Lam, D.M.K. (2015) Contriving authentic interaction: Task implementation and engagement in school-based speaking assessment in Hong Kong. In G. Yu & Y. Jin, *Assessing Chinese learners of English: Language constructs, consequences and conundrums* (pp.38-60). Basingstoke: Palgrave Macmillan. Reproduced with permission of Palgrave Macmillan.

Contriving authentic interaction: Task implementation and engagement in school-based speaking assessment in Hong Kong[1]

Daniel M.K. Lam

ABSTRACT

This chapter examines the validity of the Group Interaction task in a school-based speaking assessment in Hong Kong from the perspectives of task implementation and authenticity of engagement. The new format is intended to offer a more valid assessment than the external examination by eliciting 'authentic oral language use' (HKEAA, 2009, p.7) in 'low-stress conditions' (p.3), and emphasizes the importance of flexibility and sensitivity to students' needs in its implementation. Such a policy has then been translated into diverse assessment practices, with considerable variation in the amount of preparation time given to students. The present study draws on three types of data, namely 1) students' discourse in the assessed interactions, 2) stimulated recall with students and teachers, and 3) a mock assessment, where the group interaction task, the preparation time, and the post-interview were all video-recorded. Results show that while the test discourse exhibits some features that ostensibly suggest authentic interaction, a closer examination of students' pre-task planning activities reveals the contrived and pre-scripted nature of the interaction. Implications for the assessment of students' interactional competence and recommendations for task implementation are discussed.

Keywords: group speaking assessment, task implementation, authenticity of engagement, interactional competence

1. Introduction

In 2007, a School-based Assessment (SBA) component combining the assessment of speaking with an extensive reading/viewing program was introduced into the Hong Kong Certificate of Education Examination (HKCEE). Having operated on a trial basis for several years, SBA is now fully integrated in the new secondary school exit examination, the Hong Kong Diploma of Education Examination (HKDSE), since 2012.

The SBA component accounts for 15% of the total subject mark for HKDSE English Language, consisting of two parts. Part A is made up of two assessments, one *individual presentation* and one *group interaction* (otherwise commonly known as the 'group discussion' task), with one to be carried out in Secondary 5 (S5) and the other in Secondary 6 (S6). The speaking tasks are based on an extensive reading/viewing program. Therefore, students engage in either an individual presentation or a group discussion on the books they have read or movies they have viewed. Part B consists of one assessment in either the group interaction or individual presentation format, based on the Elective Modules (e.g. social issues, workplace communication) taught in the upper secondary curriculum. This is to be carried

[1] This extract is taken from the author's original manuscript and has not been edited. The definitive, published, version of record is available here: http://www.palgrave.com/us/book/9781137449771

out either in the second term of S5 or anytime during S6. Thus, a total of three marks (each weighing 5%) are to be submitted by the teacher. Further details of the SBA assessment tasks can be found in the Teachers' Handbook (HKEAA, 2009) available online.

This study focuses on the *Group Interaction* task, whereby students in groups of three to five (mostly four) carry out a discussion of around eight minutes. While the peer group interaction format has been used in the public exam for many years, the SBA task differs from its public exam counterpart in that students would be interacting with their classmates rather than unacquainted candidates, and are assessed by their own English teacher instead of unfamiliar external examiners. Moreover, one of the discussion tasks would be based on a book or movie that students have experienced as part of the extensive reading/viewing program.

The objectives of the SBA initiative are to elicit and assess 'natural and authentic spoken language' (HKEAA, 2009, p.7), providing an assessment context 'more closely approximating real-life and low-stress conditions' (p.3), and for students to 'interact in English on real material' (Gan, Davison, & Hamp-Lyons, 2008). Thus, the assumption is that *authentic oral language use* constitutes the basis of the validity of the assessment task, as has been reiterated in the published guidelines (HKEAA, 2009) and in validation studies (Gan *et al*., 2008; Gan, 2010).

As an assessment-*for*-learning initiative, the assessment policy for SBA places considerable emphasis on flexibility and sensitivity to students' needs in the design and implementation of the assessment tasks, a marked departure from the public exam where standardized tasks, conditions, and practices are strictly adhered to for reliability and fairness. As stated in the Teachers' Handbook,

> the SBA process, to be effective, has to be highly contextualised, dialogic and sensitive to student needs (i.e. the SBA component is not and cannot be treated as identical to an external exam in which texts, tasks and task conditions are totally standardised and all contextual variables controlled; to attempt to do so would be to negate the very rationale for SBA, hence schools and teachers must be granted a certain degree of trust and autonomy in the design, implementation and specific timing of the assessment tasks). (HKEAA, 2009, p.4)

The recommended practice is for teachers to give students the 'general assessment task' to prepare a few days in advance, and to release the 'exact assessment task' shortly before the assessment to avoid students memorizing and rehearsing the interaction (*ibid*., p.37).

Although some recommendations for task implementation are included in the Teachers' Handbook and in teacher training seminars, the emphasis on flexibility in the assessment policy has translated into diverse assessment practices (see Fok, 2012). There is considerable variation in when the discussion task with question prompts is released to students, in other words, in the length of preparation or pre-task planning time during which students have the opportunity to talk to group members about the upcoming assessed interaction (Note: the term *preparation time* is used in official documents published by HKEAA, whereas *pre-task planning time* is used extensively in the SLA and language testing literature. The two terms are used synonymously in this chapter). Varied practices in task implementation are evident, both in previous studies and my own. Gan *et al*. (2008) and Gan (2012) reported that the specific assessment task was made known to students about 10 minutes beforehand. In the school that Luk (2010) investigated, students received the discussion prompt one day before the assessment, which was also when they were told who their group members are. Of the eight schools whose teachers Fok (2012) interviewed, four gave students the actual

discussion questions one day or more before the assessment, three gave students similar sample questions a few days before but the actual questions only minutes before the assessment, and one allowed no preparation at home but gave students the actual questions shortly prior to the assessed interaction. As for the two schools in my own study, one (School L) released the discussion prompt to students 10 minutes before the assessment, and group members were not allowed to talk to each other during preparation time. The other school (School P) released the discussion prompt to students a few hours before the assessment, and students who formed their own group could plan their interaction together.

Such variation in the pre-task planning time allowed generates group interactions that are considerably different in nature. As will be seen, students having a few hours or more to prepare display an overwhelming tendency to pre-script an interactive dialogue followed by reciting and acting out the scripted dialogue, rather than participating in a spontaneous interaction as students having only 10-15 minutes of planning time do. This chapter explores what students do during the preparation time and how it affects their subsequent group interaction; and examines whether the task, as it is implemented, elicits authentic oral language use. Before outlining the details of data and methodology, I shall review some previous research relevant to this study.

2. Literature Review

Since its implementation, there has been a growing body of research that examines different facets of SBA. One strand of research looked at perceptions towards the SBA initiative by various stake-holders, for example, teachers' and students' initial responses at the first stage of implementation (Davison, 2007); students' and parents' views (Cheng, Andrews, & Yu, 2011); and teachers' perceptions and readiness of administering SBA at the frontline (Fok, 2012). Another strand of research focused on the assessed speaking performance. Some studies engaged in micro-analysis of the test discourse and students' interaction (Gan, Davison, & Hamp-Lyons, 2008; Gan, 2010; Luk, 2010), to be reviewed in more detail below. Others compared the discourse output elicited by the two task types (Gan, 2012), and examined the extent to which students' personality (extroversion/introversion) influences their discourse and test scores (Gan, 2011). At a more theoretical level, Hamp-Lyons (2009) outlined a framework of principles guiding the design and implementation of large-scale classroom-based language assessment, drawing on the case of SBA in Hong Kong.

2.1 Validity of SBA Group Interaction

Validation studies of the SBA Group Interaction task to date have yielded mixed results regarding whether the task has achieved its aim of eliciting students' authentic oral language use. Gan, Davison, & Hamp-Lyons (2008) presented a detailed conversation analysis of one group interaction from a databank of 500, focusing on topic organization and development. They identified two types of topic shifts: 'marked topic shifts', where the speaker used particular turn design features to signal the introduction of a new topic, and 'stepwise topic shifts', where the speaker referred to the content in the previous turn and introduced new elements as something relevant. The authors concluded that the similarities in topic negotiation and development to everyday conversation serve as evidence for authenticity, hence validity, of the task.

In another study, Gan (2010) compared the students' discourse in a higher-scoring group and a lower-scoring group from the same databank of 500. He found that, in the higher-scoring group, participants responded contingently to each other's contributions. By fitting their comments closely to the previous speakers' talk, these participants displayed understanding of the preceding discourse. Participants in the lower-scoring group, by contrast,

often reacted minimally. Their discourse was more 'structured' and reliant on the question prompts, but there was also some negotiation of form and meaning, where students helped one another search for the right forms to express meaning. In alignment with Gan *et al.* (2008), he concluded that the discourse exhibited characteristics of an authentic task that 'emphasize[s] genuine communication and real-world connection' and 'authentically reflects candidates' interactional skills' (Gan, 2010, p.599).

The study by Luk (2010) painted a considerably different picture. She found the group interactions characterized by features of ritualized and institutionalized talk rather than those of everyday conversation. In her discourse analysis of 11 group interactions involving 43 female students in a secondary school, participants were seen to engage in orderly turn-taking practices with turns passed on in an (anti-)clockwise direction, and to front those speaking turns in which each member delivered extended, pre-planned speech before the whole group started giving responses. There was little evidence of on-line interaction and contingent responses to previous speaker contribution, manifested in the frequently deployed surface agreement (e.g. 'I agree with you') that came without further elaboration, therefore appearing superficial and possibly perfunctory. Students also avoided seeking clarifications from each other, but concealed problems instead. These findings mirrored those of He & Dai (2006) on the group discussion task in the College English Test in China, where candidates were observed to exploit the time when others were speaking to organize and formulate their own ideas in upcoming turns, and accordingly, to focus on expressing their own ideas rather than responding actively and relevantly to previous speakers' talk. With students' interview responses as supplementary evidence, Luk (2010) concluded that students were engaging in the endeavor of managing an 'impression of being effective interlocutors for scoring purposes' rather than in 'authentic communication' (p.25).

As shown above, the findings and conclusions about the validity of the SBA Group Interaction task in terms of the authenticity of students' discourse elicited are mixed. It is not difficult to note a marked difference in the amount of preparation time between the first two studies and Luk's (2010) study, although none of them investigated in detail what students do during the planning time, or attributed the observable interactional patterns to students' pre-task planning activities. However, as will become evident in Spence-Brown's (2001) study (reviewed below) and my own, there are cases where the candidates' discourse ostensibly suggests authentic language use, but close inspection of their task engagement during the planning stage yields contrasting evidence.

2.2 Effect of Pre-task Planning Time on Task Performance

On the question of whether pre-task planning time benefits subsequent task performance, studies in testing and non-testing contexts to date have also produced different results. As reviewed in Nitta & Nakatsuhara (2014), previous research on TBLT (task-based language teaching) has found planning time beneficial from a cognitive perspective, having a positive effect on subsequent task performance most notably in fluency, and to a lesser extent in terms of accuracy and complexity (see Ellis, 2009, for an overview of these studies). However, as pointed out by Nitta & Nakatsuhara, these studies focused primarily on the cognitive complexity and linguistic demands of the task, and did not investigate the interactional aspects of the task performance.

According to Wigglesworth & Elder (2010), evidence that pre-task planning time benefits subsequent task performance in language testing contexts is less clear. While a few studies attested to a positive impact on accuracy (Wiggleworth, 1997), complexity (Xi, 2005), or both, along with 'breakdown' fluency (Tavakolian & Skehan, 2005), others found little or no benefits on test scores or the discourse output (Wigglesworth, 2000; Iwashita, McNamara, & Elder, 2001; Wigglesworth & Elder, 2010). Again, the overwhelming majority of the

studies have focused on proficiency measures – accuracy, fluency, and complexity – of the discourse output. This can be readily accounted for by the fact that testing studies on the effect of pre-task planning time to date have been exclusively on monologic rather than interactive tasks (Nitta & Nakatsuhara, 2014).

Nitta & Nakatsuhara's (2014) pioneering study of the impact of planning time on performance in a paired speaking test revealed a potentially detrimental effect on the quality of interaction. Analysis of the candidates' discourse showed that the interactions without the three-minute planning time were characterized by collaborative dialogues, where candidates engaged with each other's ideas and incorporated their partner's ideas into their own speeches. In contrast, the planned interactions consisted of more extended monologic turns where candidates only superficially responded to their partner's talk and concentrated on delivering what they prepared. The significance of the study is that, while the planning time was found to be slightly beneficial to candidates' test scores, the qualitative analysis of interactional patterns indicated that planning time might inhibit the task from tapping into the construct that the task is meant to measure: the ability to interact collaboratively.

Evident from the above review is that, in both SLA and testing research, the focus of pre-task planning effects has mostly been on proficiency measures in the discourse output; and in testing studies, there is a gap in looking at pre-task planning effects on candidates' performance in interactive (paired or group) task formats. Further, there seems to be a general lack of studies which investigate what candidates actually do during the pre-task planning time (Wigglesworth & Elder, 2010), let alone drawing links between the planning activities and the extent of candidates' authentic engagement in the subsequent dialogic task. This is perhaps because in most high-stakes assessment contexts, candidates are not given extended preparation time or the opportunity to talk to fellow candidates in the same pair/group before the assessment. Therefore, the classroom-based assessment situated within a high-stakes examination in the present study, with the assessment task implemented in such conditions that follow from a flexible assessment policy and engender particular kinds of pre-task planning activities and strategies, creates a unique, interesting context for the study.

2.3 Call for Research on Task Implementation

Given the mixed results on the authenticity of the SBA Group Interaction task in previous studies, and the possible detrimental effect of pre-task planning time identified by Nitta & Nakatsuhara (2014), the importance of investigating how the assessment task is implemented and engaged in by student-candidates is becoming apparent. In the language testing literature, several authors have called for studies on task implementation. In concluding her study on the effect of planning time on subsequent speaking performance, Wigglesworth (1997) recommended looking into what candidates actually do during pre-task planning time in future studies. Building on earlier arguments by Messick (1994), McNamara (1997) asserts that validity cannot be achieved through test design alone, but needs to be established with empirical evidence from actual test performance 'under operational conditions' (p.456). Applying this to the case of SBA Group Interaction, validation studies need to include an examination of students' activities during the preparation time, which is a non-assessed yet integral part of the assessment task. How important it is for test validation studies to look at task implementation and authenticity of engagement is most elaborated and empirically attested to in Spence-Brown (2001).

The assessment task that Spence-Brown (2001) examined involved students in a Japanese course at an Australian university conducting tape-recorded interviews with a Japanese native speaker whom they had not previously met. Data comprised students' discourse in the interview, scores and raters' comments, and retrospective interviews with students incorporating stimulated recall. The analysis identified several aspects of students'

task engagement which posed threats to the authenticity and validity of the task. Besides selecting a known informant and pretending otherwise, as well as rehearsing and re-taping the interview, students approached the task by preparing questions, predicting answers and appropriate responses to them. This enabled students to appear to be engaging in authentic interaction without actually taking the risk of doing so. In a particularly noteworthy case, a student predicted the informant's answer to a question and pre-planned his response to the answer. The surface discourse in the interview suggested successful interaction, with the student giving an appropriate response. However, the stimulated recall revealed that the student did not actually understand the informant's answer, but drew on a rehearsed response that suggested he did. Based on such findings, Spence-Brown (2001) challenged the validity of the task: while the task is designed to engage students' use of 'on-line' linguistic competence, it in fact does not. She cautioned that because the nature of task engagement is not always transparent in the task performance (the taped interview in this case), it is more meaningful to examine authenticity from the view of implementation rather than task design alone.

2.4 The present study

Informed by the findings and recommendations from the previous research outlined above, the present study sets out to examine the validity of the SBA Group Interaction task by looking at aspects of task implementation and student-candidates' engagement. Specifically, it seeks to answer the following research questions:

1) Does the SBA Group Interaction task elicit authentic oral language use from students in accordance with the task's stated aim?
2) What do students do during the pre-task planning time, and how does this affect their discourse in the group interaction?

3. Data and Methodology

The data reported in this chapter comes from a larger study, in which three types of data were collected: 1) video-recordings of test discourse, 2) stimulated recall with student-candidates and teacher-raters, and 3) mock assessments. This section provides details of the data collected for the entire research project and the data selected for in-depth case study in this chapter.

First, video-recordings of the group interaction task completed by 42 groups in two secondary schools (School P and School L) were obtained. Among them, 23 were from Part A of the SBA, and 19 were from Part B, with some of the Part B group interactions conducted by the same students as Part A in either the same or different grouping. To explore how extended preparation time as a task implementation condition might impact on the subsequent assessed interaction, this chapter focuses on the case of School P, where students were given a few hours of preparation time (cf. 10 minutes in School L). In the following section, two extracts from two different group interactions in School P will be presented. They were selected on the basis that, at first glance, the students appeared to be engaging in authentic interaction, while close analysis and additional data (explained below) revealed the contrived nature of their interactional exchange. The first extract was part of a group interaction for Part A in which students were asked to talk about the misunderstanding between the two main characters in the movie *Freaky Friday*. In the second extract, students in a group interaction for Part B assumed the roles of marketing team members, and the task was to choose a product to promote and discuss the promotional strategies. The interactions were transcribed in detail following Jefferson's (2004) conventions (see Appendix 7.1 for

additional transcription symbols used), and analyzed following a conversation analytic approach.

To supplement the test discourse data, retrospective interviews incorporating stimulated recall were conducted for 15 assessed interactions (8 from Part A, 7 from Part B) with the relevant student-candidates and teacher-raters in the two schools who were available at the time of data collection. Depending on the mutual availability of the participants and the researcher, the time gap between the assessment and the interview varied between a few days and two months. During the interviews, the video-recordings of the assessed interactions were played and paused at intervals for the students/teachers to comment on. Additional questions about particular parts of the interactions (e.g. episodes which appear to be authentic interactional exchange) and the participants' views about the assessment in general were also asked. The stimulated recall procedure enabled me, as the researcher, to gain insights on the kinds of pre-task planning activities student-candidates engaged in, and how the interactional exchanges were perceived by the teacher-raters. All interviews were conducted in Cantonese, and the interview transcripts were translated into English. The only exceptions were two interviews (for Part A and Part B respectively) with one teacher-rater, conducted in English in accordance with her preference. The following section presents the relevant stimulated recall data for the group interaction extracts analyzed.

The third type of data was from a mock assessment, where the whole assessment process from preparation time to the assessed interaction, as well as the post-interview immediately after the assessment, was video-recorded. This was to capture the fine-grained details of students' pre-task planning activities and allow closer inspection of such activities in subsequent analysis. The limitations were that, due to constraints on the participants' availability, it was possible to carry out the mock assessment with only two groups, and with reduced preparation time. These two groups of students (four in each group) were selected from the 19 group interactions for Part B, where ostensibly authentic episodes of talk exchange were found in the initial analysis of their test discourse. The two groups were each given a discussion task adapted from their Part B assessment. One group was given approximately one hour of preparation time, and the other group approximately 10 minutes as part of an investigation of whether and how the amount of preparation time impacts on the subsequent group interaction. In the post-interview, students were asked to compare their experience in the mock and the actual assessments, in particular what preparation work they did for the actual assessment and what they were unable to do before the mock assessment, and these responses were taken as complementary evidence to the video-recording of the preparation time. Extracts 4 to 6 in the section below illustrate some of the planning activities engaged in by the student group with approximately one hour of preparation time.

4. Data Analysis

4.1 Discourse in Assessed Interactions

I begin by presenting a conversation analysis of two extracts from two group interactions, where the discourse ostensibly suggests authentic interaction among the student participants.

Extract 1 (PA11: 48-60)

Lam, D.M.K. (2015) Contriving authentic interaction: Task implementation and engagement in school-based speaking assessment in Hong Kong. In G. Yu & Y. Jin, *Assessing Chinese learners of English: Language constructs, consequences and conundrums* (pp.38-60). Basingstoke: Palgrave Macmillan. Reproduced with permission of Palgrave Macmillan.

1	W:	Do you remember there is a scene showing that the door of
2		Anna's- (..) bedroom had been removed by Mrs Coleman¿ ((R nods
3		and turns her head to N just before N begins her turn))
4	N:	Yeah. I can even \\remember the phrase on her room's
5		\\((R looks briefly at W))
6		door. Parental advisory, uh keep out of my room. So::, what
7		you're trying to say i::s
8	W:	>What I'm trying to< say is privacy. ((R turns to D))
9	D:	I see what you mean. I think: (.) privacy is::- should be: (.)
10		important to anyone. Uhm just like me, if my right (.) if my
11		right to play computer game is being >exploited by my mom<, I
12		think I will get mad on her.=So, I think: lack of (.) privacy
13		is the main cause.

The group has been talking about the various aspects of misunderstanding between the mother, Mrs. Coleman, and the daughter, Anna, in the movie *Freaky Friday*. This extract shows a sequence where the group discusses another cause of misunderstanding between the two characters.

In lines 1-2, W asks the co-participants if they recall a particular scene from the movie. This takes the shape of a pre-telling, whereby W checks the requisite condition for a forthcoming telling. The next speaker, N, offers an affirmative 'yes', and provides further recalled details showing the condition has been met (lines 4-6). The sequence does not immediately proceed to W's telling, however. In lines 6-7, N issues a clarification request in the 'fill-in-the-blank' format ('what you're trying to say is'). This displays her orientation to W's prior turn as projecting more talk – the thrust of the telling sequence for which W's recall question has been laying the groundwork. Interestingly, on the one hand, N's clarification request displays her alignment with the trajectory of a telling W has been setting up, amounting to a 'go-ahead' for W to make her point. On the other hand, N modifies this trajectory by opening up another sequence, of which the clarification request is the first-pair-part (FPP).

Note how W's following response (line 8) displays sensitivity to the contingency of the unfolding sequence. Instead of staying on her own course and designing her turn like the FPP of the main telling sequence following the pre-telling, W aligns with the new trajectory of talk set up by N through formatting her turn as the answer second-pair-part (SPP) to N's question, with the preface 'what I'm trying to say is' mirroring the shape of the question FPP. Throughout these three turns (lines 1-8), then, both participants construct their responses in ways which are sensitive to and contingent on the previous speaker's talk. In other words, they seem to engage in each other's talk and develop on each other's contribution, showing evidence of authentic interaction.

Rather strikingly, however, the main telling towards which all the previous interactional work seems to have been building ends up with one word, 'privacy' (line 8). This main telling sequence that is anticipated to be making the point about privacy issues as a cause of misunderstanding, yet blatantly underdeveloped in W's talk, is then expanded in D's response

(lines 9-13). Here, he acknowledges receipt and claims understanding of W's telling, provides an affiliative assessment of the point about privacy, offers an example from his personal experience, and finally formulates the upshot of the whole sequence ('lack of privacy is the main cause'). Interestingly, then, W is seen to leave it for D to spell out the thrust of the sequence.

Thus, we see a rather odd sequential development in which W seems to (willingly) relinquish the rights to making her point, after all the preliminary interactional work that has built towards it and would have sequentially ratified an extended telling turn on W's part for such purpose. The task of bringing home the point about privacy as a main cause of misunderstanding is conveniently re-allocated to another participant, D. This raises questions as to whether this has truly been how the interaction has unfolded, or something pre-planned prior to the assessment.

Indeed, close examination of co-participants' non-verbal behavior yields preliminary evidence that this interactive sequence has been pre-scripted. In lines 2-3, towards the end of W's question, R nods and turns her head to N just before N commences her turn. Meanwhile, despite generally being the most active participant, R does not even offer a minimal verbal response such as 'mm' or 'yes' here, let alone elect herself to answer W's question. As N begins answering W's question, R glances at W again (line 5) instead of focusing her gaze on N to display listenership. Finally, in line 8, R turns to D right at the end of W's turn and just before D's, as if she has already known that D would be the next speaker.

Students confirmed in the stimulated recall that this sequence (and the whole interaction) was pre-scripted, and R explained that this was to create an opportunity for a group member who wouldn't have spoken for a while to take a turn.

Extract 2 below shows another group interaction, one that simulates a marketing team meeting for the promotion of a new product. The discourse in this episode, with reference to turn design and sequential development, gives some indication of students' authentic engagement in the simulated interactional context, and in challenging each other's ideas.

Extract 2 (PB14: 10-25)

1	L:	Mm. Yes, our company has just released (.) our beauty products
2		in- eh- uhm the teenagers. Mm:: (.) mm:: (1.9) uhm: so: are you
3		guys clear about the special features of the product?
4	K:	°Mm.° I've heard that the new products .h are composed of a
5		traditional Chinese medicine. That is quite special.
6		(..)
7	T:	Uhm:: but, do you think that the traditional Chinese
8		medicine .h have strong and strange smell? Many people may
9		refuse to use our ↑pro↓duct.
10	S:	Hey. You've missed out a ↑po↓int. That is our product also
11		includes (.) natural ingredients (.) li:ke lavender (.) which
12		is successfully cover (.) the:: ↑smell brought by the
13		traditional Chinese medicine.
14	L:	Mm::. (.) It's one of the fo- ma- m- main focus, that uh to
15		promote our product. .h Uhm, it is not smelly even if we have
16		added the traditional Chinese medicine into it. .....

The sequence begins with L, who assumes the role of team leader, initiating the topic about special features of their skincare product (lines 1-3). She discursively constructs her authoritative role through announcing the release of their product, and asking other team members if they are 'clear about the special features', thereby claiming epistemic superiority over other group members. K responds by introducing the feature of traditional Chinese medicine as product ingredient, and adds a positive assessment (lines 4-5). In providing this answer to L's question, she ratifies and co-constructs L's role as team leader. The turn design of prefacing her response introducing the Chinese medicine with 'I've heard that...' also displays K's commitment to their contextual roles as marketing team members (as people who should know about the product's features but did not create the product themselves).

K's positive evaluation of Chinese medicine as product ingredient is then met with a disagreeing response from T (lines 7-9). This begins with prolonged hesitation 'uhm', followed by a negative assessment of the Chinese medicine framed as a question. Neither K nor T orients to the question as projecting an answer, as T continues to offer a further account for disagreement predicting negative consumer reactions. The turn shape of T's disagreeing response in itself is noteworthy, indeed striking. It differs markedly from formulaic disagreeing responses such as 'I'm sorry I can't agree with you' that feature an explicit disagreeing component, and which frequently occur in other group interactions in the data.

Equally striking, perhaps, is the following response by S, which counters T's disagreement by commenting that T has 'missed out a point' – another feature of their product (line 10). This type of sequential development, where a disagreeing response is followed by another disagreeing response countering the first, is rarely observed in the data. However, S is then able to conveniently introduce this neglected feature both as a counter argument and as a new idea that she contributes on the topic, as she elaborates on how other natural ingredients such as lavender can solve the problem of the smell brought by Chinese medicine. Such a design enables S to both topicalize previous speaker's idea of Chinese medicine and make her own contribution about other ingredients.

During the stimulated recall, the teacher-rater paused the video and gave her positive evaluation on this episode of talk exchange:

Extract 3 (PB-TR-B stimulated recall, English original)

((TR pauses the video after line 9 in Extract 2))
TR: Uh I like it how she **responded to something that K said**. So **rather than say something else**..... **she asked about it**.

The teacher-rater positively remarked that T raised a question about K's idea in her response, topicalizing the previous speaker's contribution rather than focusing on delivering her own idea. Subsequently, the teacher-rater also gave a favorable evaluation of S's response, in which she further topicalized the feature of Chinese medicine and elaborated on how the problem with its smell could be solved. Throughout the stimulated recall, the teacher-rater commented several times that this group's interaction was 'authentic'.

Nevertheless, the stimulated recall with students again revealed that the entire interaction was pre-scripted and rehearsed. Within the test discourse, students' intonation and the strangely 'neat' speaker transition without many gaps and overlaps might have been a giveaway. More importantly, the students' unique ways of doing disagreement (cf. using formulaic expressions), which ostensibly suggested authentic interaction, was precisely one of the clues to a pre-planned, contrived interaction. Though performed in a playful tone here,

the kind of unmitigated negative comment directed at a co-participant (line 10) rarely occurs in spontaneous assessed interactions, as it would probably constitute a direct face threat to a co-participant.

4.2 Pre-task Planning Activities

Further insights about the kinds of pre-task planning activities students engage in, including pre-scripting, were gained through close examination of the video-recorded one-hour preparation time for one of the mock assessments. Figure 7.1 below is a schematic representation of the planning activities carried out during the preparation time.

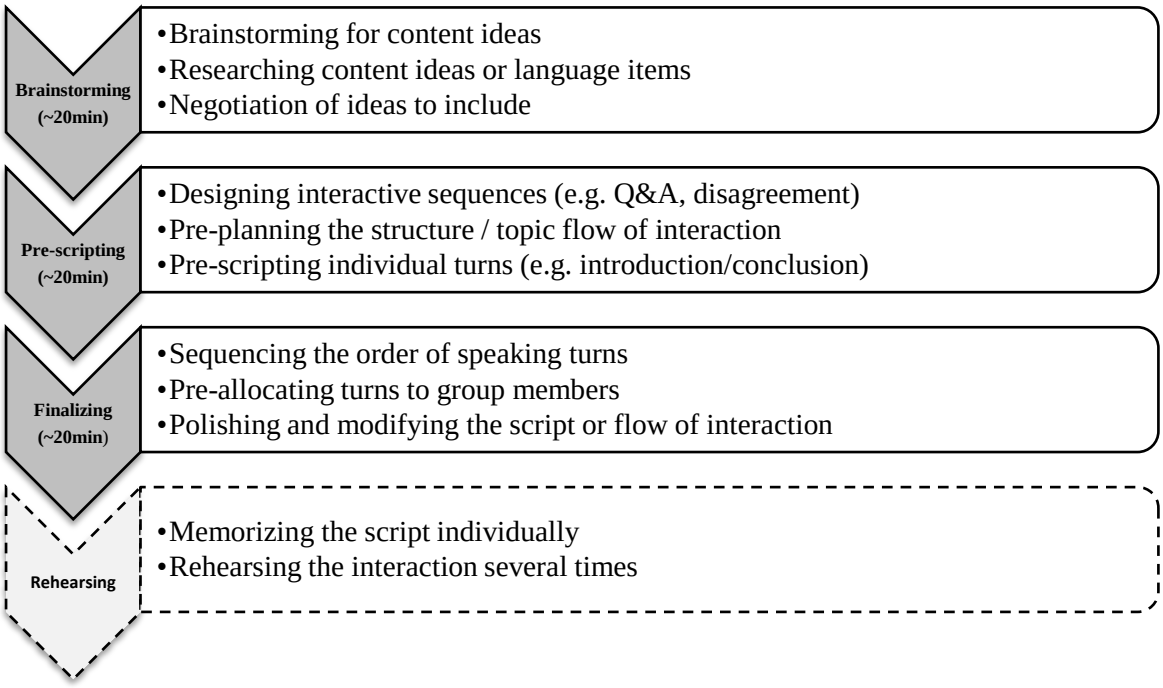

Figure 7.1 Students' pre-task planning activities

As shown in Figure 7.1, students' pre-task planning for the mock assessment can be roughly divided into three stages (represented in solid lines). The first stage involves students brainstorming for ideas about the discussion topic, researching information and relevant vocabulary items with their smartphones, and negotiating what ideas to include and exclude in the assessed interaction. In the second stage, students decide together on the structure or topic flow of the interaction. They also design interactive sequences such as question-and-answer or disagreement, and pre-script particular speaking turns such as the opening and concluding turns. In the final stage, students fix the sequence of speaking turns and assign each turn to a group member. Any final touch-ups to the script or flow of interaction are also done at this time.

It should be noted that these activities are not actually carried out in a strictly linear sequence, and are only presented in approximate order. For instance, form-focused planning activities such as looking up vocabulary items and English translation of brand names, and checking them with others in the group, are recurrent and interspersed throughout the preparation time. In the post-interviews with the two groups participating in the mock assessment, supplementary information about students' pre-task planning activities was gained regarding what they did before the actual assessment and, correspondingly, what they did not manage to do during the preparation time for the mock assessment. Students reported not having enough time for pre-scripting the interaction *verbatim* before the mock assessment.

They also reported an additional stage before the actual assessment (in dotted lines) that involved memorizing the script individually and rehearsing the interaction (referred to as 「試演」'trial acting') several times.

In the following, I discuss three types of pre-task planning activities which pose threats to the authenticity of the assessed interaction.

First, students were observed to pre-negotiate the pros and cons of certain ideas in the brainstorming stage, with differences of opinion dealt with and consensus reached. Consider the following extract of students' pre-task planning discussion:

Extract 4 (PB11MockPrep 24:00)

((Previously, someone suggested hiring three spokespersons for their three target age groups of customers))

Y: But have you guys considered the cost? It's very expensive, if we get three spokespersons.
K: Well, so maybe we can *ban* the idea of three spokespersons. *Ban* three spokespersons.
R: No. We should first have someone say let's get one spokesperson, then someone else *ban* the idea, and say we actually have three *target* groups, so why don't we have one spokesperson for each *target* group.
S: But it's mainly adults who would buy [vitamin pills] after all. Isn't one spokesperson enough?
Y: Wait. Let's get a 'mum'. Getting a 'mum' [as the spokesperson] will work!
K: We can say it's usually housewives who buy [vitamins for the whole family]. It's not the children who would buy them.

Instead of having it as a point for debate in the assessed interaction, the group pre-determined their final decision of having only one spokesperson, and pre-planned how they would work their way through the different proposals to reach such consensus in the assessed interaction. This pre-task discussion therefore eliminates the information and opinion gaps that could create a genuine need for communication and negotiation in the group interaction task proper.

Related activities which threaten the authenticity of the assessed interaction include students pre-scripting interactive episodes, pre-sequencing their turns and assigning them to individual group members. Extract 5 below shows the final stage of pre-scripting the discussion on the 'spokesperson' topic.

Extract 5 (PB11MockPrep 55:45)

S: ((points to Y)) She will introduce [the topic of] *spokesperson*
K: OK. So I'll then suggest three. ((writing on note card simultaneously)) I'll say since we have three *target groups*, why don't we get three *spokespersons*.
R: ((points to K)) You say that, you'll suggest that, right? So you suggest having three spokespersons. And then who's gonna *ban* the idea? You *ban* it, S.
S: Sure, I'll *ban* it. I'll *ban* it.
R: And after *ban*ning it I'll lead to [the topic of] '*place*'. Alright, let's do it like this.
S: ((writing simultaneously)) I'll do the banning. The cost is too high.

Lam, D.M.K. (2015) Contriving authentic interaction: Task implementation and engagement in school-based speaking assessment in Hong Kong. In G. Yu & Y. Jin, *Assessing Chinese learners of English: Language constructs, consequences and conundrums* (pp.38-60). Basingstoke: Palgrave Macmillan. Reproduced with permission of Palgrave Macmillan.

R: ((writing simultaneously)) 'Three *spokespersons'* is by K, and then S *bans* the idea, because the cost is too high. And then I'll agree with her, and afterwards I'll introduce [the topic of] '*place*'.

As seen in the transcript, the students are assigning roles and finalizing the interactive sequence where they would propose having three spokespersons, challenge the idea, then agree on the alternative of having one only, and shift to another topic. The sequence of assigned speaking turns, and the order of proposing, disagreeing, and finally reaching consensus on an idea, were all written down on their note cards as what the students themselves called the 'route map' (「路圖」) of the assessed interaction.

Finally, there was an instance of a student helping a less capable group member (Y) pre-script her turns:

Extract 6 (PB11MockPrep 41:40)

K: Oh so you can also mention this. You say 'let's start with "*product*", but I can't think of promotional ideas because it's difficult when there're so many *competitors*, so what ideas do you guys have?' And then we'll respond to her.

Thus, what Y eventually said in that turn during the assessed interaction was not even entirely her 'original work', let alone a spontaneously produced contribution.

On scrutinizing students' pre-task planning activities, we now have good evidence that what might appear as authentic exchange in the assessed interaction can in fact have been contrived. Overall, the data in School P indicates an overwhelming tendency of students engaging in contrived rather than spontaneous interaction, supported by the fact that all students in School P interviewed admitted having pre-scripted the assessed interaction. As a result of the aforementioned pre-negotiation of ideas and the subsequent pre-scripting of the relevant discussion, what the students perform and are evaluated on during the assessed interaction is, at best, a re-presentation of their pre-task interaction conducted in L1. It is not an authentic and spontaneous interaction conducted in L2 spoken English, the target of the assessment. Instances of authentic, spontaneously produced exchanges were found in interactions with only 10 minutes of preparation time (in School L and in one of the groups in the mock assessment), but are beyond the scope of this chapter. These cases and their comparison with contrived exchanges warrant equally detailed analysis and discussion, and will be taken up in future published work.

5. Discussion and Conclusion

5.1 Findings and Implications

This chapter has sought to contribute to the body of validation work for the SBA Group Interaction task, and to reveal some of the complexities in ensuring the task's validity implicated by the 'flexibility' element in the assessment policy and the corresponding practices. A main objective of this study was to examine whether the Group Interaction task, in the way it is implemented, elicits authentic oral language use. Previous studies have gauged the task's (lack of) construct validity mainly in terms of *authenticity* and its real-world connection with everyday conversation. Indeed, the relationship between authenticity and validity of a task has long been an issue in theoretical debates. Bachman (1990) attributed the preoccupation with authenticity to 'a sincere concern to somehow capture or recreate in

language tests the essence of language use' in the target domain (p.300). However, Spolsky (1985) contended that test behavior can never be an entirely authentic reflection of non-testing behavior, as interactions in testing and non-testing situations follow different rules. Some authors (e.g. Widdowson, 1979; van Lier, 1996) distinguish between *genuine* – employing texts used by native speakers for everyday communication in pedagogic tasks; and *authentic* – related to processes of engagement. Building on this distinction, Spence-Brown (2001) introduced the notion of *authenticity of engagement* in evaluating the validity of assessment tasks.

In answer to the research questions of this study – whether the SBA Group Interaction task elicits authentic oral language use, and how it is affected by students' pre-task planning activities – we can conclude that, while the task has authenticity in terms of task content, it has questionable authenticity of engagement by students. The discussion tasks do have some real-world connection, with students interacting on 'real material' (movies), or simulating real-life situations (work meetings). Students' discourse yielded ostensible evidence of authentic engagement in interacting with each other, for instance, modifying one's response to align with previous speaker's talk (Extract 1), and natural, non-formulaic ways of doing disagreement (Extract 2). Some of these were recognized and favorably evaluated by the teacher-rater. Nonetheless, stimulated recall with the students and video-recording of preparation time before the mock assessment revealed that these interactive episodes were part of a staged performance of pre-scripted dialogues.

Therefore, what the assessed interactions showed was essentially the product of students acting out a composed dialogue based on their knowledge and perceptions of what interactional competence is, rather than students' spontaneous performance of the competence that involves moment-by-moment monitoring of and contingent reaction to each other's talk in real time. Several authors have included this element of 'spontaneity' in defining competence in interaction. Bachman (1990) describes 'communicative language ability' as 'consisting of both knowledge, or competence, and the capacity for implementing it, or executing that competence in appropriate, contextualized communicative language use' (p.84). Barraja-Rohen (2011) asserts that interactional competence involves, among other skills, 'precision timing and a quick analysis of speakers' turns' (p.482). Spence-Brown (2001) questions the validity of the tape interview task based on its failure in eliciting learners' 'on-line linguistic competence' (p.471). Similarly, what can be observed in the SBA assessed interactions is often not students' *in situ* execution of interactional competence in L2, but a 'canned' product of students' execution of the competence *prior to* the assessed interaction *in L1* during pre-task planning. Furthermore, Kramsch (1986), in her seminal work on interactional competence, describes interaction as relative and unpredictable in nature, and it is on this premise that talk exchange takes place, with the objective of reducing uncertainty of 'intentions, perceptions, and expectations' (p.367). However, we have seen evidence of pre-task planning activities closing the information or opinion gap for interaction, with aspects of uncertainty and unpredictability (otherwise matters to deal with in the assessed interaction) being reduced or eliminated.

Some of the key emphases of the School-based Assessment policy, as outlined in the Introduction, were on flexibility, sensitivity to students' needs, and low-stress conditions, all constitutive of an explicit departure from standardized language assessments. In a way, the face of the assessment practices matched the policy. First, as seen in previous studies reviewed and my own, diverse practices in task implementation, rather than standardized tasks and task conditions, were found across different schools. Moreover, extended preparation time given in some schools catered for weaker students' needs, as it could reduce anxiety in the otherwise highly stressful assessment situation (Wigglesworth & Elder, 2010),

as well as enable prepared speech for those who lack confidence in spontaneous L2 interaction. The greatest tension, then, is perhaps not just about aligning policy and practice, but lies between some of the above principles behind this set of policy and practice, and the target L2 interactional competence by which the validity of the assessment task is determined. This competence, as argued above, entails spontaneous production of talk exchange in L2 predicated on genuine needs for communication (information/opinion gaps to bridge).

The findings of this study also bring to light the immense difficulty to reconcile the formative and summative elements of an assessment-for-learning initiative such as the SBA in Hong Kong. This is best summarized in Hamp-Lyons's (2009) remark that it needs to be 'meaningful at the level of the individual school and classroom', and at the same time, 'be accountable territory-wide' and 'meet the traditional expectations of rigour for summative reporting' (p.525). The current practices in task implementation by teachers and task engagement by students, as reflected in this study and some of the previous research (Fok, 2012; Luk, 2010), seem to primarily serve the aim of creating optimal impressions of performance for scoring purposes (Luk, 2010). As it stands, the English SBA has yet to accomplish being a valid assessment that fully reflects the L2 interactional competence the task is designed to assess, and to serve the pedagogical goal of developing students' competence in conducting spontaneous L2 interaction with peers. More research is needed to refine the implementation of assessment for learning, both in the Hong Kong context and in general, in order for it to truly fulfill its purpose.

Based on the findings from this study, and subject to further empirical validation, the following recommendations for the assessment policy on task implementation can be made. Students can be given an amount of preparation time just enough to brainstorm ideas and research on language items, but not for pre-scripting the interaction. Alternatively, aligning with the assessment-for-learning initiative, teachers can allow pre-planning and pre-scripting the interaction in practice assessments at early stages of the upper-secondary curriculum to accommodate weaker students, with a goal of gradually moving students towards spontaneous interaction in the graded assessments.

5.2 Limitations and Future Directions

This investigation of task implementation and engagement is necessarily exploratory. Given a small sample and the known diversity of assessment practices, I do not claim extensive generalizability of the study results. However, there is reason to believe that aspects of task implementation and engagement shown in this study are representative of a common practice in Hong Kong schools, as Fok (2012) and Luk (2010) have also provided evidence of pre-scripting. Furthermore, the mock assessment data can be considered a faithful reflection of the pre-task planning activities students engage in before the assessed interaction. Students were cooperative and did not exhibit any behavior that oriented to the mock assessment as anything less serious than the actual assessment. As acknowledged before, preparation time was reduced, and some differences in the planning activities were thus inevitable, but these were addressed in the post-interview. Future studies can, where practical conditions allow, gather larger samples of mock assessments for more generalizable results about pre-task planning activities. Controlled experimental studies would also be useful to determine the optimal pre-task planning time and conditions for the assessed interaction.

ACKNOWLEDGEMENTS

I would like to thank Prof John Joseph, the two Editors of this volume, and an anonymous reviewer for their very helpful comments on earlier drafts of this chapter. All remaining errors are my own.

REFERENCES

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

Barraja-Rohen, A-M. (2011). Using conversation analysis in the second language classroom to teach interactional competence. *Language Teaching Research*, *15*(4), 479-507.

Cheng, L., Andrews, S., & Yu, Y. (2011). Impact and consequences of school-based assessment (SBA): Students' and parents' views of SBA in Hong Kong. *Language Testing*, *28*(2), 221-249.

Davison, C. (2007). Views from the chalkface: School-based assessment in Hong Kong. *Language Assessment Quarterly*, *4*(1), 37-68.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics*, *19*(4), 474-509.

Fok, W. K. (2012). *HKCEE English Language school-based assessment: Its implementation at the frontline*. Unpublished doctoral thesis, Durham University.

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. (2011). An Investigation of Personality and L2 Oral Performance. *Journal of Language Teaching and Research*, *2*(6), 1259-1267.

Gan, Z. (2012). Complexity measures, task type, and analytic evaluations of speaking proficiency in a school-based assessment context. *Language Assessment Quarterly*, *9*(2), 133-151.

Gan, Z., Davison, C., & Hamp-Lyons, L. (2008). Topic negotiation in peer group oral assessment situations: A conversation analytic approach. *Applied Linguistics*, *30*(3), 315-334.

Hamp-Lyons, L. (2009). Principles for large-scale classroom-based teacher assessment of English learners' language: An initial framework from school-based assessment in Hong Kong. *TESOL Quarterly*, *43*(3), 524-530.

He, L., & Dai, Y. (2006). A corpus-based investigation into the validity of the CET–SET group discussion. *Language Testing*, *23*(3), 370-401.

HKEAA (2009). 2012 Hong Kong Diploma of Secondary Education Examination English Language: School-based Assessment Teachers' Handbook. Retrieved February 21, 2014, from http://www.hkeaa.edu.hk/DocLibrary/SBA/HKDSE/Eng_DVD/doc/SBA_handbook_2012_ENG_240709.pdf

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning*, *51*(3), 401-436.

Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In G. Lerner (Ed.), *Conversation analysis: Studies from the first generation* (pp. 13–31). Amsterdam: John Benjamins.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal*, *70*(4), 366-72.

Luk, J. (2010). Talking to score: Impression management in L2 oral assessment and the co-construction of a test discourse genre. *Language Assessment Quarterly*, *7*(1), 25- 53.

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

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13-23.

Nitta, R. & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-task planning effects on paired oral test performance. *Language Testing*, *31*(2), 147-175.

Spence-Brown, R. (2001). The eye of the beholder: Authenticity in an embedded assessment task. *Language Testing*, *18*(4), 463-481.

Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing*, *2*(1), 31-40.

Tavakolian, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239-273). Amsterdam: John Benjamins.

van Lier, L. (1996). *Interaction in the language curriculum. Awareness, autonomy and authenticity.* London: Longman.

Widdowson, H. G. (1979). *Explorations in applied linguistics.* Oxford: Oxford University Press.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing*, *14*(1), 101-122.

Wigglesworth, G. (2000). Issues in the development of oral tasks for competency-based assessments of second language performance. In G. Brindley (Ed.), *Studies in immigrant English language assessment* (Vol. 1. Research Series 11, pp. 81-124). Sydney: National Centre for English Language Teaching and Research, Macquarie University.

Wigglesworth, G. & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly*, *7*(1), 1-24.

Xi, X. (2005). Do visual chunks and planning impact performance on the graph description task in the SPEAK exam? *Language Testing*, *22*(4), 463-508.

APPENDIX

Appendix 7.1 Additional Transcription Symbols

\\words \\((actions))	beginning of non-verbal action simultaneous with speech
first letter underlined	sequence of words each uttered with hearable effort or emphasis
.....	rest of the turn omitted