Self-assessment Accuracy in Group Orals

Tobias Long

ABSTRACT

In an assessed group oral activity, such as a discussion, where instructor assessment measures the use of prescribed discourse functions, there may be obstacles to accurate self-assessment (SA). This paper seeks to investigate the extent to which students are able to accurately self-assess in a particular context. Discussions were video recorded and reassessed to compare student and teacher scores. The results of data analysis initially show that student SA accuracy was inconsistent.

INTRODUCTION

The role of self-assessment has become increasingly significant not only in education in general, but also in the fields of language learning and teaching. Douchy et al. (1999) refer to the 'era of assessment' where assessment and instruction are integrated and where students are seen as active, responsible, reflective collaborators in the learning process. In their review of SA literature, they found that there is 'much evidence' to support consistency between student and teacher assessment (Douchy et al**.**, p. 347). Boud and Falchikov (1989) make the distinction between SA and self-marking, where SA involves students in 'the identification of criteria or standards to be applied to one's work, and the making of judgements about the extent to which work meets these criteria' (p. 529). Although they caution that studies using self-marking should not be so readily generalized to the realms of SA, there is perhaps considerable overlap between the two. Black et al. (2003), in their discussion of SA, state that '[I]t is very difficult for students to achieve a learning goal unless they understand that goal and can assess what they need to do to reach it' (Black et al., p. 49). Clearly, accurate SA requires a solid understanding of the parameters of assessment. Oscarson states that '[t]he general pattern of research results reviewed … warrants an optimistic view of SA' (p. 8). However, he cautions that future investigations into 'teacher training focused on student-centered formative assessment, [and] investigation into the effects of SA procedures on motivation and achievement…' are in need of development (Oscarson, p. 16). Much of the research on SA accuracy focuses on higher education, and much of that is not specifically focused on second language SA accuracy. Some studies that have examined SA accuracy include Chen's (2008) study of Chinese speakers of English, where students did SA on oral presentations, and Munoz and Alvarez's (2007) study which measured communicative effectiveness, grammar, pronunciation, vocabulary, and task completion in individual performances. Harris (1997) looked at the role of self-assessment in formal settings where EFL classes are often compulsory, and Ross (1998) carried out a thorough meta-analysis of ten studies on SA and language proficiency. An interesting conclusion in one of the studies he surveyed was that achievement proficiency SA may be inferred to be more accurate than more abstract, proficiency-based criteria (Ross, 1998, p. 14). Also interesting was the conclusion that 'learner self-assessment of course objectives…may better assess the learner's confidence in the degree of content mastery' while teacher assessment may be better focused on the more abstract, proficiency measures (p. 16). With these studies in mind, and the general consensus that 'there is often a good deal of agreement between learners' self-assessments and external criteria' (Oscarson, 1984) the context of this study will be described.

CONTEXT

In the English Discussion Class (EDC), students are assessed mainly on the use of discussion skills and communication skills. In other words, it is a functional-notional syllabus, where functions are performed using simple phrases that help students in the task of exchanging opinions. Discussion

skills are 'lexical clusters, presented as interactional, formulaic function phrases' (Young, 2017). One example of such a discussion skill, 'Sources of Information', includes declarative phrases such as 'I read that…' which marks the introduction of a source of information into a discussion. The same skill features interrogative phrases, such as 'Where did you hear that?'. The importance of marking utterances is presented to students as a way to assist others in following the flow of a discussion. Discussion skills are separated into interrogative and declarative skills. This separation emphasizes the interactivity of discussion, which is explained in Hurling's introduction to EDC (Hurling, 2012). The importance of 'marking' utterances with these phrases is made clear to students. A discrepancy between students and teachers access to meta-knowledge regarding the grading rubric is important and will be discussed as it relates to SA accuracy later. Review lessons are held three times per semester to review these skills and additionally focus student's attention on particular communication skills, the other set of skills students are assessed on, which are intended to assist students in negotiating meaning in their discussions. In the lessons following review lessons, students are tested in groups of three to five students. These discussions are assessed based on a detailed instructor rubric. This rubric assists instructors in the interpretation of potential ambiguities related to the what constitutes Discussion Skill use.

In the context of the EDC, students are commonly given a variety of tools to self-assess. Davies (2012) outlined some of the more popular SA tools and included questionnaires, target language check-sheets, mini journals, best idea or contribution SA, and content or skill use self-grading (Davies, p.4-93). He found that 'quantitative SA questionnaires had a more positive effect on the frequency of communication skill use than qualitative SA questionnaires' (Davies, p. 4-95).

Quantitative Assessment, gamification, transparency and interpretation

The use of quantitative assessment was chosen as a way to simplify SA for comparison with teacher assessment, as well as a way to focus students' attention on the performance goals of the course while acknowledging the value of appealing to extrinsic motivation. By emphasizing a certain number of skills to be used in a discussion, it was hoped that anxiety could be reduced and teamwork could be increased towards the goal of successful use of Discussion and Communication Skills in group discussions. The quantification of SA, rather than a more qualitative SA found in other activities, was intended to provide transparency into the assessment criteria. The requirement that students use certain phrases presents challenges when discussions are timed and when discussion prompts do not always necessitate the use of particular skills. A kind of 'linguistic gymnastics' is sometimes needed to use particular skills. This has the potential to be frustrating and by extension, demotivating. Quantified SA brings some elements of gamification into the classroom. The use of quantified SA can reduce pressure and bring an element of fun into an otherwise potentially stressful exercise. In addition, it gives students very clear targets that align with class requirements. By giving students clear targets, they can rely on a simple rubric to achieve them.

Some studies on the effects of gamified activities have supported the use of quantitative SA as a way to address externally motivated students. In McEntee's (2017) investigation into the merits and implications of gamifying language practice, it was found that the use of a quantified leaderboard which measured the use of specific function phrases led to an increase in their use (p. 88). He cites Kim (2015) in a study that examined player types. Some players are willing to play for extrinsic reward, while others are not. For the extrinsically motivated player types, quantified SA may resemble an extrinsically focused gamified activity, thus showing potential to increase, at least short-term, motivation. Reddy and Andrade (2009) reviewed rubric use in higher education (citing Lapsley and Moody, 2007) and state that for traditional (in other words younger) students, channeling a more typical extrinsic motivation may lead to enhanced performance. The students

in EDC often admit to being motivated in this way. As it is a compulsory course, it can be assumed that there are some students whose main goal is simply to pass the course.

A major impediment to accurate SA in the present context is found in the interpretation of what qualifies as a pragmatically marked phrase. The instructor rubric, which is revised continuously to reflect development of the curriculum and accurately assess students, goes into a level of detail not found in the textbook, and arguably would be problematic to explain to students of lower proficiency. It is not uncommon for EDC students to misinterpret target-language use. Students who answer interrogative 'If' questions often do not use specific pragmatic markers in their declarative responses. These students may incorrectly assess themselves as having used a particular skill. Ross suggests that 'can do' functions may be more easily recalled, thus increasing the accuracy of SA (p.16). An EDC Discussion Skill can be seen as a 'can do' function, and in the present study, they are treated as such. For the purposes of simplifying the interpretation of alternate phrases which may accomplish the 'can do' purpose of a specific function, the question is phrased as a 'did you' question. More specifically one would look like this: 'Did you ask an 'if' question?' and 'How many times that you can remember?'. This shift in emphasis is an attempt to include low-proficiency students who may have trouble with metalanguage, which may ultimately lead to errors in SA. The main issue of concern using 'can do' assessment prompts such as 'Can you talk about possibility?' as opposed to 'Did you use 'if' in a sentence?' is that 'can do' prompts leave more room for interpretation, while 'did you' prompts make it clear that a specific phrase is being assessed.

METHOD

An attempt was made to choose students from the middle tiers of proficiency. Students are grouped into 4 levels, with level 1 being the highest. With no level 4 classes, the decision to use both level 2 and 3 classes provided an opportunity to look at differences in assessment accuracy between different proficiency levels. One review lesson from the first semester and three review lessons during the second semester were chosen to do recordings. To check students' SA, self-check sheets were distributed after group discussions on pre-test review lesson days (Appendix A). As these review lessons occur prior to discussion tests, they are a good opportunity to measure student progress, but also it was felt that these lessons were the best choice as two Discussion Skills had previously been covered in classes and there were no new Discussion Skills to introduce during the review. Students were introduced to the check sheets and instructed on how to complete them.

After the completion of their timed discussions, the check sheets were distributed and students were given time to complete them. For the first-semester check sheets, students tried to remember and write the number of times they used a particular skill. Students wrote their names and the date on their sheets, and they were collected for analysis. The second, and longest discussions from each of the classes were recorded and analyzed to ensure that complete attention could be given to the reassessment process. During the reassessment, video could be slowed or repeated to more closely assess student utterances. Interestingly, it was not unusual to miss student utterances that were simply drowned out by the noise of other speakers or spoken with hesitation. In an exploratory study on instructor positioning for assessment in multiple-group discussion classes, Long (2019) discussed the challenges of assessing students accurately. Video enabled a much more accurate instructor assessment. With audio recordings alone, the potential for misattribution of utterances may increase, as it is not always easy to know who is speaking.

Small action cameras, similar to GoPro action cameras, were chosen due to their small form factor. Audio was recorded with Zoom H2n recorders, which were used for their superior recording quality and as a backup in case there were any problems with video recording. These video and audio files were processed using Shotcut, a freeware video editing program. The

checking of these video-recorded discussions was carried out using a standard instructor check sheet on a Microsoft Surface Pro computer running OneNote 2016 (See appendix B).

Four classes from the first semester and 3 classes from the second semester were observed. With 51 students, a total of 22 discussions were recorded. This brought the total number of self-assessments to 80. As there were 2 Discussion Skills for each participant to assess, (1 skill can be divided into interrogative and declarative or listener and speaker roles—bringing the total of categories to 4 SA marks), students had to assess themselves on 4 separate skills. This makes the total number of single skill assessments 320.

Comparisons of teacher assessment with student SA numbers were divided into four categories. The first category, ESA, represents a perfectly matching SA when compared to video reassessment. The student either used or did not use the skill, and assessed this correctly. The second category, OSA+, describes a case of at least one instance of the skill being used, but with an underestimation on the part of a student. The third category, USA+, describes a case of overestimating the number of times the skill was used, but with at least one instance of the skill used in the discussion. The final category, OSA-, describes no use of the skill and an overestimation of the skill's use. Including these categories helped to gain insight into potential problems with student comprehension of the rubric, and provided some potential solutions to future attempts at examining SA in a group oral task activity. Additionally, further groupings of data were created that simplify its analysis. For each individual student, a score was calculated to reflect the overall accuracy of their SA. This was achieved by looking at the categories applied to each of the four skills on their SA sheet. Skills which fell into ESA, OSA+ or USA+ categories were considered to be generally accurate, despite some errors, while skills which fell under the OSA- category were considered to have been inaccurately assessed. Then, students were given percentage scores which indicated the extent to which their SA was accurate: O correct (0%) to 4 correct (100%). For example, a student who was accurate in assessing their use of two out of the four skills would receive an accuracy score of 50%. This allowed for a simpler understanding and presentation of the data. Both results will be examined in the discussion.

RESULTS

On average, students were 66% accurate in their self-assessment, which suggests that students are not generally accurate in their SA of skill-use following a group discussion. When the data for classes are separated, it is clear that more proficient level 2 classes had better SA accuracy. Level 2 classes' average accuracy score across both semesters came to 71%. Level 3 classes from both semesters were, on average, 38% accurate in their self-assessments. If a future study were to examine SA in this context with proficiency as a main variable under consideration, it may be of interest to explore the hypothesis that more proficient, motivated or confident students are more accurate in their own assessments of specific phrase or skill use.

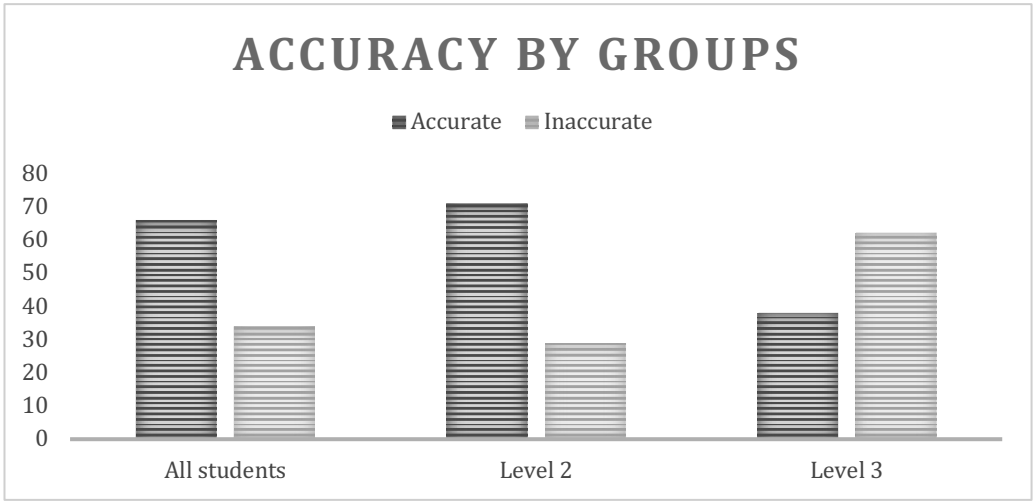

Figure 1. Assessment Accuracy Percentages by Proficiency Level

Measuring each instance of SA (320), the total number of exact SA, or ESA, came to 127, or 39%. The next category included overestimation, or OSA+. The total OSA+ came to 57, or 17.8%. The total USA+ came to 14, or 4%. These SA scores indicate that students were more likely to overestimate their skill-use rather than underestimate. The final category examined was no-use overestimation, or OSA-. Students mistakenly remembered or claimed to have used a skill 122 times, when in fact they had not used it at all in the discussion. This represented 38% of all skill assessments. If a student is unable to separate the first discussion from the second, or misunderstood the explanation of the procedure, they may be including the first discussion's skill use, or may be inadvertently combining the two discussions. It is quite likely that the instructions to count only skills used in the second in-class discussion may have been misunderstood. If this is not the explanation, it is troubling to think that students might be inflating their SA in what is arguably a low-stakes class.

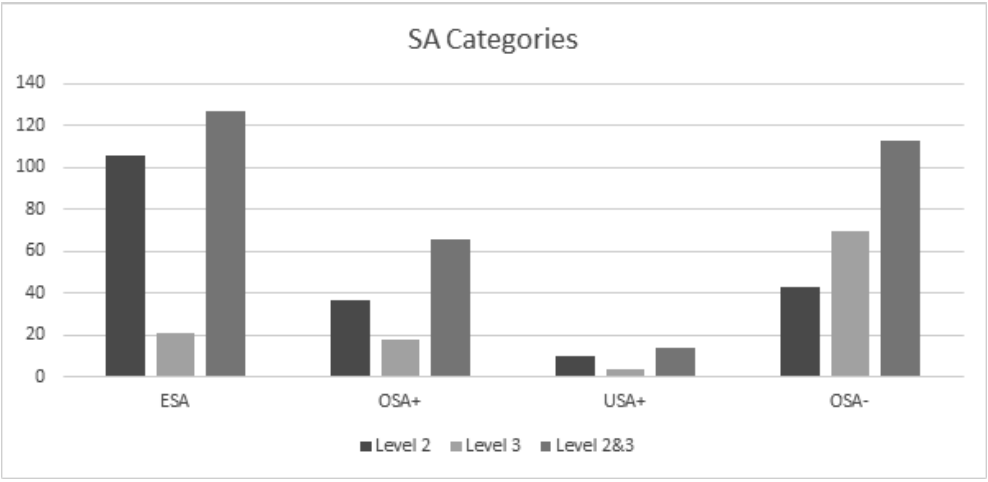

Figure 2. ESA, OSA+, USA+, and OSA- total scores (320 skill assessments).

DISCUSSION

Grouping the data by semester and proficiency levels shows some unsurprising results. The first-semester accuracy average for four classes, consisting of two level 2 and two level 3 classes came to 74%. The level 2 classes mean SA score came to 85%, while the level 3 classes came to 63%. In semester 2, both level 2 classes SA accuracy scores came to 71%. The level 3 class from that same semester was surprisingly low, at 21%. This data had some effect on bringing total SA scores for the two semesters down.

One possible explanation for the difference in accuracy levels between the first and second semesters could be that, not only do students in the first semester have fewer skills to use, but the first-semester skills tend to be easier to use, as second-semester skills are more focused on critical thinking and analysis. It could be that the additional cognitive burden of thinking critically is an obstacle to correct SA. These observations suggest that it is unsurprising to see higher levels of SA accuracy in the first semester of EDC.

Previous experience with SA in this particular course suggests that there may have been a failure to train students to self-assess accurately. It could also be that some students may simply not be invested in the process and approach it with a cavalier attitude or have a lack of motivation to really think hard or try to assess honestly. With a nearly equal percentage of somewhat correct assessments as completely incorrect assessments, there was clearly some difficulty for some students in assessing themselves accurately. Interestingly, when the ESA, OSA+ and USA+ data are combined, the total of correct SAs comes to 61%. This could be considered to be approaching a reasonable level of accuracy. At the conceptualization stage of this project, it was hoped that an SA average of over 80% was possible, perhaps even closer to 90%. This seemed reasonable considering the relatively small number of skills, and course-specific details such as a strong focus on skill use during discussions.

To understand the SA accuracy more completely in the present context, a multi-faceted analysis of the data would have been more useful in dealing with differences in SA accuracy by considering gender, student motivation, class level, average TOEIC scores or majors The present study is limited by the use of descriptive statistics, but the data collected may suggest that with careful training, students may be able to conduct SA more accurately, specifically eliminating most of the overestimation errors where no skill was used in a particular discussion.

In the context discussed in the present paper, perfect accuracy may not be necessary, or possible, to be pedagogically valuable in SA. If students in oral discussion tasks are able to accurately report their own use of Discussion Skill phrases, there may be support for a role of SA as a supplement to teacher assessment, especially in larger classes where teacher attention is more divided. Additionally, it may prove valuable to students to understand the differences between qualitative and quantitative SA as a means to devising new approaches to increasing participation and retention of new language skills. Finally, it could be interesting to consider the use of gamified quantitative SA as a path to harnessing the power of extrinsic motivation.

One positive pattern in the data shows that many students were able to assess themselves quite accurately. Out of 80 assessments, 32.5% assessed themselves correctly in every category. 18 students, or 22.5%, assessed themselves correctly in 3 categories, 21% correctly assessed 2 of the 4 categories, nearly 11.5% correctly assessed one category, and 12.5% incorrectly assessed in all categories. If this kind of quantitative SA is to be considered valid, there will need to be better ways to teach students how to assess their own performances in group oral tasks. While it is ambitious to expect perfect accuracy, in a group oral task that measures the use of phrases, it would be hoped that students could reach a much higher accuracy level.

The combination of SA categories did show some promising patterns. If perfect accuracy in SA is not expected, then there may be some benefits to allowing for particular errors in a

definition of "correctness". That is, if a student overestimates or underestimates skill use in group oral tasks, they are still making a correct SA, albeit an inaccurate one. Visualizing the data in this way can help instructors to hone in on the particular skills that may be more difficult to self-assess. In future studies, perhaps there will be some indication of the particular types of function phrases that evade recall, and this can be taken into consideration when SA accuracy is desired.

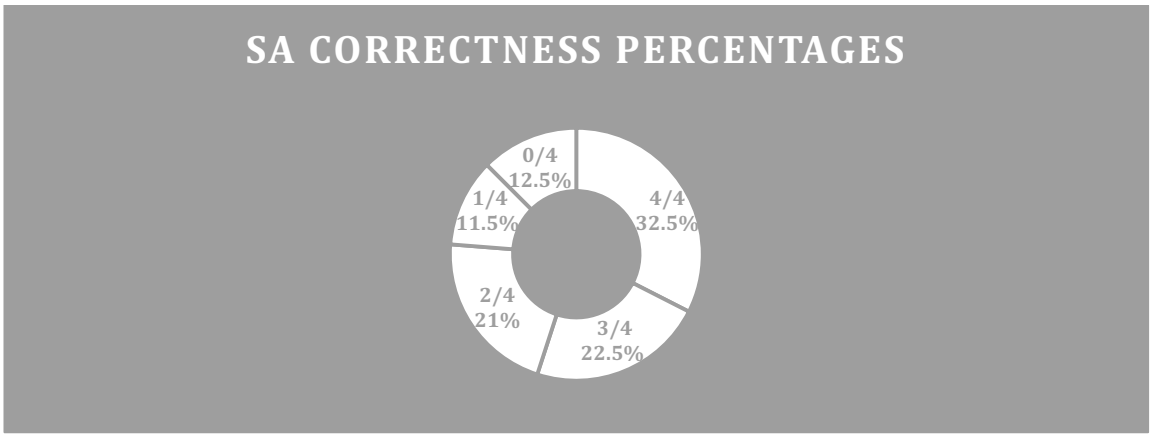

Figure 3. SA correctness percentages using combined ESA, OSA+, USA+, and OSA- scores.

CONCLUSION

As the results of this classroom research study do not fully support SA accuracy in the EDC context, it can be suggested that a more robust repetition of the present study could result in supporting the notion that students are able to accurately self-assess, but this will have to be attempted before more confident conclusions can be made. The results may be in agreement with other studies that indicate a lack of SA training may explain inaccuracy (Ross, 1998, p. 8). There may be other factors, such as a lack of motivation participate in the study. Past experience with student SA in this context suggests that the results included here may not be representative of the majority of students who have taken, or will take, this course in future. As SA is such an integral part of many EFL courses, in particular in the context under examination in this paper, accuracy of SA should be a continuing goal to improve learner autonomy, performance, and motivation. It will be interesting to explore the topic more deeply and more precisely so that different variables can be taken into consideration. The use of video recordings is also an area that could be interesting to explore. It can be used as a means to achieving not only better SA comparisons with teacher assessment but also providing students with additional feedback in the form of recordings of their own discussions to be revisited and reassessed in their own time. It is hoped that the use of quantitative SA can be used alongside qualitative SA to not only achieve practical aims such as aligning students with assessment rubrics, but also as a means to gamify activities for less motivated students. If the particular local conditions allow, an exploration of the potential merits of using student SA to supplement instructor assessment as a way to create the fairest assessment in larger classes would be an interesting area to explore. One final conclusion might be drawn that, as many supporters of SA as a classroom tool have suggested, interpretation of rubrics might be brought into the realm of collaborative effort, allowing students more of a role in the creation of assessment parameters.

REFERENCES

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice.* Berkshire: Open University Press.

Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: a critical analysis of findings . *Higher Education, 18*, 529-549.

Davies, B. (2012). Comparing the effectiveness of different kinds of self-assessment questionnaires. *New Directions in Teaching and Learning English Discussion, 1*(1), 93-98.

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education, 24*(3), 331-350.

Hurling, S. (2012). Introduction to EDC. *New Directions in Teaching and Learning English Discussion, 1*, 1-9.

Long, T. (2019). Strategic instructor positioning for accuracy in assessment. *New Directions in Teaching and Learning English Discussion, 7*, 231-239.

McEntee, P. (2017). Game of functions. *New Directions in Teaching and Learning English Discussion, 5*, 88-94.

Munoz, A., & Alvarez, M. (2007). Students' objectivity and perception of self assessment in an EFL classroom. *The Journal of Asia TEFL , 4*(2), 1-25.

Oscarson, M. (1989). Self-assessment of language proficiency: rationale and applications. *Language Testing, 6*(1), 1-13.

Ross, S. (1998). Self-assessment in second language testing: a meta-analysis and analysis of experiential factors. *Language Testing, 15*(1), 1-20.

Appendix A - Student self-check sheet example

Discussion Skills		Name:	Date:
1. Connecting Ideas			
Asking Others to Connect	Connecting to Others	**Number of Uses**	**Number of Uses**
Do you agree with me? What do you think of (my) idea?	As you said… You said… But I think…	⓪①②③④⑤	⓪①②③④⑤
2. Closing Topics			
Checking for more Ideas	Summarizing	**Number of Uses**	**Number of Uses**
Is there anything to add? Is there anything more to say?	So we agree that… So some/most of us think…	⓪①②③④⑤	⓪①②③④⑤
Communication Skills		**Number of Uses**	
Reactions		⓪①②③④⑤	
Showing If you Understand		⓪①②③④⑤	
Checking Understanding		⓪①②③④⑤	
Paraphrasing Others		⓪①②③④⑤	
Paraphrasing Yourself		⓪①②③④⑤	
Asking for Repetition		⓪①②③④⑤	
Asking for Explanation		⓪①②③④⑤	

Appendix B - Instructor discussion check-sheet example

Discussion Skills	**Interrogative**	**Declarative**
Possibilities		
Choosing Topic		
Communication Skills		
Reactions		
Showing		
Checking		
Para Other		
Para Self		
Asking Explain		
Asking Repeat		