Strategic Instructor Positioning for Accuracy in Assessment

Tobias Long

ABSTRACT

This paper presents a two-pronged exploratory study designed to examine one instructors' methods of classroom observation and assessment in EFL discussion classes. In the English Discussion Class (EDC), class sizes of seven to nine students mean that two or three group discussions of three to four students take place simultaneously. Students are graded during discussions on their use of target language, which can be difficult do to for multiple simultaneous discussions. To investigate the best possible way to position myself in relation to a group, lessons were recorded with video cameras and audio recorders. These recordings were then reassessed using the notes taken in the original, recorded lessons. The second part of this study consisted of an instructor survey on styles of observation positioning used by instructors. The results suggest that central positioning is generally more reliable for multiple group assessment.

INTRODUCTION

During my time as an instructor at Rikkyo University's English Discussion Class (EDC) I have consistently felt challenged to both assess students fairly and accurately. This challenge perhaps comes from the fact that in any given class in EDC, there are usually two groups of four students carrying on a discussion at the same time. Two discussions are held during each lesson with the first lasting at least ten minutes and the second lasting at least 16 minutes. Particularly with more enthusiastic discussion groups but also quieter groups or students, I thought that I was not consistently able to accurately record instances of target expression use. I hypothesized that constraints on my capacity to dual or multi-task, maintain focus on a particular speaker or group and on my ability to separate and/or locate utterances in a multi-speaker class setting could be mitigated to a degree by a more systematic movement pattern around the classroom. By comparing assessment scores while positioning myself closer to one group with scores assessed in a more central position, I hoped to get some ideas on better ways to observe and assess more accurately.

In EDC, students are assessed in regular and review lessons as well as in discussion tests held three times per semester. These discussion tests are different from regular and review lessons in that instructors assess only one group at a time. High levels of rater reliability for discussion tests was shown to be achievable though departmental training sessions (Doe, 2012). For regular lessons, review lessons and discussion tests, a clear grading rubric is available to all instructors. The rubric is provided to all teachers in an Instructor Handbook, and teachers are trained on how to apply this rubric in a uniform fashion through faculty development. Another simple reason for this reliability is that discussion tests consist of one group which the instructor has to assess. In this project, I hoped to assess my own assessment in two-group simultaneous discussion classes, and along the way, perhaps find some better ways to deal with difficult to assess students and groups as well as ways to mitigate external factors that make accurate assessment a challenge.

Some of my classes seem to get along quite well. Their discussions have an academic tone, but their good rapport leads to excitement and an enthusiasm to share their ideas. With up to nine students in a classroom, the noise can be distracting. The cocktail party problem, first discussed by Cherry (1953), gave me some insight into the trouble I was experiencing: it is simply the human ability to "recognize what one person is saying when others are speaking at the same time" (975-976). In his original study, Cherry mentions several factors that enhance listening, including directional clues, lip reading or gesticulation, voice clues like tone, pitch or speed, accents, and transition possibilities. This final factor led him to suggest that the human ability to use these kinds

of clues means that we have a "store of possibilities" and that this store "enables prediction to be made, noise or disturbances to be combatted, and maximum-likelihood estimates to be made" (1953, p. 977). Important to the purposes of this exploratory study was the logical assumption that attention is a major factor in predicting successful assessment in a multi-speaker environment. There are some factors that affect successful listening comprehension. In Cherry's dichotic listening tests, participants had very limited accuracy of recall in unattended audio channels. In other words, they could not remember or notice what they were not actively paying attention to. The information they did recall was more physical, for example the gender of the speaker or that the speech was human speech (p. 977). The implication for teachers who assess multiple students simultaneously is that although we may decide where our attention is at any given time during a lesson, we may not be noticing instances of the language we are assessing. This could present problems when accurate assessment of multiple speakers is the goal.

Since Cherry first described this phenomena, many studies have examined the limitations of human ability to recognize speech in noisy situations. The cocktail party problem "turned out to be a highly complex one, involving not only attention but also acoustic phenomena, masking, binaural processing, and all effects contributing to our ability to segregate (speech) signals, sometimes referred to as auditory scene analysis" (Bronkhorst, 2000, p. 117). In the context of a discussion class consisting of two groups speaking simultaneously, the auditory scene presents a host of factors that may affect accurate and complete assessment as well as the quality and credibility of feedback.

A common classroom scenario can be seen in the case of a particularly enthusiastic class. Their discussions are noticeably livelier than other classes, leading to a cacophony of voices: opinions, reactions, rejoinders, and follow-up questions. This presents a dilemma for the assessor. When the objective is to assess students on the use of certain discourse markers or questions, some students may not be awarded proper marks for their use unless noticed by an instructor. A study on listening in multi-talker environments described the necessity in command and control situations, where a listener must focus on multiple channels, as "some channels are likely to contain more useful information than others", and that in "high-context situations there are instances in which a listener's attention can be drawn to highly relevant information originating from an unexpected source" (Brungart & Simpson, 2007, p. 80). In other words, while listening to one group or a particular student, our attention shifts involuntarily.

METHOD

The data used in this exploratory study came from two assessment styles: Centrally positioned assessment and peripherally positioned assessment. In two-group classes, I set up chairs and desks for both groups. The group's tables were set up in such a way that all students could see the whiteboard or screen fairly easily. To ensure the best chance of reassessing accurately, I wanted to use video as well as audio recordings. The audio device could be placed in the center of the discussion tables to ensure voice amplification. The Zoom H2n recorders I chose for this study offer a spatial recording feature. With stereo speakers or surround sound audio, spatial recording allowed for more accurately pinpointing who was speaking.

Collecting data was fairly straightforward. In lessons where I observed from a central position or when I moved freely about the classroom, I simply recorded instances of student use of target skills. The discussions were recorded and after editing work was completed, saved to a hard drive with folders indicating which group was recorded, the lesson date, class, and discussion number. These edited discussions were then watched with a copy of the class notes taken during the corresponding discussion. Any instances of target skill use were checked against the original notes, and correct or missed assessments were noted.

For peripheral/rotation observation discussions, I used a timer application that includes an interval alarm. This way I could know when to move from group to group. I stayed closer to one group for one quarter of the discussion, then moved to the second group. The discussions I recorded were of a duration of 16 minutes or longer, so the usual time spent closer to one group was four minutes.

Skill check sheets were printed in a grid. A total of eight boxes, one for each student, made it possible to use the check sheet easily from a different vantage points within the classroom. The skills assessed were encoded using abbreviations. For example, "asking about viewpoints" was shortened to "AAV". The grid and encoding allowed for faster notetaking without having to decide who was speaking, finding their name on the check sheet, and then marking a skill use instance.

On the check sheets, the type of observation style (centrally positioned or peripherally), the interval to spend with each group (peripherally observing), impressionistic notes about the atmosphere and noise level, and any other information to identify the class, the date, and unit being studied were also noted. After the discussions were assessed, the score sheets were filed and the video and audio data were combined and edited. The errors were tallied and separated into categories reflecting the type of assessment positioning, the level of classes, and the subjective loudness of the discussion. By watching the recorded discussions with class note sheets from the corresponding lesson I was able to reassess the discussions and tally missed opportunities for an affirmative assessment of student skill use instances (MSKI).

RESULTS

By separating the MSKI data by instructor assessment position, student proficiency level, and group volume, I was able to get some ideas about the effects these may have on assessment accuracy. Speaking volume was entirely subjective and based on my experience teaching at EDC. The categories are "quiet to normal", "normal", and "normal to loud". I recorded at least two discussions per class with two video cameras and two high fidelity audio recorders. Five of the classes videos (10 total discussions) were observed from a central position in between both groups and the other five classes were observed on a rotation system, where I placed myself on the periphery of the classroom, closer to the group being observed. I moved from one group to the other, dividing the discussion into four intervals. During the time observing one group, I was still able to assess the second group and did so when possible. However, as can be seen in the results, more skill use misses occurred.

Table 1. Instructor Positioning MSKI.

Assessment Position	**Total Number of MSKI**
Central	36
Peripheral	45

*Central= between two discussion groups, Peripheral= Closer to one group

Table 2. Student Volume MSKI.

	Total MSKI	**Per Class MSKI**	**Per Group MSKI**	**Per Student MSKI**
Quiet	22	7.3	3.6	1
Normal	41	10.3	5.1	1.3
Loud	18	6	3.3	0.8

Table 3. Student Level MSKI.

	Total MSKI	Per Class MSKI	Per Group MSKI	Per Student MSKI
I	28	9.3	4.7	1.2
II	31	7.8	3.9	1
III	22	7.3	3.7	0.9

*Levels in EDC are stratified from I to IV, with Level I being the highest proficiency, and Level IV the lowest.

DISCUSSION

This limited data suggests sitting in a central position is best if the goal is to get the most valid and accurate assessment for the majority of students. However, it should be noted that this may differ from rater to rater. The total number of logged MSKI from a central position came to 36. Five classes were observed, meaning 20 discussions in total. Compared to the 45 logged MSKI in peripheral observation positions, it is not surprising that being further away from a group will lead to more MSKI.

Sitting closer to one group, an instructor may be able to more accurately assess that particular group, as the pressure of listening to two groups simultaneously is reduced. However, this claim cannot be proven without more refined data collection and analysis. For example, during recorded assessments an instructor could clearly track exactly where they were positioned during a discussion and then compare this data with video recordings. This could allow for a comparison of accuracy between assessment on groups that are actively attended to and those that are not.

As for the relative volume of discussion groups, a more rigorous measurement using an audio device or software to measure decibels would add richness to this data and analysis. Is my subjective categorization, three classes fell into the quiet to normal range, and the average number of MSKI came to around 1 per student. In the normal range, the number increased to 1.3 per student. In the classes I assigned to the normal to loud range, the number was lower at 0.9. Perhaps the louder groups negate the effects of instructor positioning during assessment.

Looking at class level data suggests that there may be some effect of English language proficiency on individual instructors' rating reliability. In the highest level classes, the per class number of MSKI was highest. This could have to do with the speed and relative fluency of the students or the fact that they may be using different lexical chunks to mark the skill being assessed, as the rubric allows for any number phrases besides those that are included in the textbook. So long as the language used accomplishes the goal of the skill, the language is acceptable and scored as target skill use. More proficient students have a larger available set of phrases with which they can satisfy this rubric, which can make the instructors' task of attending to and assessing skill use more challenging.

For lower proficiency classes, assuming students are more reliant on the exact textbook target language, the job of assessing students may be easier. Looking at the data, Level III classes had the lowest per class MSKI scores, at 7.3. The relatively predictable nature of their discussions and slow speed are the most probable reasons for increased accuracy of assessment scores. There are other factors, of course, such as the relative volume of individual student voices, the level of participation in the groups, and the rapport of the groups, all leading to potentially less acoustic activity for instructors to process. It has been my experience that level III classes are the easiest to assess, due to the slower speed of interaction and on their reliance on textbook phrases.

The data for level II classes is not drastically different to the data for Level III classes. With an average of 7.8 MSKI per class and 3.9 per group discussion, the number of assessment errors

is only slightly higher. It has been my experience that Level II classes are generally easier to assess, but they can occasionally be quite proficient, speaking a relatively fast rate, making accurate assessment more of a challenge.

Instructor Survey

Another part of this exploratory study included a survey of EDC instructors on assessment styles and challenges. Of the 28 respondents, 71 percent of instructors reported a tendency to monitor two groups simultaneously during two-group discussion classes. In the rarer case of three-group classes, instructors still seemed to prefer observing two groups at the same time. This makes sense intuitively, considering that part of the responsibility of an EDC instructor is to observe all the students in a class. One might assume that this central position would lead to fairer, or more accurate assessment. One instructor suggested that listening to one group at a time might be unfair. Others stated that listening for skill phrases is not a particularly difficult task, and that sitting between two groups made this much easier.

There are drawbacks to this central approach, however. One pertains to with particularly lively classes. Sitting so close to two excited groups can lead to involuntary shifts in the rater's attention. One instructor stated: "Usually my attention bounces around the room haphazardly picking up on student utterances which pierce through the din of talking students."

Another reported: "I can't properly monitor two or three groups at a time so I monitor one at a time. I know that I am missing things from the groups that I am not monitoring so I listen to one question/topic from a group and move on." On the topic of fairness, it could also be said that listening to two discussions simultaneously could be seen to be unfair as well, depending on the level of detail we want to capture in addition to our assessments. If it is indeed true that our attention can be captured involuntarily, perhaps we would be better to maximize attention to one group at a time. In a study on the control of auditory attention in dichotic listening situations, Koch found "clear evidence for a temporal limit in the flexibility of auditory attention […] in selective attention situations" (Koch et al., 2011). In other words, we cannot always decide when we want to shift our auditory attention to another group or speaker.

It may be that any information that is relevant to a listener can cause a shift of attention. For example, hearing one's own name in the context of a social gathering may result in a shift of attention towards the speaker who spoke it. The EDC syllabus prioritizes certain phrases over others, and instructors assess them during every lesson. We are, in a sense, primed to pick out these phrases as they are relevant to us, and to our student's grades. This priming can lead to unintended attentional shifts away from the target of our auditory attention. In other words, these target phrases are attention grabbing. Sitting between two discussion groups can be considered to be something not unlike a social gathering, but with listener's attention focused on everyone in the room and being shifted involuntarily about by various speakers. In response to a question of the kinds of classroom events that cause attentional shifts, instructors mentioned the volume of another group, reactions, behavior, inappropriate language, heavy Japanese use, ideas that would be useful for feedback or class written comments, breakdowns in communication, creative uses of target language, or just interesting ideas. More significantly, 75 percent of respondents said that their attention was drawn away from the group they were trying to observe always or often in classes.

Besides shifts of attention, another concept central to the discussion around the cocktail party problem is known as masking. Masking occurs when the perception of a sound is affected by the presence of another sound. There are various forms of masking, including effects having to do with timing and sound frequency. As humans have the ability to listen with both ears, we are able to pick up auditory input from, say, one speaker or group to our left and another to our right.

The problem, however, is the things we are trying to listen for can be drowned out, or masked, by things we don't want or need to hear. In my own experience, I have noticed that some voices are quite hard to distinguish from another student's, or very quiet, making them harder to hear and thus harder to assess. When listening to assess students' speaking performance, it seems more likely that masking effects will be greater when positioned between two discussion groups.

Sitting closer to one group gives us much more information about the group we are observing. The gestures, body language and lip movement all help us to know who has spoken, is speaking, or intends to speak. Given that we can more clearly hear our students due to the simple fact that they are closer to us is an obvious advantage as well. In addition to this, by sitting closer to one group, we can minimize the effects masking and attentional shifts.

The disadvantages to this approach of observation in two-group classes should be clear. The further away I move from one group, the more difficult it is to assess them accurately. I tried to mitigate these effects by timing my movement around the classroom and spending equal amounts of time with both groups. In my experience, sitting closer to one group does allow me to assess that group more accurately and to take more in-depth notes to be used as feedback. In the EDC, we emphasize the use of particular discourse markers. Consideration of the appropriateness of students' uses of these markers is an important aspect of the instructor's duties. Besides this, as a teacher, I strive to help my students use language as appropriately as possible, and therefore look for opportunities to give meaningful, actionable feedback. I found that sitting closer to one group and giving them as much of my attention as possible, I could come closer to this goal. This is not to say that sitting in a central position makes this impossible, rather it can make it more difficult, especially in some of the particularly difficult to assess classes.

Another approach to assessment positioning that I have tried in EDC classes is a more flexible approach to positioning. By moving around the room, I can find acoustic (and visual) vantage points that help me assess more accurately. The advantages of such an approach shares features of assessing from a central position and from a more peripheral one, closer to a particular group. The main benefit of flexible assessment positioning is that we can follow the action (or lack of it) or go where we guess we need to be to make an assessment while still keeping ourselves in a better position to assess our students. We can work with the acoustics of the classroom and with our students. Assuming that students would like to score well but are quiet, particularly shy, or that their voices are simply drowned out by the other speakers in a class, I believe the burden of assessment is not theirs to shoulder. By moving around the classroom more flexibly, without any limits to the time spent with any one group, I can focus on students who may not be scoring particularly well, and determine more accurately whether their uses of target expressions are being missed, if they simply aren't using them, and if they are using them appropriately. Of course, this approach shares drawbacks with a one-group, peripheral assessment strategy. Namely, that I may miss significant uses of target skills by the group I am further away from and not actively paying attention to.

It seems to me that when instructors choose between an assessment position closer to one group or in a more central position between two groups, they are making a choice between accuracy for one group or a more reliable assessment for all students. The first choice is not only more accurate assessment of individual students, but a better contextual understanding of the discussion and the use of skills within it and possibly more accurate and detailed feedback. The second choice is a more general understanding of the types of phrases used by the class, the kinds of ideas expressed, and most likely, a more reliable assessment of a greater number of students.

I had hoped that a timed, rotational assessment strategy would help me assess more accurately. Within the EDC rubric, a missed opportunity (or more than just one) to correctly assess the use of a target skill or phrase can have a relatively significant effect on a student's grade, and

possibly on their confidence as well. A surprising result of the instructor survey had to do with missed utterances of skill use by instructors. Initially I thought that most instructors were fairly confident about their listening ability in these two- or three-group discussion classes. However, in two-group classes, 14.3 percent of instructors thought that they "always" missed skill use utterances, 64.3 percent felt that they "often" or "sometimes" missed skill use utterances, and the remaining 21.4 percent answered that they "rarely" did. Not one instructor answered "never".

One of the questions I hoped to explore was whether spending a predetermined amount of time with each group could be as accurate as the centrally positioned observation style. Only two instructors in the survey alluded to using timers to assure equal time spent with both (or all three) groups. Of the instructors who did mention moving around the room or spending time observing one group, most seemed to do so more flexibly.

The final question of my survey asked about the most significant challenges of assessing students in EDC. Quite a few instructors felt that assessing multiple groups simultaneously is one of the biggest challenges we face. Other instructors suggested that giving accurate grades was a concern, or giving equal attention to all students. Of all 28 responses, 17 answers to this question touched on the difficulty of assessing two groups simultaneously. Some worried about the accuracy of assessment scores, others expressed concern about feedback, and others were more focused on the attention we can give to students. One response that deserves mention suggests that there may not be a better way to observe and assess larger groups of students, particularly discussion classes with two or three groups. These were the same ideas that led me to start this exploration of the topic from the start. Despite my own difficulties with assessment, I am optimistic that the discussion class assessment system can continue to improve.

CONCLUSION

Assessing multiple students can be challenging, and there are many factors to take into consideration when deciding how to position oneself relative to students. Are there multiple groups of speakers? What are the lesson aims? What is the assessment criteria? How proficient are the students? How well do the students get along? Are they engaged with the material? What kind of feedback does an instructor hope to give them?

The hypothesis that strategically positioning oneself around the room during discussions would lead to more accurate assessment and better feedback could not be supported based on the data I collected. The initial hypothesis of this study was, perhaps, too ambitious for an exploratory study to answer. The major weakness of my data collection and analysis missed two key data points: firstly, when positioned closer to one group, the data collected for that group would, according to my initial hypothesis, be more accurate than data collected sitting centrally between two discussion groups. This seems to be an obvious point. However, I was unable to look at a crucial piece of the data due to an oversight in my design of the check sheets as well as a lack of awareness of a tool built in to the audio recording devices. By including notes to indicate exactly which group I was with during each interval, I would have been able to determine the accuracy of those assessments compared with the non-attended groups over a discussion, for a class, and for the entirety of the semester. I also would have been able to compare this data with that of the centrally-positioned observation data. I discovered another way to track my movement in class too late: the audio recorders I used have remote control devices that allow time-stamping the recordings. I was unaware of this feature for much of the time I was collecting data. Secondly, it is perhaps impossible to objectively measure the quality of feedback. To compare the quality of feedback given after close observation of one group with that given after observing two groups simultaneously would require more time and a more sophisticated design of data collection and measurement, likely one using a rubric to assess feedback quality.

As mentioned previously, EDC rater reliability is high in discussion tests. Teachers can focus completely on one group, and mistakes of the sort that this study looked at are presumed to be rare. That is not to say that it doesn't happen. Even observing one group at a time, as in during discussion tests, it is entirely possible to see masking effects where quiet voices can be silenced in the milieu of voices. Generally speaking, however, it is not nearly as challenging to listen to three, four, or five speakers having a discussion and assess use of set target phrases as it is to listen to two groups at the same time. This study's results do offer some value for improving my own assessment strategies and I think also present some interesting areas to investigate further.

As for improving my assessment, I am more aware of the types of factors that make observing two groups difficult. I can react more positively to acoustically demanding situations by moving into different positions and I can feel more confident in my choices as to the costs and benefits of choosing them. As far as future investigations, it would be an interesting challenge to adapt the data collection methods of this study to compare the overall effects on student grades of different assessment patterns. For example, how do peripheral and central observation strategies affect student scores? Perhaps looking at self-assessment as a supplement to instructor assessment could be examined? The challenge of assessing multiple students seems to best be dealt with by a combination of approaches, with central positioning seemingly best for speaking skill assessment. Nevertheless, if our goal is the most accurate assessment possible, there is always room for improvement.

REFERENCES

Bronkhorst, A. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica United with Acustica, 86*(1) 117-128.

Brungart, D., & Simpson, B. (2007). Cocktail party listening in a multitalker environment. *Perception and Psychophysics, 69*(1), 79-91.

Cherry, E. C. (1953). Some experiments in the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975-979.

Doe, T. (2012). Assessment: Improving Rater Reliability on the EDC Discussion Test. *New Directions in Teaching and Learning English Discussion*, *1*(1), 1.11-1.18.

Freyman, R., Balakrishnan, U., & Helfer, S. (2004) Effect of the number of masking talkers and auditory priming on informational masking in speech recognition. *Journal of the Acoustical Society of America, 115*(5), 2246-2256.

Koch, I., Lawo, V., Fels, J., & Vorlander, M. (2011). Switching in the cocktail party: Exploring intentional control of auditory selective attention. *Journal of Experimental Psychology: Human Perception and Performance, 37*(4), 1140-1147.