**Title of Project:**
Exploring Construct and Practicality of an Interactional Test for L2 Oral Pragmatic Performance

**Researcher:**
Naoki Ikeda  
University of Melbourne  
na02ikeda@hotmail.com

**Research Supervisors:**
Dr. Carsten Roever  
University of Melbourne  
Naoki Ikeda

Dr. Ute Knoch  
University of Melbourne

---

**Project Summary**

**Motivation for the Research**

Pragmatic competence is one of the components of communicative competence models (Bachman & Palmer, 2010; Canale 1983; Canale & Swain, 1980; Purpura, 2004), and accordingly is a crucial part of human communication in real life (Roever, Fraser, & Elder, 2014). It should therefore be a key domain in language assessment (Kasper & Ross, 2013). Assessment of second language (L2) pragmatics has been developed by conceptualizing the test construct, as well as how it is conceptualized in task formats. Traditionally, the focus of pragmatics has been on offline knowledge of speech acts isolated from interaction. Despite considerable groundwork in pragmatics research, the findings in the literature are constrained by the narrowly defined construct and task design that restrict participants’ performances to pre-planned scenarios, limiting conclusions about how test takers pragmatically perform in authentic discourse contexts.

Recent arguments (Kasper & Ross, 2013) and empirical studies (Grabowski, 2009; Youn, 2013) in the field, however, have highlighted performance-based assessments eliciting L2 speakers’ oral extended discourse, which would allow for a broader conclusion about test takers’ L2 pragmatic performances and for a stronger extrapolation from task performance to reality. Although informative about test takers’ pragmatic abilities, performance-based assessments would negatively impact practicality (McNamara & Roever, 2006), for which suggestions should be made for actual practitioners of the assessments. Another recent activity worthy of note in the field is validation, which examines the extent to which a proposed score interpretation and use is justified and how so. Recent studies on validation in the field of L2 pragmatics assessment (e.g., Roever, et al., 2014), although still very limited, showed a stronger awareness of test validation, which informs test users of what the test scores mean and how useful the test scores are (Chapelle, 2008; Kane, 2006). With the insights provided by the previous studies, the researcher designed and evaluated an oral performance-based assessment instrument of L2 pragmatics, following an existing argument-based framework (Chapelle, 2008; Kane, 2006; Knoch & Elder, 2013). Unlike most validation studies, this study also addressed how the designed instrument can be implemented practically while avoiding construct under-representation (Messick, 1996).
Research Questions

1. Investigation of measurable construct of oral pragmatics: What abilities of L2 oral pragmatics utilized in extended discourse contexts are measurable in inter-personal settings at university?

2. Test validation: To what extent is the procedure for assigning test scores appropriate? To what extent does the assessment yield test results consistent across assessment contexts (raters, test tasks, and test sets)? To what extent can test takers’ test results be attributed to the construct of pragmatic abilities utilized for language activities at university?

3. Practicality: How and to what extent can monologue tasks serve as an alternative to dialogue tasks?

Research Methodology

This study involved multiple phases of research activities and a range of analytical methods to address the research questions. First, the test tasks were developed through five stages of drafting and pilot studies, including discussions with domain experts. The tasks simulated university situations where a student addresses professors, administrators, and classmates to obtain assistance for the student’s academic work. Two parallel test sets (with each set comprising three dialogue tasks and three monologue tasks) were created. The dialogue tasks simulated conversational situations where a student (test taker) physically interacts with an addressee. The monologue tasks simulated situations requiring a student to leave a voice-message to the intended addressee. Once the test tasks were finalized, the test administration in the main study was subsequently administered to 67 L2 students in Australia comprising university students (N=44) and pre-entry students (N=23), each of whom were assigned either Test Set 1 or Test Set 2. The pre-entry students were studying at a language school to satisfy the English language requirement for university admission. They did not have experiences at an English-medium university and had little exposure to English-speaking environments. The collected data also included the test takers’ perceived similarity between their task and authentic performances and their self-assessed task performances (both of which were elicited by Likert scale questionnaires) in addition to their audio-recoded task performance samples.

Prior to the rating, the researcher developed the rating criteria based on the researcher’s review of related literature and qualitative discourse analyses of the test takers’ collected speaking samples, which identified features frequently seen in the test takers’ performances and discriminating them. The findings were used to inform the rating rubrics, which comprised the six criteria (a) to (f) (described in the next section) with four band levels for each. The features summarized in the six criteria addressed research question 1. Generally, the lower scores (i.e., 2 and 1) were awarded to performances where negative features outweighed positive features, whereas the higher scores (i.e., 4 and 3) were awarded to those showing the reversed case. The rating was conducted by three raters, including the researcher himself. The test scores allowed for descriptive statistics, reliability estimates, multi-faceted Rasch analysis, and correlational analysis. The test takers’ perceptions were analyzed quantitatively. The results of the quantitative analyses were evaluated to address research Questions 2 and 3.

Summary of Findings

The qualitative discourse analyses of the test takers’ task performances led to setting six criteria (a) to (f) summarized below, each of which discriminated the test takers in the current study. Performances of pragmatically competent test takers were constructed by (a) adequate social actions tailored for the context from opening through to closing; (b) smoothly and clearly delivered contents with sound variation and controlled repair; (c) linguistic resources varied and employed naturally to deliver intended meanings, minimizing the addressee’s effort to understand what the speaker was trying
to say; (d) linguistic resources varied and employed naturally to mitigate imposition; (e) understanding of the discourse context and use of varied patterns of evaluative feedback well-tailored for the context without noticeable scaffolding from the interlocutor, and (f) naturally taking and releasing conversation turns so that the interlocutor understands when to take turns. Criteria (e) and (f) were used exclusively for dialogue performances. Generally, pragmatically competent test takers showed stronger evidence for their competencies. However, performances of less competent test takers were characterized by inconsistent evidence of the positive features described above and/or negative features undermining their performances.

The Rasch analysis showed the raters’ internal consistency in rating. Although they showed a difference in severity in rating the test takers’ performances, the severity difference was much smaller than the Rasch-estimated separation of the 67 test-takers, implying that the impact of raters’ severity difference on test-taker scores was small. It was also confirmed that the raters operated the assessment band levels appropriately. For the tasks, inter-item correlations were high. The two paralleled test sets showed comparability in characteristics of test takers, raters, and rating criteria. These results, combined with the results of the raters’ performances, suggest that the assessment can yield test results consistent across assessment contexts. Another important result was found regarding the test takers’ Rasch-estimated pragmatic ability. Some university students, who were deemed pragmatically more competent than pre-entry students because of their experiences at English-medium universities and because they had higher proficiency levels, were outperformed by pre-entry students. However, overall the test takers were separated according to their proficiency levels and exposure to the language activities in the target domain as identified in the literature. The test takers perceived their language use in the tasks to be similar to that in reality. Highly proficient university students in particular perceived a stronger similarity between their task and authentic performances. Their self-assessment of task performance was moderately correlated with their Rasch-estimated ability, as expected for a non-test criterion. The test takers’ Rasch-estimated abilities under the dialogue and monologue conditions were highly correlated, suggesting that test takers who performed positively under the monologue condition would perform just as positively under the dialogue condition.

Implications

Although the methodology and the rationale are different among the existing studies on assessing L2 pragmatics, they have a common role of addressing a crucial component of human communication, which should ultimately contribute to enriching L2 learners’ language activities. The current study has addressed the same goal as the literature, from the perspective of language assessment, by targeting L2 students’ pragmatic abilities for language activities at an English-medium university. First and foremost, the constructs of pragmatics targeted in the study can be assessed, as they discriminate the test takers. Overall, the test scores generated from the designed assessment were indicative of L2 students’ oral pragmatic abilities for university activities.

Secondly, the test takers’ task performances, which have been qualitatively and quantitatively described in the study, have highlighted the meaningfulness of pragmatics assessments at both university pre-entry and post-entry stages. In the study, 19 university students were regarded as pragmatically competent as their average scores were 3 or above (out of four band levels). For the remaining 48 test takers (including 25 university students and 23 pre-entry students), their average scores were below 3, which indicated that their pragmatic performances were more negative than positive, according to the rating criteria. The results also imply that being a university student in an English-medium context does not necessarily mean that their pragmatic abilities are sufficient for language activities in that domain. Based on the qualitative and quantitative analyses, the study has
found that performances of all of the pre-entry students, as well as the substantial number of university students, reserve much room for improvement. Students could be aware of the need for improvement through a pragmatics assessment.

For the practice of a performance-based assessment of oral pragmatics, the study suggested a possible use of a monologue assessment. As confirmed by the high correlation between test takers’ abilities estimated from the dialogue and the monologue data, both types of assessment instruments functioned similarly, in separating and ranking L2 students according to the pragmatic abilities. Dialogue assessments would allow for providing more comprehensive diagnostic information, including interactional criteria, but at the expense of increased resource intensiveness in its test administration. Because available resources and what to prioritize differ depending on individual assessment contexts, it would not be possible to provide an absolute solution for the appropriate balance between instrument practicality and construct coverage of pragmatics. However, if diagnosis of test takers’ features that are exclusively seen in dialogue performances is not a central concern, the monologue assessment could be an alternative to the dialog assessment.
References


Xi, X. (2015, September). Revisiting constructs- Recent theoretical advances and applications to practical test design, development and validation. Workshop conducted at University of Melbourne, Melbourne, Australia.