Novitas-ROYAL (Research on Youth and Language), 2011, 5 (1), 91-120.

MANAGING TASK-RELATED TROUBLE IN L2 ORAL PROFICIENCY TESTS: CONTRASTING INTERACTION DATA AND RATER ASSESSMENT

Erica SANDLUND* and Pia SUNDQVIST**

Abstract: The present study takes as an empirical point of departure the nature of interaction in second language speaking tests. We examine the relationship between ratings of students' performance in an oral proficiency test and the social practice of conducting 'test talk'. Using conversation analysis, our focal point is how students in peer-driven test interactions manage trouble related to the task-at-hand. Given that students were assessed not only on their linguistic skills, but also on their interactional ability and treatment of topics assigned, our emphasis on task management stems from a hypothesis that orientation to the test task is intimately connected to overall test outcome. We demonstrate that different types of task-related trouble (TRT) reveal diverse understandings of the test task and that 'doing-being a successful task manager' is connected to a moderate orientation to the task and test format. Students displaying such task management strategies were also assessed as highly proficient, whereas other task management strategies identified in our study correlated with low scores and grades. However, the relationship between subskill ratings and task management was not always clear-cut. We argue that the diverging understandings of the test task that learners display become part of how they are assessed and that certain task management strategies are rated less favorably than others. Our study holds promise for the fine-tuning of oral proficiency subskill ratings and raises questions as to the framing of test tasks, since this appears to have implications for student performance and evaluation.

Keywords: Conversation Analysis, speaking tests, assessment, oral proficiency, task management

Özet: Bu çalışma deneysel bir çıkış noktası olarak ikinci dil konuşma testlerindeki etkileşimin doğasını temel almaktadır. Öğrencilerin sözlü yeterlilik testlerindeki performansı ile 'test konuşması' yönetimi sosyal uygulaması arasındaki ilişkiyi incelemekteyiz. Temel odak noktamız konuşma çözümlemesi kullanarak öğrecilerin akranlarıyla yürüttükleri test etkileşimlerinde aktiviteye bağlı problemlerle nasıl başa çıktıklarıdır. Öğrencilerin sadece dilbilimsel yeteneklerinin değil ayni zamanda etkileşimsel yeteneklerinin ve kendilerine verilen konuyu ele alışlarının da ölçüldüğünü göz önünde bulundurarak, etkinlik yönetimi üzerine olan vurgumuz 'test etkinliğine yönelim genel test neticesine derinlemesine bağlıdır' varsayımından yola çıkmaktadır. Aktivite-ilintili Sorunların (AİS) farklı çeşitlerinin muhtelif test aktivitesi anlayışlarını ortaya çıkardığını ve 'iyi bir aktivite yöneticisi olabilmenin' aktiviteye ve test biçimine ölçülü bir yaklaşıma bağlı olduğunu göstermekteyiz. Bu aktivite yönetme stratejilerini uygulayabildiğini gösteren öğrenciler aynı zamanda yüksek derecede yetkin olarak ölçülmüşlerdir, ancak çalışmamızda belirlediğimiz diğer aktivite yönetim stratejileri düşük notlar ile ilişkilendirilmiştir. Fakat, alt-beceri değerlendirmeleri ile görev yönetimi arasındaki ilişki her zaman çok açık olmamıştır. Öğrencilerin gösterdiği farklı test anlayışlarının nasıl değerlendirildiklerinin bir parçası olduğunu ve belirli aktivite yönetim stratejilerinin daha az uygun olarak ölçüldüğünü savunmaktayız. Çalışmamızın sözlü yeterlilik alt-beceri ölçülerinin geçerliğini sağladığına ve öğrenci performansı ve değerlendirmesi ile alakalı sonuçları olduğu için de test görevlerinin tasarlanması konusunu sorguladığına inanmaktayız.

Anahtar sözcükler: Konuşma Çözümlemesi, konuşma testleri, ölçme, sözlü yeterlilik, etkinlik yönetimi

Introduction

John Dewey, the famous American philosopher and psychologist, once stated that "[t]here is all the difference in the world between having something to say and having to say something"

* Senior Lecturer, Faculty of Arts and Education, Karlstad University, Sweden, erica.sandlund@kau.se.
** Senior Lecturer, Faculty of Arts and Education, Karlstad University, Sweden,pia.sundqvist@kau.se.

(Dewey, 1976, p. 35). In the context of a speaking test in one's second language – not knowing what to say (but having to say something) might force a learner to produce utterances that are not well-formed (cf. Romova & Neville-Barton, 2007), to uncomfortable pauses, or to no output at all. Awareness that one is being graded on output adds particular pressure on producing talk on topics pre-set by others, something which has been addressed in language testing research (e.g. Fulcher & Márquez Reiter, 2003). Task construct, instructions, and topics all seem to interact when it comes to the establishment of conversation in second language oral tests (cf. H. D. Brown & Abeywickrama, 2010). In this paper, we take a closer look at the relationship between ratings of students' performance in a second language (L2) speaking test and the social practice of conducting 'test talk', with a particular focus on the emergence of interactional 'trouble' displayed in connection with the test task. Our focal point is how students in peer-driven test interactions manage problems related to the task-at-hand. Given that students were assessed not only on their linguistic skills, but also on their interactional ability and their treatment of topics assigned, our emphasis on task management stems from a hypothesis that successful management of test tasks is intimately connected to overall test success. Student grades and test scores set by teachers and external raters are contrasted with particulars of the test talk interaction, and we discuss possible explanations for peculiarities in the quantitative and qualitative data comparison. Our analyses join the growing body of studies that seek to bring together the strengths of L2 testing research and conversation analysis (CA).

By contrasting analyses of sequences of interaction data with rater assessment of oral proficiency (OP) in L2 English, the aim of the present study is to answer the following research questions: (1) Is there a connection between testees' displayed management of interactional trouble related to managing the task of test-taking in paired OP tests and their assessed L2 oral proficiency? (2) What other aspects of the interaction appear to play a role in testees' management of task-related trouble (TRT)? Data was originally collected for a study on the effects of extramural English on Swedish ninth graders' OP and vocabulary (Sundqvist, 2009). Twenty informants making up ten dyads constitute the empirical material for the present study. The interactions were first analyzed using conversation analytic procedures (Sacks, Jefferson, & Schegloff, 1974; Wagner & Gardner, 2004) and later compared with the existing assessment data. Our reasons for the separated analyses are grounded in the diverging foundations of oral language testing research and conversation analysis; a dilemma we also address.

Conversation Analysis in Second Language Research

The success of conversation analytic research in a wide array of institutional contexts has occasioned a joining of research interests of interactionally oriented scholars and researchers in the field of second language acquisition (SLA). The grounding in how learners display their orientations of actions in ongoing talk can, as Hellermann (2009, p. 96) remarks, "uncover the aspects of language that participants (language learners in this case) produce to accomplish their social interactions (inside or outside the classroom) and offer indigenous or participant-defined phenomenon for language researchers to study". In terms of classic SLA matters like OP, fluency, and L2 tasks, Kasper (2006) proposed that CA findings can provide insights that support reconsideration of existing SLA understandings. For example, in a study of institutional first/second language speaker interactions between clients and secretaries, Kurhila (2004) demonstrated that the interactional relevance of identities as 'first' or 'second' language speakers is by no means given. Instead, she noted that L2 speakers' orientations to their non-native status were carefully fitted into the ongoing activity, and that what at first glance appeared to be speech

perturbations related to non-nativeness could actually reflect speakers' awareness of expected linguistic behavior. Kurhila also noted that first language speakers rarely made relevant their linguistic identity. On the contrary, orienting to their institutional roles blocked the activation of language learner/teacher identities, something which could clash with the institutional constraints of a service encounter. Another example where assumptions regarding a straightforward relationship between speech perturbations and OP are called into question is Carroll's (2000) study of novice L2 speakers' entry into conversation. Not only were the L2 speakers capable of timing their entry into conversations (i.e. by orienting to transition-relevance places, cf. Sacks, et al., 1974), there was also some evidence that when they did not, the inter-turn gaps were occasioned by dysfluent characteristics of a *preceding* turn by an interlocutor.

Studies such as Kurhila's and Carroll's show that there is much to gain in SLA research by examining more carefully interactions where L2 speakers participate. Furthermore, the growing body of research on L2 interactions has not yet identified any practices that are entirely unique to L2 talk; particular features may, however, be deployed more or less frequently in comparison with L1 conversations (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Wagner & Gardner, 2004). As such, L2 conversations are 'normal' conversations in that participants are able to partake in sophisticated interactional activities despite varying linguistic proficiency. However, attempts at marrying SLA research with CA have not been unproblematic; something which we address in relation to our study in the *Data and Methodological Considerations* section.

As our title indicates, our study is situated at the intersection of social interaction and language testing, and builds on two main areas of existing research: the assessment of OP and the notion of interactional trouble. We begin by reviewing work on the assessment of OP. Then, we examine some CA studies on language testing, and finally, we carve out our application of the notion of interactional trouble.

Assessment of Oral Proficiency

Assessment of OP has proven to be complicated, since various factors interact to ultimately render a holistic score or a grade. One common option is to assess learners' spoken language in oral proficiency interviews (OPIs), where a testee is alone with an examiner during the test. Another possibility is to assess two (or more) learners at the same time. The use of dyads in OP tests has, however, been questioned. It is possible to argue that the outcome of such a paired test becomes a 'blend' – of two testees, the task(s), and the examiner – something which makes it difficult for raters to make fair assessments (Ducasse & Brown, 2009; Fulcher & Márquez Reiter, 2003; O'Loughlin, 2002; O'Sullivan, 2002). Even so, dyadic setups are commonly used in educational settings, mainly because such a format resembles natural conversation (Ducasse & Brown, 2009).

Needless to say, a speaking test should be well-structured and have good face value, but there are inevitably a number of factors that, nevertheless, may influence student performance in a negative fashion just because it is a test, such as test-anxiety (Pappamihiel, 2002), the interlocutor (cf. Davis, 2009; Iwashita, 2001), gender (O'Loughlin, 2002; O'Sullivan, 2002), and commonplace problems such as tiredness, hunger and seating arrangements (Sundqvist, 2008). As for paired dyads, Davis (2009) found that interlocutor proficiency may influence grades for some individuals. In his study, lower-proficiency testees produced more words when they had a higher-proficiency interlocutor. Davis (2009, p. 386) also found that higher-scoring testees said

more (produced more words) regardless of their interlocutor. Furthermore, lower-scoring testees tended to take on a passive role in dyads: "passive because they were unable (rather than unwilling) to contribute more fully to the task at hand" (Davis, 2009, p. 387-388).
Riggenbach (1998) emphasizes the importance of the topic and maintains that great flexibility in topic choice may lead to a better reflection of learners' interactional skills than prearranged, traditional OPIs do. With regard to interactional skills, Naughton (2006) suggests that rules of socially acceptable behavior may have a greater influence on interaction than learners' lack of ability. Consequently, delays and silence in learner speech may depend on context (e.g. a sensitive or unfamiliar topic) or on what is considered accepted behavior in a particular setting, rather than lack of interactional skills.

OP involves several aspects of language: vocabulary, grammar, fluency, etc. There is also a social dimension of OP, realized as interactional skills (see e.g. McNamara & Roever, 2006). Arriving at one holistic measurement of OP may not always be enough and, therefore, it is common to also score learners on subskills. As our interest lies in how interactionally troublesome sequences are managed, we are also interested in the assessment of subskills that, possibly, can be related to this interactional practice. Hasselgren (1997) investigated oral test subskill scores and found some of learners' subskills to be highly influential in raters' decisions on final overall grades and referred to them as core linguistic subskills, one of them being 'to keep going and contribute in interaction'. She also found that the raters gave learners' "'language' performance" precedence over their "'message and fluency' performance" (p. 250). Interestingly, when examining dimensions of OP in an L2, de Jong and van Ginkel (1992) found that in holistic assessment, fluency dominated the evaluation and discriminated effectively between learners who received lower and higher overall ratings.

Oral Proficiency Testing - Insights from Conversation Analysis
In the period 1990-2010, the contribution of discursively oriented methodologies to the area of OP assessment has allowed an inside perspective on the micro-processes of language and interaction in situ (see e.g. Kormos, 1999; R. Young & He, 1998; R. F. Young & Milanovic, 1992). A growing number of studies have applied CA to OP tests and Schegloff, Koshik, Jacoby, and Olsher (2002) claim that CA is highly suitable for gaining insights about interaction in oral assessment contexts. Similarly, Lazaraton (1992) argues that CA should be applied to widely-used speaking tests to evaluate the utility of such instruments in eliciting conversational interaction.[1] Without claiming to do justice to the full body of interactional studies on OP testing, a few findings deserve particular attention in relation to our study.

A number of studies have taken an interest in what characterizes speaking test interactions and how they differ from other institutional or ordinary conversations. Seedhouse and Egbert (2006) examined 137 recorded IELTS[2] Speaking Tests and noted that the interactional organization of the test differs significantly from interaction in classrooms or university settings in that the tests show very few repairs on part of the examiner, even in cases where candidates produce incomprehensible turns. This can be explained by institutional goals; in classrooms, the goal of "transmission of knowledge or skills from teacher to learner" calls for repair to be deployed more frequently in order to ensure intersubjectivity, whereas in speaking tests, the main goal is to

[1] See also Lazaraton (2002) for an extended overview of qualitative input in oral language testing research.
[2] International English Language Testing System, see www.ielts.org .

assess candidates' utterances in terms of what the test standards prescribe (pp. 191-192). The IELTS Speaking Tests also show a high degree of pre-allocation of turns and considerable asymmetry between candidate and examiner since the candidate is only allowed to initiate repair in specific prescribed formats. In an earlier study, Egbert (1998) compared data from OPIs with data from (everyday) native speaker conversations and noted that interviewers explicitly explained repair types. Moreover, interviewers also initiated repair in ways that the native speakers in her other data set never did. In terms of assessment of OP, Seedhouse and Egbert (2006, p. 193) indicate a correlation between test scores on the one hand and the occurrence of other-initiated repair, i.e. candidate displays of trouble in hearing or understanding questions or prompts.

As noted earlier, OP tests vary in their setup and recent studies address test formats such as *peer-peer tests*, *paired tests* (two examiners, two testees), and *oral language assessment in groups* (Gan, 2010); i.e., test formats that differ from that of traditional OPIs (an examiner and a candidate). In addition to examining test formats, topic negotiation and the effect(s) of the interlocutor are other matters that have been investigated. Gan, Davison and Hamp-Lyons (2008) examined peer group discussions as an oral assessment format. They conclude that the group format has the potential of providing 'authentic' talk where students had ample opportunity to display both linguistic and interactional competence. Lazaraton and Davis (2008) examined the interlocutor effect in speaking tests by comparing test scores with features in peer-peer interactions that supported these scores. They found that testees' OP is "fluid" (p. 331) and shifts on a turn-by-turn basis throughout the test, something which could be partly attributed to the type of language *identity* that testees made relevant in the test talk (i.e., as highly proficient or less proficient speakers of the L2).

Studies informed by CA have also dealt with the validation of assessments or rating scales. One such example is Galaczi (2008), who analyzed data from peer-peer interactions in the First Certificate in English speaking test with the two-fold purpose of describing patterns of interaction in paired tests, and of examining the relationship between observed patterns of interaction with scores on 'Interactive Communication (IC)' which had been awarded by two examiners. She argues that the CA analyses "provided some validity evidence for the IC scores" (p. 112) as testees who scored high on IC also displayed highly *collaborative* (as opposed to *parallel, blended,* and *asymmetric*) patterns of interaction, whereas testees with low IC scores generally oriented to a parallel pattern of interaction. Galaczi concludes that findings from the study could be applied to rater training and to the construction of more fine-tuned assessment scales based on empirically observed discourse features.

Finally, there are CA studies that in various ways relate to task construct and management in language learning in general and in OP tests in particular. Seedhouse (2005) critically examines research on tasks and argues that there is a important distinction between what Breen (1989) referred to as *task-as-workplan*, i.e. the intended pedagogy of a particular task, and *task-as-process* (i.e. what actually happens with the task as it is tackled by participants). This, according to Seedhouse, is a validity problem since conceptualizations of tasks are based on the workplan level, whereas data is obtained from the process level. After empirically demonstrating how the two levels diverge, Seedhouse concludes that the research focus needs to shift to task-as-process. Similarly, Hellermann and Pekarek Doehler (2010) examined task accomplishment in language classrooms and demonstrated that the same task-as-workplan resulted in rather different task

accomplishments by different student dyads. Analytically, they focused in particular on the transitioning from instructions to performance of a task as "much of the student's orientation to the task crystallizes and is negotiated at the very start of task accomplishment" (p. 28). The authors demonstrate how different orientations to the same task results in different task trajectories and that the shift from task-as-workplan to task-as-process occurs in the moments of transitioning into the task. By studying student orientations to the task, is it thus possible to observe and describe tasks from an *emic* (participant-relevant) rather than an *etic* (external analyst's) perspective (Seedhouse, 2005, p. 535).

The above mentioned studies constitute a few examples of how CA-informed/inspired studies of speaking tests have offered insights into testee conduct, examiner influence, and the relationship between features of interaction on the one hand, and ratings on the other.

Interactional 'Trouble'

In language testing research, it is often assumed that there is a causal relationship between OP level and certain displays of interactional trouble such as dysfluencies, or possibly, noticeably longer silences (cf. Foster & Skehan, 1996; Iwashita, Brown, McNamara, & O'Hagan, 2008; Lennon, 1990; Levis, 2006). However, CA studies have generated reconsideration of such general assumptions (e.g. Carrol, 2000). From a CA perspective, 'dysfluent' characteristics of talk are not per se markers of trouble; instead the notion of interactional trouble is grounded in observations on how interactants *display their noticing and management of* some problem in the ongoing interaction. Problems can be evident in, for example, producing or understanding relevant (and timely) contributions to the ongoing talk. In our material, we focus specifically on interactional trouble made relevant in the context of opening, 'properly' treating, and closing of tasks in the speaking test, where interactants are faced with a range of options for moving beyond the trouble, such as abandoning a particular topic, codeswitching, initiating repair, and so forth.

The notion of interactional trouble is closely linked to repair, i.e., the "practices for dealing with problems or troubles in speaking, hearing, and understanding the talk in conversation" (Schegloff, 2000, p. 207). The mechanism of repair functions to maintain a shared understanding of what is going on rather than to uphold linguistic standards, and a 'repairable' is not always attended to. It is often assumed that L2 interaction is particularly laden with repair associated with language form. However, CA work on repair in talk involving L2 speakers has indicated that the organization of repair in non-native speaker interaction is very much like that of native speaker talk and that speakers try to maintain the sequential flow by minimizing the imposition of repair (see Wagner & Gardner, 2004). Even though repair organization of L2 interaction seems to follow the same patterns as those in L1 talk, other-initiation of repair occurs, not surprisingly, more frequently in language teaching contexts (Rasmussen & Wagner, 2000). Similarly, Plejert (2004) observed that an increased knowledge of the L2 correlated with an increase in the range of actions that repair organization performed (i.e. on both linguistic and social levels). In terms of assessment of OP, as noted above, studies of OPIs have revealed certain particulars of repair forms and their deployment in speaking tests by candidates and examiners respectively (Egbert, 1998; Kasper & Ross, 2003; Seedhouse & Egbert, 2006; R. Young & He, 1998). Below, we outline the type of test data used in the present study, as well as our methodological considerations.

Data and Methodological Considerations

Data was originally collected for Sundqvist (2009), a study focusing on the impact of extramural English on OP and vocabulary. In total, 80 students (aged 15-16) participated and they belonged to four classes taught by three teachers (all women) at three schools. For the purpose of the present study, we made a selection of 20 students, namely the ones who were awarded the five highest and five lowest mean OP grades overall and their respective interlocutors (see also Sundqvist, 2009, p. 138). In our sample, all but one had L1 Swedish.[3]

Speech data in Sundqvist (2009) was collected from five interactional English speaking tests, spread out over a school year, and the students were assigned to random dyads on each test occasion. The researcher was the test instructor in the first four tests, where video recording was used. The students' teacher was the instructor in the fifth test, from which speech data for the present study was drawn. This particular test was the mandatory national test of English whose guidelines stipulate the use of audio recordings only. The national test guidelines ask teachers first to read specific test instructions out loud (e.g. "Be active and speak English all the time!") and then to leave the floor to the testees. However, the guidelines do not explicitly forbid teachers to intervene; thus, some might choose to do so (cf. A. Brown, 2003). The national test of English used in Sweden has been thoroughly evaluated and has high validity and reliability (Erickson & Börjesson, 2001).

In total, 199 tests were included in Sundqvist (2009). Based on the recordings, the students were assessed by four external raters, using written instructions and assessment forms adapted from Hasselgren (1996). The raters worked independently. On each test, the student performance was first evaluated (1-5) with regard to ten subskills (a-j); then the raters decided on two factorial grades (1-6), one for *message and fluency* and another for *language structures and vocabulary*. Finally, students were awarded an *overall grade for oral proficiency* (the *OP grade*, 1-6). Three of the raters assessed each student on each test; i.e., at the end there were 15 OP grades per student.[4] The OP grade was used as a measurement of the students' level of OP, defined as 'the learners' ability to speak and use English in actual communication with an interlocutor'.

In the present paper, the following three subskills are analyzed: (a) *overcoming difficulties in communication,* (i) *interactional ability*, and (j) *treatment of topic,* formulated as follows in the assessment form:

[3] In Table 1, student 'J1' has L1 Kurdish; all the others have L1 Swedish.

[4] The Pearson correlation coefficient (*r*) was used to measure interrater reliability, which ranged from .451** to .703**. Hasselgren (1997, p. 243-244) considers a minimum value of *r* at .4 as "reasonable", and so do we.

Subskill (a). *'Overcoming difficulties in communication'.*

When difficulties in communication arose, did the pupil make an independent attempt to overcome these, in English?

virtually always	5☐
	4☐
sometimes	3☐
	2☐
rarely	1☐

Subskill (i). *Interactional ability.*

The student's ability to interact with the other student was

excellent (takes initiatives, adapts speech to suit partner/situation etc)	5☐
	4☐
acceptable	3☐
	2☐
very poor (e.g. doesn't respond to cues etc)	1☐

Subskill (j). *Treatment of topic.*

The way the student treated the subject/topic was

excellent (focussed/in depth, with rich content)	5☐
	4☐
acceptable	3☐
	2☐
very poor (brief/shallow)	1☐

These three subskills were selected because of their possible connection with task management. Subskills (a) and (i) are clearly related to students' ability to interact with each other in the test, whereas subskill (j) is linked to the task-as-workplan, i.e., the ability to develop assigned topics. The remaining subskills were considered to be of less interest as they primarily targeted linguistic proficiency.

As mentioned, we use data from test 5 (the national test) in our analyses. The test, called 'The world around us', consisted of three parts. In part one, testees were instructed to take turns and ask each other questions about where they live. In part two, cards with various statements (such as "Money makes people happy") were used. Testees were to draw a card and discuss the statement, state whether they agreed or not, and explain why. Part three had a similar format but instead of statements there were questions. The average length of a recorded test was 15 minutes (including instructions). In addition to assessment data from the external raters, the students' final grades in the school subject English were also collected in Sundqvist (2009); these final English grades, set by the students' own teachers, are also included in our analyses.

The national test aims to elicit a 'natural' conversation between two students but, admittedly, it is nevertheless a test-situation, and as Wagner (1998) demonstrated, elicited speech data differs in several ways from naturally occurring talk (see also Seedhouse & Egbert, 2006, p. 169). Also, it is possible that in an L2 speaking test, the motivation for upholding intersubjectivity may differ from other types of interaction. That is, students who are aware that their *individual* level of OP will be graded and that it is important to 'just keep on talking' may be primarily focused on their own production rather than on alignment with their interlocutor.

The contrasting of data collected and analyzed on rather different epistemological grounds and for various purposes does present dilemmas, as CA work focuses on actions on a turn-by-turn basis, whereas language testing research data often presents more holistic measures of *entire samples of speech*. Since the present study has a comparative aim, we wish to emphasize that we by no means are oblivious of the problems associated with comparing shorter sequences of interaction that have been selected through analytically driven sampling, with data collected with the purpose of providing holistic measures of the interactions. Marrying the two approaches is not unproblematic. For example, Mori (2007) noted that while many SLA scholars agree with the need for increased attention to social and contextual factors of language learning and development, CA findings may, in the eye of the uninitiated, seem too small-scale or context-specific to provide answers to "what has been learned, when it has been learned, and why it has been learned" (2007, p. 853). Critics of CA approaches to SLA problems have argued that CA needs to venture beyond analysis of manifest conduct and toward sociocognitive theories of language learning (see Mori, 2007). Similarly, CA researchers, whose work has been less than heartily welcomed when "trespassing" in the SLA domain (Firth & Wagner, 1998), may hesitate to combine their modes of analysis with analytic procedures that are more or less incompatible with the basic assumptions and emic perspective of CA. In our view, the combination of methodologies raises new questions for what is compared, and how.

Earlier, we used 'marriage' as a metaphor for CA/SLA combinations and it appears as if a prenuptial agreement is in order, one which clearly defines the expected and practical contribution of each party and keeps assets of each partner as separate properties in the union. Such an approach would not require any compromises in terms of core assumptions of CA, as CA is "not built to answer theoretically motivated research questions of the type that applied linguists often ask" (Schegloff, et al., 2002, p. 14). Instead, the detailed workings of the CA approach can raise new questions in language learning and testing, which in turn can be taken to a large-scale level using other methods. We argue that comparative efforts can provide possible explanations for rater assessment of particular testees and shed light on practices in L2 test talk

that may promote or hinder testee performance, such as topic development and joint accomplishments of overcoming or abandoning interactional trouble.

Task Management in an Oral Proficiency Test – Analyses of Interaction

We refer to task management as a broad concept which can consist of many different ways for accomplishing a task. Our analyses focus on task management in terms of *task-as-accomplishment* (i.e. *process*) rather than *task-as-workplan* (cf. Breen, 1989; Hellermann & Pekarek Doehler, 2010; Seedhouse, 2005) and include a range of practices for accomplishing the task, such as developing topics, asking and answering questions, closing topics, allocating turns, and marking topical boundaries, to name a few. As we note in our analyses, ambiguity regarding the task at hand constitutes a site for potential trouble in test interaction. From our analyses of the ten tests, we show selected fragments where task-related trouble (TRT) arises that in different ways reflect recurrent patterns in the sample. After presenting the analyses of five fragments, we compare the interactions with the assessments that external raters plus their teachers provided. Our presentation of fragments aims to demonstrate how students' differing orientations to the test task yields rather different management of the task as well as of TRT, and how these orientations also appear to be treated as more or less appropriate ways of managing tasks by raters and teachers (see Appendix for a transcription key).

Same Task - Different Task Management

We begin by examining Fragments (1) and (2), where two student dyads manage the same task of discussing statements from written prompts rather differently. Both task accomplishments appear to be 'productive' in the sense that the students use the topic cards as resources for establishing and maintaining a discussion that results in assessable output, but the task management strategies reveal different understandings of how to accomplish the task.

In Fragment (1), student H1 draws a topic card about the danger of mobile phones ("Mobile phones are dangerous and disturb people"). The organization of the task at hand is topicalized in line 28 by H1, who requests direction from the teacher (here, 'teacher 2', i.e. T2) as to whether to pull a new card after the previous topic has been brought to a close. After confirmation from T2, H1 reads the card and embarks on a commentary within the same turn:

Fragment (1)

[530423031] T2: teacher 2, H1: female student, H2: male student

28	H1	yeah hhuhm (..) should I take (.) >the next<?
29	T2	mhm?
30		(2.1) ((*shuffling sounds*))
31	H1	((reads)) e:h (.) mo:beel phones are dangerous and disturb
32		(.) people (1.8) ((stops reading)) °mobile phones° (.) no?
33		(.) well (.) everyone has a (0.5) cellphone and hhhHUH.hh
34	H2	uh (.) and I have heard that the radiation from the::
35		mobile phones (0.8) are not that dangerous as: .hh (.)
36		people tend to believe
37		(0.4)
38	H1	[no]

39 H2 [that] you practically have to (.) walk around with it on
40 all day (1.6) u:h in (.) a long time for you to notice any
41 (1.4)
42 H1 yeah I don't know (what's happening but)
43 H2 any::-
44 H1 =I don't think it's (0.8) that (.) big (hh) huh (.) >I
45 mean< everyone has its ow- (.) u:h no one will notice hhuh
46 hhUH
47 H2 yeh I suppose it's quite annoying if you:: (.) if it calls
48 in the middl::=
49 H1 =of a convers-
50 H2 of a con[versation or some[thing]
51 H1 [aah] [yeah]
52 H2 but- .hhh
53 H1 well- *ja*
54 H2 but but that that'll be only annoying not dangerous
55 H1 yeah (.) mm (.) precisely (.)I think e:h (.) you're very
56 (.) u:h addicted to your phone (.) when you don't have it?
57 (.) you're (0.5) you feel like you're missing something
58 well it's quite u:hm hard *eller* (.) u::h to (.) contact
59 everyone (1.6) hhHUH (.) if you if you want to call someone
60 it's (.) you really need (0.6) the cellphone just (.)
61 otherwise you feel-
62 H2 hhRM (.) it makes it much easier to (1.6) e:h talk with
63 your friends whenever you (.) want to or're bored or
64 H1 exact (1.6) and if you're in danger you can call 911? hhhH
65 and ( )
66 H2 yeah?
67 (3.3)
68 H2 u:h ((*reads*)) (.) the sound level at concerts and discos
69 are dangerous (.) uh (.) if you're close to the ((cont.))

Although H1 initiates this new topical sequence with a teacher-directed request for permission to switch to a new topic, the initiation of a topical sequence begins immediately upon finishing reading with her own "*no?*". She offers a general comment on the topic (*everyone has a (0.5) cellphone and hhhHUH.hh*), displaying a creative vocabulary use by not recycling "*mobile phones*". The laughter at the end of her possibly incomplete turn in line 33 is treated as a transition relevance point by H2, who brings a more specific point about cellular phones: "*u:h (.) and I have heard that the radiation from the:: mobile phones (0.8) are not that dangerous as: .hh (.) people tend to believe*". His turn-initial "*and"* displays a linking to H1's turn and marks his contribution as an additional point, which is further elaborated in lines 39-40. Without awaiting completion of the trailed-off "*any*" in line 40, H1 agrees in line 42 (*yeah I don't know (what's happening) (.) but*). H2 repeats his "*any*" in line 42 in what seems as a search for a word or phrase describing the effects of radiation; however, as H1 continues in line 44, the search is

abandoned. H1's turn is somewhat incoherent and contains laugh particles, but appears to be, at least in part, related to the topic development (that pretty much everyone has a cellular phone). Despite the last rather incoherent part of the turn, H2 does not orient to the turn as problematic; instead, he treats the turn-final laughter as a transition point and brings, in line 47, a new point initiated with an agreement token seemingly directed at H1's prior turn (*yeh I suppose it's quite annoying if you:: (.) if it calls*). As he continues, H1 offers a possible completion (49) and acknowledgement tokens (51, 53) in overlap. The contributions of both parties show context-sensitivity and are fitted to preceding turns. H2 also orients back to his own initial turn on this subtopic (line 54) and also to the task formulated line 38: "*but but that that'll be only annoying not dangerous"*. His turn is thus task-oriented in that he acknowledges that his point about phone calls in the midst of a conversation may not correspond to what the task stated, i.e., that phones are dangerous. H1 agrees (line 55) and they exploit the topic for another few turn shifts before H1 finalizes her last point with a rising intonation (*call 911*?) and H2 mirrors the rising intonation (*yeah?*). A silence of 3.3 seconds follows, after which H2 unpromptedly draws another card. He reads a new statement and begins commenting on the topic immediately upon finishing reading.

As we can see, H1 and H2 display understanding of the task at hand in various ways: H1 seeks confirmation from the teacher regarding the 'rules' of the test, they manage the topic movements in a 'stepwise' manner (Gan, et al., 2008; Sacks, 1992); i.e., they display orientation to the prior turn elements and introduce related content, and when both parties seem to have reached a shared understanding of when the topic is exhausted (rising intonation in two subsequent turns), H2 draws a new card. There are few items identified as repairable (only one self-repair on pronunciation, line 32) and they collaboratively shape the topic buildup. The students, then, display an understanding of what the test talk entails and manage the discussion smoothly. Their respective turns are not extremely long and the overlaps and agreement tokens give the conversation a natural-sounding air. The timing of a new card, then, is a joint achievement. The topic cards are utilized as resources, both for showing sensitivity to the teacher's instructions, and for safeguarding the conversation from awkward pauses as a topic has becoming exhausted.

In our material, there are also occasions where 'excessive' orientation to the task instructions seems to overshadow the development of topics. In Fragment (2), the students' orientations to the test task seem to focus specifically on reaching agreement or disagreement on the statements on the topic cards, so that the standard responses 'agree' or 'disagree' work as closing devices instead of as elicitators of talk. Below, they move through two topic cards in just a few turns each, and once the agreement/disagreement options are produced, there is immediate topic decay. The extract begins with the closing of the previous topic (lines 09-14), a topical sequence that lasted only a few turn shifts. The new topic task is introduced in line 16, and this topical sequence is closed in a similar fashion using explicit agreement as topic boundary signal:

Fragment (2)

[531523172] T2: teacher 2, F1: female student, F2: female student

```
09    F1 →      .hh (0.5) so: we:: (0.8) (pt) (.) disagree? (.) or ˚agr↓ee˚
10              (2.1)
11    F2 →      u:hm (2.1) disagr↓ee.=
12    F1 →      =I >disagree there too<
```

13			(4.1)
14	F2		°ohkay°
15			(2.9) ((*scribbling sound*))
16	F2		.hh u::h ((*reads*)) the sound level (.) at concerts (.) and
17			discos (.) isu:h dangerous ((*stops reading*))
18			(3.5)
19	F1		yeah bu:dh (.) y- you could [buy (.) thisu:h earplugs or=
20	F2		[(the music and)
21	F1		=something >I know what it mean(hh)s< hhuhHE:H (.) mm?
22			[so- .hh
23	F2		[yeah
24	F1		the THING? [is that it=
25	F2		[sho-
26	F1		=SHOuld? be loud .hh (0.6) that's=
27			the thing [you know=
28	F2		[yeah
29	F1		=if you wanna listen to your favourite band it should be
30			>'*ksom*[5] ↑WHO::U↓< (.) Hhuh huh [huh
31	F2		[ye:ah,
32	F1		kind of like [th(hh)a:t (hh)
33	F2		[it's your own (.) choice (.) if you want tu:h
34			have it loud [or (.) rr- low
35	F1		[y↑eah
36	F1	→	y↑eah (.) so we:u:h (.) disagree
37	F2		disagree (0.6) again hhhHUh HUH [HUH
38	F1		[HUH
39	F1		okay ((reads))living in the countryside is better than
40			living in a town ((*stops reading*))(.)

In line 09, we see the how F1 sums up the dyad's opinions brought forth on the preceding topic, prefaced by "*so*", and formulates her action as a question: "*so: we:: (0.8) (pt) (.) disagree? (.) or °agr↓ee°*". The alternatives for explicit agreement/disagreement thus function as pre-closings and make relevant one of the options as a response. H2 offers, after what appears as some time for contemplation, "*u:hm (2.1) disagr↓ee*." The falling intonation at the end marks the alternative as 'final'. F1 responds with a third-position agreement (*I >disagree there too<*). After a four-second silence, F2 produces a quiet "*ohkay*", which acknowledges the closing of the topic. Arriving at consensus on one or the other option, thus, appears to be the F-dyad's displayed interpretation of proper task management.

A new topic card is introduced (line 16) and F1 disagrees with the topic sentence with an initial agreement token (Pomerantz, 1984). During a series of turns, F1 continues with additional

[5] Abbreviated form of the Swedish *liksom* meaning "sort of", i.e. an example of codeswitching.

support for her disagreement in lines 26-27 and 29-30, while F2 offers repeated acknowledgement and support in overlap. In lines 33-34, F2 adds her conclusion: "*it's your own (.) choice (.) if you want tu:h have it loud [or (.) rr- low*". In line 36, F1 initiates the closing with the gist of their shared response: "*y↑eah (.) so we:u:h (.) disagree*". F2 confirms F1's summary with "*disagree (0.6) again hhhHUh HUH [HUH*". The use of "*again*" and the subsequent laughter can be heard as F2 somehow marking that opting for the same alternative once more would somehow be problematic in relation to the test task. This is speculative, as they rapidly move on to the next topic (line 39); however, it appears as if the students' focus on discussing a topic only until it has been demonstrated that they share similar views and are able to provide an 'answer' actually hampers topical elaboration and causes fast topic decay. In short, Fragment (2) shows how students orient to the task in a different way than students in Fragment (1); i.e. as a matter of reaching agreement on each topic and then moving on to the next. However, the task accomplishments appear 'productive' in both excerpts since participants are clearly monitoring both the task and the interaction, and manage the tasks with few displays of trouble.

Task Abandonment – a Strategy for Managing TRT

As a contrast, our third example of task treatment is more closely related to the management of TRT. In Fragment (1), the students displayed ease in elaborating on the topic of mobile phones and had, in Dewey's (1976) words, "something to say". In Fragment (3), F1 and F2 run into the other part of Dewey's famous quote, i.e. "having to say something". The new topic is drawn without any explicit orientation to a finished previous topic; however, the same matched intonation rise (lines 100-101) as noted in Fragment (1) is also present here. The topic card, reading "How can men and women be more equal?", creates immediate trouble as F2 does not understand the core component "equal":

Fragment (3)

[531523172] T2: teacher 2, F1: female student, F2: female student

100	F2	it's just the way it is?
101	F2	.hhhyeah? .hhh
102		(5.0)
103	F1	˚hm hm hm:?˚ ((rhythmic, high-pitch singsong voice))
104	F1	((*reads*)) how can men and women:? (.) be more e:qual?
105		((*stops reading*))
106		(0.3) ((*chair scraping*))
107	F2	˚equal (.) what's- does that mean˚
108	F1	it's u::hmm *jämlikhet* ˚hhhHHuh˚
		It's u::hm equality ˚hhhHHuh˚
109	F2	˚>$ohkay$<˚.hh
110	F1	r(hh)ight?
111	T2	yeah? (.) Mm?
112		(1.6)
113	F1	uhm:: (.) (pt) maybe the::u:h (.) sssalary?
114	F2	˚yeah˚

115 (1.9)
116 F1 no the pay ( ) *JAH?* (pt) .hh(.) that's:: (0.5) no: (0.4)
117 it's [not equal bu::t=
118 F2 [˚should be˚
119 (1.8)
120 F1 it SHOUld be (0.3) I think
121 F2 yeah (.) ˚okay-˚ ((*whispered*))
122 F1 bu:t↓ .hh
123 (5.2)
124 F1 ((*sniffle*))
125 (1.5)
126 F1 >I don't ↑kno:w< I ↑think (3.2) ↑hmmm↓ ((*high-pitched*))
127 (2.1)
128 F1 *ja:a?*
129 F1 it- was- ha:rd- to::- ((staccato))
130 F2 yeah itws:::
131 F1 let's skip it
132 F2 ˚oh(hh)k(hh) hhhH[HUH˚
133 F1 [Hhhehhheh
134 T2 [hhuhhuhhuhhuh
135 F1 [hhumhhuh
136 F2 [(Christians)
137 F1 ˚huhhuh[huhhuh˚ ((*giggling*))
138 T2 [take another (0.4) card then
139 F2 is it my turn

Repair of the trouble, i.e. the lack of understanding of "equal", is initiated by F2 in line 107 with a quietly produced question: "˚*equal (.) what's- does that mean*˚". F1 offers a translation: "*it's u::hm jämlikhet*" (*it's u::hmm equality ˚hhhHHuh˚)*. F2 acknowledges receipt of this new information (line 109) in a smile voice, but since comprehension of the topic is essential for task management, F1 checks off her translation with a request for confirmation from T2 (*r(hh)ight*?). After T2 has provided affirmatives (line 111), F1 attempts to formulate a response to the topic card, with some difficulty (line 113): "*uhm:: (.) (pt) maybe the::u:h (.) sssalary*?". Her turn displays uncertainty as to how to address the topic, evidenced by delayed turn beginning, the use of the tentative "*maybe*", the prolonged 's' sound in "*salary*", and the rising intonation at the end. Her turn, though, is packaged in a way that fits as a response to the question in the topic card, i.e., that maybe salary differences may be one way of coming to terms with gender inequality. Despite agreement from F2, F1 appears unsatisfied and repairs her use of "*salary*" with "*pay*" (line 116). The remainder of her turn is fragmented with repeated restarts.

F2 appears to perceive F1's problems in formulating a topical point (line 117) and overlaps with "*should be*", produced quietly, which F1 recycles in line 120, adding "*I think*". F2 makes no verbal attempt to contribute further and F1 adds "*bu:t↓ .hh"*. The falling intonation does not

indicate a trailing off or a request for help; rather, it comes off almost as a 'shrug'.[6] There is a longer silence, a sniffle, and in line 126 F1 makes yet another attempt to find something to say on the topic, where her voice becomes high pitched and with rising intonation and emphasis on "*know*". Her turn-final *"↑hmmm↓*" and subsequent Swedish "*ja:a*?" (*well*) signal that she is persistently trying to keep the conversation going despite of lack of topical input. In line 129, she acknowledges the problem (*it- was- ha:rd- to:)* and F2 agrees, without specifically stating what was hard. F1 then offers a solution: "*let's skip it*"; i.e., that they agree on abandoning the topic. F1 acknowledges this proposal with an "°*oh(hh)k(hh) hhhH[HUH*°" produced through laughter – the laughter in itself perhaps marking the proposal as somewhat against the rules of the task. Both T2 and F1 join in, and in line 138, T2 confirms the abandonment as acceptable by asking the students to draw another card. F2, then, orients to the test format in her request for clarification as to whether it is her turn to draw a card.

It is obvious that F1 and F2 were having problems with this topic. T2 remains passive throughout their struggles and it is F1 who straightforwardly proposes the solution of moving on to another topic, whereas F2 seems unsure as to the appropriateness of such a strategy. Given the repeated signals of trouble and persistent attempts to perform the test task, F1's solution is a 'successful' one in terms of trouble resolution, and they transition away from what could possibly impinge negatively on their OP evaluation. It should be mentioned that there is nothing in the test instructions that hinders students from using 'skipping topics' as a test strategy. What is interesting for this study, then, is whether the students' task management appears to affect their subskill ratings, which is an issue for our comparative analysis presented further down. In any case, the task management strategy deployed here can be observed in F1's orientation to possibilities inherent in the task and the utilization of these as social resources for moving beyond the TRT. By doing so, it is possible that she displays a high degree of *test-wiseness* (cf. Bachman, 1990, p. 114).

Task-as-process: 'Interview'style

Yet another task orientation can be documented in Fragment (4). In this sequence, the teacher (T1) has an impact on the interactional trajectory. The topic, from part one of the fifth test, is the students' home environments and testees are instructed to talk about where they live and take turns to tell each other about their respective house/apartment, surroundings, neighbors, etc. After having listened to B2's account, B1 poses a question in line 31, "*whatu:h (.) do: you think about the place. shh?"*. B2 replies , after some delay, and restarts the projected action after "*I think it's very good when you (.) when you like*", followed by a 1.7 second pause. She restarts her attempt and self-repairs her use of "*stay*" with "*be*":

Fragment (4)

[510411072] T1: teacher 1, B1: male student, B2: female student

```
31    B1          whatu:h (.) do: you think about the place. shh?
32                (2.5)
33    B2          oh I love the place? (1.1) I:: think it's very good (hh)
34                >whenyou< (.) when you li:ke (1.7) when you like to be in
```

[6] Speculative, as data is audio only.

35		the country then you (0.8) should stay ↑there (1.2) you
36		should be: ↑there
37		(7.1)
38	B2	((*sniffle*)) *Sk' ja också ta en nu=*
		((sniffle)) shou' I also take one now=
39	T1	=mm
40		(4.1)
41		and you re:ad the que- (.) and you re:ad the statement
42		first.
43		(7.2)
44	B1	.hhh (.) hhhhhhm
45		(0.6)
46	T1	ohkay? (.) read it (.) l(hh)oud (hhh) huh
47	B2	hhHUH
48	B1	uhm *((reads))* and there is too much noise everywhere (.) at
49		school (.) at home (.) in shops[7] *((stops reading))*
50	T1	yeah
51		(2.1)
52	B1	uo:h
53		(4.8)
54	B2	°*ska jag också prata nu*°
		°***should I also talk now***°
55	T1	yes
56	B2	*jaha* uhm (.) u::h at school? I don't think it's too much
57		noise because in school you have to lea↑rn and there you
58		should ↑speak an::' (.) so on
59		(2.1)
60	T1	what do you say
61	B1	yes:: (.) and at home (1.8) it is kind- it can be w-wery
62		nois-(.) sy
63		(3.1)
64	T1	beca:use why
65	B1	we are a big family?
66	B2	°hhuh] huh huh°(.) ma:- me and my mum (.) ba:h (1.5) (pt)
67		u:h >I mean< at home? (.)it can't be so noisy
68		because it's only me and my mum hhuh huh huh
69		(3.9)
70	B2	°.hhuh°
71	T1	°perhaps you should visit Billy sometimes°
72	B2	>yeah m(hh)aybe< ((hh) hhuh huh HEH (2.8) in the shops I

[7] Pronounced 'chops'.

73		don't (.) think it's too much noise because (0.8) when you
74		go to a:: (.) when you go into a city and w'sit a shop
75		the:re (.) you know before that there are gonna ↑be: people
76		and then you
77		(1.5)
78	T1	mm
79	B2	probably don't think it's too much noise in the shops
80	T1	what do you say Billy
81	B1	I agree:?
82	T1	what about the music in the shops
83	B2	>that's good<
84	T1	uhu:hm?
85	B2	hhhuh huh HUH
86	T1	okay

Having completed her turn, there is a pause of 7.1 seconds, where neither B1 nor she herself makes any attempt to elaborate further. She then turns toward T1 and asks "*ska jag också ta en nu*" (*should I also take one now?)* (line 38). B2 thus appears to treat the prior task as completed, evidenced in her shift toward T1, the orientation toward the setup of the activity of test-taking, and the request for permission to draw another card. T1 confirms this; however, the silence continues and T1 requests that the student read the statement (lines 48-49). T1 seems to treat the silence (line 43) as reason for clarifying the request: perhaps that the student took T1's first request as an instruction to just read the card instead of reading it aloud. The other student, B1, reads the card[8], but shows some hesitation as to how to elaborate, as his initiated "*uo:h*" (line 52) followed by a longer pause indicates. Quietly, B2 turns toward T1 and asks, by codeswitching, "*ska jag också prata nu*" (*should I also talk now*) (line 54), which indicates that B2 treated the pause as an ambiguous sign that she, and not B1, perhaps is expected to contribute instead.

The apparent uncertainty about the 'rules' of the test constitutes an occasion of TRT, which on both occasions is managed with codeswitching requests for direction from T1. As T1 has affirmed, B2 initiates a topic commentary with a receipt token seemingly directed at T1: "*jaha*" in Swedish (in this case, '*okay',* line 56) before continuing with a topic-related comment on noise in school. Without awaiting further elaboration on the other parts of the task (at home, in shops), T1 assigns the next turn to B1 (line 60). He begins with an acknowledgement, displaying alignment with B2's dismissal of noise in schools and then volunteers a comment on the second thread in the task topic. The turn is equally short and T1 appears dissatisfied with the lack of explanation as to why his home is noisy. After providing one, B2 unpromptedly offers a reflection regarding her own home (66-68). She also picks up the third topic in line 72 (noise level in stores, cf. line 49). This turn is notably longer and contains support for her argument. B2's talk also displays frequent laugh particles. Jefferson (1984) observed that laughter that is unreciprocated by co-participants frequently occurs in contexts of troubles-tellings. As B2's laughter remains unreciprocated, it is not treated as an invitation to join in, but rather as a signal

[8] We have not been able to explain with any certainty how the reading of the new card gets assigned to B1 given that our data is audio only. It is possible that embodied actions could have provided direction.

to remain troubles-receptive. In this context, B2 displays some discomfort with T1's proposal that she should visit Billy[9]; by laughing, she can be heard as distancing herself from the delicacy of the proposal by showing that she is in a position to "take the trouble lightly" (Jefferson, 1984, p. 351). Without awaiting a next contribution from B1, T1 prompts him to react and he offers the task-specific response "*I agree*" (line 81). T1 again treats the response as insufficient in the offering of a new angle: the music in shops. This time, B2 responds fast (*that's good*) and the teacher employs a transitional "*okay*" (Beach, 1993) projecting acceptance and completion of the topic.

In this sequence, the testees exhibit difficulties in elaborating independently on the topics assigned, which T1 treats as problematic (evidenced in her repeated attempts to spur the discussion). As previous research has shown, examiner (here, teacher) style differs (A. Brown, 2003), and it may also be difficult for a teacher to watch her students fail to produce 'enough' output for a passing grade without at least attempting to steer them in a positive direction. In contrast to Fragment (3), where T2 appears to await trouble resolution despite repeated signals of trouble, T1 in Fragment (4) asks questions that are sequentially positioned in slots of silences, even in cases where silences can be considered to be of normal length (lines 60, 64, 80). Thus, it appears as if T1 treats B1's lack of contributions as evidence of the fact that he needs more encouragement and new angles on the topic in order to contribute, and perhaps also that B2 would be unable to offer such elicitative actions. However, by intervening on occasions where the pupils themselves had not indicated any particular trouble (long pauses or evident production difficulties), the teacher pre-empts the pupils' opportunities for establishing a 'natural' dialogue, resulting in an interview format. The task management orientations are performed in Swedish rather than in English and short responses are treated as inadequate. There are surprisingly few linguistic errors, word searches, or pronunciation difficulties, and although the turns are short, it is not *evident* from the linguistic packaging of turns that it is, in fact, lacking L2 skills causing the problems (even though it is possible that B2's inability to venture beyond first level comments is connected to L2 proficiency). A possible explanation for the differences in teacher intervention in Fragments (3) and (4) is that the B-dyad exhibits similar problems throughout the test, whereas the F-dyad managed other topics with more ease. It is also possible that a 'halo effect' (see e.g. Bechger, Maris, & Hsiao, 2010) plays a role for teacher conduct: teachers' prior knowledge of student performance creates expectations regarding their ability to overcome TRT, which results in varying intervening actions.

Task 'Resistance'

Finally, Fragment (5), illustrates another TRT strategy, namely displaying resistance toward managing the task in a way that the teacher or task instructions indicate. The task is from part one of the test, i.e., the same task as in Fragment (4). E2 begins in line 01, but the teacher (T1) uses E2's possible completion to point out that he neglected to introduce himself, which was part of the instructions. [10]

[9] Possibly 'embarrassment-resistance', cf. Sandlund (2004).

[10] The introductions were necessary for student identification since the tests were to be assessed.

Fragment (5)

[521312082] T1: teacher 1, E1: female student, E2: male student

01	E2	u:h I live in a: kind of big house? ((*monotonous voice*))
02		(1.1)
02	T1	((*sniffle*)) but a:- (.) who <u>are</u> you hhhuH
04	E1	hhhAh .hh
05	E2	>*jaha*:< (0.9) my name is Michael and I live in a:: (0.6)
07		kind of big house? (1.5) and u:h (.) I got (1.2) two:: (.)
08		°u:h° <u>laws</u> (0.4) <u>lawns</u> (.) and (.) I can play football
09		there (.) play compu:ter (.) but- in th'house? (.) u:hhh
10		(.) yeah that's it? .hhh
11		(1.0)
12	T1	that's (.) ↓<u>it</u>↑
13	E2	*ja.hh (.) >det är det väl?< (.) eller ska jag <u>fråga</u> också*
		Yeah.hh (.) >I guess?< (.) or am I supposed to ask also
14	T1	°↑nyie::° (.) do <u>you:</u> have athing- anything to ask him
15	E1	ou:hm (.) °no:? (.) <u>I</u> don't think so (hhhh)Eh .hhh° u:h
16		(.) my name is E:lin ((*cont.*))

In response, E2 produces a codeswitched change of state token "*jaha*"[11] and after producing the requested action he recycles his first turn, but continues with additional descriptive components (lines 05-09). After a brief pause, he adds "*yeah that's it?*", explicitly signaling the completion of his task. T1 displays dissatisfaction with his contribution through her second-position repeat of E2's formulation as a question/understanding check, possibly displaying "surprise or incredulity" (Kim, 2002, p. 51) with rising intonation and emphasis on both words (line 12). E2 responds in Swedish (line 13): "*ja.hh (.) >det är det väl?< (.) eller ska jag <u>fråga</u> också"* (*yeah I guess or am I supposed to ask also?*). T1, however, does not attend to E2's question with more than an ambiguous receipt/possible negation (*nyie::*), but instead turns to E1: "*do you: have athing- anything to ask him?*". E1 responds with disagreement "*ou:hm (.) °no:? (.) I don't think so*" and begins her own introduction (line 16).

The students' treatment of the task is similar to Fragment (4) in the displayed uncertainties regarding the rules of the test, but differs in terms of their displayed perception of what is expected in terms of topical treatment. E2 explicitly marks that the topic has been exhausted on his behalf (line 10), which shows an orientation to the topics as 'questions to be answered' rather than as opportunities for developing a topic. This understanding of the task is further evidenced in his response to T1's recycling of his turn in line 12, where he yet again states that he considers the topic as closable. His uncertainty, though, can be discerned in his addition of "*I guess*".[12] E1 displays a similar orientation, i.e., she does not treat T1's question (line 14) as evidence that she

[11] Swedish equivalent to 'oh' as a change of state token, cf. Heritage (1984).
[12] Swedish *'väl'*.

is expected to appear 'interested' in E2's account and ask related questions; instead, she responds negatively, i.e., that she does not have anything she would like to know more about.

The sequence shows a relatively stable pattern of this particular dyad, where the interactants fail to expand on topics. We argue that they orient to the test task almost as if it were a written exam where there is a correct 'answer'. This understanding becomes visible in their unwillingness to align with the setup of 'feigning' authentic interest in order to produce gradable output in English. By not aligning with institutional constraints of the test procedure, the students repeatedly display resistance (cf. Hellermann & Pekarek Doehler, 2010) toward the task, something which cannot be explained solely by lacking OP skills. In this dyad, as in several others in the sample, codeswitching to Swedish occurs almost exclusively in slots where uncertainties regarding the task are being negotiated. This task-oriented codeswitching practice appears to be deployed not only as a display of lack of understanding of the task, or trouble formulating questions in their L2, but also as displays of resistance toward the test tasks. Task management is also Similarly, Ustunel and Seedhouse (2005) observed that through codeswitching, learners were able to display their alignment or disalignment with the teacher's pedagogical focus, something which may also be the case here.[13]

Comparison with Assessment Data

Is there, then, any correspondence between the different ways of managing task-related trouble in students' L2 and their assessed level of OP? The assessment data used for comparison include the students' OP grade on this particular test, the overall OP grade on all five tests, and their final grade in English. In addition, we have included the raters' assessment of the students on three subskills (a, i, and j): (a) *overcoming difficulties in communication,* (i) *interactional ability*, and (j) *treatment of topic* (see Table 1).[14]

[13] See also Fragment (1) where a similar test task understanding check was performed in English.

[14] The remaining subskills primarily targeted their linguistic competence and were considered less relevant for the scope of the present study.

Table 1

Assessment data for students in the sample.

Dyad	Stud. ID	Gender (M/F)	Subskill score (a) (1-5)	Subskill score (i) (1-5)	Subskill score (j) (1-5)	OP grade: Test 5 (1-6)	OP grade overall: Tests 1-5 (1-6)	Final grade in English*
A	1	M	4.0	4.0	3.0	3.00	1.60	G
	2	M	4.0	4.0	3.0	3.00	3.07	G
B	1	M	4.0	2.7	2.0	2.00	1.73	G
	2	F	4.0	2.7	2.7	4.00	3.87	VG
C	1	M	3.3	3.0	2.7	2.67	1.73	G
	2	F	3.0	3.0	2.7	3.33	2.80	G
D	1	F	4.0	3.0	3.0	2.67	1.87	G
	2	F	4.7	4.3	3.7	4.00	3.67	VG
E	1	F	3.5	2.7	2.3	2.67	2.00	G
	2	M	3.5	3.5	3.0	3.00	2.73	G
F	1	F	4.3	4.7	4.3	5.00	5.27	MVG
	2	F	5.0	4.7	4.0	4.67	3.47	VG
G	1	F	5.0	4.7	4.7	5.33	5.13	MVG
	2	F	5.0	4.7	4.3	5.33	4.27	VG
H	1	F	5.0	5.0	5.0	6.00	5.00	MVG
	2	M	5.0	5.0	5.0	6.00	4.27	MVG
I	1	F	4.0	3.7	4.3	5.00	4.73	MVG
	2	F	4.7	3.7	3.7	3.67	3.67	VG
J	1	F	4.3	4.0	3.7	4.67	4.67	MVG
	2	F	4.0	3.7	3.7	4.33	3.40	VG

* F = Fail; G = Pass; VG = Pass with distinction, MVG = Pass with special distinction

General observations

As a general observation, it can be noted that students in dyads A to E received a lower final grade in English as compared to students in dyads F to J (with two exceptions, student B2 and D2). The A to E dyads also scored generally lower on both test 5 and on the overall OP grade compiled from all five tests. Thus, two student categories emerge: a *Low OP group* and a *High OP group*, representing dyads A to E and F to J, respectively.[15] Consequently, students in our sample generally interact with peers of similar assessed OP levels as themselves, something which can indeed impact the emergence and management of TRT (cf. Davis, 2009).

As for subskill (a), i.e. the ability to independently overcome communication difficulties in English, the same pattern applies, even though that difference between groups is relatively small. There are also (relatively small) differences between the two groups for subskills (i), interactional ability, and (j), treatment of topic. A gender factor may also be noted in that the High OP group consists of only one male student, whereas the Low OP group has a 50/50 gender distribution. Gender might play a role in dyadic setups (cf. O'Loughlin, 2002; O'Sullivan, 2002); however, the

[15] The mean OP grade on test 5 (i.e., the test from which our five fragments stem) was 3.03 for the Low OP group, as opposed to 5.00 for the High OP group.

scores also reflect the fact that girls usually do better than boys in language studies (Björnsson, 2005; Klapp Lekholm, 2008).

Task Management and OP Scores

Fragment (1) illustrated a sequence of task management where students displayed awareness and understanding of the test task, but also allowed for their discussions to operate relatively freely, with joint topic elaboration and closing. As Table 1 shows, both H-students obtained the highest possible scores on all points of measure. In Fragment (2), students F1 and F2 display diligent orientations to the task of agreeing or disagreeing with the statements, which tended to lead to fast topic decay. These students also had high scores on test 5. However, the fact that they both scored slightly lower on subskill (j), i.e., treatment of topic (see Table 1), may reflect their explicit task orientation and the fact that the agreement instructions were deployed primarily as pre-closings rather than as elicitors of more topical talk. At this stage, we make a tentative claim that various strategies for treating the test task, and assessments of their treatment of topic, are interrelated. Interestingly, F1, who acts as the main manager of the test task in Fragment (2), scored lower on subskill (a) (ability to overcome communication difficulties) than her interlocutor F2, who, in both Fragments (2) and (3), contributes less. In Fragment (3), F1 displays repeated TRT and initiates the task abandonment strategy; however, despite the fact that she makes repeated attempts to overcome the TRT and also finally offers a solution, she is rated lower on subskill (a). Although speculative, we offer three possible explanations here. First, it is possible that the very *display* of trouble, topic-related or other, results in a lower rating, and that her persistent attempts to overcome the TRT is overshadowed by the displays of trouble. Although we cannot determine this with any certainty in our data, this speculative explanation can be compared with Seedhouse and Egbert's (2006, p. 193) observation that high scorers in the IELTS Speaking Test initiated fewer or no repairs related to comprehension problems. Second, opting for the strategy of abandoning a topic may result in a lower rating on topic management *if* raters understand subskill (j) as 'the ability to treat *any* given topic'. Third, and particularly in contrast to Fragments (4) and (5), this particular teacher (T2) awaits trouble resolution instead of intervening. Given that teachers know their students and have a general perception of each student's L2 ability and overall cognitive level, it is possible that T2's expectations on these two (High OP) students plays a central role in her choice to remain passive despite overt displays of trouble. As a contrast, teacher 1 in Fragment (4) intervenes even when there are no or few displays of trouble.

Given that teacher conduct differs (cf. A. Brown, 2003), it is remarkable to note that B1 (Fragment 4) received a high score on subskill (a) (ability to overcome communication difficulties) despite the fact that he did not contribute much except for when prompted by his teacher. Thus, he did not actually make many independent attempts to overcome trouble and hardly displayed any overt trouble, other than short turns and few initiatives. Given that B1 scored low on all other measurements, it is possible that *any* attempt to overcome trouble in a conversation with few contributions becomes overrated. As for B2, she also scored high on subskill (a), although her management of TRT generally was performed through codeswitching; however, she scored low on the other subskills, which indicates that her general management of topic tasks was unsatisfactory. Moreover, data indicates that codeswitching occurs almost exclusively in segments where there is uncertainty about what to do next.

In Fragment (4), T1's involvement results in a question-answer format where students treat the task as teacher-directed rather than peer-driven. It does appear as if Low OP students are less proficient task managers, which may be an interaction of both linguistic proficiency and general scholastic aptitude. In Fragment (5), this observation is strengthened, as two Low OP students have difficulties conforming to the test format, displaying unwillingness to 'play the test-taking game'. In their conduct, they display an understanding of the task that does not entail acceptance of the fact that it is considered important to contribute to the dialogue, ask questions, act interested in their interlocutor's contributions, etc. E1 scored particularly low on her treatment of topic, but both E1 and E2 displayed similar understandings of the task in their conduct; they treated the topic cards as something that required a (minimal) answer, rather than as opportunities for developing a topic. Gan (2010) notes that lower-scoring students may actually be restricted by pre-set test prompts, which was visible in their failure to develop topical talk on prompts. It is possible that the topic cards used in the test in our study had similar effects on the Low OP group, since they did not find "something to say" (Dewey, 1976, p. 35), content or language-wise, on certain topics.

Conclusions and Implications

Despite problems associated with comparing holistic assessment data with selected fragments of interaction, we argue that the findings accounted for do have intriguing implications. First, it appears that certain types of task management and rater assessment of L2 oral proficiency are interrelated. This may not seem surprising – it seems likely that L2 proficiency would be beneficial when it comes to comprehending and discussing a given topic. As demonstrated, different strategies for task management and the negotiation of TRT appear to reflect students' differing orientations to the test format, i.e., to the task-as-workplan (Breen, 1989; Hellermann & Pekarek Doehler, 2010). Students who were assessed as highly proficient displayed task management characterized by taking the task-as-workplan as a starting point but accomplishing the task relatively freely from the instructions. In contrast, 'excessive' task management and, on a falling scale, task abandonment and task 'resistance', appear to be rated less favorably by teachers and raters. The 'excessive' task managers also belonged to the High OP group; however, their explicit orientation to the test task and repeated commentary on the assignment is perhaps reflected in their slightly lower ratings on subskill (j) (treatment of topic). We argue that raters' and teachers' assessment of topical treatment is selective in that it involves specific understandings of what is 'good' and 'bad' task management. Thus, even though the 'excessive' task managers (Fragment 3, F1/F2) have a productive dialogue, their task-as-accomplishment is rated, relatively speaking, as less successful, even though it has been argued that meta-talk about a task (in the classroom) is indicative of "successful students" (Sarangi, 1998, p. 106). Accordingly, we want to underscore (1) that task instructions, participants' displayed management of test tasks, and raters' underlying assumptions about successful task management all merit more empirical attention, and (2) that task management appears to be linked with the students' *assessed* ability to overcome communication problems and their *assessed* treatment of topic. As an example, for weak learners 'management of task' appeared to overshadow 'development of conversation', so that the task itself became a topic.

Moreover, we noted some peculiarities regarding the ratings. Although strong task managers scored high on all scores and grades, there is by no means a clear-cut relationship between their interactional conduct and the subskill ratings. For one, subskill (a) (ability to overcome communication difficulties) tended to be awarded relatively high scores for all students,

something we did not find sufficient evidence for in the fragments examined. Moreover, the teachers' conduct elicited, but also pre-empted, student talk. Even though peer dyads differ in a number of important ways from OPIs involving one examiner and one candidate, the teacher is, as Brown (2003, p. 1) notes about OP interviewers, "intimately (…) implicated in the construction of candidate proficiency". In addition, the teachers' interventions, or lack thereof, gave us important clues as to what they treated as dispreferred actions, and we observed that 'doing-being a successful task manager' means 'playing the game' and showing willingness to 'feign' interest in topics and interlocutor contributions.

We argue that the diverging understandings of the test task displayed by learners in our study become part of how they are assessed and that certain task management strategies are rated less favorably than others. It is possible that learners' task management becomes part of the rating of *all* three subskills examined, exemplified as task-related codeswitching (subskill a), the ability to solve TRTs and help each other out (subskill i), and the ability to expand on a topic and to close it once both parties understand it as exhausted (subskill j). We also claim that the application of CA accentuated the complexities of evaluating students individually on a joint achievement. In essence, it is possible that assessments of at least *some* students reflect the interaction as such rather than students' individual achievement, which in turn should be further addressed in language testing research on the validity of assessment scales (cf. Bachman, 1990; Lazaraton, 2002). By way of example, questions for further inquiry arising from our study include the impact of being paired up with an interlocutor stronger or weaker than oneself and the evaluation of displayed trouble on topics where students really had nothing to say. Our findings on task 'resistance' provide additional support for work that has encouraged flexibility in topic choice (Naughton, 2006; Riggenbach, 1998). Furthermore, they also pinpoint the importance of how raters perceive different task management strategies, as topic abandonment appeared to be rated less favorably even though this solution was productive in terms of *accomplishment* of the task. Finally, we would like to encourage more studies using post-rating interviews with raters, as such explorations could shed additional light upon the relative role of task management for OP test success.

References

Atkinson, J. M., & Heritage, J. (Eds.). (1984). *Structures of social action*. Cambridge: Cambridge University Press.

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

Beach, W. A. (1993). Transitional regularities for 'casual' "okay" usages. *Journal of Pragmatics, 19*(4), 325.

Bechger, T. M., Maris, G., & Hsiao, Y. P. (2010). Detecting halo effects in performance-based examinations. *Applied Psychological Measurement, 34*(8), 607-619.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Björnsson, M. (2005). Kön och skolframgång: Tolkningar och perspektiv. Stockholm: Myndigheten för skolutveckling.

Breen, M. P. (1989). The evaluation cycle for language learning tasks. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 187-206). Cambridge: Cambridge University Press.

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

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment. Principles and classroom practices* (2 ed.). White Plains, NY: Pearson Education.

Carrol, D. (2000). Precision timing in novice-to-novice L2 conversations. *Issues in Applied Linguistics, 11*(1), 67-110.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26*(3), 367-396.

de Jong, J. H. A. L., & van Ginkel, L. W. (1992). Dimensions in oral foreign language proficiency. In L. Verhoeven & J. H. A. L. de Jong (Eds.), *The construct of language proficiency: Applications of psychological models to language assessment* (pp. 187-205). Amsterdam: John Benjamins.

Dewey, J. (1976). *The school and society. The collected works of John Dewey: The middle works, 1899-1924, volume 1. Edited by Jo Ann Boydston.* Carbondale, IL: Southern Illinois University Press.

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

Egbert, M. (1998). Miscommunication in language proficiency interviews of first-year German students: a comparison with natural conversation. In R. Young & A. W. He (Eds.), *Talking and testing. Discourse approaches to the assessment of oral proficiency.* (pp. 147-169). Amsterdam: John Benjamins.

Erickson, G., & Börjesson, L. (2001). Bedömning av språkfärdighet i nationella prov och bedömningsmaterial. In R. Ferm & P. Malmberg (Eds.), *Språkboken* (pp. 255-269). Stockholm: Myndigheten för skolutveckling.

Firth, A., & Wagner, J. (1998). SLA property: No trespassing! *The Modern Language Journal, 82*(1), 91-94.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18*(3), 299-323.

Fulcher, G., & Márquez Reiter, R. (2003). Task difficulty in speaking tests. *Language Testing, 20*(3), 321-344.

Galaczi, E., D. (2008). Peer-peer interaction in a speaking test: the case of the First Certificate in English examination. *Language Assessment Quarterly, 5*(2), 89-119.

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher- and lower-scoring students. *Language Testing, 27*(4), 585-602.

Gan, Z., Davison, C., & Hamp-Lyons, L. (2008). Topic negotiation in peer group oral assessment situations: A conversation analytic approach. *Applied Linguistics, 30*(3), 315-344.

Hasselgren, A. (1996). *Kartlegging av kommunikativ kompetanse i engelsk. User compendium for teachers.* Oslo: Nasjonalt læremiddelsenter.

Hasselgren, A. (1997). Oral test subskill scores: What they tell us about raters and pupils. In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment - Proceedings of LTRC 96* (pp. 241-256). Jyväskylä: University of Jyväskylä and University of Tampere.

Hellermann, J. (2009). Practices for dispreferred responses using *no* by a learner or English. *IRAL, 47*(1), 95-126.

Hellermann, J., & Pekarek Doehler, S. (2010). On the contingent nature of language-learning tasks. *Classroom Discourse, 1*(1), 25-45.

Heritage, J. (1984). A change-of-state token and aspects of its sequential placement. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action* (pp. 299-345). Cambridge: Cambridge University Press.

Iwashita, N. (2001). The effect of learner proficiency on interactional moves and modified output in nonnative-nonnative interaction in Japanese as a foreign language. *System, 29*(2), 267-287.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24-49.

Jefferson, G. (1984). On the organization of laughter in talk about troubles. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action* (pp. 346-369). Cambridge: Cambridge University Press.

Kasper, G. (2006). Beyond repair: Conversation analysis as an approach to SLA. *AILA Review, 19*, 83-99.

Kasper, G., & Ross, S. (2003). Repetition as source of miscommumication in oral proficiency interviews. In J. House, G. Kasper & S. Ross (Eds.), *Misunderstanding in social life. Discourse approaches to problematic talk* (pp. 82-106). Harlow: Longman/Pearson Education.

Kim, H. (2002). The form and function of next-turn repetition in English conversation. *Language Research, 38*(1), 51-81.

Klapp Lekholm, A. (2008). *Grades and grade assignment: Effects of student and school characteristics.* PhD, Acta Universitatis Gothoburgensis, Göteborg. Retrieved from http://gupea.ub.gu.se/dspace/handle/2077/18673.

Kormos, J. (1999). Simulating conversations in oral-proficiency assessment: a conversation analysis of role playes and non-scripted interviews in language exames. *Language Testing, 16*(2), 163-188.

Kurhila, S. (2004). Clients or language learners - Being a second language speaker in institutional interaction. In R. Gardner & J. Wagner (Eds.), *Second language conversations* (pp. 58-74). London: Continuum.

Lazaraton, A. (1992). The structural organization of a language interview: A conversation analytic perspective. *System, 20*(3), 373-386.

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests*. Cambridge: Cambridge University Press.

Lazaraton, A., & Davis, L. (2008). A microanalytic perspective on discourse, proficiency, and identity in paired oral assessment. *Language Assessment Quarterly, 4*(4), 313-335.

Lennon, P. (1990). Investigating fluency in EFL: a quantitative approach. *Language Learning, 40*(3), 387-417.

Levis, J. M. (2006). Pronunciation and the assessment of spoken language. In R. Hughes (Ed.), *Spoken English, TESOL and applied linguistics: Challenges for theory and practice* (pp. 245-270). Basingstoke: Palgrave Macmillan.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Oxford: Blackwell.

Mori, J. (2007). Border crossings? Exploring the intersection of second language acquisition, conversation analysis, and foreign language pedagogy. *The Modern Language Journal, 91*(Focus Issue), 849-862.

Naughton, D. (2006). Cooperative strategy training and oral interaction: Enhancing small group communication in the language classroom. *The Modern Language Journal, 90*(2), 169-184.

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing, 19*(2), 169-192.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing, 19*(3), 277-295.

Pappamihiel, N. E. (2002). English as a second language students and English language anxiety: Issues in the mainstream classroom. *Research in the Teaching of English, 36*(3), 327-355.

Plejert, C. (2004). *To fix what's not broken: Repair strategies in non-native and native English conversation.* PhD, Linköping University, Linköping.

Pomerantz, A. (1984). Agreeing and disagreeing with assessments: some features of preferred/dispreferred turn shapes. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action* (pp. 57-101). Cambridge: Cambridge University Press.

Rasmussen, G., & Wagner, J. (2000). Reparationer i international, interlingual kommunikation [Repairs in international, interlingual communication]. In M. F. Nielsen, G. Rasmussen & J. Steensig (Eds.), *MOVIN PUBLICATIONS No. 1.* http://www.conversation-analysis.net.

Riggenbach, H. (1998). Evaluating learner interaction skills: Conversation at the micro level. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 53-67). Amsterdam: John Benjamins.

Romova, Z., & Neville-Barton, P. (2007). "I have a lot more to say than actually I am able to." A study of oral skills development of undergraduate EAL learners. *New Zealand Studies in Applied Linguistics, 13*(2), 1-15.

Sacks, H. (1992). *Lectures on conversation. Volume 2. Edited by G. Jefferson and E. A. Schegloff*. Oxford: Blackwell.

Sacks, H., Jefferson, G., & Schegloff, E. A. (1974). A simplest systematics for the organization of turn-taking in conversation. *Language 50*(4), 696-735.

Sandlund, E. (2004). *Feeling by doing: The social organization of everyday emotions in academic talk-in-interaction.* PhD, Karlstad University Studies, 2004:36, Karlstad.

Sarangi, S. (1998). 'I actually turn my back on [some] students': The metacommunicative role of talk in classroom discourse. *Language Awareness, 7*, 90-108.

Schegloff, E. A. (2000). When others' initiate repair. *Applied Linguistics, 21*(2), 205-243.

Schegloff, E. A., Koshik, I., Jacoby, S., & Olsher, D. (2002). Conversation analysis and applied linguistics. *Annual Review of Applied Linguistics, 22*, 3-31.

Seedhouse, P. (2005). "Task" as research construct. *Language Learning, 55*(3), 533-570.

Seedhouse, P., & Egbert, M. (2006). The interactional organisation of the IELTS speaking test. *IELTS Research Reports* (Vol. 6, pp. 161-206).

Sundqvist, P. (2008). "In school, you're here for talking" - Assessment of oral proficiency in the EFL classroom. In J. Granfeldt, G. Håkansson, M. Källkvist & S. Schlyter (Eds.), *Language learning, language teaching and technology: Papers from the ASLA symposium in Lund, 8-9 November, 2007.* (pp. 251-269). Uppsala: Swedish Research Press.

Sundqvist, P. (2009). *Extramural English matters: Out-of-school English and its impact on Swedish ninth graders' oral proficiency and vocabulary.* PhD, Karlstad University Studies, 2009:55, Karlstad. Retrieved from http://kau.divaportal.org/smash/record.jsf?pid=diva2:275141.

Ustunel, E., & Seedhouse, P. (2005). Why that, in that language, right now? Code-switching and pedagogical focus. *International Journal of Applied Linguistics, 15*(3), 302-325.

Wagner, J. (1998). On doing being a guinea pig - A response to Seedhouse. *Journal of Pragmatics, 30*(1), 103-113.

Wagner, J., & Gardner, R. (2004). Introduction. In R. Gardner & J. Wagner (Eds.), *Second language conversations* (pp. 1-17). London: Continuum.

Young, R., & He, A. W. (Eds.). (1998). *Talking and testing. Discourse approaches to the assessment of oral proficiency*. Amsterdam: John Benjamins.

Young, R. F., & Milanovic, M. (1992). Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition, 14*(4), 403-424.

Appendix

Transcription key (adapted from Atkinson & Heritage, 1984, pp. ix-xvi)

:	Colon(s). Extended/stretched sound, syllable or word. Not only vowel sounds.
Underlining	Emphasis.
(.)	Brief micropause of less than (0.2) seconds
(1.8)	Timed pause, within or between turns
(( ))	Double parentheses, notation of scenic details
( )	Transcriptionist doubt
.	Period: falling pitch
?	Rising vocal pitch, not necessarily a question
↑↓	Marked rising and falling shifts in intonation
°word°	Passage of talk noticeably quieter than surrounding talk
[ ]	Overlap
!	Animated tone
Whe-	Hyphens: halting or abrupt cut off sound or word
< >	Noticeably quicker (> <) or slower (< >) than surrounding speech
hhh	Audible aspiration, possibly laughter
.hh	Audible inbreath
Mo(hh)re	Within-speech aspiration, possibly laughter
(pt)	Lip smack
Hah heh hoh	Relative open/closed position of laughter
$	Smile voice
MINE	Speech noticeably louder than surrounding speech
→	Indication to readers to pay special attention to line in transcript