Article details

Batenburg E.S.L. van, Oostdam, R.J., Gelderen, A.J.S. van & Jong N.H. de (2018), Measuring L2 speakers' interactional ability using interactive speech tasks, *Language Testing* 35(1): 75-100. Doi: 10.1177/0265532216679452

Article

LANGUAGE TESTING

Measuring L2 speakers' interactional ability using interactive speech tasks

Language Testing
2018, Vol. 35(1) 75–100

DOI: 10.1177/0265532216679452
journals.sagepub.com/home/ltj
SAGE

Eline S. L. van Batenburg
Amsterdam University of Applied Sciences, Netherlands

Ron J. Oostdam
Amsterdam University of Applied Sciences and University of Amsterdam, Netherlands

Amos J. S. van Gelderen
University of Amsterdam and Rotterdam University of Applied Sciences, Netherlands

Nivja H. de Jong
Utrecht University, Netherlands

Abstract

This article explores ways to assess interactional performance, and reports on the use of a test format that standardizes the interlocutor's linguistic and interactional contributions to the exchange. It describes the construction and administration of six scripted speech tasks (instruction, advice, and sales tasks) with pre-vocational learners ($n = 34$), and reports on the extent to which these tasks can be used to assess L2 speakers' interactional performance in a reliable and valid manner.

The high levels of agreement found between three independent raters on both holistic and analytical measurements of interactional performance indicate that this construct can be measured reliably with these tasks. Means and standard deviations demonstrate that tasks differentiate between speakers' interactional performance. Holistic ratings of linguistic accuracy and interactional ability correlate highly between tasks that focus on different language functions, and are situated in different interactional domains. Furthermore, positive correlations are found between both holistic and analytic ratings of oral performance and vocabulary size. Positive within-task correlations between analytical ratings of specific interactional strategies and holistic ratings of overall interactional ability show that analytic ratings of *meaning negotiation* and *correcting*

Corresponding author:
Eline S. L. van Batenburg, Amsterdam University of Applied Sciences, Centre for Applied Research in Education (CARE), Wibautstraat 2–4, 1091 GM, Amsterdam, The Netherlands.
Email: e.van.batenburg@hva.nl

misinterpretation provide additional information about speakers' interactional ability that is not captured by holistic assessment alone.

It is concluded that these tasks are a useful diagnostic tool for practitioners to support their learners' interactional abilities at a sub-skill level.

Keywords

Holistic and analytic measurement, interactional strategies, interactive speech tasks, pre-vocational education, reliability, validity

Theoretical background

Recent years have seen a growing interest in developing and assessing L2 candidates' ability to employ their language knowledge to achieve communicative goals in interaction. All major models of communicative language ability recognize that this ability hinges on the *knowledge* of language on the one hand and the ability to *use* this language in specific contexts on the other, mediated by some form of *strategic conduct* on the part of the user (Bachman, 1990; Bachman & Palmer, 1996; Canale, 1983a, 1983b; Celce-Murcia, Dörnyei, & Thurrell, 1995; and Celce-Murcia, 2007). According to Celce-Murcia et al. (1995), strategically competent speakers are able to resolve problems experienced in all areas of speech production, including in the interactional domain, that is–in the ability to convey and understand communicative intent by performing discourse functions–the ability to manage a conversation and to produce and interpret non-verbal communication. Meanwhile, the focus in testing oral proficiency has shifted from individual testing (e.g. the Oral Proficiency Interview (American Council on the Teaching of Foreign Languages [ACTFL], 2012) to more paired and group assessments (e.g. the Cambridge ESOL suite), which evoke more interactional and interaction management functions reflective of real-life communication than do individual tests (ffrench, 1999). This focus is further reflected in the inclusion of interactional ability in widely used assessment scales, such as the CEFR Interaction Scale (Council of Europe, 2001), the Cambridge ESOL Interactive Communication Scale (cf. Taylor, 2003), and the ACTFL Interpersonal Communication Strategies Scale (ACTFL, 2012).

Paired testing formats – where two candidates interact with each other – are especially helpful in testing conversational competence but complicate the assessment of individual interactional ability. Since interaction is reciprocal, one person's behaviour is contingent on the other, because much of the discourse is co-constructed (Kramsch, 1986). With co-constructed discourse, individual performance becomes vulnerable to interlocutor effects, which poses challenges to standardization in testing (cf. Weir, 2005). A vast body of research has reported on interlocutor effects due to factors such as acquaintanceship (e.g. O'Sullivan, 2008), gender (e.g. Brown & McNamara, 2004; O'Loughlin, 2000), native versus non-nativeness (e.g. Wigglesworth, 2000), proficiency level (e.g. Nakatsuhara, 2006; Watanabe & Swain, 2007) and speech accommodation (e.g. Ross & Berwick, 1992).

In paired formats, co-construction is inevitable. This raises two questions: is it possible (a) to disentangle individual contributions from paired exchanges, and (b) is it possible to arrive at an assessment of individual ability? The central point of contention is how the

construct of interactional performance is conceptualized. (For a historic overview of construct definitions, see Chalhoub-Deville & Deville, 2004.)

On the one hand, proponents of the interactionalist model reject the idea that interactional performance can be attributed to the (psycho-)linguistic or cognitive abilities of an individual test taker. Instead, performance is considered to arise solely from the co-constructed interaction between speakers and the specific interactional context in which they engage (cf. He & Young, 1998; Young, 2000, 2011). In this light, Galazci (2008) has demonstrated that, where conversational management is concerned, interactional performance is a joint venture from which individual contributions cannot be isolated. Dörnyei and Kormos (1998) and Kormos (2006), on the other hand, focus not on individual speakers' ability to manage a conversation, but on their ability to convey and understand communicative intent in interaction. In this, communicative success is largely dependent on the individual speakers' ability to employ linguistic resources on the one hand and, where these resources fall short, strategic resources on the other (e.g. compensation, meaning negotiation and time-gaining strategies). As such, interactional performance is not only considered to be part of the fundamental process of L2 speech production, but also as an individual trait (cf. De Bot, 1992; Poulisse, 1993).

In terms of testing, the interactionalist viewpoint that interactional performance is entirely co-constructed has led to suggestions for alternative assessment forms, for example, awarding pairs shared scores for interactional competence (May, 2009) or assessing the extent to which speakers achieve fluency *across* pairs (Ducasse & Brown, 2009). Thus far, though, reconciling paired testing with the need to obtain an assessment of individual ability has proven difficult, and in this, the usefulness of the interactionalist approach to address this issue has been questioned (cf. Chalhoub-Deville & Deville, 2004; Fulcher, 2010). However, if individual ability at least partially plays a role in achieving interactional success (i.e. if individual ability is part of the construct of interest), then there is a responsibility for language testers to explore ways in which this construct can be made measurable. Although it may not be possible to measure individual ability in all areas of interactional performance (such as conversational management), it should be possible to measure those parts that stem from individual traits, such as linguistic ability and the use of interactional strategies.

As outlined above, Dörnyei and Kormos (1998) and Kormos (2006) put forward a strong case for defining the use of interactional strategies as an individual trait. Much research has been done to uncover what strategies are employed by competent speakers. These can broadly be divided into *self-supporting strategies* and *other-supporting strategies*.

Self-supporting strategies are used to overcome problems in speech production and reception. These include compensation strategies, such as message reduction, message substitution (e.g. approximation, foreignizing), and message reconceptualization (e.g. circumlocution, word coinage) strategies: time-gaining strategies: and self-monitoring strategies (cf. Dörnyei & Scott, 1995; Færch & Kasper, 1983; Poulisse, 1993; Tarone, 1977). Self-supporting strategies also include meaning-negotiation strategies, such as checking and indicating understanding; uncertainty and incomprehension; and asking for elaboration, clarification and repetition of the message (Bygate, 1987; Dörnyei & Kormos, 1998; Dörnyei & Scott, 1995; Weir, 1993).

In interaction, speakers also must ensure mutual understanding between speech partners. This requires a set of other-supporting strategies, that is, an ability to align the message to the speech partner's need for information, topic knowledge and linguistic ability on the one hand and the ability to respond to clarification requests, indications of incomprehension and erroneous interpretations of the message on the other hand (cf. Bygate, 1987).

Chapelle (1998) posits that consistencies in language use arise from the interplay between personal traits and contextual factors. The same may be said for the employment of strategies. While the interactional context will determine which strategies are called for and which ones are not, differences in the performance of these may be considered to stem from speakers' individual ability.

The present study

Communicative success is achieved by L2 speakers who know how to employ self-supporting and other-supporting strategies that help them use their (often limited) linguistic resources effectively in interaction. However, no standardized tests currently exist that focus on assessing the use of these strategies during interactional encounters. While traditional oral proficiency tests provide a wealth of information about speakers' linguistic resources (e.g. vocabulary, grammar, pronunciation, and fluency), information about speakers' strategic ability does not come to the surface. For this reason, this study aims to explore ways in which interactional performance can be assessed in such a way that it provides practitioners with detailed diagnostic information about the areas of strategic competence that their learners need to develop in order to become more skilled in managing life-like interactional situations.

The purpose of this study was to design and construct interactive speech tasks that engage candidates in achieving real-life communicative goals in a simulated setting, evoke functional language use and directly evoke the use of (some of) the aforementioned interactional strategies in a standardized manner. As such, the study introduces a new test format specifically geared towards measuring L2 speakers' interactional ability. Since the main objective was to explore whether this new test format can be administered and rated reliably and contains sufficient potential for further development, the design was tested in a small-scale study using a variety of speech tasks.

In this study we used a test format in which one candidate's interactional performance is tested in interaction with an interlocutor and in which the interlocutor's contributions were controlled through the use of scripts. One advantage of such a scripted format is that it provides the opportunity to standardize the interactional context as well as the interlocutor's contributions. Since the context largely determines what type of linguistic and strategic performance is called for (cf. Chapelle, 1998), controlling the context increases the likelihood that individual speakers' performances can be compared, allowing us to make inferences about their linguistic *and* strategic ability to convey and understand messages in interaction.

Another advantage of this scripted test format is that interactional behaviour can be predicted a priori. Individual episodes in the test can be designed to evoke a specific interactional strategy (e.g. to respond to a clarification request). These episodes can then be rated analytically, thus providing a detailed measurement of individual speakers'

ability to employ the interactional strategies. As mentioned above, this approach differs from the one taken in many high-stakes tests, such as the International English Language Testing System (IELTS), where the focus is typically on rating linguistic categories like grammatical and lexical accuracy and appropriateness, and fluency and pronunciation. It also differs from tests in which criteria for interactional ability are in place (e.g. the Cambridge ESOL examinations and the Test of English Academic Proficiency (TEAP), but are used to rate a global impression of interactional performance.

Rating interactional behaviour during a scripted episode allows for a comparison between test-takers on the one hand, and provides diagnostic information at a detailed level on the other (cf. Taylor & Galazci, 2011). As such, this test format may provide practitioners with a useful diagnostic tool to further aid their learners' interactional abilities at a sub-skill level.

A characteristic of rating specific interactional strategies analytically is that these can be combined with holistic ratings of interactional ability. This combination may provide us with information about which interactional strategies are salient to interactional ability, which strategies are sufficiently captured by holistic assessment and which add information beyond this holistic impression.

Research questions

This article discusses the construction of six tasks situated in two different interactional domains (professional and personal) and centring on three different language functions (instruction, advice and persuasion). It subsequently explores whether interactional performance (i.e. both linguistic accuracy and interactional ability) can be assessed at an individual level by means of such speech tasks. We wished to answer the following research questions:

1. Can candidates' interactional performance on scripted speech tasks be evaluated in a reliable manner by different raters?
2. Can interactional performance be measured validly by the use of different scripted speech tasks?
3. To what extent do analytic ratings provide additional information to holistic ratings?

Method

Participants

Tasks were administered to all learners (aged 14–15) of two classes in their third year of a four-year pre-vocational Business & Administration programme from one secondary school in the Netherlands. Participants had received about three years of compulsory ESL (English as a second language) instruction. Interactional skills are not a set part of the standard ESL curriculum that they followed. Participation in these tasks was agreed upon with the school prior to the start of the academic year, and so was planned to be part of the class curriculum. As a result, learners' participation in the tasks was experienced

as "business as usual." Learners were informed beforehand that individual test performance would not be discussed with the class teacher, and would not affect their grades for English, but that their participation in the tests would be rewarded with an extra credit for the category "effort" on their report cards.

Tasks were administered individually to 34 participants (56% male, 44% female) by trained research assistants, who functioned as interlocutors. Participants were tested in three separate rounds on regular school days. Variation in school attendance and a fire alarm during the second test round caused some variation in sample size from task to task.

Materials

Six dialogic speech tasks designed for this study were used to measure participants' interactional performance. Since this study introduces a new test format, the design principles underlying these tasks, together with task construction, will now be discussed.

The tasks were designed for use with pre-vocational learners who are enrolled in an educational programme that prepares them for further vocational training and employment at middle-management level in the Business & Administration sector. Thus, in order to ensure context validity, a job analysis (McNamara, 1997) was carried out to establish the task types, task settings and professional interactional routines (Bygate, 1987) that are relevant to this sector. From this analysis, three main task types reflective of service encounters were distilled, that is, instruction tasks, advice tasks and sales tasks. For each task type, two dialogic tasks were developed (six in total) in which authentic interaction is simulated. The tasks within one task type — or task set — required the candidates to achieve the same goal (e.g. to explain a procedure) and tapped similar language functions, but differed in terms of content, audience and domain. Within a task set, one task was situated in the professional domain, with the candidate assuming the role of a hotel receptionist and the interlocutor the role of hotel guest, and one task was situated in the personal domain, with both candidate and interlocutor assuming the role of acquaintances (Table 1).

To reduce variation caused by differences in background knowledge that might influence task performance, candidates were provided with the required content knowledge for each task (cf. Bachman, 2002; Weir, 2005). As much as possible, content knowledge was presented pictorially where possible in order to minimize potential L1 interference when encoding messages, to reduce the chances of borrowing from L2 task input, and to increase chances that candidates formulate messages directly in the L2 (see Appendices I, II, and III for examples of each task type).

To meet cognitive validity demands, tasks must capture the extent to which candidates can convey and understand communicative intent, despite the linguistic limitations characteristic of L2 speech (cf. De Bot, 1992; Kormos, 2006) and while observing natural processing conditions, that is, reciprocity on the one hand and time constraints on the other. Since oral interaction hinges on both linguistic ability and improvisational ability (cf. Bygate, 1987), tasks were designed to evoke the use of functional language as well as the use of specific interactional strategies.

Brown (2003) points out that interlocutor *frames* do not control interlocutor contributions sufficiently to ensure standardization. For this reason, interlocutor *scripts* were

Table 1. Six speech tasks.

Task type	Task	Goal	Domain
Instruction	(1) Key Card	Explain to a customer how to open the door using a hotel key card.	Professional
	(2) Apple Cake	Explain to a family friend how to bake apple cake.	Personal
Advice	(3) Hotel Room	Advise a guest which hotel room to choose.	Professional
	(4) Cinema	Advise a family member which film to see.	Personal
Sales	(5) Board Games	Persuade a guest to buy a gift from the hotel gift shop.	Professional
	(6) Headphones	Persuade an acquaintance to buy your second-hand headphones.	Personal

used that fully prescribed the interlocutor's textual and interactional contribution, standardizing both linguistic (complexity, register, style) and interactional (set points requiring the use of interactional strategies) challenges posed to candidates. The scripts also specified the parameters of interactive support that could be offered.

To preserve the natural flow of interaction as much as possible, the scripts provided interlocutors with standardized alternatives in case of a more or less successful performance than expected. This enabled interlocutors to respond directly to candidate contributions. For example, in task 3 (see Appendix V), the interlocutor asks "*My nephew will come and visit me for the day. Will you have a cot for me?*" It is expected that candidates will not be familiar with the word "cot," which should evoke meaning negotiation. Several scenarios may unfold:

- In cases where candidates do know the word and provide the requested information, the interlocutor proceeds to the next question in the script.
- Where candidates negotiate for meaning, the interlocutor will provide the standardized alternative of "a small baby bed," and proceed to the next question once the requested information has been given.
- In cases where candidates do not engage in negotiation at all, the interlocutor continues to the next question.
- In cases where candidates do not engage in negotiation, but respond with "yes," the interlocutor asks: "In both rooms?" This allows candidates to negotiate for meaning.

In this way, the interlocutor responds directly to what the candidate says and at the same time is able to continue the interaction and administer the next standardized test item. Similarly, candidates may or may not ask for more information when asked for their advice. The script takes this into consideration by providing alternative routes. This preserves the standardization of interactional challenges that each candidate encounters, as well as controls interlocutor effects in such a way that is not possible in tests based on interlocutor frames.

Furthermore, candidates only received content information, and did not know beforehand how the interaction would unfold. The scripts thus created a one-sided information gap that balanced the need for standardization on the one hand and the need for interactivity on the other. Finally, interlocutors were trained to deliver the script as naturally as possible so that candidates could experience an authentic interaction.

Scripts are composed of set interlocutor contributions and include all the following: (1) opens encounter, (2) asks for an explanation, (3) asks for information, (4) asks for clarification, (5) provides an interpretative summary and (6) expresses gratitude. These interlocutor contributions ensured that each candidate was prompted to fulfil the same set of interactional functions (e.g. to greet, to inform, to clarify, to close the encounter) and that each candidate encountered identical interactional challenges that elicited the use of interactional strategies. As such, the following strategies were implemented and rated in each task:

Self-supporting strategies	a. Compensation strategies
	b. Meaning negotiation strategies
Other-supporting strategies	c. Response to clarification requests
	d. Response to misinterpretation of the message

These strategies were operationalized in a variety of ways. Self-supporting strategies were elicited by asking candidates to handle language beyond their current ability (cf. Dörnyei & Scott, 1995). For example, meaning negotiation was evoked by the interlocutor asking questions that contained low-frequency words or expressions likely to be unfamiliar to the candidates (e.g. "*My nephew will visit me for the day. Will you have a cot for me?*" [task 3] in order to elicit candidates' attempts to indicate incomprehension and requests for elaboration, clarification and repetition of the message [see Appendix V for the interlocutor script of task 3]). The use of compensation strategies was evoked by pushing candidates to convey concepts that were essential to task achievement and with which they were familiar, but which required the use of low-frequency vocabulary (e.g. the interlocutor asked, "*The Red Room looks nice. But how is this room kept warm?*" which necessitated an explanation of the concept of a *radiator* [Appendix V, episode 5]).

Many strategies commonly used to support a speech partner — such as checking common ground, checking comprehension and providing sufficient detail — are used proactively in communication and thus cannot be operationalized through the use of prompts. For this reason, design options were limited to evoking reactive strategies to achieve the same goal: responding to a clarification request and to a misinterpretation of the message. Following Weir's (1993) suggestion that the interlocutor feigning a lack of understanding may generate clarification behaviour, misinterpretation prompts and clarification prompts were used. In each task, the interlocutor provided an interpretative summary that contained a mistake (e.g. "*So if I want the Red Room, I need to pay extra for Wi-Fi?*" [Appendix V, episode 9]), and asked for clarification of a particular item (*e.g.* "*Sorry*, what *is used to heat up the room?*" [Appendix V, episode 6]). To ensure that a clarification request occurred naturally in each task, it was decided to link the clarification prompt to the episode in which the compensation strategy is evoked. The assumption was that asking for clarification of a low-frequency item already present in the test would lead to a more natural development of the sequence overall.

All tasks were piloted with 10 learners of a similar age and level. Some small adjustments were subsequently made, particularly when unforeseen conceptual problems were encountered. For example, it turned out that most learners had never heard of a *whisk* (task 2). Unfamiliarity with that object caused difficulties in offering an effective description of it. For this reason, *whisk* was replaced with the more familiar *wooden spoon*.

Procedure

Prior to participation in the speech tasks, participants' vocabulary size was measured using the Peabody Picture Vocabulary Test (Dunn & Dunn, 2007), adapted for use in an L2 setting. Task sets were implemented on three separate occasions at about eight-week intervals. On each occasion, two tasks were administered to all participants in the same order, that is, task 1 preceding task 2, and so on. For each task set, a research assistant prepared participants in groups of four. Following a protocol, the assistant familiarized the participants with the language use situation, the aim of the interaction and the criteria for task execution. About 10 minutes of planning time was used for content preparation for each task set. To prevent misunderstanding of content, the preparation was conducted in the shared public language, Dutch. To simulate natural processing conditions, where linguistic encoding of conceptualized messages takes place under time pressure (Kormos, 2006; Levelt, 1989, 1995), no planning time was given for language preparation, and questions pertaining to the use of English were not answered. Having completed the content preparation, participants were immediately escorted to separate rooms, where they carried out the tasks with another research assistant. In total, four assistants conducted tasks in parallel sessions situated in separate, closed classrooms. The tasks took about five minutes each. To reduce cognitive load, participants had access to the task input sheet during this time (see Appendices I, II, and III).

Rating

Tasks were recorded on both video and audio. These performances were rated by three undergraduate students (two L1 English speakers and one L2 English speaker) of an EFL (English as a foreign language) teacher-training programme aimed specifically at obtaining a teaching degree in (pre-) vocational education. Students on this programme carry out teaching placements in this educational sector, which ensured that they had a good understanding of the target population in this study. Raters were provided with a set of instructions that contained rating scales (see Appendix IV) and a rating sheet. This rating sheet followed the format of the interlocutor script, thus allowing raters to provide ratings per interactional episode. For each rating round, raters received training in which the rating scales were benchmarked. Benchmarking took place by rating videotaped example performances selected by the first author to illustrate various ability levels. Three videos were shown in random order, and raters were asked to rate each performance using the rating scales. This was done so that raters would become sensitized to differences in performance levels. Subsequently, raters were asked to rate another two to three randomly selected videos, and to share and explain their ratings. This led raters to formulate together a shared conception of what the different levels of performances sound like, and so reach a consensus on their interpretation of the scales. This consolidated the benchmarks.

After training, the raters rated all video performances independently. They were instructed to rate each performance, first on all analytic and then on all holistic categories for each participant. They were asked to stop or rewind the video if needed. On average, the rating process took five minutes per participant, per task, so that the total time for testing and subsequently rating one candidate was about 15 minutes per task. In principle, each candidate carried out six tasks over the course of this study, all of which were rated.

Raters were asked to provide holistic ratings on a Likert scale of 1–5 for (a) the extent to which participants expressed themselves in lexically and grammatically correct English (*Linguistic Accuracy*) and (b) the extent to which participants managed to overcome potential communication problems (*Interactional Ability*). The holistic rating scales were accompanied by brief performance descriptors (see Appendix IV). These descriptors are reminiscent of the CEFR descriptors (Council of Europe, 2001, pp. 37–38) in terms of their expectations for linguistic accuracy and interaction.

Raters were also asked to provide analytic ratings on the quality of the participants' responses during individual episodes in the interaction (i.e. the extent to which their contributions were considered to be adequate and appropriate). These contributions (*Compensation, Meaning Negotiation, Clarification*, and *Correcting Misinterpretation*), were rated on a Likert scale, ranging from 1 (very weak) to 5 (very strong). Since each episode was designed to evoke one of these strategies specifically, no additional descriptors were necessary to support the analytic ratings (see Appendix IV).

Method of analysis

The variables were examined for accuracy of data entry, distributions and missing values. Missing ratings stemming from interlocutors' failure to deliver a prompt were coded as missing in the data set. Inter-rater reliability was determined by calculating the intra-class correlation coefficients (ICCs) (two-way random model, absolute agreement) for each assessed category. This is appropriate to fully crossed designs, in which two or more raters assess all candidates' performances. Since the average score of individual raters' ratings is used to calculate correlations between tasks, average-measures ICCs are reported.

Definitive final scores of participants' performances in each category were entered by calculating, and subsequently summing, the mean scores obtained with all three raters, taking missing values into consideration.

Pearson correlations were calculated to examine the correlation between the speech tasks and vocabulary size. Pearson correlations were also used to examine the extent to which the different tasks provide similar measurements of candidates' linguistic accuracy and interactional ability, and to examine the correlation between holistic and analytic ratings within tasks. To this end, the definitive final scores (average ratings) for each category were used.

Results

Inter-rater reliability

Table 2 shows a high inter-rater agreement between raters' overall impression of participants' interactional performance (*Linguistic Accuracy* and *Interactional Ability*) and on

Table 2. Inter-rater reliability (ICC).

	Instruction tasks				Advice tasks				Sales tasks			
	N	Task 1 *Key Card*	*N*	Task 2 *Apple Cake*	*N*	Task 3 *Hotel Rooms*	*N*	Task 4 *Cinema*	*N*	Task 5 *Board Games*	*N*	Task 6 *Head phones*
Holistic categories												
Linguistic Accuracy	34	.89	34	.88	31	.84	28	.81	32	.76	32	.77
Interactional Ability	34	.87	34	.80	31	.81	28	.82	32	.83	32	.78
Analytic categories												
Compensation	34	.91	34	.94	31	.91	28	.92	31	.88	32	.92
Meaning Negotiation	34	.85	34	.85	30	.57	27	.87	32	.79	30	.76
Clarification	26	.94	19	.88	16	.89	19	.96	21	.83	16	.85
Correcting Misinterpretation	34	.95	34	.91	31	.94	28	.96	32	.70	32	.80

all episodes specifically designed to evoke interactional strategies (*Compensation, Clarification, Meaning Negotiation*, and *Correcting Misinterpretation*). With the exception of *Meaning Negotiation* in task 3, inter-rater reliability is well above .70. These results strongly suggest that, on the whole, candidates' (evoked) interactional performance can be evaluated reliably, using both holistic and analytic measures.

Means and standard deviations

Means and standard deviations of each category (Table 3) indicate that ratings are spread, and that, overall, ratings are spread in similar ways, both at a holistic and analytic level. A repeated-measures ANOVA test on *Interactional Ability* revealed significant differences in difficulty between tasks, F (1,22) = 264,077, p = <.001. Pairwise comparisons demonstrated that these differences occurred between task 6 and task 3 (mean difference .551, CI .128 to .974, p = .005) and between task 6 and task 4 (mean difference .493, CI .066 to .948, p = .015). This indicates that tasks are similar in difficulty, with the exception of task 6.

No floor or ceiling effects were found; that is, the mean scores were not placed entirely at the low end or high end of the scale for any of the categories. The descriptives showed that distributions were slightly positively skewed overall, indicating that participants' scores cluster somewhat around the low values.

Reliability of analytical categories

Measured across tasks, the Cronbach alpha coefficients for the analytical categories show that all three categories form internally consistent subscales, *Compensation* α = .901; *Meaning Negotiation*, α = .714; *Misinterpretation* α = .832. Item–rest correlations were moderate to large (cf. Cohen, 1988) for all tasks, except for *Meaning Negotiation* in task 2 (.218). Overall, the analytic categories are internally consistent and, as such, have the potential to be used as subscales in testing interactional ability at a detailed level.

Table 3. Means and standard deviations (min 1 – max 5).

	Instruction tasks				Advice tasks				Sales tasks			
	Task 1 *Key Card* (*N* = 34)		Task 2 *Apple Cake* (*N* = 34)		Task 3 *Hotel Rooms* (*N* = 31)		Task 4 *Cinema* (*N* = 28)		Task 5 *Board Games* (*N* = 32)		Task 6 *Headphones* (*N* = 32)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Holistic categories												
Linguistic Accuracy	2.71	1.05	2.62	1.09	2.94	.82	2.83	.90	2.80	.83	2.56	.97
Interactional Ability	3.00	1.06	2.67	1.05	3.08	.79	3.02	.95	2.92	.97	2.57	.98
Analytic categories[a]												
Compensation	2.82	1.14	2.80	1.25	1.52	1.05	2.63	1.12	1.95[c]	1.06	2.09	1.21
Meaning Negotiation	2.22	1.05	2.39	.98	[b]	[b]	2.50[d]	1.02	2.00	.87	2.06[e]	.79
Correcting Misinterpretation	2.35	1.52	2.29	1.43	2.51	1.42	2.70	1.39	2.28	.75	2.37	.80

[a]Means and standard deviations for *Clarification* are not presented in this table due to a substantial number of missing values on this category in the data set.
[b]Meaning Negotiation was not rated reliably in task 3.
[c]*n* = 27.
[d]n = 31.
[e]n = 30.

Table 4. Correlations between scores for linguistic accuracy across tasks (Pearson's *r*).

	Instruction tasks	Advice tasks		Sales tasks	
	Task 2 *Apple Cake*	Task 3 *Hotel Room*	Task 4 *Cinema*	Task 5 *Board games*	Task 6 *Headphones*
Task 1 *Key Card*	.90	.80	.82	.78	.80
Task 2 *Apple Cake*	–	.81	.85	.78	.83
Task 3 *Hotel Room*		–	.79	.86	.86
Task 4 *Cinema*			–	.78	.73
Task 5 *Board Games*				–	.80

All correlations are significant at $p < .01$ (one-tailed).

Table 5. Correlations between scores for interactional ability across tasks (Pearson's *r*).

	Instruction tasks	Advice tasks		Sales tasks	
	Task 2 *Apple Cake*	Task 3 *Hotel Room*	Task 4 *Cinema*	Task 5 *Board games*	Task 6 *Headphones*
Task 1 *Key Card*	.76	.72	.67	.67	.77
Task 2 *Apple Cake*	–	.72	.84	.66	.85
Task 3 *Hotel Room*		–	.67	.83	.79
Task 4 *Cinema*			–	.73	.78
Task 5 *Board Games*				–	.79

All correlations are significant at $p < .01$ (one-tailed).

Correlations between tasks

The positive and high correlations shown in Table 4 and Table 5 indicate that the six tasks produce similar ratings for candidate's *Linguistic Accuracy*, ranging from $r = .73$, $n = 24$, $p < .001$ between tasks 4 and 6 to $r = .90$, $n = 34$, $p < .001$ between tasks 1 and 2. Ratings for *Interactional Ability* also correlate positively, ranging from $r = .67$, $n = 26$, $p < .001$ between tasks 1 and 4 to $r = .85$, $n = 30$, $p < .001$ between tasks 2 and 6. Fisher's r-to-z transformations show that correlations between tasks are comparable on both measures, with the exception of tasks 1 and 2, which correlate significantly higher for *Linguistic Accuracy* than for *Interactional Ability,* $z = 1.88$, $p = .03$.

Correlations between vocabulary size and oral performance

Table 6 shows that correlations between vocabulary size and holistic measurements of oral performance are consistently positive, high and significant across tasks, ranging from $r = .58$, $n = 28$, $p = .001$ in task 4, to $r = .81$, $n = 34$, $p \leqslant .001$ in task 2. Fisher's r-to-z transformations confirm that these correlations are comparable on both linguistic and interactional measures. In tasks that centre around the same language function, correlations are highly similar. The correlation with *Linguistic Accuracy* in Instruction tasks,

Table 6. Correlations between vocabulary size and oral performance (Pearson's *r*).

	Instruction tasks		Advice tasks		Sales tasks	
	Task 1 *Key Card* (*N* = 34)	Task 2 *Apple Cake* (*N* = 34)	Task 3 *Hotel Rooms* (*N* = 31)	Task 4 *Cinema* (*N* = 28)	Task 5 *Board Games* (*N* = 32)	Task 6 *Headphones* (*N* = 32)
Holistic categories						
Linguistic Accuracy	.77*	.81*	.64*	.63*	.71*	.71*
Interactional Ability	.70*	.78*	.61*	.58*	.67*	.72*
Analytic categories						
Compensation	.82*	.84*	.45**	.52**	.44**	.63*
Meaning Negotiation	.35**	.69*	[b]	.44**	.29***	.23***
Correcting Misinterpretation	.41**	.65*	.34**	.44**	.43**	.39**

*Correlations are significant at $p < .01$ (one-tailed).
**Correlations are significant at $p < .05$ (one-tailed).
***The correlation is not statistically significant.
[b]Meaning Negotiation was not rated reliably in task 3.

Table 7. Correlation between holistic ratings of interactional ability and analytic ratings of interactional strategies within tasks (Pearson's *r*), corrected for attenuation.

	Instruction tasks		Advice tasks		Sales tasks	
	Task 1 *Key Card*	Task 2 *Apple Cake*	Task 3 *Hotel Rooms*	Task 4 *Cinema*	Task 5 *Board Games*	Task 6 *Headphones*
Compensation	.93	1.0	.45	1.0	.35[ns]	.96
Meaning Negotiation	.74	.97	[b]	.52	.67	.41[ns]
Correcting Misinterpretation	.70	.86	.54	.86	.80	.53

[b]*Meaning Negotiation* was not rated reliably in task 3.
[ns]The correlation is not statistically significant.

for example, lies around $r = .80$, in Advice tasks around $r = .60$, and in Sales tasks around $r = .70$. Furthermore, correlations at an analytic level are largely significant and consistently positive. Again, correlations in tasks that share the same language function are similar, with the exception of the Instruction tasks. Here, Fisher's r-to-z transformations reveal that the correlations are significantly higher in task 2 than in task 1 for *Meaning Negotiation,* $z = 1.9$, $p = .02$ and for *Correcting Misinterpretation,* $z = 1.34$, $p = .08$. Overall, these findings show that the speech tasks correlate with an external measure of language proficiency.

Correlations within tasks

Within-task correlations demonstrate that analytic ratings of specific interactional strategies all correlate positively and on the whole significantly, with holistic ratings of interactional ability (see Table 7). Corrected for attenuation, the *Compensation* category correlates fully with Interactional Ability in tasks 2 and 4, meaning that the analytic category *Compensation* does not provide additional information about speakers' interactional ability as measured holistically. A similar observation can be made about Compensation in tasks 1 ($r = .93$, $n = 34$, $p \leqslant .001$) and 6 ($r = .96$, $n = 30$, $p \leqslant .001$).

A different picture emerges, however, when looking at the categories of *Meaning Negotiation* and *Correcting Misinterpretation*. With the exception of *Meaning Negotiation* in task 2 ($r = .97$, $n = 34$, $p \leqslant .001$), the more moderate correlations and confidence intervals for these categories indicate that the analytic scores indeed provide additional information about speakers' interactional ability that is not captured by holistic assessment alone.

Discussion

The main objective of this study was to explore whether scripted speech tasks can be administered reliably, and whether they have sufficient potential for further development. Answers to the research questions are reported below, along with suggestions for future research.

Answers to research questions

RQ 1. Can candidates' interactional performance be evaluated in a reliable manner by different raters using scripted speech tasks?

The high ICC scores reported in this study suggest that, on the whole, candidates' interactional performance can be measured reliably with the use of scripted speech tasks. Raters showed high agreement on both their holistic judgements of participants' linguistic accuracy, interactional ability and their analytic judgements of participants' performance on turns that evoked the use of specific interactional strategies.

Means and standard deviations show that the scripted speech tasks differentiate between candidates, both at a holistic and analytic level. Tasks are similar in difficulty, with the exception of task 6. Here, the interlocutor takes on a less cooperative role than in the other tasks, and challenges the bargain that is presented to him. This "hard sale" context seems to appeal more strongly to the participants' interactional abilities. However, since the aim was not to create parallel tasks, nor to use these tasks to measure growth, this does not seem problematic. Across the board, no floor and ceiling effects were found, indicating that the tasks are designed at the right level to assess pre-vocational learners around CEFR A2 level.

Measured across tasks, the Cronbach alpha coefficients for the analytical categories show that *Compensation, Meaning Negotiation*, and *Correcting Misinterpretation* all form internally consistent subscales and, as such, have the potential to be used in testing interactional ability at a detailed level.

RQ 2. Can interactional performance be measured validly by the use of different scripted speech tasks?

The results found in this study suggest that interactional performance can be measured validly by the use of scripted speech tasks, and thus that test scores can be used to make inferences about pre-vocational learners' ability to convey and understand communicative intent in interactional settings. Our arguments (e.g. Kane, 2012) for this claim are threefold. First, tasks that claim to provide information about interactional performance must simulate natural processing conditions (reciprocity and time constraints), and tap both speakers' linguistic and improvisational ability (cf. Bygate, 1987) in a standardized, interactional setting. These demands were met in the tasks' design.

Secondly, to claim test validity, candidates' performance should be consistent across different tasks that all represent the same construct domain (cf. Messick, 1996). In this study, three different task types were designed that are representative of the Business & Administration sector (instruction, advice and sales tasks), and that were situated in two different interactional contexts (professional and personal). In this light, the positive and high correlations found between all six tasks for both *Linguistic Accuracy* and *Interactional Ability* may be considered as indications of test validity. Furthermore, these correlations are fairly stable across the six tasks, on both the holistic and the analytic scales. This seems to indicate that not only linguistic accuracy, but also interactional ability, can be measured validly as stand-alone aspects of L2 communication.

Finally, since vocabulary size is a strong predictor for proficiency in speaking (e.g. De Jong et al., 2012), task performance should correlate with vocabulary size. The positive and high correlations found between all six speech tasks and independent measures of vocabulary size thus seem to provide further validity evidence.

RQ 3. To what extent do analytic ratings provide additional information to holistic ratings?

The positive and generally significant within-task correlations obtained between the analytic ratings of specific interactional strategies and the holistic ratings of *Interactional Ability*, suggest that the interactional strategies operationalized in this study are part of the central construct: interactional ability.

Corrected correlations for *Compensation* are one or close to one in most tasks, suggesting that the ability to compensate largely determines perceptions of global interactional ability. It therefore seems realistic to conclude that testing *Compensation* analytically does not add extra information beyond holistic assessment. The moderate correlations and confidence intervals obtained in *Meaning Negotiation* and *Correcting Misinterpretation*, however, indicate that these analytic scores provide information about speakers' interactional ability that is not captured by holistic assessment alone.

The strength of the correlations varies per category and task, suggesting that task effects occur at this specific level. Further research is needed to establish whether the use of interactional strategies is mediated by task, in which case more observations are needed to come to a reliable assessment at an analytical level.

Overall, the results indicate that both linguistic accuracy and interactional ability can be measured as stand-alone aspects of L2 communication, and that there is added value in assessing specific interactional behaviour analytically to reveal strengths and weaknesses. As such, scripted speech tasks may provide practitioners with a tool that has diagnostic potential in order to gain an insight into speakers' linguistic and interactional abilities across different domains and language functions, to identify speakers who have strong self-supporting, but limited other-supporting skills and vice versa, or speakers who can produce linguistically challenging messages, but struggle to understand such messages, and vice versa.

Suggestions for future research

Interactional strategies

Previous studies (e.g. Dörnyei & Scott, 1995) have identified a plethora of interactional strategies that support L2 speakers' interactional ability, only four of which were selected to represent self- and other-supporting interaction in this study. The question arises whether these strategies are sufficiently representative for testing purposes. Furthermore, in this study, other-supporting behaviour was evoked in reaction to a prompt delivered by the interlocutor (e.g. the interlocutor feigning misunderstanding). As such, proactive interactional strategies, such as checking common ground between the speaker and speech partner (Bygate, 1987) remain untested. Future research could explore whether it is possible to operationalize more proactive strategies for the scripted format as well.

Sample size

Since this study was exploratory in nature, the design was tested with a rather small sample. This poses constraints with regard to data analysis. A larger sample size would allow for regression analysis, thus facilitating a more meaningful exploration of the relative salience of the various interactional strategies in achieving task success, and of the impact that task and task type may have on the use of interactional strategies.

Rating

In this study, holistic and analytic judgements were provided by the same rater within a short period of time, and raters were instructed to rate analytically first, and then holistically. Chances are that analytic ratings were summarized into holistic ratings, as a result of which a halo effect may have occurred.

This issue could be addressed by assigning either holistic or analytic rating to each of two independent raters, as is the case in the Cambridge ESOL Main Suite (ad hoc judgements) and in the Test of English Academic Proficiency (TEAP). While such a financial investment was not feasible in this small-scale study, it would provide a helpful measure to reduce the possibility of a halo effect from occurring in rating these scripted speech tasks.

Operationalization of the "Clarification" category

In the present design, clarification behaviour was evoked by the interlocutor asking for clarification of the same low-frequency item that the test taker had attempted to explain in the previous episode. The assumption was that using a troublesome lexical item as the trigger for clarification would lead to a more natural development of the sequence overall.

Analytic ratings for the *Clarification* category obtained high inter-rater reliability scores (ranging from .83 in task 5 to .96 in task 4). Furthermore, the correlations between *Clarification* and holistic ratings of *Interactional Ability* were positive and, on the whole, significant in all tasks. These correlations were comparable to within-task correlations for the other analytic categories, suggesting that *Clarification* is part of the *Interactional Ability* construct.

However, the amount of missing ratings for this category indicates that interlocutors struggled to deliver the clarification prompt, reportedly because asking for clarification when an item had been encoded successfully felt unnatural. While Weir's (1993) suggestion that the interlocutor feigning lack of understanding might generate clarification behaviour seems true, further research is needed to optimize ways in which such behaviour can be operationalized in scripted speech tasks.

Validity

Kane (2012) differentiates between providing validity evidence at the *developmental* stage of a new assessment form, aimed at justifying the proposed interpretations and uses of said assessment, and providing validity evidence at the *appraisal* stage, aimed at objective appraisal of the validity claims made in stage one. This study was set in the developmental stage and thus far has made a case for the use of scripted speech tasks by demonstrating that

(1) tasks centred on different language functions (instruction, advice and persuasion), and situated in different domains (professional and personal) all measured interactional performance in similar ways and that (2) performance on these tasks correlated with independent measures of vocabulary size. Within this exploratory study, however, it was not possible to compare candidate's performance on the scripted test format with performance on another validated test format for oral interaction. To satisfy the need for objective appraisal, future research could provide evidence of additional convergent validity.

Conclusion

It can be concluded that using these types of scripted speech tasks in an individual test format can help distinguish L2 speakers' individual contributions and can allow for the reliable measurement of interactional ability, both at a holistic and analytic level.

Overall, these scripted speech tasks differentiate between candidates, and do so in similar ways at the holistic and analytic level. Furthermore, analytic categories are internally consistent and have the potential to be used as subscales in testing interactional ability.

The six tasks measure the construct of *Linguistic Accuracy* and *Interactional Ability* in similar ways, regardless of the language function that the task focuses on (instruction, advice or persuasion), or the domain in which the task is situated (professional or personal). Between-task correlations furthermore suggest that, by using these tasks, both linguistic accuracy and interactional ability can be measured reliably as stand-alone aspects of L2 communication.

Within-task correlations between the analytic and holistic ratings of *Interactional Ability* indicate that analytic ratings of *Compensation* do not add substantial information to holistic assessment of interactional ability, but ratings for *Meaning Negotiation* and *Correcting Misinterpretation* do provide additional information about speakers' interactional ability. This suggests that it might be beneficial to assess specific interactional strategies analytically. However, since task effects seem to occur at this level of assessment, more observations may be needed to come to a reliable assessment of speakers' interactional ability at an analytical level. Further research will have to determine the number of observations needed to do so.

Scripted speech tasks that control both linguistic and interactional challenges can be used to isolate individual contributions to the exchange and can provide a detailed measurement of the individual speakers' interactional performance. As such, it is suggested that this format provides a reliable alternative to other validated formats, including OPI-style individual assessment, and paired assessment. Although paired assessment has the potential to evoke a wide array of interactional and interaction management functions (ffrench, 1999), it is vulnerable to interlocutor effects on the one hand (cf. Weir, 2005) and constraints posed by co-construction in discourse on the other (cf. Chalhoub-Deville & Deville, 2004). The speech tasks presented in this study are robust against the influence of both co-construction as well as interlocutor effects.

Provided that one wishes to assess learners' interactional performance at a global level, these tasks may be suitable for application in both educational and research settings. As with all individual tests, however, this test format is fairly time and labour intensive. Effective use of these tasks in either an educational or research setting requires robust testing conditions to be in place. Despite these constraints, this research adds a

new perspective to the discussion about suitable formats for assessing L2 speakers' interactional competence.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

American Council on the Teaching of Foreign Languages (2012). *ACTFL assessments*. Alexandria: ACTFL. Retrieved from www.languagetesting.com/wp-content/uploads/2014/12/ACTFL%20Assessments%20Brochure.pdf

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

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*(4), 453–476.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, UK: Oxford University Press.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, *20*(1), 1–25.

Brown, A., & McNamara, T. F. (2004). "The devil is in the detail": Researching gender issues in language assessment. *TESOL Quarterly*, *38*(3), 524–538.

Bygate, M. (1987). *Speaking*. Oxford, UK: Oxford University Press.

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

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

Celce-Murcia, M. (2007). Rethinking the role of communicative competence in language teaching. In E. A. Soler & P. S. Jorda (Eds.), *Intercultural language use and language learning* (pp. 41–57). Dordrecht, Netherlands: Springer.

Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied Linguistics*, *6*(2), 5–35.

Chalhoub Deville, M., & Deville, C. (2004). A look back at and forward to what language testers measure. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 815–831). Mahwah, NJ: Lawrence Erlbaum.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman (Ed.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). Cambridge, UK: Cambridge University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching and assessment*. Cambridge, UK: Cambridge University Press.

De Bot, K. (1992). A bilingual production model: Levelt's Speaking Model adapted. *Applied Linguistics*, *13*, 1–25.

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*, 4–34.

Dörnyei, Z., & Kormos, J. (1998). Problem-solving mechanisms in L2 communication. *Studies in Second Language Acquisition*, *20*(3), 349–385.

Dörnyei, Z., & Scott, M. L. (1995). Communication strategies: An empirical analysis with retrospection. In J. S. Turley & K. Lusby (Eds.), *Selected papers from the proceedings of the 21st Annual Symposium of the Deseret Language and Linguistics Society* (pp.155–168). Provo, UT: Brigham Young University.

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, *26*(3), 423–443.

Dunn, L. L., & Dunn, D. M. (2007). *The Peabody Picture Vocabulary Test* (4th ed.). Bloomington, MN: NCS Pearson.

Færch, C., & Kasper, G. (1983). *Strategies in interlanguage communication*. London: Longman.

ffrench, A. (1999). Study of qualitative differences between CPE individual and paired test formats. Internal UCLES EFL Report. In L. Taylor (2003). The paired speaking test format: Recent studies. *Perspective*, *55*(1), 70–76.

Fulcher, G. (2003). *Testing second language speaking*. London: Pearson Education.

Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London: Routledge.

Galazci, E. (2008). Peer–peer interaction in a speaking test: The case of the first certificate in English examination. *Language Assessment Quarterly*, *5*(2), 89–119.

He, A.W., & Young, R. (1998) Language proficiency interviews: A discourse approach. In R. Young & A.W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 1–24). Amsterdam, Netherlands: John Benjamins.

IELTS Speaking Assessment Criteria. Retrieved from https://takeielts.britishcouncil.org/sites/default/files/IELTS_Speaking_Assessment_Criteria_Public.pdf

Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, *29*(1), 3–17.

Kormos, J. (2006). *Speech production and second language acquisition*. London: Routledge.

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

Luoma, S. (2004). *Assessing speaking*. Cambridge, UK: Cambridge University Press.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing*, *26*(3), 397–421.

McNamara, T. F. (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics*, *18*(4), 446–466.

Messick, S. (1996). *Validity and washback in language testing*. ETS Research Report Series 1, 1–18

Nakatsuhara, F. (2006). *The impact of proficiency-level on conversational styles in paired speaking tests*. Cambridge ESOL Research Notes, *25*, 15–20.

Nakatsuhara, N., Joyce, D., & Fouts, T. (2014). *A research report on the development of the Test of English for Academic Purposes (TEAP) Speaking Test of Japanese University Entrants – Study 3 and Study 4*. Tokyo, Japan: Eiken Foundation.

O'Loughlin, K. (2002). *The impact of gender in the IELTS oral interview*. IELTS Research Report 3, 1–28.

O'Sullivan, B. (2008). *Modelling performance in tests of spoken language*. Frankfurt, Germany: Peter Lang.

Poulisse, N. (1993). A theoretical account of lexical communication strategies. In R. Schreuder & B. Weltens (Eds.), *The bilingual lexicon* (pp. 157–189). Amsterdam, Netherlands: John Benjamins.

Ross, S., & Berwick, R. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition*, *14*(2), 159–176.
Tarone, E. (1977). Conscious communication strategies in interlanguage: A progress report. *TESOL*, *77*, 194–203.
Taylor, L. (2003). The Cambridge approach to speaking assessment. *Research Notes*, *13*, 2–4.
Taylor, L., & Galaczi, E. (2011). Scoring validity. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (pp. 171–233). Cambridge, UK: Cambridge University Press.
Weir, C. J. (1993). *Understanding and developing language tests*. Hemel Hempstead, UK: Prentice-Hall.
Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Houndmills, UK: Palgrave Macmillan.
Wigglesworth, G. (2000). Issues in the development of oral tasks for competency-based assessments of second language performance. *Studies in Immigrant English Language Assessment*, *1*, 81–124.
Young, R. F. (2000). Interactional competence: Challenges for validity. Paper presented at the Annual Meeting of the American Association for Applied Linguistics and the Language Testing Research Colloquium. Vancouver, British Columbia, Canada.

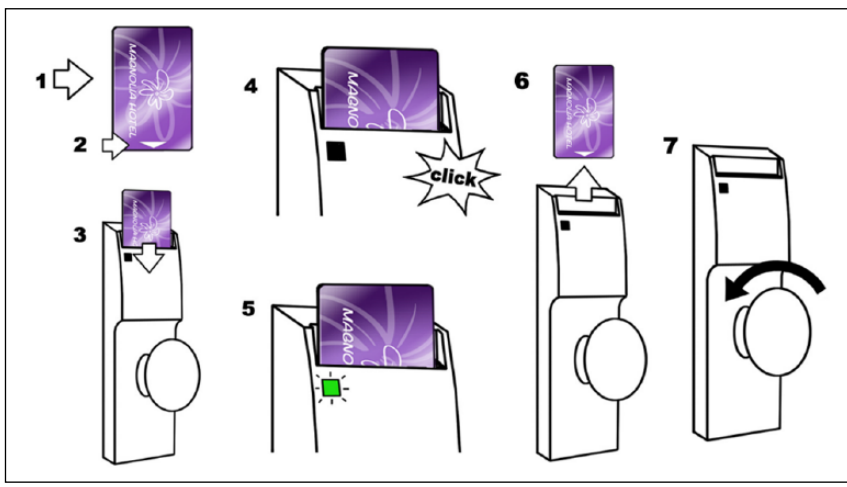
Appendix I

Instruction task (1) "Key Card"

The hotel where you work as a receptionist no longer works with keys, but with so-called key cards. One of the guests has just tried to open their door, but this did not work.

Can you explain to the guest how they can open their door, using the key card?

Task input sheet:



Appendix II

Advice task (3) "Hotel Room"

Something has gone wrong with Mr / Mrs. Jones' reservation: your colleague has booked a single room, instead of a double room. Only two rooms are still available but these differ greatly. Mr / Mrs. Jones will have to make a choice.

Can you *explain the difference*(s) between these two rooms and provide your guest with *fitting advice*?

Task input sheet:

Appendix III

Sales task (5) "Board Games"

You are working in the hotel gift shop today. One of your guests has come in to choose a present for their cousin Jada. This guest *seems* to be interested in two board games.

You don't have a permanent contract yet, so you want to show your boss that you are good at your job. You are determined to sell something today!

Which game do you think your guest should buy? Inform them of the *differences* between the games, and *persuade* them to buy this game from you.

Task input sheet:

Appendix IV

Rating scales

Scale 1: overall impression of language accuracy
On a scale of 1–5, indicate the extent to which the candidate expresses himself in lexically and grammatically *correct* English, and can be understood:
1 = barely speaks English / speaks more Dutch than English.
3 = frequently makes mistakes in sentence construction and / or lexical choice, but this does not seriously impede comprehensibility.
5 = correct sentence structure and lexical choice. Can be understood easily.

Scale 2: overall impression interactional ability
On a scale of 1–5, indicate the extent to which the candidate manages to convey his message, and to overcome potential communication problems:
1 = is barely able to convey the message. Cannot solve (mutual) communication problems independently in English.
3 = the message is not always clear or complete, but attempts to solve (mutual) communication problems in English.
5 = the message is conveyed clearly. Communication problems do not / hardly occur.

Scale 3: Analytic assessment of each marked contribution
On a scale of 1–5, indicate the extent to which the candidate provides an adequate response during each marked contribution:
1 = no or predominantly Dutch contribution; very weak / inappropriate
2 = weak / inappropriate
3 = reasonable
4 = strong
5 = very strong

Appendix V. Interlocutor script Advice task (3) "Hotel Room"

	ALTERNATIVE ROUTE	INTERLOCUTOR	EXPECTED RESPONSE	STRATEGY
(1)		Good morning / afternoon.	Greets and offers help	
(2)		I was told I need to make a choice between these two hotel rooms, is that right?	Expresses agreement	
(3)	If characteristics are given, or questions are asked, steer back to description	Then I need to know more about them. Can you first tell me a little bit about what the rooms look like?	Describes rooms	
(4)		Aha, thank you. So, tell me, on what points do these rooms differ?	Compares rooms (differences)	
(5)		OK. I thought the Red Room looked nice. But how is that room kept warm?	Explains radiator	*Compensation*
(6)		Sorry. What is used to heat up the room?	Clarifies radiator	*Clarification*
(7)	Does not seem to understand, but does not ask → continue with next question. Candidate responds with "Yes" → ask "In both rooms?" Then continue with next question.	Thanks, I have another question. My nephew will come and visit me for the day. Will you have a cot* for me? **a small baby bed*	Asks for clarification or answers question	*Meaning negotiation*
(8)		Let's see. I do want to make use of Wi-Fi during my stay. Did you say that this is possible in both rooms?	Compares rooms on Wi-Fi	
(9)		OK, so if I want the Red Room, I need to pay extra for Wi-Fi?	Corrects: Wi-Fi is free in the Red Room	*Corrects misinterpretation*
(10)	Gives uninformed advice → Continue with "Well, the thing is…"	Alright. Just trying to figure out what would be the best choice for me. Can you advise me?	Asks for relevant information	
(11)		Well, the thing is, I only have a limited budget, but I do need to make use of Wi-Fi. I'm a very light sleeper, so I'd prefer a quiet room, if at all possible.	Gives partially informed advice or asks for more information. *In case of latter:* Interlocutor provides information requested only • €80 is the limit I had in mind • I prefer a cosy room • I don't mind whether we sleep in a single bed or a double bed Silence is most important	
			Gives informed advice	
(12)	Does not fully justify→ Continue to final contribution.	And why do you think this is the best option for me?	Justifies (balancing at least cost, Wi-Fi and need for silence)	
(13)		Thank you very much for your advice.	Closes conversation	