Language Teaching (2020), **53**, 183–202
doi:10.1017/S0261444819000156

CAMBRIDGE
UNIVERSITY PRESS

PLENARY SPEECH

Considerations in assessing pragmatic appropriateness in spoken language

Andrew D. Cohen

Professor Emeritus, University of Minnesota, USA
Email: adcohen@umn.edu

Abstract

This paper is a revised version of a plenary prompted by the upsurge of interest in the role of pragmatics in teaching, learning, and assessment, and has as its purpose to take a fresh look at recent developments in the assessment of target-language (TL) pragmatics in spoken language. The first issue considered is the question of whether to attempt to assess pragmatics as it unfolds naturally in interactions, and if so, how to do it. Next, micro-level and macro-level factors in the assessment of TL pragmatics are considered. Third, a close look is given to the specific elements of TL pragmatics to assess. Fourth, there is attention to the matter of which instruments are most appropriate for collecting the desired data in the given context. Fifth, issues relating to data analysis are discussed. Finally, matters pertaining to the assessment of classroom instruction are looked at. Recommendations are given as to potentially viable directions for dealing with these issues both in terms of research studies and for assessment of classroom instruction.

Introduction

There are moments when it is crucial to be able to respond to an interlocutor in an appropriate and effective way under the pressure of real-world face-to-face communication, be it in a dominant language or in a TL. Nonetheless, assessment of this skill is not necessarily a component in language tasks, especially in large-scale assessment (Isaacs, 2016). This point should be borne in mind when looking at options for assessing speaking – the reality that spoken language has been a lesser focus in assessment over the years, primarily because of the time and effort required to do it. In addition, there may even be lack of certainty as to how to do it effectively.

This paper was motivated by the upsurge of interest in the role of pragmatics in teaching, learning, and assessment, and has as its purpose to take a fresh look at recent developments in the assessment of TL pragmatics in spoken language, both as they pertain to research and to classroom instruction. At times the focus will be more on researching the assessment of pragmatics and at times more on assessment as part of classroom instruction. The paper starts with the question of whether to assess pragmatics as it unfolds naturally in interactions, and if so, how to do it. Then we will look at micro-level and macro-level factors in the assessment of TL pragmatics. Next, we will consider which elements of TL pragmatics to assess. Then, we look at instruments intended to be used in collecting the desired data. Next, we will look at issues relating to data analysis. Finally, we will consider some issues which have specific relevance to the assessment of pragmatics as taught and performed in the language classroom.

Paper presented as a plenary lecture at the International Conference on Exploring and Assessing Pragmatic Aspects of L1 and L2 Communication: From Needs Analysis through Monitoring to Feedback. Department of Linguistic and Literary Studies, University of Padua, Italy, 25–27 July 2018.

© Cambridge University Press 2019

https://doi.org/10.1017/S0261444819000156 Published online by Cambridge University Press

1. The assessment of pragmatics in natural interaction

Work in pragmatics assessment in the late 1970s and 1980s tended to isolate and describe speech act behavior in the first language (L1) and in the TL. Somewhat limited attention was given to the discourse context within which that specific speech act or any other appeared. Likewise, limited attention was given to the clustering of speech acts within the same speech event. Consequently, assessment of pragmatic behavior did not tend to consider the larger discourse context. By the early 1990s, the focus was now on clusters of speech acts, like greetings, apologies, and requests, all performed within the same speech event (Hudson, Detmer, & Brown, 1992, 1995). Likewise in the late 1990s into the early 2000s, pragmatic behavior began to be viewed within the framework of conversational analysis (see Kasper, 2004 for one example). Shortly thereafter, the pragmatics literature began to include examples of how teachers could teach speech act performance embedded within a discourse context (see Félix-Brasdefer, 2006 for an example using a refusal to an invitation).

Concern for the larger context in which pragmatic behavior occurs has emerged as an issue to be considered in assessment. Experts would now admonish us from simply paying lip service to the notion that pragmatic ability involves knowledge of a form-function-context relationship in assessment. Rather, we need to keep in mind that the form-function-context relationship is emergent and contingent upon unfolding interaction: Interlocutors must adapt their ways of speaking consistent with the interlocutors' mutual familiarity, moment-to-moment attitudes and emotional issues that arise, and the differing aims of the discourse (Taguchi, 2018). Hence, those in favor of assessing pragmatics in natural interaction would consider it an imperative to measure a language learners' ability to adapt to the changing course of interaction and to achieve a communicative goal in collaboration with interlocutors.

Taguchi notes that by applying a bottom-up approach to analyzing naturally recorded interaction, researchers can gain evidence of how interaction unfolds turn-by-turn as participants jointly achieve a speech act. The argument is made that the analysis of conversational structure, such as turn taking, adjacency pairs, pre- and post-sequences, and preference organizations, can reveal how participants monitor the changing course of a conversation and adapt to the change by designing their contributions in a sequentially appropriate manner. The case is made that by accessing these kinds of data, it should be possible to determine the extent to which learners are able to respond in alignment with their interlocutors' conversational moves in order to achieve a communicative act.

While I applaud efforts at adding this kind of rigor to the assessment of pragmatics in spoken language, at the present time the recording and analysis of natural interactions for their pragmatics content would appear to be more appropriate for research studies than for assessment measures intended for pedagogical purposes. Since it can be a challenge to construct, administer, and evaluate results for relatively simple speaking tasks, measures which truly get at natural interaction would appear beyond the scope of classroom assessment at the present time. Perhaps this will no longer be the case with the advent of more sophisticated software designed to capture and evaluate natural interaction.

2. Determining the 'what' of pragmatics assessment

Let us look now at a sampling of directions that assessment of pragmatics could take in terms of the 'what': a consideration of the micro-and macro-level factors that are possibly involved, differences in oral and written varieties of the same language, assessing speech acts, dealing with prosody, and the role of language learner strategies.

2.1 Micro-level and macro-level factors

If the aim of a research project is to assess the pragmatics of spoken language and possibly to assess interactions in natural discourse, there are various factors to be considered at least in passing. At the micro-level there is a need to determine which pragmatic elements to focus on (e.g. speech acts, implicature, conversational turns, and other pragmatic features) within the given speech event (e.g. a specific conversation) situated within the given speech situation (e.g. a party) (Hymes, 1972). The

https://doi.org/10.1017/S0261444819000156 Published online by Cambridge University Press

description of the given situation involving pragmatics would include information on the social distance between the respondent and the other interlocutor(s), their relative power, and the degree of imposition of any speech acts being assessed.

At the macro-level there are factors relating to the participants, such as their language proficiency, ethnicity, age, socio-economic background, and gender. In a classroom setting the students may be at a relatively similar proficiency level, but their ability to perceive and perform the pragmatics of the TL may differ. And these days, gender is no longer a perfunctory variable. The issue of how to include gender in assessment, for example, warrants revisiting since it is not only a biological matter but also a social one, which is all the more important when it comes to pragmatics (Jones, 2016). It appears to be a propitious moment to conduct research on the interplay of gender and TL pragmatics.

Finally, dialect can play an important role in pragmatics, since the region represented by the respective language varieties of the interlocutors can impact the interaction pragmatically. For example, a compliment from a Kiwi for doing something well could be 'Good on ya!' (pronounced /gudon'ya/) and an American recipient of this intended praise may be wondering what it means. When this happened to me the first time, I was wondering what I had on me – perhaps some food. I totally misunderstood the intention of the remark.

2.2 Differences in the oral and written varieties of the same language

The fact that certain languages have distinct differences between spoken and written varieties can have an impact on a written discourse completion task (DCT) used as an indirect measure of speaking. One study with native Farsi speakers responding in their L1 found, for instance, that whereas the literature might suggest that written DCTs provide much the same information as oral DCTs, the reality for Farsi is that the two forms of assessment produce differing results given the marked differences between spoken and written varieties and the existence of highly complicated stylistic variations (Eslami & Mirzaei, 2014). The study showed that the oral DCTs elicited longer, more elaborate responses and more linguistic forms which represented the spoken variety of the language than did the written DCTs. In addition, on the written DCTs students mixed different styles (spoken and written) and in one situation used both formal and informal linguistic devices.

Level of formality also presents itself as an area of pragmatic concern with regard to style shifting in certain languages. Native speakers of Japanese, for example, may switch to the plain form as their primary style or use a mixture of polite and plain forms to express solidarity and closeness. When parting, however, speakers may switch back to the polite form to add formality and ritualistic tone to the closure of the conversation (Taguchi & Roever, 2017, p. 7).

In addition, there is intralingual variation according to geographic region. Barron (2019) points out that as second language (L2) speakers become more globally mobile, they become exposed to intralingual varieties across macro-level factors and quickly realize the oversimplification of the homogeneous standardized native-speaker (NS) prestige norm presented in the classroom context. One study, for instance, looked at the speech act of responding to thanking in England, Ireland, and the US in 15-year-old NSs of English (Schneider, 2017), and found some distinct differences with regard to the use of 'you're welcome,' 'my pleasure,' 'don't mention it,' 'not at all,' 'anytime,' 'that's okay,' and 'no problem.' The Australian 'no worries' did not feature in British, Irish, and American responses.

So especially in classroom assessment to what extent would a measure include any of these elements? Whereas assessing dialect differences may seem to reflect unnecessary fine-tuning, such differences may actually represent distinctions that play a meaningful pragmatic role in the local TL speech community and are thus worthy of attention in assessment.

2.3 Assessing speech acts

Currently, there is a pullback from speech acts because they only represent one area of pragmatics, albeit a highly researched one. While numerous studies have been conducted on requesting behavior,

https://doi.org/10.1017/S0261444819000156 Published online by Cambridge University Press

as well as a fair number on apologizing, complaining, complimenting, thanking, and greeting, there is still much of speech act performance that has not been adequately researched. Consequently, such speech act behavior (e.g. teasing or criticizing) has generally not been included in assessment.

It is my observation that there are more doctoral dissertations focusing on pragmalinguistics than on sociopragmatics. A logical explanation for this would be that assessing pragmalinguistics is more straightforward in that the research literature abounds with studies which involve ratings of acceptability for certain language forms in given speech act contexts. An early speech act study, for example, found that there was a pragmalinguistic difference between the use of the intensifiers 'really' and 'very' in apologies in American English (Cohen, Olshtain, & Rosenstein, 1986), with 'really' signifying regret ('I am really sorry to a put you out') and 'very' reflecting more a matter of etiquette ('I am very sorry to have to end this meeting'). Assessment of sociopragmatics, on the other hand, entails dealing with a fine line as to what is or is not acceptable in a given TL community, especially given the variation among native speakers in that community. It encompasses the bigger issues such as what you can say to whom, when, and how – for example, asking people questions that may be sociopragmatically inappropriate depending on the culture, such as how much they make a month, what their new house or car cost them, or whether they are having an in vitro fertilization procedure in order to induce pregnancy.

It is likely, therefore, that there is an under abundance of sociopragmatic tasks because they are challenging to design effectively. Perhaps those who engage in pragmatics assessment need more examples of subtle distinctions in sociopragmatics for which the assessment is not a dead giveaway. After all, a good test is not one where every item and task is so obvious that no one gets anything wrong.

In assessing speech acts, an area for assessment could be that of determining the extent of L1 cultural overlay taking place when learners are performing their TL pragmatics, especially in a foreign as opposed to an L2 context. Assuming that these differences are taught in classroom instruction, then it would be the job of assessment to evaluate how effectively the learners have incorporated these cultural differences in their performance. Of course, nonnative speakers (NNSs) may prefer not to include these features in their own performance out of a sense of agency. This then could become a constraint in determining what to assess. Ishihara (2014) would suggest not only having students perform the given oral task but also having them indicate their intentions in their response to the task. Obtaining this extra information would allow teachers to give the students credit for indicating in an assessment situation their awareness of cross-language strategy differences with regard to, say, deference to seniority to which they themselves were unwilling to conform.

For example, when I was studying Japanese, my L1 cultural overlay had me take exception to having to use honorific verbs when referring to people of higher status, especially when they were not present (e.g. *mesheagaru* instead of *taberu* for the verb 'to eat'). In light of the fact that respondents may wish to exercise their agency by refraining from engaging in the called-for pragmatic behavior out of a sense that it is inconsistent with their self-identity, respondents could get credit for indicating their awareness of the requested pragmatic behavior that they were unwilling to perform. In this fashion, the person doing the assessment would be able to find out the level of awareness respondents have about the pragmatic behavior in question.

A major reason for negative transfer from L1 culture is that there are certain speech-act-specific strategies which NSs of a given language tend to prefer in performing the speech act. Since the speakers are so accustomed to using the strategies in their L1, they may even be oblivious to the fact that the strategy use is carried over to the TL. For example, the response of Japanese speakers in an L1 apology or a refusal may be to stutter or stammer as a means for indicating proper deference or humility (see audio clips provided at http://carla.umn.edu/speechacts/japanese/introtospeechacts/index.htm). Consequently, Japanese speakers may inadvertently add this behavior to performing these speech acts in a TL. Another example would be that of American learners of Arabic employing an overt expression of apology and specific reasons for why they committed the act for which there apologizing more than native speakers of Arabic (Al Masaeed, Waugh, & Burns, 2018).

https://doi.org/10.1017/S0261444819000156 Published online by Cambridge University Press

It is most likely because speech acts provide assessors a ready source of deviations from normative pragmatic behavior that this area is featured so prominently in assessment. As research studies multiply in other areas of pragmatics, it is likely that assessment will begin to feature some of these other areas more prominently.

2.4 The issue of prosody

As suggested above, Japanese speakers may stutter or stammer slightly in their delivery of apologies or refusals as a sign of humility rather than as an indication of dysfluency. This behavior enters into the realm of prosody, an area that is often neglected in the assessment of pragmatics. Again, assuming that these areas have been taught, it would be possible in assessing oral language to look at the ways that speakers express tentativeness, politeness, or degrees of directness or indirectness, such as through pitch direction, pitch range, pauses, loudness, tempo, and voice quality.

Some years ago an effort was made to classify tones into four categories (Beebe & Waring, 2002):

1. an affective stance – emotions (personal feelings like anger or generalized moods like somber) and attitudes (feelings towards hearer or opinions about hearer like sexist or racist tone)
2. an epistemic stance (from certainty to tentativeness)
3. a relational stance (by role, level of acquaintance, or status)
4. a stance conveyed by a speech activity (e.g. threatening or intimidating, lecturing or courtroom tone).

Given that perception of tone can be a challenge even for NSs, dealing with tone would be expected to be that much more difficult for NNSs – especially for languages which deal with tone in radically different ways (e.g. American English vs. Mandarin). Perhaps it is no wonder that perception and production of tone is not that common in TL pragmatics assessment.

In a study looking at the perception of sarcasm among intermediate and advanced English as a foreign language (EFL) students (Koh, Lee, & Lee, 2018), video clips from *Friends* and from *Modern Family* were used which involved the use of both sincere and sarcastic tone. Respondents had to indicate the correct implied meaning. This was followed by a semi-structured interview in which respondents indicated how they had processed the task, what they had focused on, and the difficulties that they had had. A key finding was that if the literal meaning was negative, then the respondents had more difficulty perceiving it.

In a study looking at the production of prosody in the pragmatics performance of a nonnative English-speaking surgeon (Yates, 2017), it was found that the surgeon did not fully control the placement of stress on words that were important for her meaning, and as a result her stress on the first-person pronoun came across as too salient. She failed to use the unstressed variants in places where they would be expected in Australian communicative culture. Consequently, her pronunciation of Australian English at times gave the impression that she was inflating her own importance.

To avoid pragmatic failure, it may be crucial for speakers to perceive clues to pragmatic meaning transmitted through prosody, such as whether an interlocutor is being straightforward or sarcastic. If these features are included in the assessment of pragmatics, then presumably teachers have made some effort to include them in their instruction, which in turn prompts learners to pay more attention to them as well. Since the presence or absence of certain tonal features may have a real impact on the pragmatic import of the communication, this would appear to be an area of concern for assessment efforts.

2.5 Language learner strategies for TL pragmatics

The consideration of test-taking strategies is usually not a factor in the assessment of TL pragmatics because until now the research in this area has been scant. Fortunately, researchers are picking up on

https://doi.org/10.1017/S0261444819000156 Published online by Cambridge University Press

the challenge made in an early effort to categorize pragmatics strategies (Cohen, 2005). A study by Youn and Bi (2019), for example, looked at the pragmatics strategies which 30 respondents reported using while engaging in four pragmatics English for academic purposes (EAP) tasks. The findings indicated that the strategies found to be used included metacognitive and cognitive strategies related to task-based performance, as well as strategies specific to the performance of TL pragmatics – namely, situation-related sociopragmatic, pragmalinguistic, and interactional strategies.

It is also natural that respondents will look for whatever test-taking strategies can help them in responding to what are sometimes formidable tasks, especially since for many test takers responding orally is a major confront (Cohen, 2006). Respondents will use not only test-management strategies aimed at efficiently doing what they respondents are expected to do, but also test-wiseness strategies, which have as their goal to outsmart the measure in some way or other. In other words, respondents may purposely not engage in the activity at hand as intended, but rather deploy clever strategies in order to avoid revealing what they feel incompetent doing with regard to pragmatics. For example, as documented in a role-play study, respondents may purposely or inadvertently ignore some of the elements in the given prompt for the sake of alacrity or because they simply do not know how to deal with them (Kasper & Youn, 2017). Then it becomes the job of the raters to determine whether to rate them just for what they have done or whether to penalize them for what they did not do.

Since it can be problematic to assess strategy use on a measure based exclusively on the outcome, test-taking strategy work has been relegated more to the area of research than to that of, say, classroom assessment. My own view is that discussion about the use of language learner strategies and test-taking strategies should take place as part of instruction and possibly for test preparation since a systematic use of such strategies has been seen to enhance performance (Cohen, 2006).

Now that we have looked at what may be assessed in TL pragmatics, let us consider the instruments that could be used in an effort to assess spoken performance in TL pragmatics.

3. Determining the measures to use

What devices or instruments will be enlisted to obtain the data? What procedures, tasks, and specific items will constitute the means of gathering the data that will then be scored and in classroom assessment be graded as well?

3.1 DCT

The main frenemy in pragmatics assessment today seems to be the DCT. While this set of measures (i.e. the oral, written, and multiple-choice versions – with or without interaction; see Usó-Juan & Martínez-Flor, 2014) continuously receives criticism in the literature, it continues to be used extensively. A main advantage of DCTs is their practicality: They can be digitized, administered to many test takers, and easily modified. Moreover, compared to performance-based test items (i.e. role plays and role enactments), the scoring of DCTs can be accomplished relatively easily once the raters have been properly oriented in the criteria for rating are clear. In addition, it is worth noting that oral DCTs are now being advocated as an alternative to written DCTs as they are more practical than role plays but still directly collect oral data.

Taguchi and Roever (2017) see DCT use as defensible if the research focus is purely on describing the range of speech-act strategies that respondents have available with no claims made about their ability to use them in actual discourse. Given the current climate supporting interactional competence, Roever (2018) calls attention to the fact that DCTs do not allow for extended interaction, forcing respondents to put into a single gap what in reality might unfold over several terms with utterances being likely to constantly responsive to the interlocutor's previous utterances. Roever is also concerned that in reality conversation is likely to unfold differently from the way that it appears in oral response items, and may well not include the head acts for speech acts such as requesting, refusing, and apologizing. Similarly, Youn and Bogorevich (2019) warn that because DCT-elicited

https://doi.org/10.1017/S0261444819000156 Published online by Cambridge University Press

responses are reflective of what respondents think they would say instead of what they would actually produce in real-life discourse, depending on how plausible the DCT prompts are, they may serve as rather weak evidence in making inferences about test takers' pragmatic ability in comparison to naturalistic measures.

Despite the criticism of DCTs, an oral version in particular can at least provide a general impression of the respondents' ability to produce speech acts rather than a measure of their ability to fine-tune pragmatic distinctions appropriate for extended discourse. One way to counteract the argument that filling in a single gap is limiting would be to use a response mode involving multiple responses (Cohen & Shively, 2002/2003). In other words, the response set would entail two or more turns, thus providing a richer sampling of language than that possible with a single turn.

Whereas it has been convenient to consider written DCTs based on actual oral interactions to be an acceptable indirect measure of oral language, the study by Eslami and Mirzaei (2014) described in Section 2.2 showing distinct differences between written and oral language in Farsi should give test constructors pause for thought if they are dealing with languages that have such distinctions. Under any circumstances, respondents to written DCTs need instructions and possibly coaching in how to write responses that reflect what they would actually say.

3.2 Role play

Another popular approach to collection of spoken data for TL pragmatics analysis is role play in its various forms. Félix-Brasdefer (2018) provides a thorough description of the various kinds of role play that can be utilized – including closed role play, open role play, simulated role play (where respondents take the roles they normally assume in life), and naturalized role play (where the respondents perform in a structured setting but unaware of the specific focus of the pragmatics assessment).

Kasper and Youn (2017) did a close-order analysis of what role-play interaction actually looked like, and found that participants selected certain details from the instructions and disregarded others, and how they co-constructed responses on this basis. In fact, the role-play participants continuously generated the setting. The study was for research purposes and was not aimed at rating pragmatic performance. The selected scenario was a student's request for a letter of recommendation from a professor during an office visit, an activity that had turned up in a needs analysis as a high-priority item for stakeholders. A clear outcome of the analysis was that for the participants not all role-play specifications were equal. Given that so much role play is conducted and utilized in the collection of pragmatics data, the results of this study would suggest that role-play instructions need to be carefully piloted and respondents need to be cognizant of the importance of following these instructions closely. More recently, Youn (2018b) has suggested that as long as it is carefully designed, role play can yield valid data on task-based pragmatic performance – and that would include both structured and open role play.

One final note here is that traditionally in role-play assessment, respondents have been asked to interact with the same interlocutor who plays different roles situation-by-situation. Taguchi (2018) points out that this may cause participants difficulty in imagining themselves being in different situations. A way around this would be to show respondents pictures of who their interlocutors are supposed to be in each situation and then to have the test administrator assume the role of these various interlocutors so that the role play can be carried out.

3.3 Naturalistic data

As suggested in Section 2 above, the analysis of naturally recorded interaction can provide researchers with evidence of how interaction unfolds turn-by-turn as participants engage in conversations – the speech acts that they use, their conversational management, and so forth (Taguchi, 2018). Shively (2018) provides a series of further arguments for using data from natural settings: (1) they provide an opportunity to observe L2 speakers in real-life situations in which there are more likely to be

https://doi.org/10.1017/S0261444819000156 Published online by Cambridge University Press

more genuine than is the case in experimental tasks; (2) they can provide insights into how NNSs discursively negotiate their identities, pursue and achieve their interactional goals, recontextualize language for their own purposes, and use language to build and maintain the relationships in which they are involved; (3) they provide an opportunity to observe pragmatic features in their social and sequential context, as well as to capture the entire speech event as it unfolds, rather than isolating specific social actions as elicited methods tend to do; and (4) they reveal more variation in how participants realize speech acts, thereby providing a more complete understanding of the phenomenon compared to elicited tasks, which may only produce the most conventionalized interactional moves.

As an example of recent research involving naturalistic data, Youn (under review) used conversational analysis to look at the nature of interactional competence at various levels of achievement in the context of role-play speaking assessment. It was found that higher-level test takers used various shift markers and stepwise transitions when initiating and shifting actions, whereas the opening turn of lower-level test takers was typically without pre-sequences in which recipients would orient to a departure from a normative turn-taking structure. The researcher viewed the distinct sequential organization and interactional features found across test-takers' levels as evidence that her assessment of interactional competence was valid.

Another paper on the same database (Youn, 2019) compared the fine-grained interactional features elicited from test-takers' role-play performance with raters' scores. The corpus of 102 test-takers' role-play performance was transcribed turn-by-turn and analyzed qualitatively to identify recurrent interactional phenomena. A linear regression analysis indicated that the combination of nine significant interactional features accounted for 75% of the variance in the raters' scores. A discriminant function analysis revealed that the interactional features elicited from formal role plays were important variables in distinguishing higher from lower performance levels.

Youn would consider it a game changer to view the assessment of pragmatics in spoken language via the construct of interactional competence since this means an almost mandatory focus on co-constructed responses. If this is the case, then it becomes more of a challenge on many levels to assess TL pragmatics. For example, it becomes the raters' responsibility, then, to evaluate L2 users' ability to deploy interactional tools that lead to shared understandings (Roever & Kasper, 2018; Roever & Ikeda, in press).

Taguchi (2018) would suggest that an ideal assessment situation is one in which respondents are engaged in absolutely natural discourse so that raters can check out their facility at turn taking and at supplying appropriate forms for their contribution to each adjacency pair. She notes that whereas naturalistic interaction is uncontrolled and uncontrollable, it nonetheless gives the most informative data – how interaction unfolds turn-by-turn as participants jointly perform speech acts, monitor the changing course of conversation, and attempt to deal with these changes.

3.4 Field notes and verbal report

Field notes and verbal report techniques may help researchers better understand the results obtained from other assessment measures. While these are themselves measures, they can also serve as a way to triangulate data and help to validate the other measures, with the overall goal of more accurately determining the nature of the pragmatic intent involved.

It is a challenge to be comprehensive in describing, for example, the relevant dynamics of an interaction involving something as simple as the response to a 'thank you.' What about nonverbal responses such as a nod or a smile, which cannot easily be elicited by employing written DCTs? One pragmatics study used field notes as a data collection method for picking up on these clues to intended meaning among adults in the US and Canada (Bieswanger, 2015). The researcher also identified a vocal grunt as an additional response type among adults which he labeled 'mh-hm,' as a form of response to a 'thank you' among adults. Another study looked at rudeness in New York City through gathering examples by writing them down in a notebook every time that they were heard. Low- and high-proficiency learners of English then reacted to this material (Beebe & Waring, 2005).

https://doi.org/10.1017/S0261444819000156 Published online by Cambridge University Press

Verbal report has notably reflected a host of different approaches (see Bowles, 2010; Cohen, 2013, 2018, p. 253). For starters, there is SELF-REPORT, where respondents provide descriptions of what they do (or rather, what they think they do), characterized by generalized statements. Then there is SELF-OBSERVATION, which is the inspection of specific, not generalized, language behavior, either introspectively (i.e. within 20 seconds of the mental event) or retrospectively. A further distinction is made between IMMEDIATE RETROSPECTION (i.e. within an hour or so of the mental event) and DELAYED RETROSPECTION (say, up to a week or more after the event). Finally, there is SELF-REVELATION, typically referred to as think-aloud, which is characterized by the stream-of-consciousness disclosure of thought processes while the information is being heard and processed.

Some years ago, Elite Olshtain and I demonstrated how verbal report could be used in numerous ways to investigate how 15 EFL respondents planned apologies, complaints, and requests. The use of this technique afforded us the ability to collect data on the participants' account of their search, retrieval, and selection of language forms, and the extent to which they attended to grammar and pronunciation in their production of speech act utterances (Cohen & Olshtain, 1993). The findings included information about the retrieval process such as the languages that the respondents used in producing their speech acts. Through the use of verbal report techniques we were able to identify moments of:

- self-debate before selection of material
- afterthoughts
- awareness as to how much they were monitoring
- awareness of using formulaic speech
- instances of omissions and the explanation
- abandonment of words or expressions and why
- only partial delivery of a thought and why
- delivery of a different thought because it was doable
- lexical avoidance, simplification, or approximation.

3.5 The use of a mixed-method approach

A presumably catchall approach is that of using mixed methods in an effort to triangulate data. For example, the study could call both for the assessment of pragmatics in a robust sampling of learners, as well as measures of pragmatic abilities from a subsample, especially when using measures that are more time-consuming. Over the years I have sought to triangulate assessment in order to get a more accurate sense of respondents' handling of pragmatics – especially with the addition of verbal report to collect information from respondents on the processes involved and on the strategies that they used in producing their responses.

An example of a mixed-method approach is that of Schauer (2009), where she made use of diary data, rank-ordering tasks, multiple-choice questionnaires, multimedia instruments, production questionnaires, elicited conversation, role plays, observation of naturally occurring data, and verbal report. Though by now somewhat dated, her description serves as a testament to efforts made in pragmatics research to look carefully at various options for data collection, weighing the relative merits of each approach to measurement. A recent study by the same author (Schauer, 2017) demonstrated how a combined use of perception of appropriateness and politeness could be assessed both through a questionnaire and then an oral interview either in person or by telephone or Skype in order to obtain corroborating information from respondents.

There are numerous reasons why researchers choose to limit their instrumentation to one or two measures. Financial and human resources and time available for research are some of the constraints involved. Brown (2017), for instance, paints a rather dim picture of prospects for research given that most institutional review boards have more stringent human rights regulations in place than ever before and university budgets for such research efforts are decreasing. Despite these possible constraints, there are usually advantages in employing multiple methods for collecting data, being fully aware of considerations regarding each one.

https://doi.org/10.1017/S0261444819000156 Published online by Cambridge University Press

The desired take-away from this section on what measures to use would ideally be to make use of multiple measures. However, there is always a need to make sure that every given measure has been properly piloted on respondents similar to those in the testing pool, and there should be clear instructions as to how to perform the given tasks. In addition, it is important to be cognizant of possible differences between oral and written styles of language if indirect measures of oral performance are to be used.

4. Approaches to collecting data

It is not enough to identify which measures to use for assessment; attention needs to be given to just how these measures are used to collect the data.

4.1 Challenges in collecting naturalistic data

A significant challenge encountered by Shively when using naturalistic data to study pragmatic performance across subjects was that of getting comparable data from each of the respondents (Shively, 2008). This led her to focus on service encounter data because of their routine structure, sequence, and goal-oriented nature: (1) opening, (2) request for product, (3) optional negotiation, (4) provision of product or service, (5) payment, and (6) closing (Shively, 2018). The rationale for this was that service encounter data were found to be relatively consistent across participants as compared to other types of natural interactions such as everyday conversations.

Another issue regarding the approach to data collection is that of using video- as opposed to just audio recordings. A case in point is that of collecting data on humor. As a result of comparing the audio recordings which were the sole source of data in a 2008 study with video recordings that she used in a later study, Shively reported that the latter provided an opportunity to look at spontaneous verbal humor since smiling and facial expression were found to be important ways that speakers signaled humorous intent (Shively, 2013). She noted that there were instances of humor in her 2008 corpus that were not accompanied by laughter, but that may have been delivered with a smile. Video recordings would have allowed her to better determine whether cases of failed humor may have been related to the lack of visual cues signaling the humorous frame.

On the basis of her several studies involving naturalistic data, Shively concluded that despite the numerous challenges associated with collecting data from natural settings, if the goal of pragmatics is to study language in context, then investigating language use and learning in real-life interactions is crucial and the investment in time and resources to do so is worth it (Shively, 2018). If classroom instructors are in favor of including naturalistic data in their assessment of pragmatics, then the challenge is to find ways to do it expediently. One way would be to have students collect video data on their interactions out of class and to share them, say, on their Google drive.

4.2 Teachers as data gatherers

With regard to the data gatherers, at times it could be the students themselves as suggested just above. In research studies, outside investigators or teachers are the ones collecting data in classrooms. A concern about having teachers collecting research data – especially with younger children – is that the data may be incomplete, inconsistent, or inaccurate (Ishihara & Chiba, 2014). Having teachers as data collectors may mean fewer reactive effects from the assessment in that the respondents may be more comfortable with their presence then with unknown outsiders. There are most certainly trade-offs both ways.

4.3 Towards collecting more insightful data

There are strategies which test constructors can use in order to make oral measures of TL pragmatics more manageable (see Cohen, 2018, section 9.4 on strategies for test construction). They include keeping the pragmatics tasks realistic for the given TL community and engaging for the respondents. It can

https://doi.org/10.1017/S0261444819000156 Published online by Cambridge University Press

also be beneficial to have a discussion with the students after they have performed the tasks to see how well they understood the contextual factors, and if possible to have them identify the factors that most influenced their responses. Another would be to have the respondents compare their performance with how they think an NS would respond in order to see the extent to which they might have been expressing their agency by not engaging in the expected pragmatic performance.

In order to heighten the usefulness of the data, the tasks need to be crystal-clear to the respondents. To this end it is crucial to engage in adequate piloting with similar respondents. In addition, it helps if the tasks are meaningful and engaging for the respondents. It is probably no surprise that one reason respondents enjoy performing speech act tasks is that they have high face validity given their prevalence in interactional data. A spinoff effect of a truly enjoyable task is that the respondents may pay less attention to the fact that they are being assessed and get more into the moment of their pragmatic performance – which in turn can enhance the validity of the measure.

4.4 Coaching respondents in test-management strategies

There are likely to be strategies which learners can benefit from both in preparing for such tests and in actually taking them. For example, respondents can take practice measures involving the same kinds of tasks, paying strict attention to any instructions indicating to them the role relationship between themselves as respondents and their interlocutors.

One of the reasons for an interest in naturalistic data is that in the traditional testing situation, respondents may underperform for various reasons. One reason is that they may simply not be warmed up enough to complete the task. They may only be ready for the speaking task after it is over. A term coined many years ago to describe getting the mind ready for TL performance was DIN IN THE HEAD (Krashen, 1983). So a strategy might be to do imaginary practice beforehand just as a musician warms up a musical instrument before playing. This would be to get the din in the head going in the respondents' TL so that they do not draw a blank in the testing situation. Another reason for underperformance is that the respondents become overly self-conscious and even anxious when they are having their performance formally assessed. They may do a better job at performing their pragmatics when they are with a friend in an informal situation.

Hence, in order to collect truly reliable and valid data, attention needs to be given to prepping the respondents for the given tasks. Even with supposedly naturalistic data, it is likely that if the respondents know that audio- or video-taping of their speaking is going to be assessed, it may influence how they perform.

5. Considerations in data analysis

If the assessment is for research purposes, then there is no pressure to assign grades and to explain these grades to learners. Hence, there is no need for accountability to the respondents. If the purpose is for assessment of how effectively pragmatics is taught and learned, then attention needs to be given to just how informative the scoring is and how diagnostically helpful the results may be when shared with the learners.

5.1 Metapragmatic judgments

As far as measures of metapragmatic judgment go, there can be a problem in setting the NS benchmark since NSs may not necessarily agree on the pragmatic appropriateness of utterances. Roever (2018) contends that it is difficult to construct target utterances for metapragmatic judgment items which are clearly inappropriate but not wildly implausible. Consequently, the inappropriate items are often easy to recognize as inappropriate (the same is true for multiple-choice DCTs). Also, there does not appear to be a systematic way to identify utterances which are clearly inappropriate to NSs but less clearly so to learners. This then imposes itself as a challenge for test constructors – to design metapragmatic judgment items that work well.

https://doi.org/10.1017/S0261444819000156 Published online by Cambridge University Press

5.2 Determining the basis for norms

A current issue of concern is whether the norms being used by the raters are those being adhered to in local varieties of the given language, be it English, Spanish, or some other language that has various local varieties. If, for example, German and Japanese business associates are using English as a lingua franca, who is to determine what the appropriate pragmatics would be for such an interaction? A recent study of 125 EFL teachers in Iran found that these NNS teachers gave preference to NS linguistic norms despite emerging norms for English as an international language (EIL) which diverge from the NS norms (Tajeddin, Alemi, & Pashmforoosh, 2017). The findings from this study, then, lend support to adhering to a more traditional notion of NS norms. If there is any serious effort to diverge from this approach, then researchers need to provide an indication of just what the more local norms might look like – such as having pronounced features of one or another local variety in terms of pragmatic performance.

For instance, EIL would presumably include as acceptable (and perhaps even normative) behavior the incorporation by Japanese English speakers of their L1 pragmatics for stuttering or stammering during an apology or refusal as a sign of deference (see Section 3.3 above), as well as incorporating silence into their interactions in English. Both kinds of behaviors could have an effect on how NSs react in the interaction, especially if they are not aware of what is behind this seeming divergence from normative behavior. NS speakers of American English would tend to speak without hesitation in apologizing or refusing, and would more or less avoid what to them might seem like embarrassing silence in their interactions. For one thing, they would be likely to fill pauses.

5.3 Scales for rating role plays

Recent research would suggest including the following scales in the rating of role play:

- content delivery: smoothness and fluidity of turn initiations and transitions
- language use: range of pragmalinguistic tools in terms of structures, modals to indicate indirectness
- sensitivity to the situation: tailoring accounts and explanations to the given recipient,
- engagement with the interaction: tracking previous turns and responding as appropriate at the given moment (see Section 5.4 below for more on this)
- organization of the turns: completeness of adjacency pairs and appropriateness of pauses. (Youn, 2015)

To make these ratings work, raters have to have a sense of what fluidity means in turn taking, just as respondents need to know how they are expected to make transitions in a pragmatically appropriate way. The speech act has to be sensitive to the given situation and to the interlocutor. Raters need to have a clear sense of what conversational management in general and turn-taking ability in particular would consist of in the given situation.

Rating criteria can be devised either a priori by a group of experts in language testing or by using a data-driven approach based on qualitative analysis of performance data. In the latter approach, researchers focus on analyzing performance elicited from assessment tasks in the process of generating the criteria to better represent the context of performance (Youn & Bogorevich, 2019). While this use of data-driven criteria for rating seems like an appropriate way to establish the criteria, it is usually labor-intensive. Youn (2015) found that raters successfully used rating criteria with detailed descriptions in rating open-ended role play involving requests and decision-making. Her findings encouraged her to recommend a data-driven approach to developing rating criteria not only in making sound inferences about complex performances (such as those dealing with the TL), but also in obtaining scoring decisions that would be consistent across raters.

In a more recent study, Youn (2018a) investigated rater criteria in paired speaking assessment and resulting instances of rater variation. The main concern was whether the raters appropriately accommodated to the level of proficiency of the respondents. She provided various interpretations for what

https://doi.org/10.1017/S0261444819000156 Published online by Cambridge University Press

might have caused raters difficulty in scoring responses. For example, 9 out of her 12 raters displayed inconsistent severity patterns when rating for *language use*, which measured the ability of using appropriate linguistic expressions in making an agreement and disagreement. Of these nine raters, six raters showed more leniency than their averaged severity. Youn speculated that the severe raters might have compensated for their severity by giving greater leniency to low-proficiency students on this scale, and as a means for dealing with unintelligible utterances from these students.

In cases like this where interpreting rater behavior remains somewhat problematic, it may be valuable to use verbal report in order to collect data while raters are in the process of ruminating about their choices and getting them to indicate their rationale for assigning a given rating. Some years ago, Hamp-Lyons and Condon had raters of writing portfolios keep a written log of the criteria that they used in assessing the portfolios and the processes involved. This approach revealed numerous cases of idiosyncratic behavior among raters (Hamp-Lyons & Condon, 1993). Nowadays, there are numerous data capture programs available that could facilitate such data collection in an effort to better understand the rating process.

5.4 Rating for changes in contextual variables

Taguchi (2018) stresses that the form-function-context relationship is not stable nor predetermined, but rather is emergent, and that for this reason it is important to be aware of (and ideally capture in ratings?) the changing nature of the context, such as shifts in the interlocutors' degree of familiarity with each other, their attitudes, their emotions, and the directions that the discourse takes.

Contextual changes can include a shift in level of familiarity and in degree of politeness. There are situations in which getting some issue resolved in a service encounter calls being impolite in order to get a clerk to deal more promptly with an issue in a given situation within a TL communities. In some instances, it may be of benefit to charm a clerk by a shift to being extra friendly. The issue of attitude or mood is another interesting one. Rarely in a pragmatics test are there cues (as in a screenplay) as to the attitude the respondent should be assuming in the given situation –whether the respondent should act happy, amenable, hostile, angry, skeptical, sarcastic, and so forth. It is true that the response could vary sometimes markedly if the respondent is intended to reply cordially as opposed to angrily.

Taguchi cautions that this more informed approach to the role of context then adds another layer to the definition of pragmatic ability: the ability to adapt to the changing course of the interaction. While it sounds good to have assessment pick up on the changing course of interactive discourse, the reality of designing a measure that would tap this effectively would be challenging at best. Perhaps with ever-improving technological advances, the possibility of having respondents adequately engaged in interactive discourse is greater than it has been in the past.

5.5 Rater variability

Still on the issue of rater variability, there are studies comparing NS and NNS raters. One such study compared American English as a second language teachers' ratings of responses to a written DCT involving requesting to ratings by Iranian EFL teachers and found some marked differences (Alemi & Khanlarzadeh, 2017). Iranian EFL teachers were inclined to mention only one or two criteria in each situation, while American EFL teachers mentioned their criteria more precisely with more detail. In addition, the Iranian raters' failed to notice the use of softeners (e.g. use of modals) and query preparatory statements (e.g. giving a reason for the request). Furthermore, the Iranian teachers rated the EFL learners' answers more severely than the American raters in four out of the six situations involving a request.

There are also studies that compare NS teacher ratings to those of real-world individuals who are stakeholders in the target context, but not trained as raters. One study, for example, looking at NS teacher ratings of Japanese foreign-language (FL) students studying the TL to be tour guides found

https://doi.org/10.1017/S0261444819000156 Published online by Cambridge University Press

that the teachers paid most attention to grammar, vocabulary, and fluency in their rating, whereas raters who were actually tour guides paid more attention to the candidates' pronunciation and politeness (Brown, 1995). Just as study-abroad students may be surprised to find that people whom they interact with in the host community are not as patient or accommodating as their teachers were in class before they went abroad, it may also be the case that raters bend over backwards to give respondents the benefit of the doubt, whereas in the real world, issues of pronunciation and perceived impoliteness may be the principal contributors to a successful interaction.

But lest we lump all tour guides into one group, let us remember that there may be variation in how they rate NSs. A study of NS teacher ratings of EFL apologies found raters to exhibit different degrees of severity and leniency in their ratings, contradicting the popular view that NSs provide a benchmark in language assessment, including pragmatic assessment (Tajeddin & Alemi, 2013). A follow-up study looking at raters using a macro-level variable, namely, the dialect of the raters, revealed differences between American and British raters who were rating complaints by Iranian EFL learners (Alemi & Khanlarzadeh, 2015). It was found that American raters put the provision of a clear analysis of the complaint situation well above the issue of politeness, which was not the case with British raters.

With regard to the rating of multi-turn DCTs, Roever (2018) notes that there are some major questions as to their validity. In this procedure, the raters first assess the fit of the response to the gap, and then its idiomaticity. Roever would contend that the validity of the ratings would depend on macro-level variables associated with the raters – namely their social class, their level of education, as well as their geographic location (as already indicated above in other research). Their educational level alone, for example, could have a bearing on how appropriate they judge the pragmatics to be for any given situation.

In addition, there are chronic problems surfacing among NNS raters as well. There is evidence that rater-orientation courses can improve NNS ratings, as in the case of Iranian NNS EFL raters producing more accurate and consistent scores for responses on a written DCT involving requests, more approximating the native benchmark and also being more consistent with their nonnative peers' ratings (Tajeddin & Alemi, 2014). Yet, it was also found that L1 cultural norms and perceptions of appropriateness relating to certain refusal contexts could not be easily reshaped through a rater-orientation course. Another study provided more detail as to similarities and differences among NNS EFL raters, both with regard to sociopragmatic and pragmalinguistic dimensions of ratings apology role-play situations (Tajeddin & Alizadeh, 2015).

One of the more detailed teacher rater studies had EFL learners provide complements for seven situations, and then 60 EFL teachers were asked to rate these responses and to indicate the criteria that they used for their ratings (Alemi, Eslami Rasekh, & Rezanejad, 2014). The criteria in order of frequency of mention among the teachers were: politeness, the relationship between the interlocutors, the degree of tact, authenticity of the compliment, how sincere it sounded, length and level of formality (e.g. not too short or too formal), and linguistic appropriacy (in terms of grammar and lexical choice). The findings showed noticeable variation among NNS raters in all of the rating categories.

5.6 Rating for fluency

Finally, how important is it that the respondents be considered fluent in their responses? Traditionally, hesitancy phenomena have been considered a useful yardstick for how effective someone is in getting across the message. But there are language communities in which hesitancy phenomena are a sign of deference, humility, and overall politeness. This example was given in Section 3.3 of how Japanese speakers decline an invitation, apologize for some behavior, or perform other face-threatening speech acts: it may be crucial that they employee hesitancy phenomena – that they stutter and stammer – so that they do not come across as arrogant, pompous, or too sure of themselves.

This section has provided a considerable amount of food for thought regarding the processes involved in producing ratings of pragmatic performance. Ideally, the effect of this extended focus will be to underscore the need for care in arriving at reliable and valid scores and grades for

https://doi.org/10.1017/S0261444819000156 Published online by Cambridge University Press

respondents. Even with university-level assessment, if teachers are the raters, they may adopt a level of leniency inconsistent with the research purposes. For example, teachers may be inclined to use an idealized reference for performance not based on how NSs actually perform pragmatically, but rather on how teachers feel intuitively they would perform – as was the case in a study reported by Ishihara (2013). This then would be a motive for having empirical data from NSs as a benchmark, assuming that consensus is reached as to what that model might reflect in terms of regional variety and the appropriate level of formality, as in the case of establishing EIL norms as discussed in Section 5.2. Presumably, for classroom assessment of pragmatics any discrepancies with regard to a single measure would be offset by the results from other measures. In a research setting, on the other hand, this would often not be the case, which is all the more reason why the data from such measures need to be analyzed in a rigorous way.

6. Assessment of TL pragmatics in the classroom

Let us now take a closer look at some issues related to classroom assessment of pragmatics. Given the host of factors enumerated above that could and perhaps should go into any rigorous effort to collect TL pragmatics data for assessment purposes, it is probably no wonder that classroom teachers have tended over the years to give only limited attention to pragmatic issues in their oral assessment. If there is genuine interest in pragmatics assessment for the language classroom, then the issue of expediency comes to the fore. It can be rather intimidating for any given classroom teacher to choose from among the numerous factors those to pay attention to in preparing a given set of assessment tasks.

6.1 The washback effect

If a goal of instruction is to send the message that pragmatics is worthy of attention by learners, then an effort needs to be made to include this focus in assessment. Learners are probably more likely to study these elements of language as a consequence. Having to perform oral language tasks on a test can certainly be a motivator for learners to practice the language. In addition, depending on the personality of the students, they may well come away from the experience with a certain sense of satisfaction. However, at the same time the coverage of pragmatics needs to be fair and reasonable. Hence, fine-tuned distinctions that might be applied to assessment instruments intended for research purposes may not be appropriate for classroom use. Instead, more basic elements would need to be assessed, with the understanding that the intention is to promote basic pragmatic performance: major speech acts, implicature, and other areas.

With the advent of more sophisticated technological advances, assessment of pragmatics has become more within reach. The question is what to assess in a classroom setting and how much time and effort to put into such assessment. If the assessment is for the learners, then a premium should be on assessing TL pragmatic behavior of high value or high impact for those given students in their specific learning context – for instance, how to greet, make requests, give and respond to compliments, and deal effectively with implicature. If learners perceive that the feedback will have a high payoff, they are likely to pay greater attention and ideally benefit more from it.

A study looking empirically at the washback effect of a written DCT on FL teachers and learners found that added pragmatics tasks enhanced treatment-group learners' noticing of pragmatic elements in the listening materials and teacher talk (Tajeddin & Dabbagh, 2015). In addition, learners in the pragmatics treatment included the production of pragmatically appropriate discourse among the goals of their language learning.

6.2 A discursive approach to pragmatics

As noted extensively in this article, there is interest in assessing the development of pragmatic ability within the larger discourse context (see, for example, Al-Gahtani & Roever, 2012). Roever and Ikeda

https://doi.org/10.1017/S0261444819000156 Published online by Cambridge University Press

(in press) point out that it is a daunting challenge to produce a holistic measure of pragmatics in the classroom, especially since it is even a challenge to do so in research settings. A more feasible option may be to focus on one or more high-stakes aspects of classroom behavior, such as the ability to participate effectively in a classroom discussion. College instructors who would like their NNSs to vie more effectively for speaking time in seminars may want to teach them the pragmatics of how to break into a classroom discussion appropriately, how to hold the floor, and how adjacency pairs work in an interaction (see Caprario, 2018). The instructors might also want to help their students refine their use of speech acts to be used in getting the floor politely, agreeing or disagreeing, complementing or challenging fellow students on their ideas, hedging so that they do not sound overly authoritative, and closing the interaction. Assessment then would focus on the ability of learners to engage in conversational management in the TL.

6.3 Efforts to enhance the viability of pragmatics assessment in the classroom

Teachers are increasingly making use of corpora to obtain samples of language that more accurately represent what TL speakers are likely to say, as opposed to what appears in textbook dialogues based on the intuitions of the textbook writers (Bardovi-Harlig, 2017). Another issue is whether this classroom assessment is relatively one-shot or whether it constitutes a whole program over the course of, say, a semester. An example of an extended program of pragmatics assessment is that offered by Ishihara (2009) in her description of a freshman EFL class in Japan. The instructor used collaboratively developed authentic assessment tools, such as reflective writing, rubrics, role plays, and self/peer-assessment, and facilitated interaction and assessment in the learning process. These assessments were designed (a) to elicit learners' pragmalinguistic competence to use community norms, (b) to elicit their sociopragmatic awareness of the consequences of their own pragmatic language choice, and (c) to evaluate the extent of the match between learners' intention and the interlocutors' interpretation.

Younger learners pose a particular challenge even when more informal measures of assessment measures. But they are not necessarily complete, consistent, or accurate, as a study by Ishihara and Chiba (2014) demonstrated when teacher assessment was compared with an interactive assessment performed by one of the researchers. The use of dynamic assessment is another way to move the action along (Poehner, 2008). An issue with this approach is that it sometimes blurs the distinction between teaching and assessment. In other words, it may be difficult to tell just what parts of the oral output are appropriate for assessment and what parts should be relegated to instruction since responses are co-constructed by the teacher and the student. Whether or not this blurring of the distinction is problematic depends on the given assessment situation.

Under any circumstances, in talking about classroom assessment, there is a concern for being expedient. Classroom teachers need to be concerned about their students' time and therefore do not usually provide an opportunity for their students to do a host of different oral and written tasks in order to arrive at a comprehensively rigorous sampling of their pragmatic abilities. Again, it is possible to have students collect samples of their speaking out of class, but there are limits both to how many such samples it is feasible to have collected and to the logistics of having all or a portion of them evaluated by the teacher.

7. Conclusions

Until the advent of more sophisticated software designed to capture and evaluate pragmatic elements in the discourse of natural interaction, assessment of pragmatics will most likely entail a focus on certain elements. Predictably assessment will continue to focus on certain speech acts, implicature, and conversational turns, while appropriately taking into consideration the proficiency level of the learners, their age, their ethnicity, possible regional dialects, and the gender variable – especially with the advent of new ways of looking at gender. With regard to speech acts, ideally we will get a better handle on how

https://doi.org/10.1017/S0261444819000156 Published online by Cambridge University Press

to assess sociopragmatic awareness and performance, rather than focusing primarily on pragmalinguistics since it is easier to assess.

Whereas negative transfer from the L1 and from L1 culture is highly documented in the literature, a greater role can be afforded the issue of agency in the assessment process. Namely, respondents could be given credit for demonstrating their awareness of pragmatic behaviors that they themselves choose not to perform.

Next, assessment that uses written language as an indirect measure of speaking needs to be informed by TL-specific issues, such as in the case of languages like Farsi which have pronounced differences between spoken and written varieties, and Japanese which has style shifting in oral language from the casual to the polite form.

In addition, there is much that could be done in assessment with regard to the issue of prosody. Perhaps it would warrant having a component of a language course that just deals with issues of stress, pitch, pausing, tempo, and intonation associated with taking a particular stance in a given interaction – e.g. one of disbelief, using sarcasm as a signal. Yet another fruitful area for assessment could be that of the strategies used by respondents to perform, say, speech acts in an appropriate way – such as by knowing which strategies to use in order to make an appropriate apology (e.g. an expression of apology, an acknowledgment of responsibility, and an offer of repair). Assessment of strategies in performance would also include looking at those test-management strategies which help the respondents perform the task most expediently.

Despite the criticism of DCTs, if the interest is in getting a general impression of the respondents' ability to produce speech acts rather than obtaining a measure of their ability to fine-tune pragmatic distinctions appropriate for extended discourse, DCTs can be a viable option. Having said that, it is crucial to use this measure, and any measure, with care. In addition, with regard to role-play measures, just as with DCTs, these measures can provide a viable source of TL performance in pragmatics, as long as those doing the assessment have adequately piloted the measures and have made sure that the participants abide by the given instructions.

With regard to assessment of naturalistic discourse, it is likely that most work will need to be done by researchers, until a point at which such assessment can be brought comfortably into the language classroom as a means of assessing gains in pragmatics from TL language instruction. Again, researchers more than classroom instructors may have the luxury of using field notes and verbal report to better understand the processes involved in producing what are considered to be responses of a pragmatics nature.

As for more use of naturalistic data in assessing pragmatic performance, if the intention is to couple it with classroom instruction, then there need to be ways to accomplish this expediently, such as by having students generate their own video clips of pragmatic performance out of class. There remains the very real issue of how to analyze these video clips. In more advanced classrooms, perhaps the students themselves could be involved in this process as a group effort. Regardless of whether the data are naturalistic or elicited, it could be beneficial to include the learners in discussions before and after the collection of the data to have a better sense of what the learners' performance actually entailed. We could call these a PRE-BRIEFING and a DEBRIEFING. Such discussions would, presumably, include mention of test- management strategies.

It was noted that the basis for norms in assessing pragmatic performance is an area still in need of development. If we are to heed the call to accept more local norms, then efforts need to be made to specify what these norms would entail. Recent research on assessing pragmatic performance in role plays has improved our understanding of, for example, how respondents handle their turns. But it also shows inconsistencies across raters, due to idiosyncratic differences among raters – differences due in part to their status as NS or NNS – as well as the dialect of the TL that they speak or favor.

Personally, I would like to see more work done on how to assess for sociopragmatic control in a way that differentiates among respondents. For these items to be challenging, they must be subtle enough so that they are not obvious. Such assessment would entail, for example, probing the extent to which learners are savvy with respect to topics that are appropriate to broach in a given situation. But designing items aimed at assessing sociopragmatic awareness effectively is not easy.

https://doi.org/10.1017/S0261444819000156 Published online by Cambridge University Press

A final take-away from this paper is that giving attention to pragmatics in assessment can result in positive washback in that both teachers and learners are likely to pay more attention to pragmatics in the classroom, which many language educators would vie was a benefit. In addition, while there may be value in taking a rather broad-brush approach to assessing pragmatics, it may also be beneficial to focus on certain features, such as how to participate in classroom discussion.

References

Alemi, M., Eslami Rasekh, Z., & Rezanejad, A. (2014). Iranian non-native English-speaking teachers' rating criteria regarding the speech act of compliment: An investigation of teachers' variables. *The Journal of Teaching Language Skills*, *6*(3), 21–49.

Alemi, M., & Khanlarzadeh, N. (2015). Native EFL raters' criteria in assessing the speech act of complaint: The case of American and British EFL teachers. *Iranian Journal of Applied Language Studies*, *7*(2), 25–60.

Alemi, M., & Khanlarzadeh, N. (2017). Native and nonnative teachers' pragmatic criteria for rating the request speech act: The case of American and Iranian EFL teachers. *Applied Research on English Language*, *6*(1), 67–84.

Al-Gahtani, S., & Roever, C. (2012). Proficiency and sequential organization of L2 requests. *Applied Linguistics*, *33*(1), 42–65.

Al Masaeed, K., Waugh, L. A., & Burns, K. E. (2018). The development of interlanguage pragmatics in L2 Arabic: The production of apology strategies. *System*, *74*, 98–108.

Bardovi-Harlig, K. (2017). Acquisition of L2 pragmatics. In S. Loewen & M. Sato (Eds.), *Routledge handbook of instructed SLA* (pp. 224–245). New York: Routledge.

Barron, A. (2019). Norms and variation in L2 pragmatics. In N. Taguchi (Ed.), *Routledge handbook of SLA and pragmatics*. (pp. 447–461). New York/Abingdon: Routledge.

Beebe, L. M., & Waring, H. Z. (2002). The pragmatics in the Interlanguage pragmatics research agenda: The case of tone. The Annual American Association for Applied Linguistics Conference, Salt Lake City, UT, 6–9 April.

Beebe, L. M., & Waring, H. Z. (2005). Pragmatic development in responding to rudeness. In J. Frodesen & C. Holten (Eds.), *The power of context in language teaching and learning* (pp. 67–79). Boston: Thomson/Heinle.

Bieswanger, M. (2015). Variational pragmatics and responding to thanks – revisited. *Multilingua 34*(4), 527–546.

Bowles, M. A. (2010). *The think-aloud controversy in second language research*. Abingdon: Routledge.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, *12*(1), 1–15.

Brown, J. D. (2017). Forty years of doing second language testing, curriculum, and research: So what? *Language Teaching*, *50*(2), 276–289.

Caprario, M. (2018). Pragmatic competence in English-language class discussions. International conference on exploring and assessing pragmatic aspects of L1 and L2 communication. Department of Linguistic and Literary Studies, University of Padua, Italy, 25–27 July.

Cohen, A. D. (2005). Strategies for learning and performing L2 speech acts. *Intercultural Pragmatics*, *2*(3), 275–301.

Cohen, A. D. (2006). The coming of age of research on test-taking strategies. *Language Assessment Quarterly*, *3*(4), 307–331.

Cohen, A. D. (2013). Verbal report. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (5 pp). Oxford: Wiley-Blackwell. Retrieved from https://sites.google.com/a/umn.edu/andrewdcohen/publications/research-methodology

Cohen, A. D. (2018). *Learning pragmatics from native and nonnative language teachers*. Bristol: Multilingual Matters.

Cohen, A. D., & Olshtain, E. (1993). The production of speech acts by EFL learners. *TESOL Quarterly*, *27*(1), 33–56.

Cohen, A. D., Olshtain, E., & Rosenstein, D. S. (1986). Advanced EFL apologies: What remains to be learned? *International Journal of the Sociology of Language*, *62*(6), 51–74.

Cohen, A. D., & Shively, R. L. (2002/2003). Measuring speech acts with multiple rejoinder DCTs. *Language Testing Update*, *32*, 39–42.

Eslami, Z., & Mirzaei, A. (2014). Speech act data collection in a non-Western context: Oral and written DCTs in the Persian language. *Iranian Journal of Language Testing*, *4*(1), 137–154.

Félix-Brasdefer, C. (2006). Teaching the negotiation of multi-turn speech acts: Using conversation-analytic tools to teach pragmatics in the FL classroom. In K. Bardovi-Harlig, C. Félix-Brasdefer, & A. S. Omar (Eds.), *Pragmatics & language learning* (pp. 167–197). Honolulu, HI: National Foreign Language Resource Center, University of Hawai'i.

Félix-Brasdefer, J. C. (2018). Role plays. In C. R. Hoffmann & W. Bublitz (Eds.), *Pragmatics of social media* (pp. 303–329). Berlin/Boston: De Gruyter Mouton.

Hamp-Lyons, L., & Condon, W. (1993). Questioning assumptions about portfolio-based assessment. *College Composition and Communication*, *44*(2), 176–190.

Hudson, T., Detmer, E., & Brown, J. D. (1992). *A framework for testing cross-cultural pragmatics* (Technical Report #2). Honolulu, HI: Second Language Teaching & Curriculum Center, University of Hawaii at Manoa.

Hudson, T., Detmer, E., & Brown, J. D. (1995). *Developing prototypic measures of cross-cultural pragmatics* (Technical Report #7). Honolulu, HI: Second Language Teaching & Curriculum Center, University of Hawai'i at Manoa.

Hymes, D. (1972). Models of the interaction of language and social life. In J. Gumperz & D. Hymes (Eds.), *Directions in sociolinguistics: The ethnography of communication* (pp. 35–71). New York: Holt, Rhinehart, & Winston.

https://doi.org/10.1017/S0261444819000156 Published online by Cambridge University Press

Isaacs, T. (2016). Assessing speaking. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 131–146). Berlin: De Gruyter Mouton.

Ishihara, N. (2009). Teacher-based assessment for foreign language pragmatics. *TESOL Quarterly*, *43*(3), 445–470.

Ishihara, N. (2013). Teacher-based assessment of L2 Japanese pragmatics: Classroom applications. In S. J. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 124–148). Basingstoke: Palgrave Macmillan.

Ishihara, N. (2014). Assessment of pragmatics in the classroom. In N. Ishihara & A. D. Cohen (Eds.), *Teaching and learning pragmatics: Where language and culture meet* (pp. 286–317). New York: Routledge.

Ishihara, N., & Chiba, A. (2014). Teacher-based or interactional? Exploring assessments for children's pragmatic development. *Iranian Journal of Language Testing*, *4*(1), 84–112.

Jones, L. (2016). Language and gender identities. In S. Pierce (Ed.), *Routledge handbook of language and identity* (pp. 210–224). Abingdon: Routledge.

Kasper, G. (2004). Speech acts in (inter)action: Repeated questions. *Intercultural Pragmatics*, *1*(1), 125–133.

Kasper, G., & Youn, S. J. (2017). Transforming instruction to activity: Roleplay in language assessment. *Applied Linguistics Review*. doi:10.1515/applirev-2017-0020

Koh, W.-Y., Lee, S., & Lee, J. (2018). L2 pragmatic comprehension of aural sarcasm: Context, intonation, and literal meaning. *The Annual American Association for Applied Linguistics Conference*, Chicago, 24–27 March.

Krashen, S. D. (1983). The din in the head, input, and the language acquisition device. *Foreign Language Annals*, *16*(1), 41–44.

Poehner, M. E. (2008). *Dynamic assessment: A Vygotskian approach to understanding and promoting L2 development*. New York: Springer.

Roever, C. (2018). *Testing of L2 pragmatics and possibilities for measuring implicit knowledge* (Unpublished manuscript). Department of Linguistics & Applied Linguistics, The University of Melbourne, Australia.

Roever, C., & Ikeda, N. (in press). Testing L2 pragmatic competence. In G. Schneider & E. Ifantidou (Eds.), *Handbook of developmental and clinical pragmatics*. Berlin: Mouton de Gruyter.

Roever, C., & Kasper, G. (2018). Speaking in turns and sequences: Interactional competence as a target construct in testing speaking. *Language Testing*, *35*(3), 331–355.

Schauer, G. (2009). *Interlanguage pragmatic development: The study abroad context*. New York: Continuum.

Schauer, G. A. (2017). 'It's really insulting to say something like that to anyone': An investigation of English and German native speakers' impoliteness perceptions. In I. Kecskes & S. Assimakopoulos (Eds.), *Current issues in intercultural pragmatics* (pp. 207–227). Amsterdam: John Benjamins.

Schneider, K. P. (2017). Pragmatic competence and pragmatic variation. In R. Giora & M. Haugh (Eds.), *Doing pragmatics interculturally: Cognitive, philosophical, and sociopragmatic perspectives* (pp. 315–333). Berlin: Walter de Gruyter.

Shively, R. L. (2008). *Politeness and social interaction in study abroad: Service encounters in L2 Spanish* (Ph.D. dissertation). University of Minnesota, Minneapolis.

Shively, R. L. (2013). Learning to be funny in Spanish during study abroad: L2 humor development. *Modern Language Journal*, *97*(4), 930–946.

Shively, R. L. (2018). Naturalistic data in L2 pragmatics research: Challenges and opportunities. In A. Gudmestad & A. Edmonds (Eds.), *Critical reflections on data in second language acquisition* (pp. 192–218). Amsterdam: John Benjamins.

Taguchi, N. (2018). Data collection and analysis in developmental L2 pragmatics research: Discourse completion test, role-play, and naturalistic recording. In A. Gudmestad & A. Edmonds (Eds.), *Critical reflections on data in second language acquisition* (pp. 7–32). Amsterdam: John Benjamins.

Taguchi, N., & Roever, C. (2017). *Second language pragmatics*. Oxford: Oxford University Press.

Tajeddin, Z., & Alemi, M. (2013). Criteria and bias in native English teachers' assessment of L2 pragmatic appropriacy: Content and FACETS analyses. *Asia-Pacific Education Researcher*, *23*(3), 435–434.

Tajeddin, Z., & Alemi, M. (2014). Pragmatic rater training: Does it affect non-native L2 teachers' rating accuracy and bias? *Iranian Journal of Language Testing*, *4*(1), 66–83.

Tajeddin, Z., Alemi, M., & Pashmforoosh, R. (2017). Idealized native-speaker linguistic and pragmatic norms in English as an international language: Exploring the perceptions of nonnative English teachers. *Language and Intercultural Communication*, *18*(3), 300–314.

Tajeddin, Z., & Alizadeh, I. (2015). Monologic vs. dialogic assessment of speech act performance: Role of nonnative L2 teachers' professional experience on their rating criteria. *Journal of Research in Applied Linguistics*, *6*(1), 3–27.

Tajeddin, Z., & Dabbagh, A. (2015). Interlanguage pragmatic test tasks: Does a low-stakes test have washback to L2 teachers and learners? *The Journal of Asia TEFL*, *12*(4), 129–158.

Usó-Juan, E., & Martínez-Flor, A. (2014). Reorienting the assessment of the conventional expressions of complaining and apologising: From single-response to interactive DCTs. *Iranian Journal of Language Testing*, *4*(1), 113–136.

Yates, L. (2017). Learning how to speak: Pronunciation, pragmatics and practicalities in the classroom and beyond. *Language Teaching*, *50*(2), 227–246.

Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing*, *32*(2), 199–225.

https://doi.org/10.1017/S0261444819000156 Published online by Cambridge University Press

Youn, S. J. (2018a). Rater variability across examinees and rating criteria in paired speaking assessment. *Papers in Language Testing and Assessment*, *7*, 32–60.
Youn, S. J. (2018b). Task design and validity evidence for assessment of L2 pragmatics in interaction. In N. Taguchi & Y. Kim (Eds.), *Task-based approaches to teaching and assessing pragmatics* (pp. 217–246). Amsterdam: John Benjamins.
Youn, S. J. (Under review with the TESOL Quarterly). Interactional features of L2 pragmatic interaction in role-play speaking assessment.
Youn, S. J. (2019). Managing proposal sequences in role-play assessment: Validity evidence of interactional competence across levels. *Language Testing*.
Youn, S. J. & Bi, N. Z. (2019). Investigating test-takers' strategy use in task-based L2 pragmatic speaking assessment. *Intercultural Pragmatics*, *16*, 185–218.
Youn, S. J., & Bogorevich, V. (2019). Assessment in L2 pragmatics. In N. Taguchi (Ed.), *Routledge handbook of SLA and pragmatics* (pp. 308–321). New York/Abingdon: Routledge.

Andrew D. Cohen is Professor Emeritus from the University of Minnesota. He was a Peace Corps Volunteer with the Aymara Indians on the High Plains of Bolivia (1965–1967), professor of ESL in the English Department at UCLA (1972–1975), professor of language education at the Hebrew University (1975–1991), and then professor of second language studies at the University of Minnesota (1991–2013). He co-edited *Language learning strategies* with Ernesto Macaro (Oxford University Press, 2007), co-authored *Teaching and learning pragmatics* with Noriko Ishihara (Routledge, 2014), authored *Strategies in learning and using a second language* (Routledge, 2011), and most recently published *The learning of pragmatics from native and nonnative language teachers* (Multilingual Matters, 2018). He has also published many book chapters and journal articles. Copies of most of his papers are available for download on his website: https://z.umn.edu/adcohen.

Cite this article: Cohen, A. D. (2020). Considerations in assessing pragmatic appropriateness in spoken language. *Language Teaching*, *53*(2), 183–202. https://doi.org/10.1017/S0261444819000156.

https://doi.org/10.1017/S0261444819000156 Published online by Cambridge University Press