Article

LANGUAGE TESTING

Lasnguage Testing
28(1) 5–29

DOI: 10.1177/0265532209359514
http://ltj.sagepub.com
SAGE

Effective rating scale development for speaking tests: Performance decision trees

Glenn Fulcher
University of Leicester, UK

Fred Davidson
University of Illinois at Urbana-Champaign, USA

Jenny Kemp
University of Leicester, UK

Abstract

Rating scale design and development for testing speaking is generally conducted using one of two approaches: the measurement-driven approach or the performance data-driven approach. The measurement-driven approach prioritizes the ordering of descriptors onto a single scale. Meaning is derived from the scaling methodology and the agreement of trained judges as to the place of any descriptor on the scale. The performance data-driven approach, on the other hand, places primary value upon observations of language performance, and attempts to describe performance in sufficient detail to generate descriptors that bear a direct relationship with the original observations of language use. Meaning is derived from the link between performance and description. We argue that measurement-driven approaches generate impoverished descriptions of communication, while performance data-driven approaches have the potential to provide richer descriptions that offer sounder inferences from score meaning to performance in specified domains. With reference to original data and the literature on travel service encounters, we devise a new scoring instrument, a Performance Decision Tree (PDT). This instrument prioritizes what we term 'performance effect' by explicitly valuing and incorporating performance data from a specific communicative context. We argue that this avoids the reification of ordered scale descriptors which we find in measurement-driven scale construction for speaking tests.

Keywords

domain analysis, language testing, rating scales, performance decision trees, test design, validity

Corresponding author:
Glenn Fulcher, 162 Upper New Walk, The University of Leicester, School of Education, Leicester, LE1 7RF, UK.
E-mail: gf39@le.ac.uk

Rating scales designed for performance assessment have been classified in a number of ways, including holistic or analytic, primary or multiple trait (Hamp-Lyons, 1991), as 'real-world' or 'ability/interaction' focused (Bachman, 1990), and as oriented towards the user, the assessor, or the test constructor (Alderson, 1991). Whatever the classificatory systems used to describe rating scales in relation to their purpose, there are currently two major approaches to rating scale design. The first (and oldest) approach is to design the rating scale on the basis of a measurement model. If performance data is considered at all, it is usually in the post hoc activity of selecting performance samples that are claimed to typify a level on a scale. The second approach places value on performance data in the construction process of the scale. Performance data is either described in detail, or referred to in establishing differences between levels on a scale. The resulting scales cannot usually be used to rate performances in any other context except for the one for which they were developed, as the scales incorporate specific descriptions of performance in particular domains and genres. While this limits the generalizability of score meaning, it also makes the inferences drawn from scores sounder. In the next section, 'Two Approaches to Rating Scale Design', we outline the measurement-driven and the performance data-based approaches, briefly analysing the content of a service encounter rating scale constructed using the measurement-driven approach in order to show its weaknesses.

Irrespective of the approach to design, current rating scales share the assumption that the construct increases in a linear fashion, as described in the levels or bands of the scale. Although second language acquisition researchers have long argued that this is a simplification that distorts the known facts of how humans learn second languages (Meisel, 1980), even data-based rating scales, while claiming to describe non-linear acquisition (Fulcher, 1996), still present hierarchical levels of ability. In measurement-driven approaches there is the strongest tendency to claim that the levels represent a 'ladder' to be climbed (Westhoff, 2007), even though there is no evidence of linearity (Hulstijn, 2007, p. 666).

In order to avoid this problem, we argue that performance data-based scales need to evolve into a new type of rating instrument, which we call Performance Decision Trees (PDTs). PDTs represent an improvement on performance data-based scales in that they escape from the illusion of linear development in language use. They are based in a thorough analysis of the context of performance and the nature of interaction in specific communicative situations. The importance of context and interaction in the assessment of speaking has been widely demonstrated (Jacoby and McNamara, 1999; Jacoby and Ochs, 1995; Kramsch, 1986), and the socially constructed nature of discourse has become the focus of much research (Brooks, 2009; Chalhoub-Deville, 2003; Swain, 2001). Yet, this has not so far been exploited in scoring test performances. PDT design attempts to incorporate these insights into the design of the scoring mechanism.

To illustrate this new method of scale design we have selected the domain of service encounters, with specific reference to travel agency discourse, in order to show the importance of context and interaction in the assessment of speaking. Service encounters have been selected because of their critical importance in 'getting things done' for survival in a target language context (McCarthy and Carter, 1994, pp. 24–27), avoiding intercultural misunderstanding (Ryoo, 2005, p. 81) and managing interaction in an

academic setting (Biber et al., 2002). A scale to evaluate such interactions could be used to evaluate the ability of tourism students to successfully engage in service encounters in a second language, as well as to evaluate whether general language learners could get the services they might need when travelling.

In the third section, 'Interactional Competence in Service Encounters', we analyse the discourse competence, discourse management techniques, and pragmatic competences required to engage in successful travel agency service encounters. In order to do this we refer to previous studies of travel agency discourse, and present original data collected for this study. In the fourth section, 'Service Encounters in Marketing', we look at research from marketing that has looked at the kinds of service encounters that produce brand loyalty, which helps define pragmatic competence from an indigenous perspective. Drawing on discourse, focus group and narrative studies among others, we summarize the qualities of service encounters that have become the basis for training within the industry, and which may be utilised within a scoring instrument.

We use all these sources to produce a rich description of language use in a specific context in order to derive the categories of assessment. In the fifth section, 'A Scoring Model for Service Encounters', the rich description is used to derive a PDT that can be used to score performances on tasks designed to simulate travel agency communication. The paper concludes with suggestions for its potential use and further research that is required to operationalize the PDT presented.

Two approaches to rating scale design

Measurement-driven methods

Measurement-driven-approaches have a long history. The oldest and most prevalent is the *a priori* method, which relies upon an individual or committee who are perceived to be experts in the teaching and assessment of the construct of interest. Drawing upon experience and knowledge, a rating scale is crafted that appears reasonable to the designers. Such scales are not without theory, but theory impacts through experience and remains largely implicit (Wilds, 1975). While a scale may be refined over time, it also becomes a system with which users feel comfortable, so that its use and interpretation is a matter of socialization (Lowe, 1986).

Currently the most influential measurement-driven approach is the scaling of descriptors using Rasch measurement to create a scale in a pre-defined number of levels (Baker, 1997). This approach to scale design is primarily identified with the creation of the Common European Framework of Reference (CEFR) and the work of North (1996; 2000). The CEFR primarily consists of a set of rating scales in six levels (A1 to C2) that are widely used in Europe to guide assessment and learning. Although the scale is empirically derived, it is not based on performance data, as there is no reference to the performance of learners or test takers on specific tasks, or even perceptions of the value of performances. Rather, this methodology depends on the ability to use a measurement model to place band or level descriptors drawn from disparate sources onto a single scale using teacher estimates of descriptor difficulty as data. The measurement model – Rasch in this case – is seen as an external arbiter that decides what does and does not survive in the scale development process.

The service-encounter scale of the CEFR is typical of the kinds of scales generated by this approach. The public domain contexts in which these transactions may take place are described in a taxonomy (Council of Europe, 2001, pp. 48–49) in which 'goods' and 'services' are thrown together, there is no analysis of categories or their relationships, and no distinction drawn between qualitatively different communicative transactions, such as purchasing fish and chips, or a car (McCarthy and Carter, 1994, p. 63; Ylänne-McEwen, 2004, pp. 518–519). Still less is there any attempt to distinguish between purchasing goods, and obtaining services that are 'less tangible' (Coupland, 1983. pp. 464–465). Any section or item from this list may (or may not) be relevant to language use in a particular domain.

When analysing the CEFR's illustrative scale for transactions (Council of Europe, 2001, p. 80), we immediately face a number of problems which stem directly from its method of construction. Some descriptors refer to specific situations, while others do not. Level B2, for example, refers to getting a traffic (parking?) ticket, damaging property, and dealing with being blamed for an accident. Other levels are less specific. When a context of language use is mentioned, it is not necessarily referred to in other descriptors. Dealing with travel agents is specifically mentioned in Level B1, but not at other levels, despite references to travel. We are therefore left with the question of whether 'dealing with travel agents' is something that is suddenly possible at level B1. Furthermore, participant roles are mixed within the same level. At A2 for example, the learner can 'ask for and provide' goods and services. This seems to imply that an A2 learner would be able to function as a service provider as well as a server seeker. At level B2 would this mean that a learner could explain to a client how to seek compensation, as well as ask for compensation as a customer? The distinction between levels is unclear, with descriptors referring to the vague concept of 'complexity' at each level. At level B1 learners can deal with 'most' transactions, as well as 'less routine' situations. But there is no definition of 'less', 'more' and 'most'. A2 is characterized by 'common', 'everyday', 'simple' and 'straightforward' transactions, but we are not told what these might be.

B2 would appear to be somewhat more interpretable than the other levels because it implies the ability to deal with problems, some of which might be quite serious. However, level B1 also states that one can deal with authorities (which implies a problem with customs or whatever), make complaints, and return unsatisfactory purchases. Irrespective of whether it is possible to distinguish between levels B1 and B2 in these terms, the whole notion of using 'problems' to distinguish between levels is itself highly problematic. As we shall see below, the participants in a service encounter work together to establish a relationship that enables both to achieve their respective goals. Conflict or loss of face are avoided at all costs. Situations in which disagreement is likely to arise are very different from normal service encounters, in that the parties are likely to have different expectations about the outcome of the interaction, and power tends to reside with the service provider rather than with the customer. One such example is the service encounter between social workers and unemployed homeless clients (Spencer, 1997). In these encounters a service request is frequently met by an alternative service offer (Spencer, 1997, p. 190), which would hardly be appropriate in most contexts.

Measurement-driven scales suffer from descriptional inadequacy. They are not sensitive to the communicative context or the interactional complexities of language use.

The level of abstraction is too great, creating a gulf between the score and its meaning. Only with a richer description of contextually based performance, can we strengthen the meaning of the score, and hence the validity of score-based inferences.

Performance data-based methods

The prototypical performance data-based methodology in speaking scale development was discussed by Fulcher (1987). This requires the collection of performance samples from learners undertaking test tasks drawn from the universe of generalization, transcribing the performances, and identifying key performance features using conversation or discourse analysis. When features are isolated, the number of levels in a scale that can be empirically established using discriminant analysis; the resulting scale levels are then populated by descriptors drawn from the primary analysis (Fulcher, 1993; 1996; 2003). Although extremely time consuming, this method is currently the only one that requires the close analysis of language (North, 1993, p. 129), and is therefore grounded in performance data. The main disadvantage of the method is that it generates level descriptors that raters frequently find too complex and have difficulty using in real-time rating (Fulcher, 2003).

Another performance data-based method is the empirically derived, binary-choice, boundary definition scales (EBBs) of Upshur and Turner (1995; 1999), Turner (2000) and Turner and Upshur (2002). What distinguishes this method is that the scale – and hence the cognitive process that raters must follow – is set forth as a series of repeated and branching binary decisions. EBBs are constructed by rank ordering performances on test tasks and then identifying key features that judges use to separate the performances into adjacent levels. EBBs represent an innovation in the logic of how raters judge performance with reference to performance data in specific contexts of language use. EBBs may not contain the rich description of the previous method, but they are relatively easy to use in real-time rating, and do not place a heavy burden on the memory of the raters.

The PDT presented below takes advantage of both of these performance-based methods. While it incorporates the descriptive richness of the first, it also retains the simplicity of decision making offered by the EBB. Thus, the first step in the development process is to describe the nature of the interaction in the specific communicative context of interest.

Interactional competence in service encounters

The first element of interactional competence in this domain is *discourse competence* (Canale, 1983a; 1983b; Canale and Swain, 1980), or the ability to understand and utilize knowledge of the structure of a service encounter to provide a service, or get the service needed. A minimal requirement for successful task completion is the ability to structure questions and responses in coherent and cohesive adjacency pairs to produce interactive discourse that follows a service encounter script. We describe and illustrate this in the subsection ‘Discourse competence in service encounters’ below.

The second element is *competence in discourse management*. In the subsection ‘Competence in discourse management’ below, we are concerned to describe the kinds

of discourse management techniques that successful interactants use to move the discourse through the phases of the service encounter script.

The third element of interactional competence is *pragmatic competence*. This is the language learner's ability to interact appropriately with others in ways that preserve face in 'close encounters' (McCarthy, 2000) where temporary relationships are established as an important part of the interaction. We discuss pragmatic competence extensively in the subsection 'Pragmatic competence' below.

Discourse competence in service encounters

Any learner who completes a task by realizing the obligatory elements of their role within a service encounter would have demonstrated basic discourse competence within this domain. Mitchell (1957) was the first researcher to document the genre of buying and selling in the marketplace. Since then a great deal has been done to elucidate the language and discourse skills required (McCarthy and Carter, 1994, pp. 24–26). Hasan (1985) used service encounters as the basis for developing her seminal theory of text structure which explains how we relate context, genre and linguistic realizations (Ventola, 2005, pp. 27–28). She argued that a description of the field, tenor and mode (Halliday, 1985, p. 12) of a text provides a definition of its *contextual configuration* which links the realization of specific utterances to their social context. The configuration of service encounters is made up of nine defining discourse elements: the greeting (G), sale intention (SI), sale request (SR), sale compliance (SC), sale enquiry (SE) sale (S), purchase (P) purchase closure (PC) and the close or finis (F). SE, SR and SC are iterative and may be repeated; the elements SR, SC, S, P and PC are obligatory. The genre of the text is defined by these obligatory elements and therefore link the generic structure potential (GSP) to its context.

The GSP explains how we are able to recognize discourse as belonging to the genre of the service encounter, as the following example from a travel agency discourse test shows (Mills, 2009, p. 107).

1. S8: I need a ticket to.. osaka japan>
2. S7: <7> er….we..we have
3. three flights to osaka weekly ….monday::/ thursday::/ and
4. friday:: ∴…..the flight on monday:/ and friday:/ are direct ∴
5. and..the one..on….thursday:/…. has a stopover in
6. hong kong/∴..when were you thinking of flying to osaka:>/ =
7. S8: =….er <4> [p] sorry..can you/….can you repeat that please/ =
8. S7: = <3>thursday::/ }
9. S8: { [p] thursday: ∴ }
10. S7: { thursday:/ <4> on thursday:/..
11. <has a stopover in hongkong>/∴
12. S8: [p] ah yes <5> I would
13. prefer/….[p] thursday: ∴ =
14. S7: = [p] thursday: ∴ <4> er <5> will this be
15. round trip or one way/ =

16. S8: = round trip <3> round trip
17. returning….following - ..monday∴ =
18. S7: = er..how would you like to
19. fly/ economy:/ busine:ss/ or first class/ =
20. S8: = business class
21. please - =
22. S7: = <3> and..will anyone be traveling with you/ =
23. S8: = <5> no..I'm traveling alone∴ =
24. S7: = <3> er <3> OK then∴ ….
25. <please give me a minute while I check price and availability>
26. <17> the flight/..the flight/..departs at/..nine/ [<forty am∴>
27. <and arrives in osaka:>/ <3> osaka:/..at.. < five thirty
28. pm>∴….local time∴….the price/ ..i:s/ one thousand/∴ ..four
29. hundred/∴…twenty:/ dollars∴ <4> <shall I book it for you>/ =
30. S8: = ….not yet∴ ..I'll get back to you∴ <thank you very much for
31. your help∴>

This dialogue was generated by a simulation task in which Korean students of English for tourism were asked to act out a service encounter (see Appendix 1 for the task). The exchange clearly begins with a sales request in line 1, followed by sales compliance, and a sequence of iterative sales enquiries. The sale is initiated, but the sale purchase is turned down, before the closure takes place. While not a complex interaction, this sample of student discourse shows mastery of the genre, which provides evidence of interactional discourse competence. We argue that an ability to produce the basic obligatory elements of the genre is the first criterion for the evaluation of service encounter interaction.

Competence in discourse management

Coupland (1983) has argued that discourse is variably managed in travel agency encounters and has attempted to identify features that make discourse management smoother. The first feature is *marking a transition boundary in speech*. Using a full lexical marker is typical of the most proficient speakers (e.g. 'now, the next thing I have to find out …'). Although not as efficient, the use of a filled pause is an implicit marking of a transition boundary; and no marker use (an extended pause) has been found to cause communication problems.

The second feature is the *elicitation of identification of purpose* and *the provision of an explicit response*, as in: 'can I help you?' or 'mmm I was just wanting some information on the ferries please.'

The third feature is *identifying participant roles* (Coupland, 1983, p. 470), which relates to the opening questions from the client. Coupland discovered that the most explicit interrogatives (i.e. that were closer to a fully specified illocutionary act) were most successful, as in: '...have you any idea how much it would be for two adults and a child in Ostend at about the end of August?' A reference to the illocutionary act in the question (tell-give, let-know, let-have) is the most explicit way of identifying role, while questions that make explicit the participant roles but do not mention the act of transferring information are slightly less explicit.

Fourth, Coupland discusses the *management of closings* (1983, pp. 471–473) where the most competent speakers perform explicit transaction closings, after which is the use of a single word bridge (e.g. 'okay', 'alright', 'fine', etc.) to act as a closing, while the least proficient speakers do neither. A transaction closing is typically constructed of a pre-closing move such as an encounter evaluation (e.g. 'tremendous', 'quite painless wasn't it?') a proverb or aphorism (e.g. 'oh well I'll have that much more to spend'), identifying the next activity (e.g. 'oh I'll try next Friday then'), reformulating conclusions (e.g. 'so I'll look it up when I go back') or reformulating purposes (e.g. 'well that's the only reason I came in'). This pre-closing is followed by leave taking.

The following extract from our data illustrates a very successful leave taking following a discussion in which the sale and purchase are not completed. These are situations in which the pragmatics require face-saving moves with an extended closing.

Managing closings (original data)
(TA = Travel Agent; C = Customer)

1	TA:	the only thing with that is that probably won't be there till Thursday [because the se- are
2	C:	I know (that's it) that's (why I'll speak)
3	TA:	(.) we are not the only agents that deal with Company A
4	C:	that's why I'll speak to her today I'll [speak to( ) =
5	TA:	[yeah see ( )
6	C:	=pop in [( )
7	TA:	I'll [give you my=
8	C:	[if we ca-
9	TA:	=number
10	C:	right=
11	TA:	='cause I mean I can always get you something else [anyway
12	C:	[okay 'cos that's what I was thinking yeah so
13	TA:	I'm just really looking at price range=
14	C:	we-
15	TA:	=and things [like that
16	C:	[yeh so (if worst comes to worst [(we'd probably=)
17	TA:	[but you will defini- i will defini [tely get
18	C:	=[get] some[thing (if) we come and sat down and we look through [( )
19	TA:	[something ( )
20	TA:	[course you can yeah] I'm in all week anyway so:::
21	C:	e:r okay
22	TA:	but it will be (Thursday just not too sure exactly about that [one)
23	C:	[okay] that's all right I ([ )
24	TA:	all right then ((louder + raised pitch)) if you need any questions just ask (me when you bring it in) and I'll talk you through it all

25 C: (right)
26 TA: all ri:::ght
27 C: thanks for your [time
28 TA: [see] you later
29 C: cheers
30 TA: ‘bye ‘bye
31 C: ‘bye
32 TA: (‘bye)
33 C: bye

This is a very explicit transaction. Exchanges 1–10 contain the first pre-closing sequence, in which the agent tells the customer that the holiday may disappear if he does not make a quick decision; he identifies his next activity, which is consulting his travelling partner and ‘popping in’. The travel agent gives him an alternative, which is calling. The second pre-closing sequence occurs in exchanges 11–20, where the travel agent reassures the customer that if the holiday has gone she will be able to find something else. We see the use of an aphorism in exchange 16, and a repetition of the next activity in exchange 18. From the second half of exchange 20 to exchange 25 we see a third pre-closing sequence establishing when the travel agent will be available, ending in an offer of help. The participants then begin extended leave taking.

A fifth feature that has been identified with the quality of service encounter discourse is the *use of backchannelling* in the speech of the client. McCarthy (2003, p. 52) characterizes this as a distinctive feature of small talk that enhances ‘congenial relationships’.

From conversational analytic studies of service encounters and the analysis of our own data, we therefore arrive at five criteria for discourse quality in service encounters:

- transition boundary markers
- explicit expressions of purpose
- identification of participant roles
- management of closings
- use of backchannelling.

Pragmatic competence

If learners are able to establish basic discourse competence and manage it reasonably well, we become interested in the more complex aspects of service encounter interactions. The pragmatics of the context take us beyond Hasan’s notion of sales enquiry to look at relational management, or how the interactants use the non-obligatory elements of the GSP to establish and develop a temporary relationship that makes the encounter both successful and pleasant. While these elements are not essential in a service encounter, they almost always occur in real-world interactions.

The weakness of Hasan’s (1985) GSP for complex service encounters lies in the unanalysed nature of the Sales Enquiry (SE). This single element accounts for anything not directly related to a greeting or the sale itself. However, in more complex service encounters the SE covers a very wide range of interactions. Of particular importance are elements

that resemble 'borrowing' from other genres, particularly casual conversation. Ventola (1987, p. 83) refers to this as 'side-sequencing' in which another genre is embedded within the structure of the service encounter. Such side-sequences maintain field and mode, and so are always related to the topic of the service encounter, but allow fluctuations in tenor. Temporarily the roles of the participants as vendor and client are suspended as they 're-align' themselves. Side-sequences provide insights into the 'unfolding of social interaction in stages' (Ventola, 2005, p. 25). However, Ylänne-McEwen (2004, p. 521) interprets such sequences somewhat differently, seeing them not as stepping outside the role, but as an integral part of the role. She accounts for the sequences by identifying different goals that participants bring to a service encounter:

- the overt, instrumental goal of the encounter, such as making travel arrangements;
- identity goals, such as self-presentation or fulfilling an institutional role; and
- relational goals, such as establishing and maintaining social relations.

Ylänne-McEwen provides the following example of side-sequence that could not be easily accounted for by the notion of sales enquiry.

Side-sequencing

SI	=	Agent:	can I help you?
SR	=	Customer 1:	Kusadasi in er Turkey
??	=	Agent:	ah! I'm going there in the summer
		Customer 1:	Eh?
		Agent:	I'm going there in the summer
		Customer 1:	are you?
G		Customer 2:	hiya
G		Agent:	hiya
		Customer 1	er
??		Customer 2:	it's nice in Turkey been in Turkey?
		Agent:	I haven't been before no
		Customer 2:	it's lovely
		Agent:	is it?
		Customer 2:	yeah

It is argued that in this encounter the agent 'is aligning herself not to the role of staff/server, but rather to the role of a fellow traveler, who has not visited the particular destination and wants to find out something about it' (Ylänne-McEwen (2004, p. 523). She goes on to ask the customers about how cheap food and drink is in Turkey before returning to the institutional genre. The change in tenor places the participants in equal power roles, and the interaction looks more like casual conversation. However, the fact that these side-sequences are so ubiquitous in service encounters raises the possibility that they serve a specific purpose, despite their optionality. As Ylänne-McEwen (2004, p. 533) may partially agree, this is what Fairclough (1995, p. 138) terms 'synthetic personalization', or engaging in 'relational talk' to further the institutional

goal of selling by showing interest through the use of personal, non-institutional, language. While critical discourse analysis sees ulterior motives at play, the real purpose of these side-sequences is to establish 'rapport' in the management of a temporary relationship in the gaps between the transactional elements of the discourse. This 'facilitates [transactional episodes] and enhances their efficiency and threads them into the socially recognizable fabrics that constitute our everyday spoken genres' (McCarthy, 2003, p. 35). This is supported in our own data on service encounters, as in the following extract.

Establishing rapport through relational talk (original data)

1	TA:	=yeah. [an] ything else you need sorting.=
2	C:	[Um]
3	TA:	=have you got accommodation then in Barcelona ↓yeah.
4	C:	u::m well the conference is (.) at a particular [hotel (.) so that's]=
5	TA:	[Oh right, so they]
6	C:	=arranged by the [org]anisation.
7	TA:	[yeah]
	TA:	oh that's lucky then yeah.
8	C:	so um:: (1.0)
9	TA:	that's [all] sorted=
10	C:	[er]
	C:	=that's that's sorted. [er::]
11	TA:	[just got] to make your own way [there.]
12	C:	[It may be that] (.) for the couple of da::ys sightseeing that we tag o::n (.) that I ask you a[bout somewhere.]
13	TA:	[yeah that'll be okay.] have [you been] to Barcelona=
14	C:	[er::m]
15	TA:	=↓before.
16	C:	↓no:::.
17	TA:	it's a fantastic city, you'll [love it.]
18	C:	[No, I] haven't done.
19	TA:	brillian- If you love art, you'll love Barcelona.
20	C:	u::m yes I do.
21	TA:	yeah. [you'll love it.]
22	C:	[I do. and] erm my sister's been but I haven't [(.) so::]
23	TA:	[I re]ally liked it. all the Gaudi and=
24	C:	=yeah=
25	TA:	=it's just beautiful. (.) lovely. [( )]
26	C:	[yeah. she's]=
	C:	= shown me lots of pho[tos before and (it's beautiful)=
27	TA:	[yeah. (.) it's lovely (.)=
28	C:	And (.) the architecture's (.) fantastic seems amazing

In this extract the travel agent is projecting herself as a friend, sharing her holiday experiences and interests. The tenor of the interaction has temporarily changed, although the field has not. Of particular interest in this extract is the lexically cohesive chain of 'fantastic-love-like-lovely-beautiful-lovely', as the two speakers echo each other's assessment of the city from direct experience on the part of the travel agent, and from photographs on the part of the customer.

We can see from this extract why this genre has been termed a 'socially expanded service encounter' (Ryoo, 2005, p. 93), among which the only non field-related admissible topic is the weather (Coupland and Ylänne-McEwen, 2000), while the integration of personal accounts and stories related to the field is common (Bastos, 1996, pp. 161–162). This interaction fulfills both an interpersonal and a goal-related role; as McCarthy (2003, p. 34) argues, this 'small talk' is 'anything but superfluous, frivolous, secondary, or irrelevant to the analysis of the main stream of talk'.

We therefore arrive at another criterion for successful performance in service encounters. This may be stated as the degree to which participants are capable of embedding relational talk within transactional talk in order to establish the kind of rapport that helps participants achieve their goals more effectively. The nature of this pragmatic rapport can be further defined by looking at indigenous criteria, which we investigate in the next section.

Service encounters in marketing

Lumley et al. (1994) and Jacoby and McNamara (1999) have raised the question of whether the criteria by which language testing specialists judge successful communication matches what is valued by professionals from the field. This has been termed 'indigenous assessment', in which 'professionals typically call upon a rich inventory of tacitly known criteria in order to determine whether and to what extent some particular performance is competent or falls short of the mark' (Jacoby and McNamara, 1999, p. 224). In this section we look at research that may inform the construction of assessments and scoring in this domain.

In the last two decades communication and interaction in service encounters has become the focus of attention for marketing researchers. It is now widely accepted that failure to provide satisfaction in fleeting encounters with customers causes them to switch to other service providers (Keaveney, 1995). This represents a shift in marketing research from a transaction-oriented model to one that focuses on relationships (Grönroos, 1993). Service encounters with customers have even been termed 'rites of integration', in which temporary relationships are managed in a ritualized social interaction designed to result in the customer leaving with a sense of well-being (Shiel et al., 1992). Research in this area is openly directed at increasing profits by learning how to establish a lasting relationship between service provider and customer (Storbacka et al., 1994; Zeithaml, 2000). For example, Stern, Thompson and Arnould (1998) use interviews and narrative analysis to discover how customers perceive the relationship they have with a service provider to discover what leads to brand loyalty. From customer narratives they discover that perceptions of sensitivity/insensitivity to needs or concerns rank highly in estimates of the success of the encounter.

In the marketing literature it is also recognized that there is a tension between *efficiency* and the *personalization* needed for the customer to feel that his or her needs are being taken into account. In addition to meeting basic needs, the supplier therefore develops additional strategies, such as: offering advice, making small talk, or taking a personal interest in the customer (Suprenant and Solomon, 1987). The response of the customer is interpreted in terms of *service quality*, dictating whether or not the customer will use these services again. Service quality has become so important that discourse analysis tools have been widely used in an attempt to define this illusive construct.

The first attempt to produce a model of service quality was that of Parasuraman et al. (1988), but it has since been expanded as researchers have become more interested in the role of interaction. For example, Chandon et al. (1997) argue that ceremonials and rites are important in the scripts of encounters, and that short interactions create ties that are linked to a sense of satisfaction. They therefore include 'interactivity' and 'ritual' in part of their definition, which are aspects of the structure of the encounter and its discourse. Chandon et al. (1997) present a number of dimensions relative to successful interactions:

Traditional categories

- *Effectiveness* concerns the evaluation of the aim and the result of encounter.
- *Materiality* includes the sub-dimensions of service employee appearance, equipment and physical facilities of the agency.
- *Accessibility* refers to the ease of access and contact.
- *Agent satisfaction* measures the employee's professional satisfaction with the encounter.

Discourse related categories

- *Interactivity* encompasses the service relations at work during the encounter. It includes the following six secondary dimensions: responsiveness; listening; ability to explain; understanding; personalization; and psychological proximity.
- *Rituality* includes all the ceremonial and contextual aspects which shape the 'climate' of encounter: courtesy of each individual; confidence; security; attitudes of receptionists; waiting time; and perceived competence of the contact personnel.

It is clear that many of the categories investigated in the marketing literature, through the use of customer narratives, focus groups, questionnaires/interviews and discourse analysis, relate to customers' *perceptions* of the quality of the service and the degree to which their needs are being taken into account. It also seems that at the heart of this is establishing a relationship in very much the same way that has been described by conversation analysts. However, in the marketing literature this is often described as establishing 'rapport' (Gremler and Gwinner, 2000) in the same way that Fairclough (1995) uses the term: to engage in personal interaction with the intention of positively influencing the customer. However, the description of the construct has varied across studies. In a meta-analysis of marketing studies, Gremler and Gwinner (2000, pp. 84–89) argue that rapport can be defined in terms of two factors: *enjoyable interaction* and *personal connection* (2000, p. 92).

We may therefore add to the indigenous criteria for assessment:

- *Rapport* 'is a customer's perception of having an enjoyable interaction with a service provider employee, characterized by a personal connection between the two interactants' (Gremler and Gwinner, 2000, p. 92) and involves feelings of the following: warmth; a harmonious relationship; humour; comfort in the interaction; bonding; care; personal interest; closeness; and similarity.

A particularly clear example of how rapport is achieved by the participants comes toward the end of an interaction from our data, presented in example 4.

Achieving rapport (original data)

1	TA:	=there's (w-) definitely have a couple of days shopping] because (.) er sightseeing because it's lovely. (.) it's really [nice.]
2	C:	[and shopping.
3	TA:	shopping yeah. ((laughter)) well that's a must isn't it. I mean we are female. ((laughter)) that's just got to ↑be. ((laughter)) yeah it's a great city.
4	C:	[great.]
5	TA:	[really] really [nice.]

The explicit reference to being the same gender in exchange 3, and sharing a common interest in shopping, draws the customer and agent into a common bond in which they can share the humour of a stereotype. This creates a harmonious relationship.

Finally, marketing researchers have also taken an interest in non-verbal communication. Gabbott and Hogg (2000, p. 385) treat a service encounter as action which 'takes place in a theatre (servicescape) and the performance requires actors, audience, script, setting, rehearsal, appearances, and importantly, authenticity.' Seen as part of 'dramaturgy', they used a quasi-experimental method using an actress in a hotel reception to show that body language, including facial expression, eye contact, posture, and gesture have a profound impact upon customer service evaluation. They argue that smiling (especially when listening) and fidgeting as little as possible contribute to enhancing rapport with the customer, as does 'attractiveness', which involves similarity, familiarity and liking for the person, as well as 'physical pleasantness'. They also discovered that periods of extended silence resulted in a more negative evaluation of the service encounter. Similar findings have been made in investigations into rater perceptions of the quality of interaction in language testing, especially with regard to non-verbal behaviour and eye contact (May, 2007). We may therefore extend the definition of maintaining rapport to include the following:

- maintaining regular eye contact especially when listening;
- smiling regularly;
- avoiding fidgeting;
- giving the customer a comfortable personal space;
- appearing well dressed and groomed;
- filling extended silences.

Embedded relational sequences initiated by the customer tend to occur at places where otherwise there may be extended silences, particularly when travel agents engage in computer searches or filling details into an online form. However, travel agents tend to describe what they are doing for the customer when they are engaged in computer related tasks if the customer does not initiate an embedded sequence. This is illustrated in the following example 5 from our data.

Filling extended silences (original data)

1	C:	um flights for next June to Barcelona =
2	TA:	to [Barcelona]
3	C:	=[I'm interest]ed in. um the dates (.) a::re (.) the eighth to the sixteenth or seventeenth.
4	TA:	°Okay°
5	C:	erm
6	TA:	°(I'll just have a look on the) system.°
7	C:	it's for a conference and (.) u::m (.) those are the (3.0) those are the dates of it.
8	TA:	°yeah.°
9	C:	um. (.) adding on the (.) sort of weekend for a bit of sightseeing as well.
10	TA:	°(okay)° ((Begins typing)) °just trying (to get onto the) system.° how many people is it for.
11	C:	two
	(6.5)	((TA trying to access information on computer))
12	TA:	°can't get on at the moment I'm afraid° (.hhh)
	(12.0)	((TA looking at screen, waiting for page to upload))
13	C:	you haven't been open here very long have you.
14	TA:	no just over a month now. (.) where would you like to fly from?
15	C:	um (.) somewhere local.
16	TA:	so Birmingham if we can.
17	C:	Birmingham or (.) I don't know [if there] are flights from East=
18	TA:	[East Mids]
19	C:	=↑Midlands
20	TA:	yeah usually I'll just see who does it. (1.5) er:: they don't do Barcelona A-Airline. °I'm sure B-Airline do it ( )° (2.0)
21	C:	Barcelona only has one airport, [doe]sn't it.
22	TA	[yeh].
	(4.5)	((TA working on computer))
23	TA:	this takes a little whi- ooh sorry (got) straight on today. so what t- (hhh) what date did you want to go out on?

In exchanges 7 to 9 the customer initiates a relational embedded sequence immediately after she has been informed that the travel agent is going to use the computer search system. In exchange 10 the agent returns to transactional language to ask how many people the booking is for, but then in exchange 12 indicates that there will be a delay, as

she can't get onto the system. As a reaction to the extended period of silence, the customer initiates another relational sequence in exchange 13, this time relating to 'local change' (Coupland and Ylänne-McEwen, 2000, p. 166); but once again the travel agent returns to transactional language after one response in exchange 14. The embedding comes to an end in exchange 23 when the travel agent indicates that she has got into the system and she asks for the information provided in exchange 3 again. In exchanges 6, 10, 12 and 22–23 the travel agent cannot engage further in relational language because she has to concentrate on getting into the computer system, and so gives a commentary on what she is doing with the computer to the customer. These behaviours are the observable linguistic components of silence avoidance that lead to a perception of care and personalization.

A scoring model for service encounters: The performance decision tree

We are now in a position to outline the content and nature of a scoring method for complex service encounters which, being directly linked to performance and task, would form a critical part of the architecture of a complete service encounter test (Fulcher and Davidson, 2009). The model needs to account for intended or expected test outcome in an effect-driven manner as well as for the following performances, which mark this type of encounter as a specific interactional genre:

Interactional Competence in a Service Encounter

A. Discourse Competence
    1. Realization of service encounter discourse structure
    2. The use of relational side-sequencing
B. Competence in Discourse Management
    3. Use of transition boundary markers
    4. Explicit expressions of purpose
    5. Identification of participant roles
    6. Management of closings
    7. Use of backchannelling
C. Pragmatic Competence
    8. Interactivity/rapport building
    9. Affective factors, rituality
    10. Non-verbal communication

These elements can be transformed into the Performance Decision Tree (PDT) presented in Figure 1. This is a prototype scoring system for tasks, as well as providing a clearer picture of what competencies and skills are required to successfully engage in complex service encounters. It is intended to be used with a paired task similar to that devised by Mills (2009), as discussed earlier in the subsection 'Discourse competence in service encounters'.

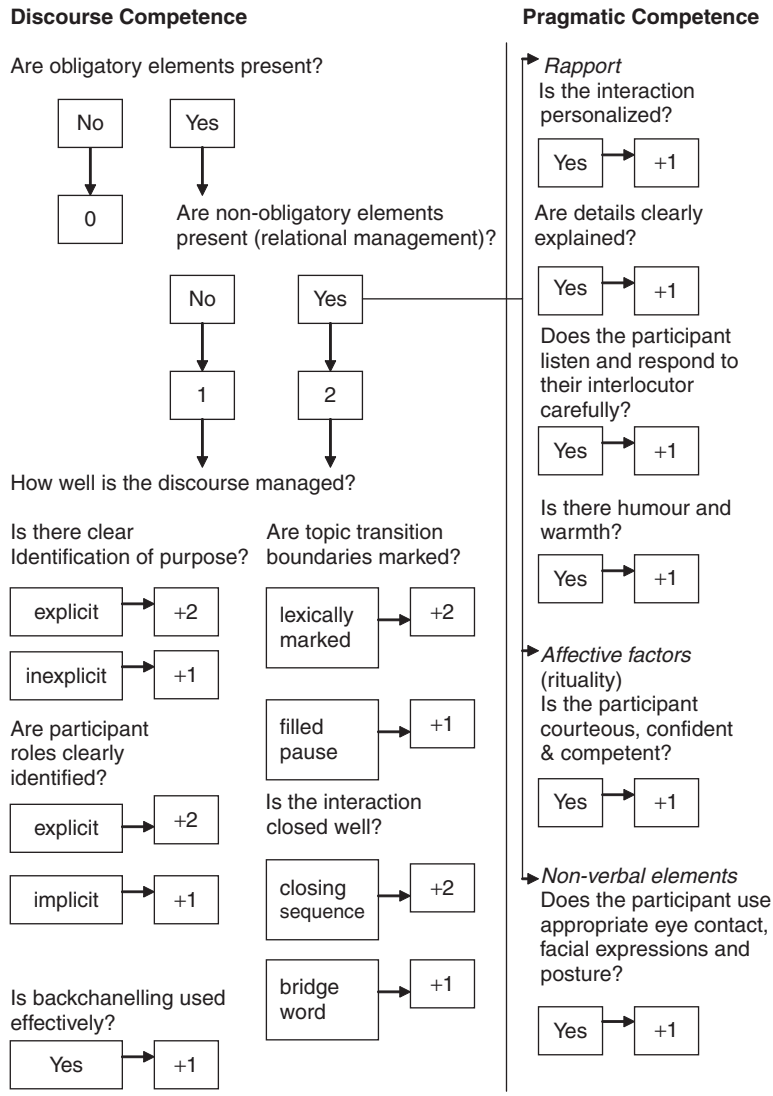

Figure 1. The Performance Decision Tree: Constructs and variables for a test of service encounters

Context, interaction and information

It should be noted that some of the elements may be more or less relevant to a particular role in the interaction. While the elements on the left side of the PDT could be used to score the performances of both test takers, the right side of the PDT is more appropriate

for the role of the service provider. In fact, it is hard to see how a single rating instrument could be used to evaluate both roles either together, or independently. Mills (2009) for example, has shown that there is a difference in the discourse produced by the learner playing the role of the travel agent and the discourse of the learner playing the role of the customer. The travel agent tends to speak more, ask more questions, and has to use a wider range of politeness strategies. Only the customer is allowed to interrupt or use other discourse features associated with a higher level of power. It may be that an 'interaction score' for both candidates is therefore not appropriate, but a separate score for each participant is needed to reflect how they contribute to the co-construction.

This also raises the question of whether learners need to be given the opportunity to perform in both roles in tests where the social status and power of the two interactants is not equal. For many years it has been commonplace to argue that the kinds of interactions in tests do not reflect the 'realities' of normal conversation (van Lier, 1989). The resulting literature on speaking tests has held up casual conversation as the gold standard of spoken communication to be emulated in tests. However, with the new focus on interactional competence, we begin to see that most communication is not between two participants with completely equal rights and roles, particularly in business, educational and service encounter settings.

One solution would therefore be to have both learners take on both roles, to score the service provider on the entire scale, and the customer only on discourse competence. However, such decisions would depend upon the purpose of the test. In a test for learners who are going to be entering the travel industry it would clearly be preferable for all learners to be scored on the entire scale in the role of the service provider.

In the role of the service provider it would in theory be possible for a test taker to get a score between 0 and 20, depending on how completely and well they realized their discourse and pragmatic competence in the interaction. In this sense a scale is produced, but not of the traditional linear type. Specifically, there is no implication that ability is uniform across descriptors, or that a particular score is arrived in a uniform manner. The PDT brings together the description of performance-data based rating scales and the EBB methodology in a system that offers rich description behind the scale, but provides raters with a much simpler set of binary decisions that may be much easier to use in live rating.

We would also argue that the PDT allows the creation of a diagnostic profile that would define the meaning of each score. Any particular score can be constructed in a number of different ways. For example, a score of 7 could be arrived at in (at least) the following two ways:

> 7 (obligatory elements present; inexplicit identification of purpose; explicit identification of role; interaction closed using a bridge word; personalized interaction; careful listening and response to interlocutor).
>
> 7 (obligatory elements present; clear identification of purpose; lexically marked boundaries; interaction closed using a bridge word; non-verbal communication).

Feedback to participants may then include advice on what aspects of their participation would need to be improved for a higher score to be awarded. Alternatively, in the context

of a tourism language class, a teacher could focus instruction on (for example) the use of closing sequences in interaction in order to bring conversations to a more natural close, or the use of side-sequences to personalize interaction. The PDT may therefore bring us a step closer to integrating the outcomes of classroom assessment into more targeted instruction.

Conclusions

With reference to the research literature and original data, in this article we have argued that the definition of service encounters in measurement-driven scales does not provide an adequate basis for test design or learner assessment because the descriptors cannot be related to context, performance conditions, quality of performance, or interactional competence. The descriptive content lacks the richness of actual performance. The driving force behind measurement-driven scales is a psychometric model rather than an understanding of human communication and language use.

PDTs, on the other hand, grow out of a careful analysis of the discourse of the type of encounter to which we wish score meaning to generalize. To this extent it involves 'thick description' from a data-based approach while incorporating elements of the empirically derived binary-choice boundary definition scales, as operational usage involves a series of binary yes/no decisions. Some rating events will move very quickly through the system, and others will take longer. Unlike the static 'levels' of traditional scales, this scale is flexible, because it is related directly to a context of language use. It makes no assumption about the linear and hierarchical relationship of descriptors, and provides observable behaviours as counterparts of constructs.

A major part of a validity claim for a PDT would rest upon the comprehensiveness of the description upon which it was generated, and the relevance of the assessment categories to current theories of 'successful interaction' within a particular context. As such, other PDTs must be developed through a careful analysis of communication in context, and a theoretical description of the constructs that underlie successful interaction, in order to generate context sensitive assessment categories.

Further research will be needed. In particular, although we have hypothesised that raters will find the PDT easier to use than complex performance data-based rating scales and traditional rating scales, this requires empirical investigation. Prototyping the PDT presented in this paper is a next step in the research. We also acknowledge that this PDT has been built primarily on data from native speaker interactions, on the assumption that these represent competent exchanges. We recognize, however, that native speaker models may not be the most appropriate, and studies of service encounters between non-native speakers may provide more appropriate PDTs for other contexts.

We have argued that an analysis of how people use language in actual communicative contexts can form the basis for more dynamic and contextually sensitive approaches to rating that help to define the nature of interactional competence in context. Performance Decision Trees are more flexible and do not assume a linear, unidimensional, reified view of how second language learners communicate. They are also *pragamatic*, focusing as they do upon observable action and performance, while attempting to relate actual performance to communicative competence.

Acknowledgements

We are indebted to the three anonymous reviewers of this paper. Their insightful comments and advice on the first draft was invaluable in helping us to introduce clarity and focus. However, any remaining muddle remains our responsibility alone. We are also grateful to Samantha Mills, a student of the first author of this paper, for permission to reproduce data and a task from her dissertation.

References

Alderson JC (1991). Bands and scores. In Alderson JC, North B (Eds.), *Language testing in the 1990s*. London: Modern English Publications and the British Council.

Bachman LF (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Baker R (1997). *Classical test theory and item response theory in test analysis*. Lancaster: Department of Linguistics and Modern English Language.

Bastos LC (1996). Power, solidarity and the construction of requests in service encounters. *ESPecialist, 17*(2), 151–174.

Biber D, Conrad S, Reppen R, Byrd P, and Helt M (2002). Speaking and writing in the university. A multidimensional comparison. *TESOL Quarterly, 36*(1), 9–48.

Brooks L (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing, 26*(3), 341–366.

Canale M (1983a). From communicative competence to communicative language pedagogy. In Richards C, Schmidt RW (Eds.) *Language and communication* (pp. 2–27). London: Longman.

Canale M (1983b). On some dimensions of language proficiency. In Oller JW (Ed.), *Issues in language testing research* (pp. 333–342). Rowley, MA: Newbury House.

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

Chalhoub-Deville M (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*(4), 369–383.

Chandon JL, Piere-Yves L, and Philippe J (1997). Service encounter dimensions – a dyadic perspective: Measuring the dimensions of service encounters as perceived by customers and personnel. *International Journal of Service Industry Management, 8*(1), 65–86.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching and assessment*. Cambridge: Cambridge University Press. Available online at: www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf.

Coupland N (1983). Patterns of encounter management: Further arguments for discourse variables. *Language in Society, 12*, 459–476.

Coupland N, Ylänne-McEwen V (2000). Talk about the weather: Small talk, leisure talk, and the travel industry. In Coupland J (Ed.), *Small talk* (pp. 163–182). London: Longman.

Davidson F, Fulcher G (2007). The Common European Framework of Reference (CEFR) and the design of language tests: A matter of effect. *Language Teaching, 40*(3), 231–241.

Fairclough N (1995). *Critical discourse analysis: The critical study of language*. London: Longman.

Fulcher G (1987). Tests of oral performance: The need for data-based criteria. *English Language Teaching Journal, 41*(4), 287–291.

Fulcher G (1993). The construction and validation of rating scales for oral tests in English as a Foreign Language. Unpublished PhD thesis, University of Lancaster, UK.

Fulcher G (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, *13*(2), 208–238.

Fulcher G (2003). *Testing second language speaking*. London: Longman/Pearson.

Fulcher G (2008). Criteria for evaluating language quality. In Shohamy E (Ed.), *Language testing and assessment*. Encyclopedia of Language and Education, Vol. 7 (pp. 157–176). Amsterdam: Springer.

Fulcher G, Davidson F (2007). *Language testing and assessment*. London and New York: Routledge.

Fulcher G, Davidson F (2009). Test architecture, test retrofit. *Language Testing, 26*(1), 123–144.

Gabbott M, Hogg G (2000). An empirical investigation of the impact of non-verbal communication on service evaluation. *European Journal of Marketing*, *34*(3/4), 384–398.

Gremler DD, Gwinner KP (2000). Customer-employee rapport in service relationships. *Journal of Service Research*, *3*(1), 82–104.

Grönroos C (1993). Toward a third phase in service quality research: Challenges and future directions. In Swartz AT, Bowen DE, and Brown SW (Eds.), *Advances in service marketing management*, Vol. II (pp. 49–64). Greenwich CT: JAI Press.

Halliday MAK (1985). Context of situation. In Halliday MAK, Hasan R, *Language, context, and text: Aspects of language in a social-semiotic perspective* (pp. 3–14). Victoria, Australia: Deakin University Press.

Hamp-Lyons L (1991). Scoring procedures for ESL contexts. In Hamp-Lyons, L. (Ed.), *Assessing second language writing in academic contexts* (241–276). Norwood, NJ: Ablex.

Hasan R (1985). The structure of a text. In Halliday MAK, Hasan R (1985). *Language, context, and text: Aspects of language in a social-semiotic perspective* (pp. 52–69). Victoria, Australia: Deakin University Press.

Hulstijn JA (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal, 91*(4), 663–667.

Jacoby S, McNamara, T. (1999). Locating competence. *English for Specific Purposes, 18*(3), 213–241.

Jacoby S, Ochs E (1995). Co-construction: An introduction. *Research on Language and Social Interaction*, *28*(3), 171–183.

Keaveney SM (1995). Customer switching behavior in service industries: An exploratory study. *Journal of Marketing, 59*(2), 71–82.

Kramsch C (1986). From language proficiency to interactional competence. *The Modern Language Journal, 70*(4), 366–372.

Lowe P (1986). Proficiency: Panacea, framework, process? A reply to Kramsch, Schulz, and particularly Bachman and Savignon. *Modern Language Journal*, *70*(4), 391–397.

Lumley T, Lynch B, and McNamara T (1994). A new approach to standard-setting in language assessment. *Melbourne Papers in Language Testing, 3*(2), 19–40.

May LA (2007). Interaction in a paired speaking test: The rater's perspective. Unpublished PhD thesis, School of Languages and Linguistics, University of Melbourne, Australia.

McCarthy M (2000). Captive audiences: The discourse of close contact service encounters. In Coupland J (Ed.), *Small talk* (pp. 84–109). London: Longman.

McCarthy M (2003). Talking back: 'Small' interactional response tokens in everyday conversation. *Research on Language and Social Interaction, 36*(1), 33–63.

McCarthy M, Carter R (1994). *Language as discourse: Perspectives for language teaching*. London: Longman.

Meisel JM (1980). Linguistic simplification. In Felix S (Ed.), *Second language development: Trends and issues* (pp. 13–40). Tübingen: Gunter Narr.

Mills S (2009). An evaluation of a paired format oral test for Korean learners of English for Hotel and Tourism. University of Leicester: Unpublished MA dissertation.

Mitchell TF (1957/1978). The language of buying and selling in Cyrenaica: A situational statement. *Hespėris* XLIV: 31–71. Reprinted in TF Mitchell (1978). *Principles of Firthian linguistics* (pp. 167–200). London: Longman.

North B (1993). *Scales of language proficiency: A survey of some existing systems*. Strasbourg: Council of Europe, Council for Cultural Co-operation, CC-LAND 94 24.

North B (1995). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System, 23*(4), 445–465.

North B (1996). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. In Huhta A, Kohonen V, Kurki-Suonio L, and Luoma S (Eds.), *Current developments and alternatives in language assessment*. Proceedings of the LTRC 1996 (pp. 423–447). Jyvaskyla: University of Jyvaskyla Press.

North B (2000). *The development of a common framework scale of language proficiency*. New York, Peter Lang.

North B, Schneider G (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*(2), 217–263.

Parasuraman A, Zeithaml AV, and Berry LL (1988). SERVQUAL: A multiple-item scale for measuring customer perceptions of service quality, *Journal of Retailing*, *64*, 12–40.

Ryoo H-K (2005). Achieving friendly interactions: a study of service encounters between Korean shopkeepers and African-American customers. *Discourse and Society, 16*(1), 79–105.

Shiel C, Bowen DE, and Pearson CM (1992). Service encounters as rites of integration: An information processing model. *Organizational Science, 3*(4), 537–555.

Spencer WJ (1997). 'We don't pay for bus tickets, but we can help you find work': The micropolitics of trouble in human service encounters. *The Sociological Quarterly*, *38*(1), 185–203.

Stern BB, Thompson CJ, and Arnould EJ (1998). Narrative analysis of a marketing relationship: The consumer's perspective. *Psychology and Marketing, 15*(3), 195–214.

Storbacka K, Standvik T, and Grönroos, C. (1994). Managing customer relationships for profit: The dynamics of relationship quality. *International Journal of Service Industry Management, 5*(5), 21–38.

Suprenant CF, Solomon MR (1987). Predictability and personalization in the service encounter. *Journal of Marketing, 51*(2), 86–96.

Swain M (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing*, *18*(3), 275–302.

Turner CE (2000). Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *Canadian Modern Language Review, 56*(4), 555–584.

Turner CE, Upshur J (2002). Rating scales derived from student samples: Effects of the scale marker and the student sample on scale content and student scores. *TESOL Quarterly, 36*(1), 49–70.

Upshur J, Turner CE (1995). Constructing rating scales for second language tests. *English Language Teaching Journal*, *49*(1), 3–12.

Upshur J, Turner CE (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse.' *Language Testing, 16*(1), 82–111.

van Lier L (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral Proficiency Interviews as conversation. *TESOL Quarterly, 23*(3), 489–508.

Ventola E (1987). *The structure of social interaction*. London: Francis Pinter.

Ventola E (2005). Revisiting service encounter genre: Some reflections. *Folia Linguistica, 39*(1–2), 19–43.

Westhoff G (2007). Challenges and opportunities of the CEFR for reimagining foreign language pedagogy. *The Modern Language Journal, 91*(4), 676–679.

Wilds C (1975). The oral interview test. In Jones RL, Spolsky B (Eds.), *Testing language proficiency* (pp. 29–44). Arlington, VA: Center for Applied Linguistics.

Ylänne-McEwen V (2004). Shifting alignment and negotiating sociality in travel agency discourse. *Discourse Studies, 6*(4), 517–536.

Zeithaml VA (2000). Service quality, profitability, and the economic worth of customers: What we know and what we need to learn. *Journal of the Academy of Marketing Science, 28*(1), 67–85.

Appendix 1: A travel agency simulation (Mills, 2009)

Sample Item

One student is the travel agent and one is the customer. The customer wants to know the price and availability of seats.

Rubrics

The customer will receive a role card stating the destination, preferred travel day, and class of ticket. The travel agent will receive one of the following flight information tables.

Customer

City	Travel date	Round trip/ one way	Class	Companions
e.g. Osaka	Thursday	round trip	business	No

Travel Agent

Flight Days	Direct or stopover	Departure time	Arrival Time	Price
e.g. Monday Wednesday Friday	Direct Stopover	11:20am	6:40pm	$750

		Starlight Airlines Fares and Flight Schedules					
Destination		Schedule	Stopover	Departs	Arrives	Price (Economy/ Business/First Class)	
						One way	Return
South Korea	Incheon	Monday	no	9:20 AM	4:30 PM	$477 / $753 / $1240	$877 / $1453 / $2240
	Incheon	Wednesday	Beijing	9:20 AM	6:30 PM	$427 / $723 / $1140	$827 / $1383 / $2140
	Incheon	Friday	no	9:20 AM	4:30 PM	$477 / $753 / $1240	$877 / $1453 / $2240
Japan	Osaka	Tuesday	no	10:15 AM	5:40 PM	$476 / $752 / $1100	$776 / $1352 / $2100
	Osaka	Thursday	no	10:15 AM	5:40 PM	$476 / $752 / $1100	$776 / $1352 / $2100
	Osaka	Saturday	Tokyo	10:15 AM	7:40 PM	$426 / $722 / $1100	$726 / $1252 / $2000

		Candidian Airlines Fares and Flight Schedules					
Destination		Schedule	Stopover	Departs	Arrives	Price (Economy/ Business/ First Class)	
						One way	Return
South Korea	Incheon	Tuesday	no	7:50 AM	3:00 PM	$480 / $800 / $1400	$880 / $1500 / $2700
	Incheon	Wednesday	Manila	7:50 AM	4:10 PM	$460 / $780 / $1300	$860 / $1450 / $2600
	Incheon	Friday	no	7:50 AM	3:00 PM	$480 / $800 / $1400	$880 / $1500 / $2700
Japan	Osaka	Monday	no	9:40 AM	4:10 PM	$515 / $780 / $1540	$915 / $1480 / $2540
	Osaka	Thursday	Hong Kong	9:40 AM	5:30 PM	$505 / $780 / $1540	$905 / $1420 / $2500
	Osaka	Friday	no	9:40 AM	4:10 PM	$515 / $780 / $1540	$915 / $1480 / $2540