Questions and answers about language testing statistics:

Testing intercultural pragmatics ability

James Dean Brown
brownj@hawaii.edu
University of Hawai'i at Mānoa

Question:

What sorts of tests have been developed and used for testing intercultural pragmatics ability? What do we know about such testing? And, how have those tests been analyzed statistically?

Answer:

The literature on developing intercultural pragmatics tests has (a) found that different testing formats vary in their effectiveness for testing pragmatics, (b) discovered that certain variables are particularly important in testing pragmatics tests, and (c) relied on increasingly sophisticated statistical analyses in studying pragmatics testing over the years. I will address each of these three issues in turn.

Different Testing Formats Vary in Their Effectiveness for Testing Pragmatics

Starting with Hudson, Detmer, and Brown (1992, 1995), six testing methods have been prominent to varying degrees in the literature to date (as shown in Table 1):

- *Multiple-choice Discourse Completion Task* (MDCT) – requires examinees to read a situation description and choose what they would say next.
- *Oral Discourse Completion Task* (ODCT) – expects examinees to listen to an orally described situation and record what they would say next.
- *Discourse Role-Play Task* (DRPT) – directs examinees to read a situation description and then play a particular role with an examiner in the situation.
- *Discourse Self-Assessment Task* (DSAT) – asks examinees to read a written description of a situation and then rate their own pragmatic ability to respond correctly in the situation.
- *Role-Play Self-Assessment* (RPSA) – instructs examinees to rate their own performance in the recording of the role play in the DRPT.

Hudson et al. (1992, 1995) created initial prototype tests and validated them for EFL students at a US university. They noted that the MDCT did not work particularly well for them. Yamashita (1996) then created Japanese versions of those same tests and verified that all but MDCT worked reasonably well for Japanese as a second language (SL). Enochs and Yoshitake (1996) and Yoshitake (1997) verified that the six assessments worked well for Japanese university EFL students. Ahn (2005) created Korean versions for all but the MDCT and verified that they worked reasonably well for Korean as a FL. Liu (2007) reported on developing a MDCT that worked, which he accomplished by using students to generate the speech acts and situations that were used.

Hudson et al. (1992, 1995) and a majority of the other researchers have used paper-and-pencil testing formats. However, other formats have also been used. Tada (2005) was the first to create computer-delivered tests with video prompts. Roever (2005, 2006, 2007, 2008) was the first to develop and use web-

based testing followed by Itomitsu (2009). Rylander, Clark, and Derrah (2013) focused on the importance of video formats. And, Timpe (2013) was the first to use *Skype* role-play tasks.

Certain Variables Are Particularly Important in Testing Pragmatics

In creating their first prototype tests, Hudson et al. (1992, 1995) identified a number of variables that have proven important across many of the subsequent studies, but to varying degrees. These variables are labeled across the top of Table 1. The first was the six ***testing methods*** discussed in the previous section. The second variable was ***speech acts***, which initially included three key ones: (a) *requesting* (i.e., asking another person to do something or for something), (b) *refusing* (i.e., rejecting another person's request), and (c) *apologizing* (i.e., acknowledging fault and showing regret for doing or saying something). The third variable was ***contextual conditions***, which initially included three key conditions: (a) *imposition* (i.e., the degree of inconvenience to the listener of the request, refusal, or apology), (b) *power difference* (i.e., the degree and direction of differences in power or position between the speaker and listener), and (c) *social distance* (i.e., the degree of shared social familiarity or solidarity between the speaker and listener).

Other variables were added as research continued. For example, Roever (2005, 2006, 2007, 2008) added the assessment of idiosyncratic and formulaic implicatures, as well as situational routines in addition to speech acts. He also added rejoinders after the response slot in designing his items. Tada (2005) specifically examined perception versus production of pragmatics to his study. Liu (2006, 2007) innovatively used speech acts and situations generated by students. Grabowski (2009, 2013) examined the relationship between grammar and pragmatic knowledge (which he further subdivided into sociolinguistic, sociocultural, psychological knowledges). Itomitsu (2009) also studied grammar and three aspects of pragmatics (appropriate speech acts, routines, and speech styles) and used requests speech acts, but also added offers and suggestions. Roever (2013) focused on implicature, but also considered vocabulary, collocations, idiomatic word meanings, and morphology. Rylander et al. (2013) added a number of speech acts using refusals and apologies, but also compliments, farewells, greetings, introductions, invitations, suggestions, offers, and complaints. Timpe (2013) included new speech acts: in addition to requests, she used offers, and also examined routine phrases, and phrases/idioms. Youn 2013 added speech acts of expressing opinion and giving feedback on email and compared role-plays with monologic speaking and pragmatics tasks. And finally, Youn and Brown (2013) compared heritage and non-heritage KFL students' performances on such tests.

Increasingly Sophisticated Statistical Analyses have Been Used to Study Pragmatics Tests

A quick glance at the second to last column in Table 1 will reveal that all of the studies have used classical testing theory (CTT), which involves traditional descriptive statistics, reliability estimates, correlation coefficients, and in some cases item analyses. However, as time went by, researchers increasingly used three more complex analyses:

- *Rasch analysis* allows researchers to put items and examinees on the same logit scales.
- *FACETS analysis* is a variation of Rasch analysis that allows researchers to put a variety of different facets (e.g., items, raters, rating categories, etc.) on the same logit scale and, among other things, allows simultaneous display of whatever facets are selected so they can be compared to examinee performances (for instance, examinees can be represented on the same scale as raters and rating categories, as in Brown, 2008).

Table 1

Pragmatics Testing Projects (Quantitative) Described in Terms of Testing Methods, Speech Acts, Contextual Conditions and Value Added to the Knowledge of Pragmatics Assessment[1]

Testing Project; L2 Being Learned and where	Testing Method[1]						Speech Acts				Conditions[2]			Test Type[3]	Statistical Analyses[4]	Value Added to Knowledge of Pragmatics Assessment
	WDCT	MDCT	ODCT	DRPT	DSAT	RPSA	Requesting	Refusing	Apologizing	Other	Imposition	Power	Social Dis.			
Hudson et al. 1992, 1995; ESL in US	X	X	X	X	X	X	X	X	X		2	2	2	P&P	CTT	Created the initial tests and validated all but the MDCT for EFL students at a US university.
Enochs & Yoshitake, 1996; Yoshitake 1997; Both EFL in Japan	X	X	X	X	X	X	X	X	X		2	2	2	P&P	CTT	Verified that six assessments worked reasonably well for Japanese university EFL students; scores also compared to 3 TOEFL subtests.
Yamashita 1996; JSL in Japan	X	X	X	X	X	X	X	X	X		2	2	2	P&P	CTT	Created Japanese versions and verified that all but MDCT worked reasonably well for Japanese as a SL.
Ahn, 2005; Brown 2008; Brown & Ahn, 2011; All KFL in US	X		X	X	X	X	X	X	X		2	2	2	P&P	CTT, G theory, FACETS	Examined the effects of numbers of raters, functions, item types, and item characteristics on reliability and difficulty/severity in several combinations.
Roever 2005, 2006, 2007; ESL/EFL in US/Germany/Japan		S					X	X	X		2	1	1	WBT	CTT, FACETS, DIF	Assessed idiosyncratic and formulaic implicatures, situational routines, and speech acts; formats similar to MDCT, but speech acts added rejoinders after the response slot.
Tada 2005; EFL in Japan		S	S				X	X	X		2	2	1	CLT, Video	CTT	1st to be computer delivered with video prompts for tests similar to MDCT and OPDCT (specifically examined perception vs. production of pragmatics)
Liu 2006; EFL in PRC	S	S			S						2	2	2	P&P	CTT, Rasch	Speech acts and situations were generated by students.
Liu 2007; EFL in PRC	S										2	2	2	P&P	CTT, Rasch	Focused on developing a MDCT that worked; Speech acts and situations were generated by students.
Roever 2008; ESL/EFL in US/Germany/Japan		S					X	X	X		2	1	1	WBT	CTT, FACETS	Speech acts section only; rejoinders after the response slots; examined effects of raters and items.
Youn 2008; KFL in US	X		X	X			X	X	X		2	2	2	P&P	CTT, FACETS	Examined the effects of test types and speech acts on raters assessments.
Grabowski 2009, 2013; ESL in US				S								1	2	P&P	CTT, G theory, FACETS	Speaking tests similar to DRPT; rated and examined the relationship between grammar and pragmatic knowledge (further subdivided into sociolinguistic, sociocultural, psychological knowledges).
Itomitsu, 2009; JFL in US		S					X			X				WBT	CTT	Grammar and three aspects of pragmatics (appropriate speech acts, routines, and speech styles); three not distinguishable; only total scores validated; speech acts included requests, offers, suggestions.
Roever, 2013; NS & ESL in Australia		S												P&P	CTT, FACETS	Focuses on implicature (along with subtests on vocabulary, collocations, idiomatic word meanings, & morphology)
Rylander, Clark, & Derrah, 2013; EFL in Japan							X		X	X				P&P, Video	CTT, Rasch	Focuses on importance of video: added speech acts (refusals & apologies , but also compliments, farewells, greetings, introductions, invitations, suggestions, offers, & complaints).
Timpe, 2013; EFL in Germany		S			S		X			X		2	2	WBT	CTT, Rasch	Focused on American English self-assessment, a sociopragmatic comprehension test, and *Skype* role-play tasks. Sociopragmatics test include speech acts (requests and offers), routine phrases, and phrases/idioms
Youn 2013; KFL in US				S			X			X		2		P&P	CTT, FACETS	(a) based on needs analysis, developed open role-play tasks similar to DRPT but more interactive; (b) added speech acts of expressing opinion and giving feedback on email; (c) compared role-play with monologic speaking and pragmatics tasks; &(d) exceptionally thorough reliability & validity study based on Kane's (2006) argument-based approach.
Youn & Brown, 2013; KFL in US	X		X	X			X	X	X		2	2	2	P&P	CTT, FACETS	Focused on comparison of heritage and non-heritage KFL students

[1] X = adapted same test; S = Similar test

[2] Number of levels (1 or 2) of each condition, e.g, Imposition high or low would be 2 levels

[3] P&P = Paper & Pencil test; CLT = Computerized Language Testing; WBT = Web-based Language Testing

[4] CTT = Classical Test Theory; G-theory = Generalizability theory; Rasch = Rasch analysis; FACETS = Multifaceted Rasch analyses; DIF = Differential Item Functioning

[1] Only quantitative research studies are considered here. In addition, whenever multiple publications appeared to be based on the same data, I grouped them as one project.

- *Generalizability theory* (G theory) allows researchers to study and minimize multiple sources of error in two stages: (a) a *Generalizability study*, which is used to estimate variance components for whatever facets the researcher wishes to study and thereby to understand the relative proportions of variance accounted for by the object of measurement (usually variance due to examinees) and other facets that are sources of variance (for example, raters and rating categories) (note that this can be done for either norm-referenced or criterion-referenced tests by using different procedures) and (b) a *Decision study*, which is used to estimate the appropriate generalizability coefficients (analogous to reliability estimates) for different numbers of levels in each facet (e.g., estimates can be provided for 2 raters or 3, 4, 5, etc. while also examining what happens simultaneously if 2 rating categories are used or 3, 4, 5, 6, etc.). For an example of this entire process, see Brown and Ahn (2013).

These analyses and others have been applied in various ways with generally increasing levels of sophistication in the pragmatics testing literature. Hudson et al. (1992, 1995) created the initial tests and validated all but the MDCT for EFL students at a US university using CTT. Enochs and Yoshitake (1996) and Yoshitake (1997) verified that the six assessments worked reasonably well for Japanese university EFL students using CTT. Those scores were also compared to the three sets of TOEFL subtest scores available at that time. Yamashita (1996) created Japanese versions and verified that all but MDCT worked reasonably well using CTT. Ahn (2005), Brown (2008), and Brown and Ahn (2011) used FACETS and G-theory analyses to examine the effects of numbers of raters, functions, item types, and item characteristics on reliability and difficulty/severity in various combinations. Roever (2005, 2006, 2007) used FACETS and differential item functioning analyses. Liu (2006) used Rasch analysis to study the effectiveness of speech acts and situations that had been generated by students. Liu (2007) also used Rasch analysis but focused on developing a MDCT that worked. Roever (2008) applied FACETS analysis to study the effects of raters and items. Youn (2008) used FACETS analysis to examine the effects of test types and speech acts on raters assessments. Grabowski (2009, 2013) used both G theory and FACETS analysis in the process of examining speaking tests similar to DRPT with a focus on the relationship between grammar and pragmatic knowledge. Roever (2013) used FACETS analysis in his study of implicature. Rylander et al. (2013) used Rasch analysis in their study testing many different speech acts while using video formats. Timpe (2013) also used Rasch analysis in her study of American English self-assessment, a sociopragmatic comprehension test, and *Skype* role-play tasks. Youn (2013) relied on Rasch analysis in her elaborate validity study (based on Kane's (2006) argument-based approach) of role-plays with monologic speaking and pragmatics tasks. And finally, Youn and Brown (2013) used FACETS analysis in their comparison of heritage and non-heritage KFL students' performances.

Conclusion

Different testing formats (including the original WDCT, MDCT, ODCT, DRPT, DSAT, RPSA, and a number of variations on those themes) have been shown to vary in their effectiveness for testing pragmatics depending on the context and the variables involved. In the process, a wide range of variables have been studied in the literature to date (especially, testing methods, speech acts, and various conditions). In addition, CTT, Rasch, FACETS, and G theory have been the major forms of analysis in the increasingly sophisticated pragmatics testing literature in a variety of different ways.

In all probability, pragmatics testing will continue to grow in the future. No doubt additional tests will be developed (a) to assess pragmatics in additional languages, (b) to accommodate new additional variables as the subfield of intercultural pragmatics continues to expand, and finally, (c) to adjust to refinements in pragmatics constructs and testing formats. It will be interesting to see what impacts all this activity will have on the teaching and testing of English and other languages around the world—and of course in Japan.

References

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

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). Cambridge: Cambridge University.

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

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

Grabowski, K. C. (2013). Investigating the construct validity of a role-play teste designed to measure grammatical and pragmatic knowledge at multiple proficiency levels. In S. J. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 149-171). London: Palgrave Macmillan.

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.

Itomitsu, M. (2009). *Developing a test of pragmatics of Japanese as a foreign language*. Unpublished Ph.D. dissertation, Columbia University.

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

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

Roever, C. (2007). DIF in the assessment of second language pragmatics. *Language Assessment Quarterly, 4*(2), 165–189.

Roever, C. (2008). Rater, item, and candidate effects in discourse completion tests: A FACETS approach. In E. Soler & A. Martinez-Flor (Eds.), *Investigating pragmatics in foreign language learning, teaching, and testing* (pp. 249-266). Bristol, UK: Multilingual Matters.

Roever, C. (2013). Testing implicature under operational conditions. In S. J. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 43-64). London: Palgrave Macmillan.

Rylander, J., Clark, P., & Derrah, R. (2013). A video-based method of assessing pragmatic awareness. In S. J. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 65-97). London: Palgrave Macmillan.

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

Timpe, V. (2013). *Assessing intercultural language learning.* Frankfurt am Main: Lang.

Yamashita, S. O. (1996). *Six measures of JSL Pragmatics* (Technical Report #14. Second Language Teaching and Curriculum Center). Honolulu, HI: University of Hawai‘i.

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

Youn, S. J. (2008). *Rater variation in paper vs. web-based KFL pragmatic assessment using FACETS analysis.* Unpublished MA thesis, University of Hawai‘i at Mānoa.

Youn, S. J. (2013). *Validating task-based assessment of L2 pragmatics in interaction using mixed methods.* Unpublished PhD dissertation, University of Hawai‘i at Mānoa.

Youn, S. J., & Brown, J. D. (2013). Item difficulty and heritage language learner status in pragmatic tests for Korean as a foreign language. In S. J. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 98-123). London: Palgrave Macmillan.

Where to Submit Questions:

Your question can remain anonymous if you so desire. Please submit questions for this column to the following e-mail or snail-mail addresses:

brownj@hawaii.edu.

JD Brown
Department of Second Language Studies
University of Hawai‘i at Mānoa
1890 East-West Road
Honolulu, HI 96822
USA