pragmatic knowledge. Pragmatics tests that are linked to curricula will be further guided by the course goals, objectives, and student learning outcomes in ways that stand-alone tests developed for research purposes are not.

In the following sections we report on a variety of item formats taken directly from the interlanguage pragmatics literature. Although designed for the detailed study of interlanguage knowledge, each of these tasks has potential in testing by virtue of the degree of control by the test designer and reasonable ease of scoring. Testing different areas of pragmatics renders the test more discrete and easier to score, thus making it easier to assess development. For example, a learner may not perform a specific speech act or use a specific conventional expression for many reasons. Scaling tests back to be less than comprehensive allows us to assess what learners still need to know and how we may be able to help them instructionally. Before we examine potential test formats, we review basic principles in testing as they have been applied to pragmatics.

3. Key Testing Principles Guiding Development and Evaluation of Pragmatic Assessment

Language assessment is developed and used to gather information about test takers' language abilities to make informed decisions about them. It is, thus, important for us to decide which assessment tools are the most appropriate for our particular score interpretation and use, especially given an array of assessment tools available for language assessment practices. Typically, the qualities of language assessment are often determined by the following three important criteria: reliability, validity, and practicality. We will discuss a basic concept of each criterion, and how these issues have been dealt with in a pragmatic assessment context below.

3.1. Reliability in pragmatic assessments

Reliability refers to the consistency of test scores across different testing circumstances. For example, if a student takes a test over time, and his or her score has not changed significantly, we can be assured that a given test is a reliable tool reflecting a stability of the construct being measured. This cannot be easily examined because it requires students to take the same given test repeatedly over a period of time. Thus, internal consistency is often used to examine the reliability of a test by looking at the correlations of each part of the test to the entire test typically based on classical test theory approaches such as KR20, KR21, or Cronbach alpha (Brown, 2005). As summarized in Brown (2008, p.234), the reliability estimates of self-assessment, written discourse completion tasks (WDCTs), and oral discourse completion tasks (ODCTs), and role-play tasks turned out to be acceptably high; the exception was the multiple-choice discourse completion tests (MDCTs). However, previous research has reported conflicting results for internal consistency reliability of the multiple-choice DCTs. Like Brown, Enochs and Yoshitake-Strain (1999) and Yamashita (1996) reported that their Multiple-Choice DCTs had

relatively low reliabilities ranging from .47 to .56; in contrast, Liu (2006, 2007) and Roever (2005, 2006) reported that their reliabilities for the Multiple-Choice DCTs were acceptably high at .83 and .91, respectively. This discrepancy in reliability estimates may be due to the fact that items and distractors in the Multiple-Choice DCT are simply more difficult to construct than other measures of pragmatics (Liu, 2007; Roever, 2006) or the fact that each test consists of varying number of items and test takers of different proficiency because internal consistency reliability estimates are significantly affected by test length, difficulty of test, and test score variance (Bachman, 1990).

When students' performance is directly observed and scored by human raters as in the written DCTs, oral DCTs, and role-play tasks, inter-rater reliability and intra-rater reliability also come into play. The former is related to the extent to which two or more raters agree with each other on the score they award to each test taker (or performance), and the latter is concerned about the extent to which the same rater awards the same score to the same individual (or performance) over a period of time. Inter-rater reliability of test scores in pragmatic assessment has been relatively well established (Brown, 2008), indicating that multiple raters tend to rate the test takers in the same order. However, high inter-rater reliability alone does not guarantee that raters award the same scores to each test taker: it is likely that some raters consistently award scores more harshly or leniently than others.

In order to deal with rater bias and intra-rater reliability issues, the Multifaceted Rasch model has been applied in many studies (Brown &Ahn, 2011; Liu, this volume; Roever, 2008; Tajeddin & Alemi, this volume; Youn, 2007) in the pragmatic assessment literature. They have suggested that different raters showed varying degrees of severity on their rating across different task types but rater training can minimize rater biases and enhance consistency in their rating. Taguchi (2011) reports on inter-rater reliability in a small, regionally and culturally diverse set of raters (n=4). Pragmatic variation can be found even within members of the same speech community (Félix-Brasdefer & Koike, 2012). Such variation complicates the notion of inter-rater reliability for pragmatics assessment (Ishihara, in press).

3.2. Validity in pragmatic assessments

Compared to reliability, validity has not yet been fully investigated in the pragmatic assessment literature. Validity refers to the degree of appropriateness of interpretation and use of test scores. Depending on the type of evidence required to support intended test score interpretations, validity is often divided into construct, content, and criterion-related evidence of validity (Bachman, 1990).

Construct validity is at the heart of language assessment because it pertains to "the meaningfulness and appropriateness of the interpretations that we make on the basis of test scores" (Bachman & Palmer, 1996, p.21). Construct validity can be established based on differential group comparison which seeks for the statistically significant mean difference in a test score found between different proficiency level groups (Brown, 2005). Correlational approaches are also commonly used to examine the construct validity of a test. Roever (2006) used both approaches to show that his 36 items measuring (12 each of) implicature, routine, and speech acts reflect the construct of pragmalinguistic knowledge given the fact that native-speakers significantly scored higher than non-native speakers, and subsection scores are moderately correlated suggesting that three distinct pragmalinguistic constructs exist with some

overlap. In contrast, Rose (1994) found that a DCT may not be a cross-culturally valid means of assessing pragmatic knowledge. He argued that the assumption that the same DCT scenarios can be used in both Western and non-Western contexts to elicit the same speech acts, just as they were in the cross-Indo European elicitations of the CCSARP (Blum-Kulka, House, & Kasper, 1989), did not hold. In standard DCT format which does not allow opting out, respondents would be coerced into performing speech acts that they would not perform in such contexts. Rose argued that multiple choice DCTs which included opting out as one response could be more revealing (see Bardovi-Harlig, 1999a on scenario construction). Brown (2001) also reported that DCTs and other pragmatic assessment tools failed to elicit consistent performance from different tasks tapping into the same aspects of pragmatic competence. Based on exploratory factor analysis, he found that pragmatic assessment tools that require production such as written DCTs, oral DCTs, and role plays, on the one hand, and pragmatic assessment instruments that are receptive such as multiple choice DCTs and self-evaluations, on the other, constituted unique factors as an indication of strong test method factors. Similarly, Hudson (2001) also found a strong method effect among his three measures (the written DCT, the language lab DCT, and the role play) of the speech acts of requests, refusals, and apologies. He showed that there was a relatively high correlation between the written DCT and the language lab DCT for the speech acts of request and refusal. In contrast, the role play did not correlate highly with them for the same speech acts.

Despite such findings, validation studies in pragmatic assessments are still underrepresented and in their initial stage. In order to fully investigate construct validity, further research is needed, including application of confirmatory factor analysis models, and more specifically, multi-trait and multi-method (MTMM) methods, which would provide possible explanations for varying observed performance across different tasks tapping various aspects of pragmatic knowledge.

Content validity can be established by experts' subjective judgment about the degree of representativeness of the test content compared to a target language use (TLU) domain which could be real world tasks or curriculum. Thus, the authenticity of a test is related to content validity in that it is defined as "the degree of correspondence of the characteristics of a given language test task to the features of a TLU [target language use] task" (Bachman & Palmer, 1996, p. 23). Compared to many previous studies investigating the construct validity of pragmatic assessments, to date, very little research has been conducted to systematically establish the content validity of a pragmatic assessment. This is partly because pragmatic assessment has typically been used as a research instrument, not as an assessment tool, and this does not require test developers to demonstrate whether test content matches the content of the non-testing situation. Among the few studies that have been done to examine the content validity of a pragmatic assessment, Brown (2001) argued that the content validity had been established by Hudson, Detmer, and Brown (1992, 1995). Brown argued that the six types of pragmatic tests (oral, written, and multiple-choice DCTS, and role plays, plus two self-assessments), planned and designed with care, captured learners' sociopragmatic and pragmalinguistic abilities broadly because they covered three different speech acts (requests, refusals, and apologies) considering three contextual factors: relative power, degree of imposition, and social distance. Nonetheless, the content validity of the six tests in their studies of EFL and Japanese as a second language has not yet been systematically investigated based on real data such as expert judgments on tasks relevance or corpus analysis, which makes their argument for content validity relatively weak.

Liu (2007), on the other hand, established content validity of the situations and options of his multiple-choice DCT of apologies in his study through exemplar generation, situation likelihood investigation, metapragmatic assessment, and verbal protocol analysis by Chinese EFL learners.

Criterion-related validity involves testers comparing their developed assessment tool to an external criterion to see whether there is positive correlation between one test and others meant to measure the same construct. Criterion-related validity is subdivided into concurrent validity and predictive validity (Bachman, 1990). Concurrent validity can be examined by comparing the scores of a given test to that of similar tests. If the same test takers score within the same range consistently on both tests, concurrent validity could be said to be high. Brown (2001) found that scores obtained from six measures of pragmatic competence used in Hudson, Detmer, and Brown (1992, 1995) did not correlate highly enough with each other to claim that the pragmatic tests are concurrently valid. Similarly, Sasaki (1998) demonstrated that the correlation between the appropriateness scores of production questionnaires and role plays was too low to warrant concurrent validity of those two pragmatic measures.

Another aspect of criterion-related validity is predictive validity which is defined as how well a given test will predict future performance in a given domain. To see whether predictive validity holds for a given test as a means of valid measure of pragmatic competence, it would be necessary to compare pragmatic assessments in question with naturally occurring data. There are a few previous studies investigating this issue, mostly resulting in the low predictive validity of DCTs. For example, Beebe and Cummings (1996) found that a DCT and real telephone requests elicited different level of elaboration of refusals, suggesting that DCTs provided only limited data. Likewise, Hartford and Bardovi-Harlig (1992b) compared semantic formulas for refusals elicited from DCT scenarios of academic advising sessions with those resulting from authentic advising sessions, and found that the DCT failed to elicit the broad range of semantic formulas and extended negotiations which are common in naturally occurring data.

3.3. Practicality in pragmatic assessments

Lastly, practicality needs to be considered to determine the test quality. It is defined as "the relationship between the resources that will be required in the design, development, and use of the test, and the resources that will be available for these activities" (Bachman & Palmer, 1996, p. 36). If required resources of a given test are greater than available resources, no matter how reliable and valid it may be, it would not be practical and consequently, would be unlikely to be implemented. Thus, it is important to estimate the human and material resources, time, and cost associated with developing or administering a given test at the initial stage. The practicality of pragmatic assessment has been discussed in terms of test development, administration, and scoring in previous studies (Brown, 2001; Liu, 2007; Roever, 2006). However, it would be very difficult to determine test practicality in advance without allowing for its intended purpose and use. For example, although a multiple-choice DCT would be time-efficient and cost-effective in terms of scoring, and thus make a large-scale test possible, at the same time, as has been argued in previous studies (Liu, 2007; Roever, 2006), it is difficult to come up with reliable and valid options and distractors for every item. In contrast, written and oral DCTs are relatively easy to create and could elicit actual written or oral responses from test takers, but could be more expensive to administer and score. (But see Bardovi-Harlig, 1999a, 2009 for the ethnographic

requirements of developing a DCT, and Rose, 1994 and Liu 2006, 2007 for exemplar generation.)

3.4. Beyond traditional measures of pragmatic competence

As can be seen above, pragmatic assessment types with long traditions of use such as multiple-choice DCTs, oral DCTs, written DCTs, role plays, and self-assessments have turned out to differ in their reliability, validity, and practicality. Most of them still fall short, unable to demonstrate the qualities that make tests satisfactory for high-stakes decisions. However, these criteria cannot be satisfied to the same degree; we need to strike a balance among reliability, validity, and practicality depending on the intended purpose and use of a given test (Bachman, 2005). Thus, increasing attention needs to be paid to a test's intended uses and consequences (Bachman & Palmer, 2010) which will help us to develop, select, use, and evaluate a pragmatic assessment tool in a purposeful manner.

4. Assessing Interest in Pragmatics Assessment: Interviews with Language Experts

After evaluating existing tests of pragmatics, we interviewed colleagues who are experts in interlanguage pragmatics to ascertain the level of interest in additional formats for testing pragmatics and to learn what language programs were doing that has not been reported in the literature. The colleagues who agreed to be interviewed were either language coordinators or had curricular responsibilities and were knowledgeable about the curriculum in their respective language programs, Russian, Spanish, and Swahili.

The open-ended interviews highlighted three recurrent themes relevant to further test development: 1) testing is an integral part of teaching; 2) teachers are implementing a variety of assessments; and, 3) there is an interest and need for pragmatics tests.

4.1. Testing is an integral part of teaching

What is taught should be tested. Testing what is taught is no different in pragmatics than any other area of language instruction. Our colleagues reminded us that testing pragmatics helps students take the teaching and learning of pragmatics seriously. They reported that there is a difference in pragmatic performance when students are tested because students are obligated to pay attention. These teachers rely on testing to achieve planned positive washback effects.

Nevertheless, such planned assessment has often been absent from materials development in pragmatics (e.g., Bardovi-Harlig& Mahan-Taylor, 2003). An explanation may be found in early papers on pedagogy in pragmatics which sought solely to raise awareness and increase comprehension of what native speakers might say, mean, or expect, and advocated individual learner choice when it came to performance. However, early proponents for the teaching and learning of pragmatics advocated pragmatics testing by the late 1980s and early 1990s. Lawrence Bouton, Professor Emeritus and co-founder of the International Conference on Pragmatics and Language Learning at University of Illinois at Urbana-Champaign, was an early proponent of testing pragmatics. The close relationship between teaching and testing has recently has been rearticulated by Ishihara and Cohen (2010).

4.2. Teacher implementation of pragmatic assessment

We learned that three different levels of testing are in use in our cohort language departments: on-the-spot tests, distributed graded assignments, and examinations. One language coordinator reported using what might be called on-the-spot assessment: Students are challenged to greet appropriately in Swahili whenever they see an instructor. In contrast to the relatively short greetings for Indo European languages, Swahili greetings can take up to 10 turns to signal interest in further communication, an aspect of conversation that American learners must master. Students can expect a "pop quiz" by engaging in greetings whenever they meet an instructor. Speaking Swahili in an English-speaking context whenever one encounters another speaker of Swahili is similar to an impromptu role play for the students.

A second practice is in use in Russian, with five conversation assignments distributed throughout the semester. The language coordinator advocates the integration of sociocultural and speech act knowledge into instruction throughout the semester, with lots of practice. Students conduct five recordings of themselves with a partner throughout the semester. They are given descriptions, narrations, speech acts, and genres. Low-level students combine open conversation with a list of formulas and expressions. 10% of the grade comes from the conversational recordings. Students receive feedback on the conversations for formative assessment which includes feedback on the pragmatics such as "you didn't use a formula to say goodbye or [didn't use] an opening." Because these conversations can be planned and lists of expressions are provided, students can draw on explicit as well as implicit knowledge to prepare for them. The assignments are less like tests in the classic sense, and more like the learning activities advocated by Ishihara and Cohen (2010).

A third approach is used in Spanish, where students are exposed to pragmatics in upper-level undergraduate linguistics classes and are evaluated through examinations. Speech acts are introduced as communicative actions covering both pragmalinguistic and sociopragmatic aspects, and oral practice includes online role play for communicative practice. Role plays performed in class receive formative feedback. The final exam has a separate section on pragmatics which includes identifying and recognizing speech acts, sequential analysis of speech acts, and cross-cultural comparison (one role play in Spanish and one in English). These assessments include both activities which assess L2 performance—by use of the classic role play format—and L2 metapragmatic knowledge—by comparison and analysis of conversation samples.

4.3. An interest in and need for pragmatics tests

Teacher implementation of pragmatics assessment as part of teaching L2 pragmatics underscores an interest in pragmatics testing. The fact that, as of yet, no pragmatics items are used in placement tests in any of the programs underscores the need for the development of new formats in the testing of pragmatics which can be implemented with current testing formats. Along with practicality, reliability of new formats must be established. Two of the instructors compare pretest-posttest scores to assess the impact of instruction, although pretests do not impact grades. This suggests that although pretest-posttest comparisons are being used to assess teaching effectiveness to inform teacher-decisions, they are not well-enough established for teachers to use them for summative assessment in student grades.

Our small, informal survey encouraged us to continue to the third step of test development which is to propose new test formats that are both revealing of pragmatic knowledge and are practical to implement.

5. Exploring Tasks from Pragmatics Research for Use in Testing

In this section we explore six possible item formats for testing second language pragmatics. These test formats are derived from the published literature on pragmatics and have been used in research studies. They present tasks which address aspects of pragmatics which have been investigated independently of language assessment, and they are all highly controlled. They can either be easily scored or can be modified to be easily scored. We evaluate each task for its potential content validity and practicality. An assessment of reliability will have to wait until we run the candidate formats in testing conditions.

By format, the tasks that we present for consideration include oral production (oral for oral), written production (written for written), and audio and/or audio-visual conversational excerpts with written/read interpretations (Table 1). By area of pragmatics, the tasks cover conventional expressions, pragmatic routines, conversational implicature, pragmaticality judgments, sociopragmatic judgments, interaction of grammar and pragmatics, and speech act identification tasks. The production tasks simulate turn taking by providing unanticipated turns through computer generation or audio presentation, requiring responses from the test takers.

Table 1. Sources, pragmatic focus, and format of candidate tasks from pragmatics research

Task	Component of pragmatics targeted	Format
Conversation simulations (a) (Bardovi-Harlig, et al, 2010)	Responding to turns; use of conventional expressions	Aural input/oral production
Conversation simulations (b) (Bardovi-Harlig, Mossman, &Vellenga, in press)	Responding to turns; Academic discussion; use of pragmatic routines	Aural input/oral production
Written exchange (Kuha, 1997)	Responding to computer-generated turns, speech act (originally complaints); multiple turns	Written
Conversational Implicature Task (Taguchi, 2005, 2008)	Interpretation of conversational implicature; listening comprehension (needed for conversation); processing of prosodic information	Aural presentation with written/read interpretation
Speech Act interpretation task (Koike, 1989, 1996)	Intended (illocutionary) and perceived (perlocutionary) force of utterances; grammar and pragmatics	Audio and Audio-Visual presentation with written/read interpretation

Pragmatic acceptability judgment tasks (Bardovi-Harlig&Dörnyei, 1998)	Sociopragmatic knowledge; pragmalinguistics; interaction of grammar and pragmatics	Audio-Visual presentation with written/read interpretation (two presentations)
Prediction task (based on Koike, 1996 and Bardovi-Harlig, 2009)	Sociopragmatic knowledge (which contexts require what speech act), independent of linguistic form (pragmalinguistics)	Written, audio, or audio visual input
Aural multiple choice task (Teng & Fei, 2013)	Sociopragmatic and pragmalinguistic knowledge	Aural presentation of distractors

In the following sections, each task type is reviewed in turn, and modifications are suggested where relevant. We begin by considering the production tasks first, then move on to consider interpretation, judgment, and prediction tasks.

5.1.Conversation simulations: Oral DCTs with turns

Oral production and simulated turn taking are features of the computer-delivered timed aural-oral DCT. Following a very brief scene setting statement (called a *scenario* in pragmatics research) participants hear a conversational excerpt over headphones and rapidly respond to an aural turn (which they do not read). Listening to an interlocutor, and then responding rapidly simulates conversation. With an interlocutor turn, there's no need for extensive scenario explication (which slows down the testing and involves more reading or listening than speaking skills). The interlocutor's turn constrains the learner's response, even in DCTs (Bardovi-Harlig& Hartford, 1993). Examples (1) and (2) were developed to investigate the use of conventional expressions in L2 pragmatics (Bardovi-Harlig, 2009; Bardovi-Harlig et al, 2010). Examples (3) and (4) were developed to test the effectiveness of teaching pragmatic routines for academic discussion (Bardovi-Harlig, Mossman, &Vellenga, in press). Computer-delivery keeps the pace and the short spoken and read scenarios help learners who are not strong readers. The short scenarios also keep the focus on conversation. Learner responses are recorded into mp3 files on the computer. Learners were given 7 seconds to respond with conventional expressions (Examples 1 and 2) and 10 seconds to respond to academic discussion items because they were asked to assimilate opinions, given to them either as statements "you have the same opinion" or through their classmate's spoken turn (Examples 3 and 4).

(1) Contexts for conventional expressions (introduction)
Your friend introduces you to his new roommate.
(Aural only): "This is my new roommate, Bill."
(On screen only) You say:
[Target: *Nice to meet you*]

(2) Contexts for conventional expressions (declining assistance)
You go to a clothing store and you need to find a new shirt. A salesperson approaches you. You don't want the salesperson's assistance.
(Aural only): "Can I help you?"
(On screen only) You say:
[Target: *No thanks, I'm just looking* or <*thanks*>*I'm just looking* <*thanks*>]

(3) Context for academic discussion: Agreement item
Narrator (visual and audio):
Your group is discussing the way that people communicate. You have the same opinion as your classmate.
Classmate's turn (audio only): People spend too much time talking on the phone these days.
[Screen only] You say:
[Target: Agreeing expressions including *I agree (with), Good point, That's right, You're right, That's true*]

(4) Context for academic discussion: Disagreement item
Narrator (visual and audio):
Your group is talking about the news and media. You think that newspapers like *The New York Times*and *The London Times* are still very important.
Classmate's turn (audio only): Nobody reads newspapers these days.
[Screen only] You say:
Target: Disagreeing expressions including *Yeah but, Okay but, I agree but*]

These items can be analyzed from a number of perspectives. At the broadest, they use spoken turns to constrain learner answers rather than lengthy descriptions. This increases the approximation to conversation in that the items are aural-oral. They seek to elicit specific expressions from respondents. Examples (1) and (2) were developed from extensive field work and piloting in a specific speech community in the American Midwest. The expressions were selected because native speakers in the speech community used one expression to the exclusion of others, making the target very clear. Evaluation of these items can be done with a check list marking whether the expression is present or not. (This is how the initial analysis for Bardovi-Harlig & Vellenga, 2012, was conducted).

The pragmatic routines for the academic discussion contexts were identified in an academic corpus, the Michigan Corpus for Academic Spoken English (MICASE) and are less constrained by context than the conventional expressions. There are multiple correct answers as long as the response carries an overt agreement (in agreement contexts) or disagreement (in disagreement contexts). A scoring complication arises in the use of low-use disagreement markers: The corpus reveals that *I disagree* and *I don't agree* are clearly dispreferred by speakers in the academic setting recorded by MICASE, occurring only 4 times/million words, whereas *Yeah but* and *Okay but,* appeared in MICASE 120+ times/million words and 90+ words/million words, respectively. We would propose giving less credit or no credit for these routines.

The simulated conversation format of these oral DCTs can be adapted to a range of speech acts, conversational structures, and pragmatic routines. They are timed, constrained, and oral for oral. Using timed turns not only asks learners to respond at a conversational rate, but it also keeps the length of responses relatively short (in keeping with a conversational turn, in contrast to written responses) and is likely to call on implicit knowledge. Oral production, time pressure, and the opportunity to produce a response to a turn increases the content validity of simulated conversations. Practicality is high because once authentic models are found, the task is easy to make and can bedelivered via computer for clear listening and recording of responses. Scoring can either be done after transcribing (which is time consuming) or by check list which eliminates transcription.

5.2. Written turn-taking production tasks

An automated computer-delivered interactive task called the Interactive DCT was designed by Kuha (1997) to investigate complaints in an interactive task which read typed responses and allowed respondents to perform multiple turns. The Interactive DCT offered a technology-based solution to the problem of turn-taking. Kuha structured her computer program to simulate a disagreeing interlocutor so that it would be possible for the respondent to complain in each of the three turns allowed. The computer is programmed to look for key words. If a respondent begins with a greeting, the computer greets too. Then the computer begins a hierarchical search. It searches first for threats. For example, in the barking dog scenario (Example 5), it searches for strings such as "shoot it," replying "I don't like threats." It next searches for requests for redress. Because the computer always disagrees, requests for redress receive an argument against the premise in the first turn, and a refusal to comply in the second turn. Problem statements also receive an argument against the premise or the computer will shift the blame. Requests for information receive the response "Why do you say that?" Finally, if no key words are present, the computer will reply with the default response "Is something wrong?" An example of a completed computer "conversation" appears in Example (5).

5. Interactive Computer Task (NS respondent)

First Screen Your neighbor's dog has been barking all night for the past week, and you are not getting any rest. This morning, you and your neighbor happen to come out at the same time to pick up the morning paper.

You say:

Second Screen You say: Hey, doesn't your dog ever sleep?

Your neighbor says: "Why do you say that?"

You say:

Third Screen You say: Well it just happens that I hear him barking quite a bit through the night. Don't you hear him?

	Your neighbor says: "This is a pretty noisy neighborhood anyway."
	You say:
Final turn	You say: Not that noisy. Maybe you should get him some doggy sleeping pills.

The Interactive DCT eliminates the need for an interlocutor. It also promotes consistency in interlocutor turns, because the computer-as-interlocutor has a limited repertoire.

With the advent of computer- and device-mediated contexts in which interlocutors communicate via writing, it is now possible to simulate written communication in written-for-written format. Such adaptations will retain the advantage of scoring written responses, and if done right, will gain the advantage of authenticity which written-for-oral tasks lack. Modifying this task for written-for-written rather than written-for-oral would be to change the scenario from meeting face-to-face to meeting online: "This morning, you are online (or on Skype or G-Chat or SMS) and notice that your neighbor has just logged on too." Written-for-oral tasks (the most notable of which is the written DCT in which participants write what they believe they would say in response to a scenario they have read) dominate written tasks in pragmatics. Of 152 published articles surveyed by Bardovi-Harlig (2010), 57 studies or 37.5% used a written task exclusively. Of those, only six examined authentic written events. (A new written-for-written test which evaluates EFL teacher-trainees' recognition of a writer's stance on a controversial issue in newspaper editorials and their ability to explicitly identify and label linguistic cues to stance was developed by Ifantidou & Tzanne, 2012). The remaining 51 of 57 studies, or 89.5% of the written studies, used writing to explore characteristics of spoken language. Cohen and Shively (2007, p. 196) elegantly described this practice as "an indirect means for assessing spoken language in the form of a written production measure."

To test social language use, we can a) set up a context which is credible (SMS, emailing, chat), b) control the turns of the interlocutor to create highly similar contexts for test-takers, and c) score by computer. Carr and Xi (2010) demonstrated that key expressions for short-answer tasks can be searched and scored by automated scoring systems. Written-for-written tasks have much higher potential for content validity than written-for-oral tasks do. Written data have long been valued in pragmatics for their practicality in avoiding transcription, and now technology has progressed to the point where machine scoring could further enhance practicality.

5.3. Audio and video presentations responded to in paper and pencil format

Koike (1989, 1996) investigated the intended illocutionary force conventionally associated with syntactic forms. Illocutionary meaning and its relation to grammar/form is investigated in these tasks. Focusing on English-speaking learners learning Spanish suggestions (other speech acts are also included) Koike investigated learners' ability to recognize the negative question in Spanish as a neutral suggestion instead of the English interpretation of the form as a criticism (roughly equivalent to the distinction in English between *Why didn't you X*? vs. *Why not X?*). In Koike's (1996) task, students were given a scenario, following which they watched a short video.

Upon viewing, they answered three questions. The second question—identification of the speech act—is discussed here. The video presentation was used to capture body movements and facial expressions as well as audio information (Koike, 1996) and built on the audio presentation of the earlier study which delivered verbal and prosodic cues only (Koike, 1989).

6. Situation (L1): You go to see your Spanish instructor in her office because you are having some trouble.

(View: Professor sitting in office. Female-Colombia)

(Presented in L2) Professor: Hola! Pasa. Sí, tengo tiempo para hablarte ahora. ¿Está preocupado con la nota que sacaste en el examen parcial? Pero no fue un examen muy difícil. Em—¿no has pensado en estudiar junto con tus colegas de la clase ahora? ["Hi! Come in. Yes, I have time to talk to you now. Are you worried about the grade on the midterm exam? But it wasn't a very hard exam. Um—have you thought about studying together with your fellow students in the class now?" Translation is not provided to the students]

What was the main goal of what the speaker said? (Circle one)

A. Invitation	E. Order
B. Apology	F. Information question
C. Request	G. Mild rebuke
D. Suggestion	H. Other _____

The task was developed when Koike identified a specific problem for native speakers of English learning Spanish (i.e., nonequivalence of negative questions as suggestions), but other trouble spots can be identified. In English, for example, one could test for the illocutionary force of conventional expressions which are either multifunctional in American English such as *I'm sorry* which functions as apology, condolences, and an alerter ("excuse me") or which are perceived to be multifunctional by learners as in *You're welcome* which learners identify both as a response to *Thank you* and as a greeting "You are welcome to my home" (Bardovi-Harlig, 2014). The video presentation contributes to the content validity in that test-takers assess the illocutionary force of utterances using both language and non-verbal cues. The scoring of the multiple choice test and the wide availability of digital recording and playback devices speaks to high practicality.

5.4. Tests of implicature

Conversational implicature (Grice, 1975; Levinson, 1983) is one component of pragmatics that received relatively little attention in early interlanguage pragmatics research, but has received increasingly more attention in the past decade. Bouton's (1992, 1994) early work on L2 implicature by advanced (college-matriculated) learners of English employed written multiple-choice tasks. Like Bouton, Roever (2005, 2006) used written multiple-choice tasks in the interpretation of implicature, moving from pencil and paper format to a web-based test.

7. Jack is talking to his housemate Sarah about another housemate, Frank.
Jack: 'Do you know where Frank is, Sarah?'
Sarah: 'Well, I heard music from his room earlier.'

What does Sarah probably mean?
1. Frank forgot to turn the music off.
2. Frank's loud music bothers Sarah.
3. Frank is probably in his room.
4. Sarah doesn't know where Frank is."

Taguchi's oral comprehension task, called the *pragmatic listening task* (Taguchi, 2005, 2008) simulates online conversational processing of implicature. As in the earlier written versions, Taguchi (2005) used a multiple-choice format with items adapted from earlier work on implicature (Bouton, 1992, 1994; Holtgraves, 1999). Each item contained a short dialogue spoken by male and female native English speakers. The final reply of the sequence provides an answer to the speaker's question which learners were asked to interpret (Example 8).

8. Indirect Opinion (negative); Taguchi (2005)

Ben: Good morning, honey. I can't believe I fell asleep in the middle of the movie last night. Did you watch it till the end?
Barbara: Yeah, I did.
Ben: How was it? Did you like it?
Barbara: Well, I was glad when it was over.
Question: What did Barbara think about the movie?
(a) She thought the movie was good.
(b) She didn't enjoy the movie. (Correct answer)
(c) She thinks Ben should have watched the movie.
(d) She liked the end of the movie.

Taguchi's (2008) study adapted the instrument from her earlier study (2005); in the later version each dialogue was followed by a yes-no question to check the participants' comprehension of the speaker's intention (Examples 9-10). In the Yes/no version of Example 8 the learners are asked "Q: Did the woman like the movie?"

9. Indirect Refusal

John: Hey, Mary, where are you?
Mary: I'm in the kitchen.
John: Hey, ah . . . . . could you clean the house this weekend? I cleaned it the last two weeks, and this weekend I have plans.
Mary: Oh, ah . . . I'm going to see my parents this weekend. I won't be back until Monday.

Question: Can the woman clean the house?

10. Indirect Opinion (positive)

Jane: Dr. White, do you have time?
Dr. White: Sure. Come in.
Jane: Ah. . . . did you have a chance to read my book report? It was my first time to write a book report, so I'd like to know how I did on it.
Dr. White: Oh, it's exactly what I wanted.

Q: Does Dr. White like the book report?

The yes/no format is even easier to score than the original multiple choice options used by Bouton, and subsequently by Roever, although both Bouton and Roever's items can be delivered aurally with written options as Taguchi (2005) did.

Similar multiple-choice items testing implicature are found in the TOEFL listening section (*TOEFL Planner*, 2010, pp. 52-55). In one example, a 30-turn conversation is played, then test takers are asked to complete 5 multiple choice comprehension questions, the last two of which are pragmatics items. In Example (11), test-takers are asked to interpret a conversational implicature and Example (12) a closing which is a conversational structuring move. In both cases, the relevant turns are repeated for interpretation. The larger context comes from the full conversation which the test takers heard at the beginning of the section.

11. Read part of the conversation again. Then answer the question.
(Female student) I'm sorry I had to miss practice, though. I feel bad about that.
(Male coach) Family's very important.
What does he mean when he says "Family is very important."
a. He hopes the woman's family is doing well.
b. He would like to meet the woman's family.
c. The woman should spend more time with her family.
d. The woman had a good reason for missing practice.

12. Why does the coach say: "Good. That's all the news there is. I think that's it for now."
a. He wants to know if the woman understood his point.
b. He wants the woman to act immediately.
c. He is preparing to change the topic.
d. He is ready to end the conversation.

The issue of written versus aural modes is as important an issue in interpretation as in production. Conversations are spoken, not read, and conversational implicature will most likely be encountered aurally. Outside of the ease of data collection, there is nothing "neutral" about favoring literacy over oracy. The preference for written presentation may be a result of populations previously tested, but other learner populations may do better with aural stimuli. The aural presentation of conversations used by Taguchi (2005) has greater potential for increased content validity because both are in the oral mode and because all test-takers respond to an identical signal in real time, whereas the written conversations used by other researchers in implicature tasks have lower content validity due to mismatched mode and the possibility for

imagined delivery of the speaker turns. As with other multiple choice formats, practicality is high for scoring.

5.5. Pragmaticality judgment tasks

A pragmaticality judgment task is akin to a grammaticality judgment task better known in other areas of SLA research. A judgment test is thought to access a learners'/speakers' underlying system without requiring production. A pragmatics acceptability judgment task asks learners to assess the appropriateness of another speaker, thus not involving a learner's own confidence (as in the self-assessment tasks used by Hudson, Detmer, & Brown, 1995). The follow example is taken from a video judgment task used by Bardovi-Harlig and Dörnyei (1998) to test pragmatic awareness. The video judgment task gives learners access to facial expressions and relative location of speakers as well as prosody, notably stress, intonation, rate, and loudness. The original task asked learners to rate the acceptability of the item holistically as shown in Example (13). (13) is grammatical but pragmatically inappropriate given the student's statement "I would like you to fill this out" addressed to her teacher.

13. [Student hears]: Anna goes to ask her teacher to fill in a questionnaire. She knocks on the office door.
[Student sees: The teacher is seated at her computer in her office typing. Anna knocks.]
A: (knocks on the door)
T: Yes, come in.
A: Hello. My name is Anna Kovacs. If you don't mind, I would like you to fill this in for me. [Anna hands her teacher the questionnaire]

[answer sheet]

Hello. My name is Anna Kovacs. If you don't mind, I would like you to fill this in for me.

Was the last part appropriate/correct? Yes No
If there was a problem, how bad do you think it was? Not bad at all __:__:__:__:__:__ Very bad

In a replication of Bardovi-Harlig and Dörnyei (1998), Niezgoda and Roever (2001) reported the reliability of the pragmatic severity ratings to be satisfactory at α=.73. The data from the higher level Czech learners was more reliable (α=.78) than that of lower level ESL learners (α=.55), indicating that proficiency level is likely to be a factor. There was a considerable investment in making the original video instrument, but both task administration and scoring have high practicality. It should be noted that video production and editing has become much easier since the original video was recorded in 1996, thus increasing the overall practicality. As noted earlier, judgments of audio-video conversations have the potential for higher content validity than written judgment tasks.

Parts of an utterance can also be identified as having an error as in TOEFL grammar items. This may inherently link grammar and pragmatics in a way that other tasks do not. (See Grabowski, 2008, for a recent admonition to link grammar and pragmatics in testing.)

For a test item of this type, Example (13) could be converted into Example (14) which emulates early TOEFL format for grammaticality judgments. Example (14) shows an item that native speakers and ESL students reported as unacceptable; Example (15) is an acceptable item.

14. [Student hears]: Anna goes to ask her teacher to fill in a questionnaire. She knocks on the office door.
[Student sees: The teacher is seated at her computer in her office typing. Anna knocks.]
A: (knocks on the door)
T: Yes, come in.
A: Hello. My name is Anna Kovacs. If you don't mind, I would like you to fill this in for me. [Anna hands her teacher the questionnaire]

[On response sheet] How do you assess Anna's request?
Hello. My name is Anna Kovacs. If you don't mind, I would like you to fill this in for me
A B C D

Where is the problem? No problem b) B c) C d) D e) A-D

15. [Student hears]: Peter's teacher wants to talk to Peter about the class party. Peter makes arrangements to come back.

[Student sees]: The teacher seated at his desk, speaking to Peter in the front row of class.

T: Peter, we need to talk about the class party soon.
P: Yeah, if tomorrow is good for you, I could come any time you say.

[On response sheet] How do you assess Peter's reply?

Yeah, if tomorrow is good for you, I could come any time you say.
A B C

Where is the problem?
a. No problem b) A c) B d) C e) A-C

Example (15) is appropriate, although it is anticipated from learner self-report that *could* may be understood as marking past time rather than as a mitigator (Bardovi-Harlig, 1999b).

The pragmaticality judgment task offers learners both visual and aural input. The video presentation situates the speech acts in a context that included speakers as visible actors, thereby enhancing content validity. The original task was timed to increase the likelihood that learners would respond by feel rather than by rule. The additional presentation of a written target sentence (which could increase the potential for explicit reflection) is balanced by the timed task and locating the infelicity by its general position in the utterance rather than requiring a close analysis.

5.6. Sociopragmatic prediction tasks

A sociopragmatic task can be created from a hybrid of production tasks and Koike's (1996) list of speech acts from her identification task. This task tests learners' sociopragmatics (knowledge of what should be said in a situation) independently of their knowledge of how to say it. Saying the same thing as proficient speakers in the target language depends on how a situation is interpreted (Bardovi-Harlig, 2009). For example, a scenario developed from an American academic context (Example 16) yields expressions of gratitude from native-speakers (in both authentic interaction and in response to the item derived from it), but yields both expressions of gratitude and apology from learners (Bardovi-Harlig, 2009; Bardovi-Harlig, Rose, & Nickels, 2008). The ability to choose (c) "thank her" over (b) "apologize" would demonstrate a learner's alignment with the sociopragmatics of the American Midwest where thanking is a common move in closing university office hours (see also Hartford &Bardovi-Harlig, 1992a).

This prediction format would test learners' sociopragmatic knowledge, without requiring speech act production.

16. You stop by your teacher's office to ask a question about the assignment. She takes time to answer your question. You know she is very busy, so before you say good-bye, you…

A. make another appointment
B. apologize
C. thank her
D. offer to help her
E. leave

There are many examples throughout the literature where learners and native speakers produce different speech acts or semantic formulas. Most of them can be described with no technical language and could be used to model this type of items as well as others. This format is highly practical for both presentation and test-taker multiple-choice response, and can be easily scored. Because the answer is a sociopragmatic assessment of what is called for in the situation rather than a performance of a conversation turn, the written format does not negatively affect content validity, but using a timed version of the test would increase likelihood of response by feel (as in a conversation) than by explicit knowledge, thus enhancing content validity.

5.7. Oral multiple choice tests: A new twist on an old standard

A new twist on an old standard, the multiple choice task, was introduced for the teaching of Chinese by Teng and Fei (2013). In the oral multiple choice task, learners listen to the options and select the most appropriate for the situation. As can be seen in Example (17), Teng and Fei provided written responses as well (in both pin-yin and Chinese characters) for the purpose of teaching. For the purposes of testing, we would recommend aural answers only. The format would include the scenario followed by the letters of the four distractors and the icon which is clicked by the respondent in order to hear the spoken utterance.

17. You can't meet the deadline for a term paper and want to ask your professor for an extension. Which of the following is/are acceptable? (Note: There might be more than one answer for this situation. Choose all you think are appropriate.)

A Lǎoshi, wǒnéngbùnéngwǎnjǐtiānjiāo? 老师，我能不能晚几天交？

ISSN 2251-7324

B Lǎoshi, wǒwǎnjǐtiānjiāo, xíngbùxíng? 老师，我晚几天交，行不行？

C Lǎoshi, wǒháiyàojǐtiānshíjiāncáinéngxiěwán. 老师，我还要几天时间才能写完。

D Lǎoshi, wǒzhèjǐtiānbìng le. Wǒhòutiānjiāo, xíng ma?老师，我这几天病了。我后天，行吗？

This format is highly practical for both presentation and test-taker multiple-choice response, and can be easily scored. Because the learners listens to the spoken options which include the nuances of spoken language are thus richer than written responses, the aural format greatly enhances the content validity of the items. The items format presented here is set up for self-paced response (because it comes from a practice activity), but this can be converted to a timed task by presetting the presentation of the distractors.

This section presented seven different assessment formats which could be used to test different aspects of pragmatic knowledge. This set is not exhaustive. Others may wish to suggest additional types of items for consideration. We believe that it is important to broaden the inventory of types of pragmatics tests. This paper has undertaken three steps toward the development of new pragmatics test: evaluation of existing means of assessment, establishing need and interest in such tests, and suggestion new formats based on tasks in publish pragmatics research. The fourth step will be to pilot the new formats as tests to better examine their potential as assessment tools in light of empirical reports on their reliability, validity, and practicality.

6. Concluding Remarks

Traditional measures of L2 pragmatic competence have been in use since their introduction in 1992, but studies have found that they tend to be less reliable, less valid, and more limited than is ideal for assessment. Repeated testing of the same measures without resolution suggests that we must consider new item formats or resign ourselves to the problems we now face. Many new item types have been successfully used for L2 pragmatics research, and they have the potential to enhance task authenticity and practicality for testing, and to broaden our construct representation. However, in order to use them to make decisions about test takers, these items also have to be evaluated for reliability, validity, and practicality according to their intended purpose. For constructed response tasks, such as conversation simulations and written exchanges, scoring rubrics need to be carefully developed and inter/intra-rater reliabilities should also be investigated (when the yes/no scoring outlined earlier is not sufficient). Selected response tasks such as those testing conversational implicature, speech act interpretation, pragmaticality judgments, and ability to predict appropriate speech acts need to be checked for their internal consistency and validated through differential group comparison or correlational approaches.

In consideration of practicality for program-level assessment, the workload can be divided between class-level and program-level evaluation to achieve construct representation, with more time-intensive assessment taking place in class (see Ishihara & Cohen, 2010) and more controlled assessment at the program level. For example, two student learning outcomes in our program are to be able to express agreement and disagreement in an academic discussion, and to lead an academic discussion group for 10 minutes. Expressing agreement and disagreement can be assessed via a computer-delivered conversation simulation in the program assessment, whereas leading a discussion group can be assessed in class.

The increased interest in pragmatics among testers and the corresponding adoption of pragmatics assessment by teachers in classrooms show that it is time to move ahead toward development of new tests which go beyond the traditional measures. Integrating measures from pragmatics research provides a good starting point. Such new pragmatics measures would help testers to assess learners' pragmalinguistic knowledge for speech acts and semantic formulas, as well as sociopragmatic knowledge used in a variety of target language contexts; with additional refinement after implementation, a new bank of pragmatics test will allow assessment in a more reliable and valid manner. These improved L2 pragmatics testing practices will in turn lead to greater use of assessment of pragmatic ability at both classroom and program levels, resulting in more meaningful interpretations of learners' pragmatic knowledge, and hopefully better decisions about teaching practices to enhance learners' pragmatic ability.

Acknowledgments

We wish to thank César Félix-Brasdefer, Alwiya Omar, and Maria Shardakova for agreeing to be interviewed for this paper and sharing their expertise with us.

References

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

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*, 1–34.

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

Bachman, L. F., & Palmer, A.S. (2010). *Language Assessment in Practice*. Oxford: Oxford University Press.

Bardovi-Harlig, K. (1999a). Researching method. *Pragmatics and Language Learning*, *9*, 237-264.

Bardovi-Harlig, K. (1999b). The interlanguage of interlanguage pragmatics: A research agenda for acquisitional pragmatics. *Language Learning*, *49*, 677-713.

Bardovi-Harlig, K. (2009). Conventional expressions as a pragmalinguistic resource: Recognition and production of conventional expressions in L2 pragmatics. *Language Learning, 59*, 755-795.

Bardovi-Harlig, K. (2010). Exploring the pragmatics of interlanguage pragmatics: Definition by design.In A. Trosborg (Ed.) *Pragmatics across languages and cultures (Vol. 7 of Handbooks of pragmatics*; pp. 219-259).Berlin: Mouton de Gruyter.

Bardovi-Harlig, K. (2014). Awareness of meaning of conventional expressions in second language pragmatics. *Language Awareness, 23,* 41-56.

Bardovi-Harlig, K., Bastos, M.-T., Burghardt, B., Chappetto, E., Nickels, E., & Rose, M. (2010). The use of conventional expressions and utterance length in L2 pragmatics. *Pragmatics and Language Learning, 12,* 163-186.

Bardovi-Harlig, K., &Dörnyei, Z. (1998). Do language learners recognize pragmatic violations? Pragmatic vs. grammatical awareness in instructed L2 learning. *TESOL Quarterly*, *32*, 233-259.

Bardovi-Harlig, K., & Hartford, B. S. (1993c). Refining the DCT: Comparing open questionnaires and dialogue completion tasks. *Pragmatics and Language Learning*, *4*, 143-165.

Bardovi-Harlig, K., Mossman, S., &Vellenga, H. E.(in press).The effect of instruction on pragmatic routines in academic discussion. *Language Teaching Research.*

Bardovi-Harlig, K., Nickels, E., & Rose, M. (2008). The influence of first language and level of development in the use of conventional expressions of thanking, apologizing, and refusing. In M. Bowles, R. Foote, S. Perpiñán, & R. Bhatt (Eds.) *Selected Proceedings of the 2007 Second Language Research Forum* (pp. 113-130). Somerville, MA: Cascadilla Proceedings Project. **(**also available: http://www.lingref.com/cpp/slrf/2007/index.html)

Bardovi-Harlig, K., &Vellenga, H. E. (2012). The effect of instruction on conventional expressions in L2 pragmatics. *System, 40*, 1-13.

Beebe, L. M., & Cummings, M.C. (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: Mouton de Gruyter.

Blum-Kulka, S., House, J., & Kasper, G. (1989).*Cross-cultural pragmatics: Requests and apologies*. Norwood, NJ: Ablex.

Bouton, L. F. (1992). The interpretation of implicature in English by NNS: Does it come automatically-without being explicitly taught? *Pragmatics and Language Learning*, *3*, 53-65.

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

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment.* New York: McGraw-Hill.

Brown, J. D. &Ahn, R.C. (2011). Variables that affect the dependability of L2 pragmatic tests. *Journal of Pragmatics, 43,* 198-217.

Brown, J. D. (2001). Six types of pragmatics tests in two different contexts. In K. Rose & G. Kasper (Eds.), *Pragmatics in Language Teaching* (pp.301-325). New York: Cambridge University Press.

Brown, J. D. (2008). Raters, functions, item types and the dependability of L2 pragmatics tests. In A. Martinez-Flor & E. Alcon (Eds.), *Investigating pragmatics in foreign language learning, teaching, and testing* (pp.224-248). Clevedon: Multilingual Matters.

Carr, N., & Xi, X. (2010). Automated scoring for short-answer reading items: implications for constructs. *Language Assessment Quarterly, 7*, 205-218.

Cohen, A. D., & Shively, R. (2007). Acquisition of requests and apologies in Spanish and French: Impact of study abroad and strategy-building intervention. *The Modern Language Journal*, *91*, 189–212.

Crystal, D. (Ed.). (1997). *The Cambridge encyclopedia of language* (2nd ed.).New York: Cambridge University Press.

Educational Testing Service. (2010). *TOEFL Test Prep Planner.*

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

Félix-Brasdefer, J. C., & Koike, D. (Eds.). (2012). *Pragmatic variation in first and second language contexts: Methodological issues*. Amsterdam: John Benjamins.

Grabowski, K. (2008). Measuring pragmatic knowledge: Issues of construct underrepresentation or labeling? *Language Assessment Quarterly, 5*, 154-159.

Grice, P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics* (Vol. 3, pp. 41–58). New York: Academic Press.

Hartford, B. S., & Bardovi-Harlig, K. (1992a). Closing the conversation: Evidence from the academic advising session. *Discourse Processes*, *15*, 93-116.

Hartford, B. S., & Bardovi-Harlig, K. (1992b). Experimental and observational data in the study of interlanguage pragmatics. *Pragmatics and language learning, 3*, 33-52.

Holtgraves, T. (1999). Comprehending indirect replies: When and how are their conveyed meanings activated? *Journal of Memory and Language, 38*, 519-540.

Hudson, T., Detmer, E., & Brown, J.D. (1992). *A framework for testing cross-cultural pragmatics* (Technical Report 2). 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* (Technical Report 7). Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Hudson, T. (2001). Indicators for cross-cultural pragmatic instruction: some qualitative tools. In K. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp.80-102). Cambridge: Cambridge University Press.

Ifantidou, E., & Tzanne, A. (2012). Levels of pragmatic competence in an EFL academic context: A tool for assessment. *Intercultural Pragmatics, 9*, 47-70.

Ishihara, N. (in press). Teacher-based assessment of L2 Japanese pragmatics: Classroom applications. In S. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 155-195). Basingstoke, UK: Palgrave Macmillan.

Ishihara, N., & Cohen, A. D. (2010). *Teaching and learning pragmatics: Where language and culture meet.* Harlow, UK: Longman/Pearson Education.

Kasper, G. (1996). Introduction: Interlanguage pragmatics in SLA. *Studies in Second Language Acquisition, 18*, 145-148.

Koike, D. A. (1989). Pragmatic competence and adult L2 acquisition: Speech acts in interlanguage. *The Modern Language Journal, 73*, 279-289.

Koike, D. A. (1996). Transfer of pragmatic competence and suggestions in Spanish. In S. M. Gass, & J. Neu (Eds.), *Speech acts across cultures: Challenge to communication in a second language* (pp. 257-281). Berlin: de Gruyter.

Kuha, M. (1997). The computer-assisted interactive DCT: A study in pragmatics research methodology. *Pragmatics and Language Learning, 8*, 99-127.

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

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

Liu, J. (2007). Developing a pragmatics test for Chinese EFL learners. *Language Testing, 24*(3), 391-415.

Niezgoda, K., &Roever, C. (2001). Pragmatic and grammatical awareness: A function of the learning environment? In K. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 63-79). Cambridge: Cambridge University Press.

Roever, C. (2005). *Testing ESL Pragmatics: Development and validation of a web-based assessment battery.* Berlin: Peter Lang.

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

Rose, K. R. (1994). On the Validity of Discourse Completion Tests in Non-Western Contexts. *Applied Linguistics, 15*, 1-14.

Sasaki, M. (1998). Investigating EFL students' production of speech acts: A comparison of production questionnaires and role plays. *Journal of Pragmatics*, *30*, 457-484.

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

Taguchi, N. (2011). Rater variation in the assessment of speech acts. *Pragmatics, 21*, 453-471.

Teng, C., & Fei, F. (2013). A consciousness-raising approach to pragmatics teaching: Web-based tasks for training study-abroad students. *Journal of Technology and Chinese Language Teaching, 4*, 50-63.

Trosborg, A. (Ed.) *Pragmatics across languages and cultures (Vol. 7 of Handbooks of pragmatics*). Berlin: Mouton de Gruyter.

Yamashita, S. (1996). *Six measures of JSL pragmatics* (Technical Report 14). Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Youn, S. (2007). Rater bias in assessing the pragmatics of KFL learners using facets analysis. *Second Language Studies, 26*, 85-163.