SLA Second Language Acquisition

# Measuring Second Language Pragmatic Competence

## A Psycholinguistic Perspective

Rod Ellis, Carsten Roever,
Natsuko Shintani and Yan Zhu

# Measuring Second Language Pragmatic Competence

## SECOND LANGUAGE ACQUISITION

*Series Editors*: **Professor David Singleton**, *University of Pannonia, Hungary* and Fellow Emeritus, *Trinity College, Dublin, Ireland* and **Professor Simone E. Pfenninger**, *University of Zurich, Switzerland*

This series brings together titles dealing with a variety of aspects of language acquisition and processing in situations where a language or languages other than the native language is involved. Second language is thus interpreted in its broadest possible sense. The volumes included in the series all offer in their different ways, on the one hand, exposition and discussion of empirical findings and, on the other, some degree of theoretical reflection. In this latter connection, no particular theoretical stance is privileged in the series; nor is any relevant perspective – sociolinguistic, psycholinguistic, neurolinguistic, etc. – deemed out of place. The intended readership of the series includes final-year undergraduates working on second language acquisition projects, postgraduate students involved in second language acquisition research, and researchers, teachers and policymakers in general whose interests include a second language acquisition component.

All books in this series are externally peer-reviewed.

Full details of all the books in this series and of all our other publications can be found on http://www.multilingual-matters.com, or by writing to Multilingual Matters, St Nicholas House, 31-34 High Street, Bristol, BS1 2AW, UK.

SECOND LANGUAGE ACQUISITION: 166

# Measuring Second Language Pragmatic Competence

## A Psycholinguistic Perspective

**Rod Ellis, Carsten Roever, Natsuko Shintani and Yan Zhu**

MULTILINGUAL MATTERS
Bristol • Jackson

DOI https://doi.org/10.21832/ELLIS7731
Library of Congress Cataloging in Publication Data
A catalog record for this book is available from the Library of Congress.
Names: Ellis, Rod, author. | Roever, Carsten, author. | Shintani, Natsuko, author. | Zhu, Yan, author.
Title: Measuring Second Language Pragmatic Competence: A Psycholinguistic Perspective /Rod Ellis, Carsten Roever, Natsuko Shintani and Yan Zhu.
Description: Bristol; Jackson: Multilingual Matters, 2024. | Series: Second Language Acquisition: 166 | Includes bibliographical references and index. | Summary: "Extending earlier work on grammar, this book investigates the development of tests measuring second language learners' implicit and explicit pragmatic abilities in English. It reports detailed analyses of the use of the tests with two groups of learners and in studies investigating study abroad and the effects of instruction"—Provided by publisher.
Identifiers: LCCN 2023043378 (print) | LCCN 2023043379 (ebook) | ISBN 9781800417724 (paperback) | ISBN 9781800417731 (hardback) | ISBN 9781800417755 (epub) | ISBN 9781800417748 (pdf)
Subjects: LCSH: Second language acquisition—Ability testing. | Pragmatics—Ability testing. | English language—Ability testing. | Second language acquisition—Psychological aspects.
Classification: LCC P118.75 .E45 2024 (print) | LCC P118.75 (ebook) | DDC 404/.20287—dc23 LC record available at https://lccn.loc.gov/2023043378
LC ebook record available at https://lccn.loc.gov/2023043379

British Library Cataloguing in Publication Data
A catalogue entry for this book is available from the British Library.

ISBN-13: 978-1-80041-773-1 (hbk)
ISBN-13: 978-1-80041-772-4 (pbk)

**Multilingual Matters**
UK: St Nicholas House, 31-34 High Street, Bristol, BS1 2AW, UK.
USA: Ingram, Jackson, TN, USA.

Website: https://www.multilingual-matters.com
Twitter: Multi_Ling_Mat
Facebook: https://www.facebook.com/multilingualmatters
Blog: https://www.channelviewpublications.wordpress.com

The policy of Multilingual Matters/Channel View Publications is to use papers that are natural, renewable and recyclable products, made from wood grown in sustainable forests. In the manufacturing process of our books, and to further support our policy, preference is given to printers that have FSC and PEFC Chain of Custody certification. The FSC and/or PEFC logos will appear on those books where full certification has been granted to the printer concerned.

Typeset by Deanta Global Publishing Services, Chennai, India.

# Contents

# About the Authors

**Rod Ellis** is a Distinguished Research Professor in Curtin University (Australia), a longstanding professor at Anaheim University, visiting professor at Shanghai International Studies University and Emeritus Distinguished Professor of the University of Auckland. He is also a fellow of the Royal Society of New Zealand. He has written extensively on second language acquisition and task-based language teaching. His publications include *Understanding Second Language Acquisition* (Oxford University Press, winner of the British Association for Applied Linguists best book prize) and *Study of Second Language Acquisition* (Oxford University Press, winner of the Duke of Edinburgh prize for best book in applied linguistics). His most recent books are *Reflections on Task-Based Language Teaching* (Multilingual Matters) and (co-authored) *Task-Based Language Teaching: Theory and Practice* (2020), published by Cambridge University Press.

**Carsten Roever** is Professor in Applied Linguistics at the University of Melbourne. He holds a PhD in Second Language Acquisition from the University of Hawai'i. His research interests include language testing, second language pragmatics, conversation analysis and quantitative research methods. He is particularly interested in learning of L2 pragmatics for languages other than English. He is the co-author (with Naoko Taguchi) of *Second Language Pragmatics* (2017, Oxford University Press) and (with Aek Phakiti) of *Quantitative Methods for Second Language Research* (2018, Routledge). His most recent single-authored volume is *Teaching and Testing Second Language Pragmatics and Interaction* (2022, Routledge).

**Natsuko Shintani** is a professor in the Faculty of Foreign Language Studies, Kansai University. Her research interests include the interface between second language instruction and second language acquisition, with particular emphasis on task-based language teaching, second language writing and the role of individual differences. Her work has appeared in journals such as *Language Learning*, *TESOL Quarterly* and

*Studies in Second Language Acquisition*. Her current research focuses on the role of instruction in developing L2 pragmatic knowledge.

**Yan Zhu** is an associate professor at the College of Foreign Languages and Literature, Fudan University, China. Her research focuses on curriculum innovation, task-based language teaching and teacher education. Dr Zhu is the author of *Language Curriculum Innovation in a Chinese Secondary School: A Study of Teacher Cognition and Classroom Practice*s (2018, Springer) and author/co-author of journal articles appearing in *Language Teaching Research*, *The Modern Language Journal*, *System*, etc. She is currently the principal investigator for a project supported by the National Social Science Fund of China. She has twice been awarded the 'Teaching Achievement Prize' by the Shanghai Municipal Education Commission.

# Acknowledgements

We acknowledge with gratitude financial support from the Australian Research Council for project DP180102707 and with JSPS KAKENHI Grant Number JP20H01297, which enabled us to carry out the research on which this book is based.

The following chapters drew on previously published articles and chapters. We are grateful to the publishers for permission to reproduce them in the book.

**Chapter 1**

Roever, C. and Ellis, R. (2021) 7 Testing of L2 pragmatics: The challenge of implicit knowledge. In J.C. Félix-Brasdefer and R. Shively (eds) *New Directions in Second Language Pragmatics* (pp. 142–155). De Gruyter Mouton.

**Chapter 4**

This article was published in *Research Methods in Applied Linguistics* 2 (1), Ellis, R. and Shintani N. Investigating a failed novel test of socio-pragmatic knowledge, Article 100046, Copyright Elsevier (2023).

**Chapter 5**

Ellis, R., Zhu, Y., Shintani, N. and Roever, C. (2021) A study of Chinese learners' ability to comprehend irony. *Journal of Pragmatics* 172, 7–20.

**Chapter 10**

Roever, C., Shintani, N., Zhu, Y. and Ellis, R. (2023) Proficiency effects on L2 pragmatics. In A. Martínez-Flor, A. Sánchez-Hernández and J. Barón (eds) *L2 Pragmatics in Action: Teachers, Learners and the Teaching–Learning Interaction Process* 58, 145.

**Chapter 11**

This article was published in *Applied Pragmatics* 4 (1), Shintani N. and Ellis, R. *The Roles of Language Proficiency and Study Abroad in*

*Japanese Students' Receptive Pragmatic Competence* 1–32, Copyright John Benjamins (2022). https://benjamins.com/catalog/ap.20011.shi

Finally, we would like to thank Bolormaa Shinjee, Matt Lucas, Sora Shioji, Yusuke Kuroki, Yoko Fujisawa and Atsushi Doi for their support with computerising the tests, assisting the data collection and referencing. We would also like to thank Atsushi Mizumoto, Hiroya Tanaka, Yasuhiro Imao and Todd Allen for their helpful feedback on the research materials and analyses.

# Preface

The book presents the results of a research project funded by an Australian Research Council grant awarded to Rod Ellis and Carsten Roever and by a JSPS KAKENHI Grant awarded to Natsuko Shintani. Yan Zhu also helped in the collection of data in China and in the preparation of a number of the chapters in the book. The book is, therefore, very much the collaborative endeavour of these four researchers. The purpose of the research was to investigate how second language (L2) learners' pragmatic abilities of English could be measured and, in particular, whether it was possible to design tests that would provide relatively separate measures of implicit and explicit pragmatic abilities.

By and large, existing assessments of pragmatic knowledge have paid little attention to the type of knowledge being measured. Our intention was to fill this gap. We drew on well-established methods of assessment of pragmatic knowledge (e.g. a Metapragmatic Knowledge Test and role plays), but we also looked for ways in which these tests could be scored in novel ways and also developed novel tests (e.g. an Irony Test and an Elicited Imitation Test) to measure test takers' implicit knowledge.

While the tests we developed could be used with a wide variety of second language (L2) learners, the project only collected data from learners for whom English is a foreign language. We would have liked to have also collected data from L2 learners resident in Australia, but the advent of COVID made this impossible. We collected data from samples of university students in China and Japan and also administered the tests to a small sample of adults who had learned English as their first language (i.e. native speakers).

In addition, to the test data which we used to assess learners' implicit and explicit knowledge, we also investigated three factors that mediate the development of learners' pragmatic knowledge – L2 linguistic proficiency, experience of living in an English-speaking country and formal instruction.

In the book, we explain the background to the development of the tests and report detailed analyses of the data from each test and the results of these analyses. We also report the results of a confirmatory factor analysis in order to establish whether our assumptions about what type of ability each test measures were or were not supported. We also report our analyses of studies that investigated the impact of the mediating factors.

As we carried out this research we became aware of a number of limitations. This is not surprising given the innovative nature of the project. We view the research as exploratory and point out the limitations in the concluding chapter of the book. We point to ways in which the tests could be improved and to future directions. Our hope is that the book will be of sufficient interest to motivate other researchers to undertake further research into the measurement of implicit and explicit pragmatic abilities.

Our intended readers are researchers, language testers and teacher educators interested in the assessment of L2 pragmatics. We will not consider – except occasionally – how the tests might be utilised for assessing pragmatic abilities in the classroom.

# Part 1
# Background

The purpose of the chapter in this part of the book is to provide the reader with background about the kinds of tests that have been used to measure pragmatic competence. In this chapter, we also consider which type of knowledge – explicit or implicit – the different tests are likely to be measuring. In this way, the chapter sets the scene for Part 2 of the book, wherein we explain the rationale for the new tests we developed.

It may help readers if we offer definitions of implicit/explicit knowledge, as these constructs are central to the psycholinguistic perspective that informs the book. Implicit knowledge of language is knowledge that is intuitive. It can be processed automatically and without consciousness. Explicit knowledge of language is knowledge that can only be accessed through controlled processing and is therefore conscious. It is often linked to metalinguistic terminology. Fully competent speakers of a language will draw mainly on their implicit pragmatic knowledge, but there may be occasions when they may find it necessary to access their explicit knowledge. Less competent speakers whose linguistic resources are not yet automatic will have to rely to a greater extent on explicit processing. A speaker who is fully pragmatically competent is one who has developed implicit processing abilities.

The distinction between implicit and explicit knowledge has been applied mainly to linguistic competence, but it is also central to pragmatic competence. Subsequently, however, we reframed the implicit/explicit distinction in terms of 'abilities', as we consider that 'ability' is better suited than 'knowledge' when it comes to pragmatics. This will be explained in Chapter 2. In this chapter, we will stick to using 'knowledge'.

# 1 Tests of L2 Pragmatics: What Do They Assess?

Carsten Roever and Rod Ellis

## Introduction

Testing of L2 pragmatics is a fairly recent enterprise. Apart from some early attempts (Farhady, 1980; Shimazu, 1989), systematic test development did not start until 1995, when Hudson *et al.* (1995) introduced their test battery. Although several tests following different theoretical orientations have been developed subsequent to Hudson *et al.*'s pioneering work, no large-scale language test such as TOEFL and IELTS for English, TestDaF for German, HSK for Mandarin Chinese or JPT for Japanese specifically tests pragmatics. This is an odd state of affairs given that pragmatics and social aspects of language use are part of the major constructs of communicative competence (Bachman & Palmer, 2010; Canale, 1983; Canale & Swain, 1980) on which large-scale tests are purportedly based. However, research on testing L2 pragmatics has been lively since Hudson *et al.*'s project and will be sketched out in the following section before we discuss how different measurement tools have been used, and how they could be deployed for measurement of implicit and explicit pragmatic knowledge.

## Testing L2 Pragmatics: A Brief Overview

In testing of L2 pragmatics, two major research streams exist, which have developed assessment instruments for somewhat different constructs of L2 pragmatics. The older stream focuses on pragmatic competence as conceptualised by Leech (1983, 2014) and Thomas (1983) and incorporates research on politeness (Brown & Levinson, 1987), routine formulae (Coulmas, 1981) and implicature (Grice, 1975). Testing instruments generally follow an analytic psychometric tradition (Klein-Braley, 1997), consisting of multi-item batteries, and include Discourse Completion Tests (DCTs), multiple-choice comprehension/recognition tasks and rating scales. Most measurements tap explicit processing, though some could be modified to elicit implicit processing. As most of these tests are discrete point, administration and scoring is only moderately resource intensive.

Within this first research stream, two generations of tests can be distinguished. The first generation of pragmatics assessment focuses on speech acts and politeness and takes off from Hudson *et al.*'s (1995) pioneering test battery. Hudson *et al.* (1995) developed three different types of DCTs (written, oral and multiple choice), role-play tasks, and two self-assessment scales. They investigated learners' knowledge of appropriateness for requests, apologies and refusals. Their instrument was contrastively designed for L1 Japanese-speaking learners of English, and it led to several spin-offs. Yamashita (1996) adapted the test battery for L1 English- speaking learners of Japanese, Yoshitake (1997) used it with EFL learners in Japan and Ahn (2005) adapted it for L2 Korean. Liu (2006) addressed the challenge of developing reliable multiple-choice DCTs with Chinese learners of English, while Tada (2005) used video-based scenarios to contextualise his productive and multiple-choice DCTs for Japanese learners of English.

The second generation of tests is also situated in traditional pragmatics but broadens the construct from speech acts to other aspects of pragmatics, notably implicature and routine formulae. It emphasises practicality of measurement through web-based delivery to facilitate wider uptake of pragmatics assessment. Roever (2005, 2006) developed a test of English L2 pragmatics with a pragmalinguistic focus as opposed to Hudson *et al.*'s sociopragmatic focus, assessing knowledge of implicature, routine formulae and speech acts. In the same tradition, Itomitsu (2009) developed a test of Japanese as an L2, assessing knowledge of speech acts, speech styles, routine formulae and grammar. Roever *et al.* (2014) expanded work by focusing on sociopragmatics in a test piloting new instruments, including dialogue completion, appropriateness judgements and corrections and comprehension of extended discourse. Finally, Timpe (2013) tested speech act comprehension, routine formulae and idiomatic language use supplemented by role plays.

The more recent stream encompasses the third generation of tests and focuses on interactional competence, most commonly following the conceptualisation by Hall and Pekarek Doehler (2011), which is infused with the research approach of Conversation Analysis (Sacks, 1992; Schegloff, 2007). It assesses learners' ability to manage extended conversations, create interpersonal meanings and design contributions that fit the addressee's epistemic and social status. Testing instruments are role plays and extended monologues, and scoring is done by raters. These measurements tap implicit processing far more than explicit processing, but they tend to be resource intensive to administer and score.

Two tests have been developed in this tradition which are squarely situated in the Interactional Competence paradigm. Youn (2013, 2015) developed a role play–based test and created scoring criteria bottom-up from the test takers' interactions. Ikeda (2017) expanded on Youn, also employing role plays but with an additional focus on the possible use of monologue tasks and the role of proficiency and exposure. Working outside Interactional Competence as a framework, Grabowski (2009, 2013) also employed role plays and scored them following Purpura's (2004) model of language ability. Finally, Walters (2007, 2009) attempted to measure aspects of interactional ability receptively and productively, but his test remained mostly unsuccessful.

It is apparent from this brief summary that testing of social aspects of L2 use has been conducted in quite different frameworks, leading to a variety of test instruments. These instruments will be discussed in more detail in the following sections.

## Traditional Construct of Pragmatic Knowledge: Speech Acts *et al.*

This construct of pragmatics informed the first and second tradition of pragmatics tests and is anchored in Austin's philosophical view of speech acts as ways 'to do things with words' (Austin, 1962), i.e. to impact the world by using language ('I name this ship …'). This is complemented by Levinson's (1983) linguistic perspective of pragmatics as grammaticalisation of context, i.e. the real-world context impacts language use (e.g. Japanese *desu/masu* forms or *tu/vous* distinctions), and language use cannot be entirely accounted for by syntactic and semantic rules: language users' choices are motivated by considerations beyond grammar and word meaning.

To account for the different choices language users make when speaking to different interlocutors and in different contexts, the speech act view is usefully supplemented with an anthropological perspective on human relationships, provided by politeness theory (Brown & Levinson, 1987). Politeness theory posits social context factors which drive language users' selections of linguistic expressions: relative Power, Social Distance (degree of acquaintanceship, 'in-the-same-boatness') and degree of imposition ('cost' of the speech act for the hearer in terms of money, time or social sanctions).

In conceptualising pragmatic competence, Leech (1983) brought these considerations together by differentiating between two separate but interconnected aspects: sociopragmatic knowledge and pragmalinguistic knowledge. Sociopragmatic knowledge is language users' knowledge of the social rules and norms of the target language community, e.g. who is deserving of respect or what constitutes a major or minor imposition. Pragmalinguistic knowledge is their knowledge of linguistic tools that can be deployed for pragmatic purposes, e.g. conventional indirectness with modals ('can you'/'could you') to make requests. The two types of knowledge must be mapped onto each other so that language users know what kinds of linguistic tools are relevant to different constellations of social context factors.

While the above view of pragmatics has been the most influential in L2 pragmatics assessment and research, it is strongly oriented towards speech acts, and does not account well for other phenomena, notably indirect language use and highly routinised expressions. Grice (1975) developed the concept of conversational implicature to describe how speakers convey meaning by flouting certain tacit conversational maxims, e.g. an orientation to truthfulness of contributions by exaggerating or employing irony ('How was the dentist?' – 'I had a great time'). Coulmas (1981) described a special case of speech acts: routine formulae, also known as 'conventional expressions' (Bardovi-Harlig, 2009), which are formulaic expressions tied to certain situations, social purposes and social roles, e.g. 'Can I get you anything else?' said by a waiter in a restaurant to a customer.

For second language learners, all these aspects of pragmatics are potentially challenging, and indeed early research on L2 pragmatics stemmed from interest in pragmatic errors (Thomas, 1983, 1995). L2 learners may lack sociopragmatic knowledge and cause offense by involuntarily infringing social norms of the target community, e.g. by mis-analysing a distant social relationship requiring formal language use as close and requiring informal language use, or they may lack pragmalinguistic knowledge and not control linguistic tools for displaying their orientation to sociopragmatic norms, e.g. they may not be able to use complex polite expressions ('Check my paper please' vs. 'I was wondering if you'd mind having a quick look at my paper'). Learners may also find comprehension of implicature challenging, not recognising irony or implied meaning. Finally, certain routine formulae may be learned very early on, whereas others take much longer to be understood and made available for practical use.

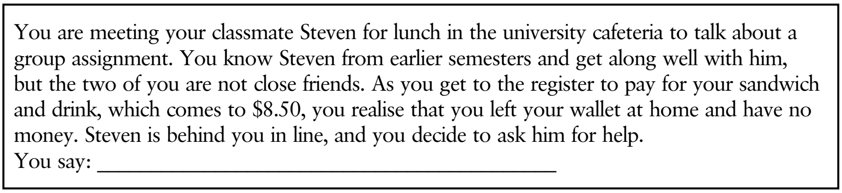
## Testing Speech Acts: Discourse Completion Tests

Tests of L2 pragmatics following the speech act/politeness construct have primarily focused on learners' ability to produce (and to a lesser extent recognise) sociopragmatically appropriate language use. This has led to the somewhat paradoxical situation in L2 pragmatics assessment that test tasks more commonly assess production than comprehension. The primary assessment tool has been the Discourse Completion Test (DCT), used by Hudson *et al.* (1995), and its spin-offs, Roever (2005, 2006), Liu (2006) and Tada (2005). In their most typical form, DCTs consist of 12–24 paragraph-length scenarios that encapsulate different settings of the social context factors (Power/Distance/Imposition). In testing of L2 pragmatics, they commonly elicit the speech acts request, apology and refusal, but they can also elicit other speech acts. Test takers are asked to imagine themselves in the scenario and produce an utterance directed at the imaginary interlocutor, as in Figure 1.1 below.

This basic DCT could be administered as a written task, where input and test-taker production are in writing; or a spoken task, where input is aural and output is oral. Test takers can also be given multiple-choice response options for the gap and choose the most appropriate one. All these varieties were used by Hudson *et al.* (1995), Yamashita (1996) and Yoshitake (1997), with the overall finding being that written and oral DCTs are reliable but that multiple-choice DCTs suffer from very low reliability (Brown, 2001a). However, Liu (2006) developed a multiple-choice DCT that functioned reasonably reliably.

Variations on the basic DCT include the use of a rejoinder (response from the imaginary interlocutor) after the gap (Roever, 2005, 2006), or dialogues with gaps where the imaginary interlocutor's turns are provided and test takers fill in the missing turns (Roever *et al.*, 2014). Both types work reliably, but like written, oral and multiple-choice DCTs, they elicit offline, explicit processing. The only DCT which might tap implicit or highly automatised processing is the production part of Tada's (2005) video-based instrument. Tada played 24 short video clips to test takers with the last utterance replaced by a multiple-choice task ('perception test'), or an instruction to produce the missing utterance orally ('production test'). He attained satisfactory reliability of his perception test ($\alpha = 0.75$), which was the focus of his study, and a reasonable correlation of 0.74 among ratings. Tada did not enforce a time limit on test-taker responses for his production part, only instructing test takers to 'start to speak immediately after the video clip finishes' (Tada, 2005: 246). Enforcing a time limit would potentially lead to this task type eliciting implicit/highly automatised processing.

Scoring of productive DCT responses has to be done by human raters. Hudson *et al.* (1995) scored their written and spoken DCT responses with a five-point Likert scale ranging from 'very unsatisfactory' to 'completely appropriate' on six criteria: ability to use the correct speech act, formulaic expressions, amount of speech used and information given, level of formality, level of directnessand level of politeness. However, other studies (such as Tada, 2005, as well as Takimoto's 2009 research study) used holistic ratings on 'overall appropriateness', while Roever *et al.* (2014) employed a two-step rating approach for their multi-turn DCTs, first having raters assess the fit of the response to the gap, and then



You are meeting your classmate Steven for lunch in the university cafeteria to talk about a group assignment. You know Steven from earlier semesters and get along well with him, but the two of you are not close friends. As you get to the register to pay for your sandwich and drink, which comes to $8.50, you realise that you left your wallet at home and have no money. Steven is behind you in line, and you decide to ask him for help.
You say: ________________________________________

**Figure 1.1** Sample DCT request item

its idiomaticity. Experience from previous studies suggests that it is generally not too difficult to rater-score DCT responses reliably, even when raters simply score them against their own impressions of appropriateness. However, it is questionable how defensible such a scoring method is, since it assumes a middle-class, educated native speaker benchmark for what is appropriate/polite in a language, whereas such norms in reality vary widely along age, class and geographical lines.

Another approach to testing speech acts receptively was undertaken by Timpe (2013, see also Timpe *et al.*, 2017). She assessed test takers' comprehension of offers and refusals by showing them an utterance or short conversation embedded in a scenario and asking them to choose one of four options for rephrasing the target utterance. While the task in Timpe's version likely elicits explicit processing, it could be redesigned as a speeded, productive task where participants watch a video vignette and, immediately after the target utterance, rephrase it orally. This would require human rating but could provide another measure of pragmatic comprehension.

## Problems of DCTs

While DCTs allow relatively rapid collection of a fairly large amount of data and the manipulation of context variables of interest, they suffer from two major problems, which call into the question the defensibility of inferences from DCT scores about real-world language performance.

For one thing, DCTs do not allow extended interaction, forcing respondents to put into a single gap what in reality might unfold over several turns, with utterances being constantly responsive to the interlocutor's previous utterances. For example, a typical DCT response in Figure 1.1 might be something along the lines of 'Hey, Steven, I forgot my wallet. Can you lend me a few dollars for lunch?' Such a response would probably receive high scores as it contains an appropriate alerter ('Hey, Steven'); a grounder providing a reason for the upcoming request; and a conventionally indirect request with a modal ('can') and mitigation ('a few dollars'). The only problem is that this conversation is likely to run differently in reality and would most likely not even involve an on-the-record request, since bald requests and refusals are relatively rare in natural discourse. It might instead run as shown in Figure 1.2.

This methodological weakness of the DCT, i.e. that it forces respondents to do what they would not need to do in reality, probably also accounts for their second major shortcoming. As Golato (2003) demonstrated empirically in her study on compliment responses, respondents use language in DCTs that they do not employ in reality, and the frequency of different response types differs appreciably (also found in a comparative study by Economidou-Kogetsidis, 2013).

You (turns to Steven with a sheepish grin): 'Oh damn, I forgot my wallet.'

Steven (grins): 'No worries, I'll spot you a few bucks.'

You: 'Thanks, man.'

**Figure 1.2** Imaginary conversation in the cafeteria

These findings advise extreme caution when drawing conclusions from DCT data. Taguchi and Roever (2017) describe two scenarios under which DCT use is defensible:

(1) The research focus is purely on describing the range of strategies/ semantic formulae that respondents have available, with no claims made about their ability for use.
(2) In both the DCT and the real world, the response is an extended written response, e.g. an email.

### Speech acts and measuring implicit/explicit pragmatic knowledge

Given the limitations of DCTs, their usefulness for developing measures of explicit/implicit pragmatic knowledge is limited. To tap implicit processing, it would be conceivable to use video vignettes like Tada's with the last utterance to be supplied by the test taker under time pressure. However, that utterance would have to be brief and not require further elaboration, such as an expression of gratitude, a compliment response, a formulaic response ('Thanks for that' – 'That's alright') or an assessment of a news telling ('That's great', 'That's too bad'). To tap explicit processing, a brief writing task might be useful, such as a one-paragraph email. For receptive tasks, variations on Timpe's (2013) task with a speeded yes/no decision would be possible.[1]

## Metapragmatic Judgements

Metapragmatic judgements, also known as 'appropriateness judgements', have not been very commonly used in pragmatics assessment (Roever *et al.*, 2014, is an exception) but occur quite frequently in pragmatics research (e.g. Bardovi-Harlig & Dörnyei, 1998; Li & Taguchi, 2014; Takimoto, 2009, 2012)[2]. Metapragmatic judgements usually present a scenario similar to a DCT scenario together with a target utterance, which is typically a request, apology or refusal. Respondents then judge the appropriateness of the utterance on a Likert scale (Roever *et al.*, 2014; Takimoto, 2009).

Variations to this standard methodology include binary judgements, which were also used by Roever *et al.* (2014) and followed by a productive correction task. Similarly, in Bardovi-Harlig and Dörnyei's (1998) research, respondents first indicated whether a target utterance was grammatically and pragmatically accurate or problematic, and if they decided it was problematic, they rated the severity of the problem on a Likert scale. While Bardovi-Harlig and Dörnyei did not require participants to indicate whether the problem was grammatical or pragmatic, Li and Taguchi (2014) required their participants to choose whether a target utterance was pragmatically inaccurate, grammatically inaccurate or accurate on both counts.

Finally, a multiple-choice DCT is also a metapragmatic judgement task as test takers are asked to choose the most appropriate of the response options (Matsumura, 2001, 2003). Response options could also be ranked in order of appropriateness, which Takimoto (2009) did with a magnitude estimation approach, by providing one response option as an anchor and then asking participants to rate other response options in relation to the anchor.

## Problems with metapragmatic judgement items

The biggest challenge with metapragmatic judgement items is scoring. Scoring usually happens against a native speaker (NS) benchmark, and native speakers do not necessarily agree on the appropriateness level of an utterance – however, they also do not vastly disagree. Roever *et al.* (2014) found with a benchmarking sample of 50 native speakers of Australian English that there was an absolute majority (50%+) for the most appropriate response for every item, and the majority response with two adjacent ratings (one level below and one level above) accounted for at least 98% of NS responses in every case. They employed a scoring protocol whereby test takers who chose the majority response received two points, and those who chose an adjacent response received one point provided that response was chosen by at least 10% of NSs. While this scoring approach was systematic and empirically based, it did not reflect the proportions of NSs who chose each option and did not take into account that appropriateness might be judged differently for utterances produced by NSs and non-native speakers (NNSs).

Another issue of metapragmatic judgement items in Roever *et al.*'s (2014) study was their lack of difficulty. Both the ESL and EFL sample scored very high on these items, leading to a section average of 79.1%, which in turn lowered reliability. This reflects a challenge in item design: it is difficult to construct target utterances for metapragmatic judgement items which are clearly inappropriate but not so wildly implausible or 'off the wall' that they are extremely easy to recognise 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 problem could be ameliorated by limiting the test to learners from one L1 background and

**Table 1.1** Appropriateness choice items from Roever *et al.* (2014: 101)

| Item | Setting | Speech Act | Appropriate | Failure |
|---|---|---|---|---|
| 1 | Home | Response to preference question | N | Too formal |
| 2 | Café | Compliment response | Y | |
| 3 | Home | Assessment | N | Too negative |
| 4 | Bar | Information request | N | Too formal, unresponsive |
| 5 | Street | Response to invite | Y | |
| 6 | Home | Response to greeting | N | Unresponsive/demanding |
| 7 | Walk | Response to news telling | N | Dismissive |
| 8 | University | Response to news telling | Y | |
| 9 | Unspec. | Response to offer | N | Too truthful |
| 10 | Uni | Response to request | Y | |
| 11 | Phone | Response to invite | N | Too formal |
| 12 | Pub | Response to offer | N | Too demanding |

exploiting transfer (as Hudson *et al.* [1995] did for their battery), but such a strategy reduces the usefulness of the test dramatically.

Interestingly, for the binary appropriateness choice items used by Roever *et al.* (2014), the difference in difficulty between ESL and EFL learners was much more pronounced than for appropriateness ratings: ESL learners scored similar as for Likert-scale based appropriateness judgements (79%), but EFL learners scored clearly lower (48%). Appropriateness choice items were not primarily based on requests, apologies and refusals but mostly consisted of responses to initiating utterances, as shown in Table 1.1 from Roever *et al.* (2014).

Going beyond speech acts and employing overly polite responses (rather than just impolite responses) might help make appropriateness judgements more difficult for test takers.

## Metapragmatic judgement tasks and explicit/implicit pragmatic processing

Metapragmatic judgement tasks are the most similar to grammaticality judgement tasks, a mainstay in explicit/implicit studies of grammatical knowledge. As Roever *et al.* (2014) showed with their appropriateness choice tasks, it is definitely possible to set them up as a binary appropriate/inappropriate decision task, which could be delivered in a speeded environment, with the interaction shown as a short video clip followed by the binary choice to be made immediately (with response time captured).

The problem of item difficulty remains but might be ameliorated by the inclusion of overly polite responses and by weakening the distinction between sociopragmatic and pragmalinguistic knowledge; it is likely that items with pragmalinguistic infelicities (such as unconventional or unidiomatic phrasing) would be more challenging for test takers than items where the target utterance is clearly rude.[3] It can be argued that conventionality of phrasing is part of displaying sociopragmatic knowledge (indeed, it is part of Hudson *et al.*'s [1995] rating criteria) and therefore should be considered part of a construct of pragmatic knowledge. The challenge in working from this assumption will be not to slide too far into general proficiency and ensure that an utterance's inappropriateness can be clearly argued to be pragmatic, rather than due to a general linguistic shortcoming.

## Implicature

Implicature requires a listener to go beyond the surface meaning of an utterance and actively construct its implied meaning based on knowledge of the immediate or larger context.

Most testing research on implicature has been based on the work of Bouton (1988, 1999), who distinguished between idiosyncratic and formulaic implicature. Idiosyncratic implicature is general conversational implicature where the implied meaning in inferable based on the preceding utterance and background knowledge, e.g. 'Do you know where Mark is?' – 'I heard music from his room earlier' to indicate that the speaker believes Mark to be in his room, or 'Can you turn up the volume a bit?' – 'I've got a cat on my lap' to refuse the interlocutor's request. Bouton's second type is formulaic implicature, which includes indirect criticism by praising an unimportant part of the whole ('How did you like his outfit?' – 'The hat was nice'), the Pope Q ('Are rents downtown expensive?' – 'Is the Pope Catholic?') and topic shift implicature ('How is your dissertation coming along?' – 'It's been really hot lately, hasn't it').

Taguchi (2011) used a different classification by differentiating between conventional and nonconventional implicatures. In her study, conventional implicatures consisted of refusals by means of explanations and accounts ('Do you want to come over for dinner?' – 'I still have a lot of work to finish'), and routine formulae in service encounters ('For here or to go?' – 'To go, please'). Non-conventional implicature was similar to Bouton's (1988) idiosyncratic implicature.

Roever (2005, 2006) used implicature as part of his pragmalinguistically oriented test and found that it was the component most strongly correlated with overall proficiency with little influence of exposure to the L2 environment (supported in a different analysis by Roever *et al.*, 2014). It appears that learners must have a certain proficiency level to be able to process implicature, possibly because they need to comprehend the utterance at surface level to realise that it requires further processing.

Implicature has been nearly exclusively researched with multiple choice tasks where respondents select the correct interpretation from the response options (Roever, 2005, 2006; Taguchi, 2008, 2011). A variation on this methodology is Taguchi (2008), who showed dialogues with the final (target) utterance a refusal or indirect opinion, followed by a yes–no question, e.g. '"A: Can you go answer the phone? – B: I'm in the bath." Can B answer the phone?' Taguchi also recorded the response time from the moment the question was displayed to when the participant pressed the button. Taguchi (2011) used multiple-choice items with response time capture, starting response-time measurement from the moment the response options were displayed. Response-time capture is otherwise uncommon in L2 pragmatics studies.

### Implicature and measurement of explicit/ implicit pragmatic knowledge

In terms of measuring implicit processing of implicature, Taguchi showed that it is possible to set up implicature tasks as speeded judgement tasks with a binary decision. Provided items are standardised in terms of length and a baseline measure of reading speed is obtained, correctness and response time measurements should provide a measure of implicit/highly automatised pragmatic knowledge.

## Routine Formulae

Routine formulae (or 'conventional expressions' in Bardovi-Harlig's [2009] parlance) are more or less fixed macrolexemes whose meaning is tied to a specific constellation of context factors. Routine formulae range from invariable ('Never mind') and somewhat variable with a limited number of possible variations ('Thanks very much/a lot/a bunch/ heaps') to allowing highly variable slots ('I was [just] wondering if …'). They can be closely tied to social roles ('Can I get you anything else?') and settings ('Order in the court') or be applicable across a wide range of settings and interlocutors ('You're welcome'). They constitute a special case of implicature and speech act, in that the meaning of the whole formula goes beyond the sum of the semantic surface meanings of its components fixed, macrolexemic nature, high frequency of occurrence, and situation-boundness make them an area of pragmatics in their own right.

In L2 pragmatics assessment, Roever (2005, 2006) tested routine formulae with multiple-choice tasks, which consisted of a scenario and four response options. Knowledge of routine formulae was more strongly affected by residence in the target language community than was the case for speech act and implicature knowledge, and showed less of a proficiency effect (Wang *et al.*, 2014).

In acquisitional L2 pragmatics research, Bardovi-Harlig (2009) used a spoken DCT instrument, delivered aurally, to elicit routine formulae and investigate learner variation in producing formulae.

### Measuring implicit knowledge of routine formulae

It would certainly be possible to measure knowledge of L2 routine formulae with a judgement task similar to Taguchi's (2008) implicature judgement tasks by asking test takers for their intuitions as to whether the target formula is likely to occur in the scenario shown. Again, response-time capture would probably provide a good measure of implicit/highly automatised processing. The focus in such tasks should probably be on the likelihood of occurrence in the situation rather than the pragmalinguistic idiomaticity of the formula, i.e. a negative exemplar should be a formula associated with another situation ('Thank you' – 'Never mind') rather than an unconventional variation on an acceptable formula ('Thank you' – 'You're much welcome').

## Testing Interactional Competence: Role Plays

Interactional competence is primarily conceptualised as interactants' ability to participate in extended conversations. Following Hall and Pekarek Doehler (2011), this requires the comprehension and production of recognisable social actions, i.e. what an utterance accomplishes in terms of the *mechanics* of a conversation, and how it contributes to driving the conversation forward. Social actions encompass speech acts (offers, refusals, requests, compliments, etc.) but also actions that contribute to the structure of the conversation, including pre-tellings ('the most amazing thing happened to me today'), signalling an incipient story; continuers ('mhm'), signalling that the recipient is listening and wanting the speaker to continue; and opening and closing moves ('So, let's get in touch again tomorrow'), signalling the beginning and end of a conversation. While such actions could theoretically be accommodated under a complex taxonomy in a speech-act paradigm, another aspect of the meaning of social actions cannot be easily thus accommodated, namely their sequential placement. To interpret utterance meanings, interlocutors take into account sequential organisation in conjunction with utterance form and content, the situational context and their common-sense knowledge of what people are likely to do. For example, interactants have implicit knowledge that requests are usually preceded by explanations and accounts, which is why the imaginary conversation in Figure 1.2 above works so unproblematically:

(1) You (turns to Steven with a sheepish grin): 'Oh damn, I forgot my wallet.'
(2) Steven (grins): 'No worries, I'll spot you a few bucks.'
(3) You: 'Thanks, man.'

'Your' turn in l.1 is interpreted by Steven in l.2 as a request to borrow money, even though you never say that you want to borrow money. However, people do not make statements to an interlocutor for no reason, and Steven's common-sense knowledge together with the sequential placement of the explanation as an opener leads him to interpret the utterance as a request, and he replies accordingly. This is followed in l.3 by a gratitude expression, closing off this sequence (known as sequence-closing third in CA). Note that Steven could have theoretically taken the utterance in l.1 as a piece of information with no requestive meaning and replied to it accordingly ('mhm', 'ok'), but since the implied requestive meaning is fairly obvious to a competent interlocutor due to the situation and the sequential placement, a response that is not responsive to this meaning could be seen as intentionally and churlishly ignoring it.

Assessment of interactional competence crucially involves interaction, so the two major studies (Ikeda, 2017; Youn, 2013, 2015) in this area have been based on data generated by role plays. While Hudson *et al.* (1995) also employed role plays in their test battery, they did not score interactional aspects but only focused on the target speech acts of request, apology and refusal. By contrast, Ikeda (2017) and Youn (2013, 2015) specifically set out to score interactional features and to determine if clear differences could be found between levels of learners' interactional ability.

From an assessment perspective, the main challenge in using role plays is standardisation. No two conversations will be exactly alike, and the role-play conductor's contributions will be affected by the test taker's utterances, so exact comparability between two role plays is impossible. In addition, role plays are resource intensive as they require one-on-one administration and scoring by human raters. However, role plays are the only somewhat structured method available for eliciting interactional competence, which is why they have been employed in assessment of interactional competence.

Youn (2013, 2015) administered role plays and monologue tasks to investigate ESL learners' interactional competence. Her role plays involved asking a professor for a recommendation letter and negotiating a meeting time with a classmate. The monologic task required test takers to give a classmate feedback on an email. Youn also included two monologic TOEFL speaking tasks as measures of speaking proficiency.

Youn attempted to counter the standardisation issue by providing test takers and role-play conductors with an outline of the role play and the social actions (pragmatic meaning) they would need to perform. In a major innovative step, Youn developed her rating criteria bottom-up, based on the role-play data, rather than deriving them from a theoretical model (as Grabowski [2009, 2013] did, see later). She had raters score her role plays on five criteria (from Taguchi & Roever, 2017: 236–237):

- Content delivery: smoothness and fluidity of turn initiations and transitions.
- Language use: range of pragmalinguistic tools in terms of structures, modals to express indirectness.

- Sensitivity to the situation: recipient-designing contributions, for example, by including accounts and explanations with a higher-power interlocutor.
- Engagement with the interaction: understanding of previous turns and active recipiency.
- Turn organisation: completeness of adjacency pairs and appropriateness of pauses.

She found that the role plays distinguished test takers' levels of interactional competence reliably. Similar to Roever *et al.* (2014), Youn also validated her test using an argument-based approach (Kane, 2012b), which offers a structured approach to confirming which kinds of inferences can defensibly be drawn from test scores.

Ikeda (2017) took off from Youn's work and also assessed interactional competence but put stronger emphasis on the impact of proficiency and exposure to the target language environment, as well as the overlap between role plays and monologic tasks. The latter research focus was implemented to investigate whether it would be possible to replace impractical and resource-intensive role-play tasks with more practical monologic tasks. While monologic tasks do not contain interactional features, the language used in them also needs to be recipient designed, taking into account the imaginary interlocutor's epistemic status (knowledge about the topic) and social status (politeness level required). Ikeda (2017) asked test takers to complete three role-play tasks with a trained interlocutor, and three monologic tasks, all set in an Australian university setting. The role-play tasks involved talking to a professor, an administrator and a classmate, and Ikeda designed two scenarios per pairing: with the professor, test takers request a signature to change classes and an extension on an assignment; with the administrator, test takers request a change of classes and help with a projector; with the classmate, test takers ask the classmate to pair up for a presentation and to re-organise the presentation. For each interlocutor type, one topic was done dialogically and one monologically.

Ikeda (2017) used the following criteria (from Roever & Ikeda, 2020: 489):

- Social actions to achieve communicative goals: take adequate actions explicitly tailored to the context to achieve a communicative goal.
- Facility with the language: deliver contents smoothly and clearly with sound variation (e.g. stress) and repair, when necessary.
- Language use to deliver intended meanings: control varied linguistic resources and employ linguistic resources naturally to deliver intended meaning, minimising the addressee's effort to understand the intention and the meaning of the speaker's utterance.

- Language use for mitigation: control varied linguistic resources to mitigate imposition.
- Engagement in interaction: engage in interaction naturally by understanding the interlocutor's turn and responding with varied patterns of responses well-tailored for the ongoing context.
- Turn organisation: take and release conversation turns in a manner that conveys to the interlocutor when to take turns.

Similar to Youn (2013, 2015), Ikeda's tasks also separated test takers clearly, but due to overlaps between proficiency and exposure, it was impossible for Ikeda to investigate the relative contributions of these background factors. However, he did ascertain a very large degree of overlap (correlation of $r = 0.94$) between test-taker scores on the dialogues and monologues based on the first four rating criteria. Ikeda (2017) also employed argument-based validation following Kane (2012b) to demonstrate that his test allowed conclusions to be drawn about test takers' interactional competence in a tertiary context in Australia.

## Other tests of interactional abilities

Besides Ikeda's and Youn's tests, which were designed under an Interactional Competence paradigm, two other tests investigated interactional abilities in different theoretical frameworks: Grabowski (2009, 2013) relied on Purpura's (2004) model of language ability, and Walters (2007, 2009) situated his test within Conversation Analysis (CA). Timpe (2013) used role plays as a productive supplement to her receptively focused test battery.

Grabowski's (2009, 2013) instrument consisted of four role plays from everyday contexts, requiring participants to make a complaint or sensitive request to an interlocutor. Grabowski's rating criteria were derived top-down from Purpura's model, and she found high reliabilities for most of the criteria. However, using criteria based on a general theory of L2 competence rather than a discourse approach precluded Grabowski from establishing what discourse features differentiated higher and lower-level performances.

Walters (2007, 2009) developed an unusual and creative test of interactional abilities, which unfortunately did not work well. Following Conversation Analysis, Walters investigated test takers' ability to predict a likely upcoming social action and to react to certain social actions appropriately. In the 10-item multiple-choice listening part of the test, test takers listened to a dialogue, and were then asked what the social action following the last turn would be likely to be, or what social action the final turn accomplished, e.g. 'Woman: "Hey, are you busy?" – Man: "No, not particularly",' with the correct interpretation of the man's utterance being that ‘It was possible to do something with the woman’ (Walters, 2004: 307).

In the productive part of the test, the tester and test taker had a conversation, in the course of which the tester included an assessment (in the CA sense, i.e. evaluation of something unrelated to the test taker), a compliment related to the test taker and a preliminary move (such as a pre-telling, ‘Hey, want to hear something funny?’). Test takers’ responses to these actions were then scored by raters.

Walters’s (2004) multiple-choice listening test was beset by very low reliabilities, and while the productive test showed reasonable interrater reliability, Walters (2004) cautions that the agreement between raters only reached 40%. The low reliabilities Walters attained were probably partially due to the homogeneity of his test-taker sample, but it is also likely that questions like the ones in his listening test are not answerable by language users. People do not reflect on the social actions that utterances like pre-tellings accomplish, and they do not deploy them consciously to accomplish these actions; in other words, they lack explicit knowledge, which is exactly what a multiple-choice test elicits. In addition, social actions are not deterministic, so even a pre-telling (‘The most amazing thing happened to me today’) followed by a go-ahead (‘What?’) does not make a telling compulsory. The telling could be delayed (‘I’ll tell you later, let’s have dinner first’) or cancelled (‘Never mind, you wouldn’t understand anyway’), or the pre-telling could be teasingly extended (‘Wouldn’t you like to know’). Given this indeterminacy, there was no clearly correct response option in the multiple-choice test, which tends to doom the reliability of multiple-choice instruments (as Hudson *et al.*, 1995, also found for their multiple-choice DCT).

Finally, Timpe (2013) included four Skype-delivered role plays in her test, two of which involved talking to a professor and two talking to a fellow student. Interlocutors followed a fairly fixed script, more detailed than the outline provided by Youn (2013). Raters scored test takers’ performance on three theory-derived criteria: how well the discourse was managed, how appropriate the performance was pragmatically (based on Hudson *et al.*, 1995, criteria) and how appropriate the performance was overall. Timpe attained high interrater reliabilities, but her fairly broad rating criteria do not provide much insight into test takers’ interactional abilities.

## Problems with interactional measures

From the point of view of drawing inferences from interactional tests in terms of real-world performance, it needs to be acknowledged that role plays are not the same as real conversations. In role plays, participants orient to the social situation that they are role playing as well as to the social situation of being in a role play. This means that they are aware of being watched, evaluated and assessed, and they will behave and talk accordingly. For example, Stokoe (2013) showed that police officers role playing interviews with suspects tended to stick to the manual, whereas real-world interviewers did not. Another difference is that there are no stakes associated with role plays, so the actual outcome does not have real-world consequences. This is supported by a study by Ewald (2012), who found that direction-giving in role plays was shorter and less precise than in real-world interactions. Overall, role-play interactions tend to be more oriented towards displaying linguistic abilities, whereas real-world interactions tend to be more oriented towards solving actual tasks. This problem could potentially be circumvented by having test takers engage in elicited conversation and solve a task together, but the lack of standardisation in this approach makes it less useful for testing, although it has been successfully used in L2 pragmatics research (Hanafi, 2015; Zhang, 2016) and also general L2 assessment (Galaczi, 2014).

Besides the aforementioned standardisation problem, a serious issue with interactional measures is their low practicality. Testing of interaction requires one-on-one sessions between testers and test takers, recording of interactions and scoring by human raters. This is logistically challenging and personnel intensive, but these are primarily concerns for large-scale testing operations, not so much for smaller-scale research projects. There is promising work in progress on computer-based testing of interaction through intelligent agents (Suendermann-Oeft *et al.*, 2015), although this technology is unlikely to be available in the very near future.

## Tests of interaction and implicit/explicit knowledge

Interaction requires online processing with no planning time as well as responsiveness to aural input, so interactional competence in conversation depends on ability for implicit processing. It is likely that it would not rely on highly automatised knowledge, as no two conversations are the same, so automatisation through practice is not really possible for interactional abilities (unless targeted practice of interactional skills is provided in a classroom, but that is so exceedingly rare as to be virtually nonexistent). While there is little opportunity to plan an individual utterance, speakers may pre-plan the overall structure of a conversation in terms of how to bring up a particular topic. This would happen in real-world interactions as well, at least where the target social action is an initiating one (in CA terms, first-pair part of the core adjacency pair), such as request or complaint. Where it is responsive, such as refusal, it would be much harder for participants to plan, especially if they do not know that they are going to have to refuse. In any case, sequential organisation of requests and refusals as well as turn design and fine-tuned recipient design are likely beyond conscious control and rely entirely on implicit processing.

Monologic tasks do afford test takers planning time, and while they also require recipient design, they do not require moment-by-moment adjustments for contributions based on interlocutor reactions. For researching the explicit/implicit processing of pragmatics, this difference between monologic and dialogic tasks might allow for interesting comparisons.

## Persistent Problems, Challenges and Research Gaps in Testing L2 Pragmatics

Construct coverage is probably the biggest challenge in L2 pragmatics testing, since the construct is potentially vast. Any type of real-world language use is influenced by context, intended meaning, social norms and preceding utterances, and eliciting test-taker knowledge about all these aspects would require a large test battery. Even modest attempts, such as the ones so far undertaken, require careful establishment of contexts and social relationships as well as role playing of interactions, all of which is resource intensive and lowers the practicality of pragmatics tests, and therefore the likelihood that they will be used.

A persistent theoretical question is the relationship between pragmatics and general L2 knowledge. Does testing of pragmatics provide sufficient value-add to justify the extra cost, i.e. does it provide extra information about test takers that tests based on the four skills or grammar/vocabulary cannot provide? Or can general L2 knowledge successfully account for and predict pragmatic performance? Studies like Ikeda's (2017) suggest that large differences in proficiency (measured without pragmatics) predict interactional performance, but smaller differences do not. Also, different areas of pragmatics are differentially impacted by overall proficiency (Roever *et al.*, 2014), and tests have probably not relied strongly enough on implicit/highly automatised knowledge and too much on explicit knowledge to really investigate this issue.

A continuing practical issue in pragmatics testing is to create items of sufficient difficulty. Most tests have been more easy than difficult for test takers, though this is of course directly related to test-taker ability level, and the inclusion of less able test takers lowers item difficulty indices. However, relatively little knowledge exists about what makes pragmatics items more or less difficult.

The particular challenge we address in this book is how to measure the ability to process pragmatic meanings implicitly. As we have pointed out, tests of L2 pragmatic competence have mostly contented themselves with measuring explicit processing. In the early days of pragmatics testing, prior to Hudson *et al.*'s (1995) battery and Roever's (2001) dissertation, it was uncertain whether it is even possible to test various aspects of L2 pragmatics reliably. We now know for certain that it is possible but, as Roever (2005) showed with verbal protocols, learners bring all their explicit and episodic knowledge to bear on answering test tasks, analysing scenarios to infer the meaning of implicature, consciously choosing politeness levels of speech acts and recalling real-world events to identify routine formulae. To measure the implicit pragmatic processing needed for real-world performances, it will be necessary to construct tests that require responses automatically without conscious deliberation. This is the goal we set ourselves.

## Notes

(1) A reviewer of this chapter pointed out that test items cannot be expected to reflect the real world. While this is clearly true, a processing perspective can inform how to design the test conditions so that they match the processing conditions found in real-world performances. This is the psycholinguistic perspective we have adopted in this book.

(2) The tests that we developed (see Chapter 2) were primarily intended for research purposes (e.g. in studies investigating the effects of study abroad or of pragmatics instruction) rather than for pragmatics assessment.

(3) In the Metapragmatic Knowledge Test we developed, we included items that were inappropriate due to being overly polite and found that these did prove to be more difficult for L2 learners (see Chapter 3).

# Part 2
# The Development of the Tests

In this part of the book, we describe the test battery and report a series of studies involving each of the tests that we developed.

In reaching decisions about what tests to include, we were cognisant of the demands that a battery of tests would make on test takers and the need to minimise these as much as possible. To ameliorate the resource-demanding nature of some of the tests, we decided to include discrete-point tests that could be easily scored alongside tests that required transcription and coding. The practicalities of administering the tests in China and Japan also influenced our choice of tests and prevented us from exploring some of the possibilities put forward in Chapter 1. For example, we were not able to collect response times for the computerised discrete-item tests, with one exception. We aimed to assess a broad range of pragmatic aspects, including speech acts, implicature and sequential organisation. We wanted to include tests of implicit/explicit abilities that measured both comprehension and production.

To the best of our knowledge, there has been no attempt to establish a psycholinguistic basis for the implicit/explicit distinction applied to pragmatics. Consequently, our project should be seen as an exploratory one. Its exploratory nature is evident in both the inclusion of novel tests and in new ways of scoring more established tests. We will openly admit to limitations in our tests and put forward suggestions for how the tests might be further developed.

In Chapter 2, we explain the theoretical basis we drew on when designing the tests in making claims about what each test measures. We will first briefly review research that has investigated implicit and explicit grammatical knowledge, as this was the starting point of the project. We then argue that in the case of pragmatics, the implicit/explicit distinction needs to be understood in terms of processing rather than knowledge. Pragmatics is not a matter of ‘rules’ but of general ‘principles’ that govern how what we say relates to context. The issue then is whether these principles are processed automatically without consciousness or deliberately with conscious effort. Our aim was to design tests that would favour implicit or explicit processing of these principles. It is more appropriate then to talk about pragmatic *abilities* than pragmatic *knowledge*.

In the same chapter, we provide general descriptions of each test and the background questionnaire and proficiency test that all the participants completed. We also provide information about the participants – a group of native speakers and two populations of undergraduate learners of English, one in China and the other in Japan. This chapter then provides the background for the whole project. However, because we anticipate some readers will elect to read specific chapters rather than the whole book and because the details of the instruments varied slightly from study to study, we will repeat information about the instruments in each chapter.

Chapters 3–8 examine each of the testing instruments in turn. These chapters have a number of purposes. One is to carry out an evaluation of the tests and also to identify ways in which they can be improved. The second is to use the data collected from the tests to assess whether there were grounds for claiming each test measured the pragmatic abilities intended (i.e. implicit or explicit). The third is to use the tests to investigate the pragmatic abilities of two sets of EFL learners and, in particular, to see whether their test scores were related to their English language proficiency and language learning experiences.

Chapter 3 addresses the Metapragmatic Test, which we designed to provide a measure of explicit pragmatic processing. We begin by defining metapragmatic knowledge and consider to what extent language users are likely to possess this kind of knowledge. We then provide a review of a selection of metapragmatic tests that have been used in L2 pragmatics research. To evaluate the test, we examine its intrinsic properties and compare the test performances of three groups of test takers (the native speakers, the Japanese university students and the Chinese university students). Despite some weaknesses, we conclude that the test has achieved its main purpose – to assess explicit pragmatic processing – but we also offer some suggestions for improving it.

The Social Variables Test (Chapter 4) was also intended to assess explicit pragmatic processing. In designing the test, we looked for a way of focusing on the sociopragmatic aspect of pragmatic competence (i.e. how social variables determine the level of mitigation needed to ensure politeness). We hit upon the idea of asking participants to decide which of three situations best matched a stimulus utterance that realised a speech act such as a request or apology. We report an evaluation of the test based on data collected from the native speakers' and the Japanese L2 learners' performance on the test. We identify a number of weaknesses in the test and conclude that it, as it stands, it is not sufficient for its purpose – namely, it does not constitute a valid way of measuring participants' explicit pragmatic processing. We conclude with suggestions for how the test might be improved for future research.

Chapter 5 reports an investigation of Chinese University students' ability to comprehend irony, which previous research has shown to be a late-developed aspect of pragmatic competence. We discuss the extent to which this ability draws on implicit or explicit processing, pointing out that this will depend on whether the ironic meaning of an utterance is processed directly (i.e. without first processing the literal meaning) or indirectly (i.e. after first processing and then rejecting the literal meaning). The Irony Test we developed asked the participants to first read information about a situation in which an utterance occurred and then to listen to the utterance. They were instructed to indicate whether the utterance conveyed the speaker's positive or negative attitude. Response times were recorded. The test items differed in terms of whether the utterances expressed literal or ironic meaning and, in the case of the latter, whether the ironic meaning was positive or negative. The Chinese learners scored at the same level as the native speakers on the literal items but had markedly lower scores on the ironic items. Also, the learners' response times were much slower than those of the NSs. We concluded that whereas NSs can access ironic meaning directly and implicitly, the learners – even those with advanced language proficiency – probably relied on explicit processing strategies.

In Chapter 6, we describe a novel use of elicited imitation as a way of measuring implicit pragmatic processing of hedges. In the Elicited Imitation Test, the participants were presented with a brief description of a situation on a computer screen, read an utterance which then disappeared, answered a question about the utterance and then reproduced the utterance orally. The utterances in the test items included hedges (e.g. 'just', 'possibly' or 'somehow'). The test was scored in terms of whether the native speakers and Chinese learners included the target hedge when they reproduced an utterance. To demonstrate the concurrent validity of the test, we showed that there was a statistically significant relationship between the hedging scores and scores derived from the Dialogic Role Play. The test had acceptable reliability, and the discriminability of the items was generally very good. There was considerable variance in the learners' hedging scores, but overall, these were much weaker than the native speakers' scores. We also examined the relationship between the test scores and the learners' language proficiency/language experience. We argued that, as the learners were very unlikely to have focused their attention on the hedges when they reproduced the utterances, the test functioned as we intended – namely as a measure of the ability to process hedges implicitly.

Chapters 7 and 8 address the role plays included in the test battery. Chapter 7 reports the results of a study investigating the Monologic Role Play and Chapter 8 the Dialogic Role Play. The approach we adopted in both chapters was to use discourse-analytic techniques to analyse the data collected from the native speakers in order to develop a scheme for scoring the learners' inclusion of specific pragmatic features in their performances of the role plays. This approach differs from the customary way of scoring role plays, which has involved the training of raters in the use of rating schemes (see Chapter 1). We argue that a points-based scoring system offers a less subjective means of assessment than rating schemes. Chapter 7 investigated the Japanese learners' performance of a monologic role play. It draws on research involving 'genre' (e.g. Halliday & Hassan, 1989) and on research investigating the pragmalinguistic features of speech acts (e.g. Blum-Kulka *et al.*, 1989). Chapter 8 investigated the Chinese learners' performance of a dialogic role play and involved the interactive negotiation of a solution to a problem. Through an analysis of how native speakers realised the problem-solution pattern, we were able to derive a set of sequential pragmatic elements and pragmalinguistic features that characterised the successful completion of the role play. In both chapters, we discuss the advantages and disadvantages of this discourse-analytic approach to assessment. Both role plays were intended to provide measures of the learners' implicit pragmatic processing.

Finally, in Chapter 9, we made use of the six tests to see to what extent they could be interpreted as providing distinct measures of implicit and explicit L2 processing. Using data from both the Chinese and Japanese learners (i.e. all 187 learners), we tested three models using confirmatory factor analysis. The results suggested that the tests may indeed be distinguishing the two types of processing. However, further analyses were less supportive. We hypothesised that language proficiency would be a better predictor of the measures of explicit processing than the measures of implicit processing on the grounds that the formal nature of the language instruction the learners had received would favour the type of proficiency required for explicit processing. Conversely, we predicted that the length of time learners had spent in an English-speaking country would predict scores on the tests of implicit processing more strongly. Correlational analyses and a multi-regression analysis did not support these hypotheses, and we suggest why. We put forward ways of revising the tests to enhance the likelihood of tapping implicit processing.

Were we successful in designing tests that could distinguish learners' implicit and explicit processing? Limitations in the design of some of the tests – in particular, the Social Variables Test – and doubts about the theoretical basis of some of the other tests – the Elicited Imitation Test, for example – make strong claims untenable. Nevertheless, the chapters in this section of the book do provide evidence supportive of the model on which the tests were based, while the individual chapters provide in-depth information about the different tests and point to ways in which they might be further developed.

Below is a list of the tests along with the abbreviations we will use to refer to them.

| | |
|---|---|
| EIT | Elicited Imitation Test |
| DRP | Dialogic Role-Play Task |
| MRP | Monologic Role-Play Task |
| IRT | Irony Test |
| MKT | Metapragmatic Knowledge Test |
| SVT | Social Variables Test |