The Bulletin of International Education Center Tokai University
Vol. 1, No. 1, 2019, pp. 23-42
DOI: 10.18995/24347337.1-1.23

Student Attitudes and Speaking Assessment

Simon WILKINS
Tokai University

Abstract

MEXT argues that amid ongoing globalization, it is essential that universities develop an educational environment where students can acquire English speaking skills. Within its *Five Recommendations on the English Education Reform Plan Responding to the Rapid Globalization,* MEXT also calls for the improvement of evaluation methods while taking into account students' motivations and attitudes. This paper details one cycle of a multi-cycle action research process that critically analyses current evaluation methods in a local context and students' attitudes towards them. Both qualitative and quantitative data were collected, analysed and triangulated via an emic perspective, with qualitative data analysed inductively and quantitative data analysed through the Rasch model. Data analysis found that students experience a range of strong emotions during conversation in the classroom, and particularly during the assessment of this conversation. The most prominent of these emotions was nervousness. Observation and reflection highlight potential sources for this nervousness and propose future pedagogical and research approaches that can help mitigate this nervousness during assessment.

Keywords: Assessment, English conversation, action research, student anxiety

1.Introduction

The desire for an increased focus on spoken output has been increasingly emphasized in Japan. Reports from the Japanese Ministry of Education, Science and Technology and Education (MEXT) on the Reform of English Teaching Methodology (1947, 1998, 1999, 2011, 2014) have consistently highlighted the need for improved speaking skills. Recent initiatives from MEXT (2014) have also called for the adoption and verification of evaluation systems that take into account contextual factors specific

to Japanese students, these include “students' motivation and attitude for active learning” (p.2), and that students should not be “afraid of making mistakes” (p.2).

This paper represents the first cycle of a larger study that adopted a critical analysis of current assessment procedures in the researcher’s own classroom. The research focusses on a class that aimed to develop students’ casual conversation skills and measure their progress. The study focused on Japanese students’ attitudes to assessment procedures in class and attempted to take into account such attitudes in future cycles of research. It is hoped that this “verification” process outlined by MEXT (2014) will hold resonance in similar language classrooms. Both qualitative and quantitative data were collected and analysed during an action research process via an emic perspective, with qualitative data analysed inductively and quantitative data analysed via the Rasch model. The results in this paper will focus on data relevant to students’ attitudes to evaluation and the assessment instrument itself.

Previous studies in the local context (Wilkins, 2018) and the wider Japanese context (Kitano, 2001; Ohata, 2005; Corlett, 2012;) reported a strong emotional response from students towards assessment of their speaking skills. Negative emotions that were reported and may require remedial attention were feelings such as anxiety, nervousness, and frustration. Reasons for these emotional responses are initially considered through reference to the literature, and the planning and action stages of this first cycle of action research reflect the findings of previous studies.

2. Literature Review

In the wider context of language classrooms, research suggests a number of explanations as to why anxiety is a notable student reaction to speaking, and in particular, assessment of speaking. Young (1991) suggests a number of sources for potential student anxiety in language classrooms found in a range of language classrooms; including students’ interpersonal skills, a gap between student and teacher expectations, classroom procedures, and feedback and assessment. These sources of potential anxiety raised by Young (1991) can be further summarised as those which emerge as a result of pedagogical tensions, and those which stem from cultural issues.

In the context of Japan, potential pedagogical tensions are highlighted by Rapley (2008), who explains that although English-speaking skills are considered important in

Japan, entrance examinations at the senior high school and university levels exert the greatest pressure on Japanese teachers of English, and as such produce teaching models that are not in accordance with the intentions of the MEXT reports. Rapley describes these "traditional Japanese methods" (p. 1) as focusing on elements of grammar and translation that are not conducive to productive language use. In the context of the present study, new assessment procedures that aim to measure speaking performance via a syllabus that focuses on a communicative approach may, therefore, enhance feelings of anxiety. In such a pedagogical context, it is claimed that the communicative approach may not lead to fluency, but rather to "desultory silence" (Scrivener 1994, p. 1), as students who attempt to develop speaking fluency in their classes fail to grasp the perceived need to speak, and thus lack the motivation or necessity to produce talk.

Established assessment procedures in the Japanese context are another pedagogical factor in considering Japanese students' attitudes to speaking and its measurement. In Japan, summative assessment procedures such as university entrance exams, or TOEIC, remain the primary recognized measurement of student achievement (Cohen and Spillane, 1992; Mulvey, 2010; Watanabe, 2004). Attempts by MEXT to address this situation in 2004 by introducing learner-centred methodologies such as Assessment for Learning (AfL), were later abandoned as "misguided" (Takayama 2008, p.3). Takayama argues, however, that this does not mean that there is no future for these assessment procedures in Japan and that further research is necessary. Takayama (2008) points out the homogenizing effect of the Programme for International Student Assessment (PISA) rankings and the strong regional competitiveness Japan has with high-ranking PISA regions such as Singapore, Shanghai and Hong Kong. Since these regions have adopted policy-supported AfL procedures, and have performed well in PISA rankings, factors such as homogeneity and regional competitiveness, which Takayama (2008) describes as being highly influential on educational policy, may compel Japan to follow suit. The Central Education Council (1999) claims that changes to the entrance examination system have been under way for some time. Mulvey (2010) predicts that changes are inevitable, due to rapidly falling admission rates for universities, which make the entrance examinations redundant.

Cultural issues are further factors to consider when taking account of Japanese students' attitudes. McDonough and Shaw (1993) explain that communicative language

teaching has led to shy or introverted students being placed at a disadvantage. According to the research, the notion of shy or introverted students has particularly resonance in Japan. It has often been observed that reticence in speaking is familiar to teachers in the English as a Foreign Language classroom in Japan (Anderson,1993; Doyon 2000; Marchand, 2012). This includes behaviour such as students who rarely volunteer answers, and protracted silences and repeated pauses during conversation. Turner and Hiraga (1996) describe Japanese students in Britain as passive and unwilling to engage in dialectic and analytic discourse. They claim that this may stem from cultural factors caused by the Japanese academic value of demonstration over transformation of knowledge. Nakane (2006) highlighted the difference between Japanese and Australian students by claiming that Japanese students commonly used silence in order to avoid mistakes, while Australian students used verbal strategies in order to find compromise over misunderstandings.

Another potential cause of anxiety during assessment of students' conversational skills lies in the nature of conversation itself, and the interpersonal features involved (Young, 1991). During conversation, the tenor of discourse describes the social roles we assign ourselves and others when we are speaking. Our role in an interaction will influence the language we use. Eggins (2004) compares "informal" and "formal" situation types. Informal interactions are those between close friends who see each other often and freely express their feelings. Meanwhile, formal interactions are between strangers from different social levels who do not meet often and are not free to express feelings openly. During an assessment of casual conversation, a formal situation is created in what is typically an informal context. Eggins divides the tenor of discourse into three different continua:

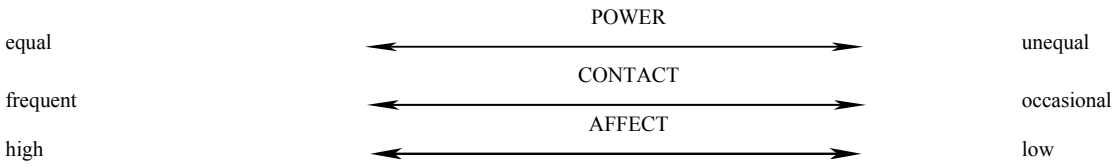

Figure 1. The continua of power, contact and affect (Eggins, 2004, p. 100)

Power refers to the extent to which the relationship between the interactants is equal (such as between two friends) or unequal (an employer and employee). Contact

means the frequency of interactions, with friends, family members and neighbours at one end, and strangers at the other. Affective involvement refers to the extent to which emotions or attitudes are expressed freely between interlocutors. This may be high between friends, spouses and children, and low between passengers on a train. The teacher–student relationship promotes unequal power. Teacher–student contact is usually occasional, and affective involvement is low. Student anxiety thus arises when students speak with a teacher.

Another reason for anxiety might be that a casual conversation includes visual and aural contact, and feedback from the teacher is immediate, as illustrated in Figure 2:

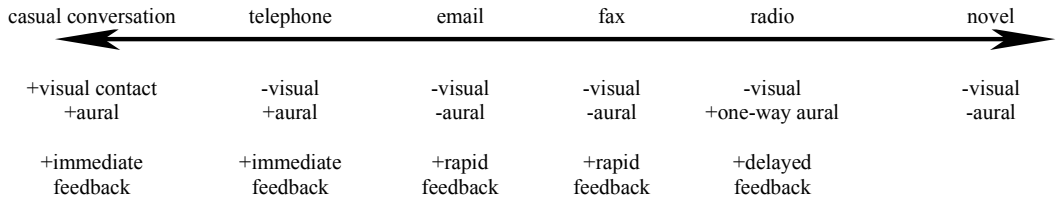

Figure 2. Representation of interpersonal distance (Eggins, 2004, p. 91)

The nature of a casual conversation means that students' use of the English language is accompanying the action. The notion of immediate feedback, which might include negative feedback from a person whom they consider a model of accurate language, could increase anxiety. In order to reduce nervousness in the context of conversation, particularly while measuring student performance, affective involvement and interpersonal distance should be taken into account during assessment design.

3. Method

3.1 Participants

Forty-five students in a compulsory English Conversation class in a university in Japan participated in this cycle of the research. Students were of mixed gender and mixed ability. The average TOEIC Bridge score of all students at the university was 117 during enrolment. All students were non-English majors.

3.2 Research Design

This study locates itself within an action research paradigm. Burns (2011) describes action research in the English language classroom as “problematising” (p. 2) teaching. The teacher becomes an investigator of a “problem” within their own personal teaching context, intervening in a deliberate way based upon systematically collected data.

Data were collected and analysed using both qualitative and quantitative techniques. Qualitative information was collected via classroom documents, student feedback, classroom observations, assessment tasks and a reflective journal. Classroom documents included all lesson plans and worksheets used in class. Students were also asked to note reflectively what they thought about class activities and assessment. Lesson plans for each stage of the syllabus included space to make detailed observations during class time. A reflective journal also recorded observations and feelings after the event. Student feedback included their own reflective journal based on what was studied, as well as anonymous feedback provided at the end of the semester.

Quantitative data drawn from assessment instruments was analysed in order to compare, contrast or develop qualitative findings further. Rasch analysis provided information on elements of the syllabus or assessment procedures that students found difficult or confusing, with data-model fit statistics, enabling the diagnosis and identification of students requiring remedial instruction, as pioneered by Engelhard (2009). Correlations of assessment tasks helped determine the gap between the aims of the syllabus and what students understood in class. The range of data collection and analysis was designed to increase the objectivity of observations through triangulation. In this way, data can be compared and crosschecked to reach valid conclusions based upon numerous sources of information.

This paper details one cycle of the action research process undertaken as part of a larger research project and focuses on data concerned with assessment and anxiety. Each research cycle contained four stages, as proposed by Kemmis and McTaggart (1998): namely, Planning, Action, Observation and Reflection. Cycle One took place in the first half of one semester, totalling 15 classes.

3.3 Planning

A syllabus was designed that adopted the major theoretical underpinnings of the genre-based approach to conversation. The genre-based approach is described by Slade and Widin (2004), who claim that spoken interactions have "identifiable generic structures(s)" (p. 9), which can be shared with students in order for them to reproduce such interactions for themselves. Building on this concept, Slade and Widin outline a range of different spoken genres, including narrative, anecdote, recount, exemplum, observation, opinion, gossip, and joke-telling; all of which have identifiable generic structures. Eggins and Slade (1997) identify highly interactive "chat" segments of talk, which are not amenable to generic analysis, but also more monologically structured "chunk" segments of talk, which have distinctive beginning, middle and end structures with predictable lexico-grammatical features that can be made explicit to students through classroom instruction. The generic structures and grammatical patterns associated with various text types help form the basis for a set of criteria that are specific to particular spoken genres and can act as the basis for assessment. The genre-based approach thus offers a methodology that addresses concerns raised by Rapley (2008). Through the integration of grammar instruction into the study of English conversation, communicative methodology can utilise grammar instruction in a way that is familiar to Japanese students.

Recount was the chosen genre for Cycle One of the research described in this paper. For illustrative purposes, Table 1 shows the deconstruction of a typical recount text with structural, lexico-grammatical and turn-taking criteria chosen as aspects for assessment:

Table 1.

Deconstruction of a recount text to inform syllabus design

Structural Criteria	**Lexical-grammatical Criteria**	**Turn-taking Criteria**
Orientation	Greeting	Asking questions
Sequence of events	Past tense verbs	Supplying helpful information
	Temporal conjunctions	Expressing surprise or support

Using the criteria illustrated in Table 1 an assessment instrument was created with a polytomous rating scale of 0-5 based upon the teacher's perceived success of a

student in accomplishing these criteria during a conversation with a partner. Further description of criteria for assessment is illustrated in Table 2:

Table 2.

Description of criteria for assessment

Criteria	**Description**
Generic greeting	Students take part in a turn-taking greeting, reciprocating and answering generic greeting phrases.
Orientation	Students develop key "WH" information at the start of their talk, to orientate the listener: e.g. "Last Saturday I went to the cinema"
Sequence of events	Students describe a number of events that occurred during a particular event, ordered in a logical time-ordered manner.
Past tense verbs	Students use past tense verbs appropriate to a recount: e.g. had, ate, went.
Temporal conjunctions	Students connect ideas and events through use of temporal conjunctions: e.g. next, then, after that.
Turn-taking	The speaker and listener ask and answer questions during the conversation.

3.4 Action

At the start of the semester students were asked to create their own spoken recount texts during a conversation with a partner and were assessed based on the criteria described in Tables 1 and 2. As the semester progressed, students were again assessed at the midway stage of the syllabus on another recount conversation after the syllabus intervention. Criteria were made explicit to students throughout the semester and reinforced through classroom instruction. During assessments, the teacher was an observer of two students engaging in a conversation and provided assessment and feedback based upon the taught criteria.

4. Results

In this section, data analysis is explained in detail, but with particular focus on findings in the larger research project that contained themes of anxiety and assessment. This explanation is followed by a reflection on the data analysis that was undertaken during the first cycle of research. The steps taken in preparing the data for analysis are also described.

4.1 Observation and Reflection

At the observation and reflection stage of the action research process, I began assembling my data and looking for broad patterns and trends. Figure 3 shows an attempt to summarize the analysis of qualitative data during the observation phase of Cycle One, including the simplification and abstraction of the raw data into broad themes with the research focus in mind, and triangulation to find patterns in the data.

A central theme that repeatedly and consistently appeared in the data were words and ideas that reflected the *emotions* of students. The word "nervous", or synonyms of "nervous", were among the most frequent words to appear in the data, and in a range of sources. Notions of frustration were also expressed in the datasets. These included comments students made about not being able to complete a task or finding a criterion or task too confusing or too difficult. However, not all instances of student emotion were negative: there were also examples of enjoyment and satisfaction. Figure 3 highlights and summarizes instances of student emotion in the data during Cycle One, as well as the data source, and the time period with which the instance is associated. Exempla are presented in the form of representative quotations.

Nervousness, or synonymous emotions, were the most frequently reported emotions in the data. Nervousness was reported by students themselves and was often visible to teachers through students' body language and speech patterns. The "Post-assessment" section of Figure 3 contains representative quotations that illustrate the nervousness students felt after teacher assessment. Or, as noted during weekly reflection in the teacher journal, when talking face-to-face with the teacher. Nervousness in a non-test situation and when talking to classmates was infrequently reported; emotions were generally positive at this time. During in-class practice, students and teachers also reported student enjoyment or satisfaction. The context in which nervousness was most apparent was during the final assessment. The main reason for anxiety might be explained by Eggins' (2004) description of tenor as being broken down into three continua: power, contact and affective involvement (see Figure 1). The teacher–student relationship promotes unequal power. Teacher–student contact is occasional, and affective involvement is low. Student nervousness arose in the data when students spoke with the teacher, even when formal assessment was not the key goal of the conversation. Another reason for nervousness might be that a casual conversation includes visual and aural

contact, and feedback is immediate, as illustrated in Figure 2 (Eggins, 2004). The nature of a casual conversation means that students' use of the English language is accompanying the action. The notion of immediate feedback, which might include negative feedback from a person whom they consider a model of accurate language, would be extremely daunting to students.

		Student Emotions			
Time period	**Source**	**Nervousness**	**Frustration**	**Enjoyment**	**Satisfaction**
Post-assessment	Student feedback after assessment	*I was very nervous*	*I was confused by unexpected questions*	*It was a little bit fun*	*I could say what I wanted to say*
	Teacher reflection after assessment	*Some students were visibly shaking*	*Students dwelled painfully on mistakes or unknown answers*		*They were clearly pleased when finished*
Weekly	Teacher journal	*When talking to me (in English) they are more nervous than with their partner*	*One student said that they were frustrated about not knowing how to formulate a response in English*	*Students seem to enjoy the opportunity to talk to each other*	*Audible sighs of relief*

Figure 3. Matrix display to examine patterns in student emotion

Figure 3 also shows that frustration was a major theme of the data. Student frustration was primarily expressed by students through the use of negative verbs such as "I can't" or "I didn't" in relation to a particular criterion or activity in the classroom. Frustration was far more apparent after the final assessment than in data collected before the assessment. One reason for this might be the anonymous nature of feedback after the final test, allowing students to be more critical. However, this criticism was not usually aimed at the syllabus itself but towards students' own performance. Student judgments on their proficiency during class-time were sometimes negative, but often tempered with positive comments.

After the final assessments, students' feedback contained a larger number of negative verbs describing their performances. A major contributor to the theme of frustration was reflection on practising the micro-turn-taking of discourse, such as

question/response. These are aspects of talk that do not contain easily identifiable generic and monologic structures and are usually spontaneous. These micro-aspects of conversation might include recasts of language; expressing support or surprise; providing helpful information or help with vocabulary choices; question and answering (see Table 2). The teachers noted that some formulaic greetings were misunderstood, for example:

Teacher: *How's it going?*

Student: *By bus.*

This interaction was noted in the teacher journal as occurring quite often, due to students attaching significance to the words "how" and "going" without recognizing it as a formulaic expression. Students expressed frustration and even embarrassment when this mistake was pointed out, or they realized it during self-reflection afterwards. Frustration was also recorded during the final assessment and during in-class activities, when conversations were ended abruptly or contained long pauses. Pauses occurred as questions were either not understood, or when students were unable to formulate a suitable English response. In peer-assessment, students often had great trouble asking and answering questions to encourage extended dialogue.

Turn-taking during conversation proved to be a large source of frustration for many students. I noted in my journal that rather than using strategies to continue the conversation, some students would take extended periods of time to ensure they gave the correct answer to whatever question or idea they were confronted with, at the expense of fluency. In a natural conversation this would usually create feelings of discomfort, so I noted in my journal my own feelings of frustration that the syllabus was not addressing issues of fluency. Feelings of frustration, therefore, developed from the gap between receptive understanding and productive ability in the target language. This frustration was exacerbated in the final assessment, due to a gap between what students thought they could achieve during class time with their peers, and what they felt they could achieve with the teacher. When micro-aspects of discourse were introduced to the conversation by the teacher in the final assessment, it created a sense of frustration. During class practice, students were unable to replicate micro-aspects of discourse by themselves, so this interaction simply did not take place. As well as nervousness, therefore, frustration also increased when students engaged in casual conversation with the teacher. This was

due to the teacher introducing micro-aspects of discourse to the conversation, which increased the difficulty level of the speaking interaction.

Figure 3 also illustrates that not all emotions experienced by students during the course of the syllabus were negative. There were also elements of enjoyment and satisfaction. Enjoyment was often closely associated in the data with terms like “opportunity” or “chance”, as students described how using the syllabus gave them opportunities to use English in class for a purpose. This sense of enjoyment appears to be closely linked to feelings of satisfaction, as students and teachers commented on being able to achieve something students had not necessarily had the opportunity to do before: to speak for an extended period of time in English. Even after assessment, many students displayed either through body language or through feedback that it was a worthwhile achievement to be able to speak beyond one or two-word answers. Some students even showed surprise at what they were capable of achieving. In future cycles of action research and syllabus design, detailed feedback that promotes the sharing of learning goals and how students were meeting them might work to lessen feelings of nervousness and frustration. Students who had also shown frustration and nervousness also indicated enjoyment and satisfaction, both during the syllabus and after the completion of the final assessment.

One student mentioned that their nervousness was so great they thought they were going to die, but were finally pleased to announce that they did not, in fact, die, and could have a conversation in English. The data seems to suggest that enjoyment and satisfaction were closely related, and that a genre-based approach afforded increased opportunity to speak in English, which students found rewarding and even surprising.

During Cycle One, whilst assessment considerations were integrated into the original syllabus design at the start of the action research process, these were largely summative in nature; and although they provided useful quantitative data for analysis, they did not necessarily adequately address the aims of the syllabus. Criteria for assessment were based upon the deconstruction of a personal recount text.

As part of the syllabus design and criteria generation for assessment, it was useful to create a latent variable map that hypothesized the difficulty students might have with various elements of the syllabus. This hypothesis could inform syllabus design and lesson content based on the assumed needs of my students. These assumptions could then be

tested, challenged and verified, or used to inform future syllabus design. For example, criteria that were hypothesized to be difficult might prove to require less attention in the classroom than was first assumed. Equally, criteria that the teacher thought were not difficult might actually prove more challenging to students than assumed. Table 3 shows the hypothesized difficulty level of criteria generated for assessment, and to inform syllabus design:

Table 3.

Hypothesized latent variable map for the generic structure and textual features of a recount genre

Logit	Student's use of structure and features	Structure and textual features
5.00		
4.00	High use of features	
3.00		**Turn-taking**
2.00		
1.00		
.00	Moderate use of features	**Temporal conjunctions**
-1.00		**Past tense**
-2.00		**Orientation**
-3.00		**Sequence of events**
-4.00	Low use of features	**Generic Greeting**
-5.00		
Polytomous rating scale used (X= 0, 1, 2, 3, 4, 5)		

Table 3 shows the hypothesis that turn-taking, such as answering questions, would prove to be the most difficult aspects of the syllabus for students to master; whereas lexical-grammatical criteria, such as using temporal conjunctions and past tenses, might prove to be easier. Table 3 shows that during assessment, scores were awarded on a rating scale of 0 to 5, depending on the fulfilment of these criteria as judged by the teacher. A score of 0 was awarded to students if they did not use that particular item in their talk; 3 indicates that students sometimes used that item or used it fairly effectively; and a rating of 5 means that they used the item effectively and often.

During assessment, students used the content provided in the planned teaching and learning cycle to individually construct their own recount texts in a conversation with the teacher in the final classes. When participating in the conversation as an interlocutor, I awarded scores for the relevant criteria based on my own judgement. The criteria for assessment are presented in Tables 1 and 2. After the assessment, the scores were prepared to form a text file for entry into a Rasch analysis using Winsteps (Linacre, 2007). The analysis in this section is for the 45 students included in Cycle One of the action research. Table 4 shows the assessment criteria ordered vertically according to their difficulty, as determined by the Rasch measure in response to student scores; these can be compared with the hypothesized difficulties presented in Table 3. At the bottom of the hierarchy of difficulty we see “Sequence”, thus indicating that students found sequencing events to be the easiest item on the test. The most challenging item on the test was “Turn-taking”, which correlates with the hypothesized order of difficulty, thus providing preliminary evidence of construct validity.

Each “X” in Table 4 represents one student, and the higher the “X” appears on the figure, the more proficient the student’s language ability, based on this measurement. “(M)” shows the location of the mean of the persons or items, and “(S)” shows one standard deviation above or below the mean. The vertical spacing is the approximate placement of the items on the linear Rasch dimension, so that “Orientation” to “Generic Greeting” has roughly the same increase in item difficulty as “Temporal Conjunctions” to “Past Tense”. In the test, 7 students are performing at least one standard deviation below the person mean, and 12 students are performing one standard deviation above the person mean. Most students taking the test performed within one standard deviation of the mean. We can see that six students have ability equal to “Turn-taking”; so Rasch calculates that they have a 50% expectation of success on this item, based on inferences drawn from response patterns in the data (Linacre, 2007). Their probability of success in easier items increases above this 50%.

From Table 4 it was possible to identify a number of problems with the assessment. Firstly, the test had been too easy for the highest-level students, who were able to tackle all items effectively; while conversely, four students could not perform effectively on any items in the test. The implications of this were either that the assessment needed more very easy and very difficult items in the test, or that some students were not learning these

items effectively in the scheme of work in the classroom. This was useful as a diagnostic tool, as it informed my next cycle of research and indicated problematic points in my syllabus; it also allowed targeted support for individual students.

Table 4.

Person-Item map showing persons on a common scale with items

Logit	Person	Item
10.00	XX	
	XXXXXX	**Turn-taking**
8.00	XXXX	(T)
	(S)	
6.00	XXXX	
	XX	
4.00		(S)
2.00	XXXXXX	
.00		(M) **Temporal conjunctions**
	XXXXXX (M)	**Past tense**
-2.00	XXX	
	X	**Orientation**
-4.00		(S) **Generic Greeting**
	XXXX (S)	**Sequence of events**
-6.00		
-8.00	XXX	(T)
-10.00	XXXX (T)	

Polytomous rating scale used (X= 0, 1, 2, 3, 4, 5)

A potentially valuable, but underdeveloped, use of Rasch analysis of classroom tests is shown by Engelhard's (2009) investigation of person fit statistics as a diagnostic tool in mixed-method research: this illustrates the duality of Rasch analysis, where the same analyses can be conducted for persons as well as items; thereby allowing mis-fitting persons to be identified and qualitative investigation to be conducted, in order to determine causes and possible remedial intervention. For illustrative purposes, the original dataset of six items and five of the 45 persons is shown in Table 5, with persons arranged in order of fit.

Table 5.

Person correlation and fit

Person No.	Score	Infit MnSq	Outfit MnSq
12	13	1.20	5.43
14	28	1.32	5.40
10	15	1.87	2.64
26	22	1.77	1.36
6	14	1.68	1.25
MEAN	21.2	.96	1.04

In Table 5, Infit MnSq, Outfit MnSq and Standard error are all "fit" statistics that indicate how accurately or predictably data fit the Rasch model. "Outfit" is an outlier-sensitive fit statistic, and sensitive to unexpected observations by persons on items that are relatively very easy or very hard for them. "Infit" is an inlier-sensitive fit statistic that is more sensitive to unexpected patterns of observations by persons on items that are roughly targeted at them (Linacre, 2007). Infit was an innovation of Ben Wright (Bond and Fox, 2007), who noticed that the standard statistical fit statistic (which we now call "outfit") was highly influenced by a few outliers (very unexpected observations). He therefore devised the infit statistic, which was more sensitive to the overall pattern of responses. Infit weights the observations by their statistical information, which is higher in the centre of the test and lower at the extremes. The effect is to make infit less influenced by outliers, and more sensitive to patterns of inlying observations.

Five persons show misfit large enough to warrant investigation. Person number 14 is extremely proficient, scoring 28 out of a possible 30, and has acceptable infit, but very large outfit, which is consistent with failing on a very easy item. This result is probably of no concern, but investigation of unexpected responses can clarify the reason for the misfit. Persons 12, 10 and 6 are of low proficiency, and have misfit that warrants further attention. Not only are they of limited proficiency, but they do not respond consistently with the other people's response patterns. Person 26 is of slightly higher than average ability, but has an infit mean-square figure of 1.77, so is also of concern. These misfitting students are deviating from the latent Rasch trait that defines the expected trajectory of this sample of persons through this curriculum; this identifies them as possible candidates for remediation.

Qualitative investigation of these students helped to identify the causes of this misfit. Person 6, for example, had very poor attendance, and was therefore unprepared for this form of summative assessment. Person 14 wrote a number of comments in their student journal that showed confusion about some of the criteria.

In this particular assessment, we see that more consideration is needed in preparing students for assessment in order for their abilities to be measured more effectively; and in this case, it appears that criteria need to be made more explicit to students, so that they know exactly what is expected of them, in order to reduce misunderstanding. During Cycle One, although criterion was discussed in class and exemplified via model texts, a detailed assessment rubric was not shared with students, and model texts proved to be confusing, or prompted students to rely on written output. In further cycles of research, it would appear that an assessment rubric might need to be constructed beyond the simple assigning of scores on a polytomous rating scale, and that this should be made explicit to students.

In Table 4, we see that criteria based on elements of generic structure and lexico-grammar in a genre are the easiest criteria for students. By contrast, criteria that are not necessarily genre-specific, but universal speaking abilities, were by far the most difficult for students. In Table 4, the criteria are separated into two quite distinct groups according to level of difficulty. It proved problematic to measure the many different aspects of speaking ability on one unidimensional line, as Rasch analysis dictates. For example, one of my students who had performed very well in a classroom environment became extremely anxious in a spoken conversation, especially in the assessment situation. In such cases, it was therefore impossible to measure any of that student's language ability, since the student was unable to produce language. Measuring students on one summative assessment did not seem to be an accurate measurement of their overall speaking ability and might equally have mirrored my own teaching ability and syllabus design. Turn-taking aspects of conversation were more difficult to assess than generic structural and grammatical elements of a recount genre. Such a speaking test, therefore, is likely to be measuring two different forms of spoken ability: mastery of structure and lexico-grammar; and mastery of spoken output, including pronunciation, fluency and turn-taking. This means that one summative test for students' speaking may not be appropriate or an accurate reflection of their skills, and that other forms of assessment are necessary.

5. Conclusions

The study shows that students can feel high anxiety when undertaking speaking assessment and that this anxiety should be taken into account when designing syllabi and assessment procedures. The study also highlights reasons for this anxiety, allowing remedial intervention in syllabus and assessment design. Based on a triangulation of both quantitative and qualitative data, it became clear that a genre-based approach must entail greater integration of formative assessments into the syllabus at various stages; it must also provide opportunities for students to experiment with language that allows them to be more spontaneous, and also less anxious about making mistakes. An assessment based solely upon the structural and lexico-grammatical elements of speaking might not be adequately assessing more universal speaking abilities; it might also prove to be too prescriptive, by not allowing students to experiment, or to have sufficient confidence to be spontaneous with language. As such, the syllabus design in further cycles of action research needed to find ways to integrate formative assessment, and also to remove teacher-dominated summative assessments that led to nervousness and frustration. Forms of self-assessment might remove feelings of nervousness and frustration by correcting the power imbalance between teacher and students in interactions, as well as creating the necessity for making criteria for assessment explicit and clear to students. Removing the teacher from the immediate assessment process might also greatly improve anxiety. Indeed, Figure 3 shows that without teacher proximity, student emotions changed to that of enjoyment and satisfaction.

6. Future Research Cycles

During the planning and action stages of Cycle Two, the teacher was removed entirely from the assessment process, as students recorded their conversations on smartphones and uploaded examples of their speaking to the teachers' secure website, with content only available to the teacher and individual student. The benefits of this were to reduce anxiety during assessment and allow the teacher greater time to perform assessment and provide more detailed feedback; it also potentially allows other teachers to perform objective assessment of students that were not their own. Cycle Two also introduced Assessment for Learning strategies, making criteria more explicit to students, and allowing opportunities for peer and self-assessment that narrowed the gap between

student and teacher expectations and gave students the confidence to experiment with language and make mistakes. Additional criteria were also developed that allowed for the assessment of fluency and pronunciation, which offered opportunities for increased test validity. Detailed discussion of further cycles of the action research project are beyond the scope of this paper and will be published and disseminated in further papers.

References

Anderson, F. (1993). The enigma of the college classroom: nails that don't stick up. In Wadden, P. (ed.) *A handbook for teaching English at Japanese colleges and universities*. New York: Oxford University Press, 101-110.

Bond, T. & Fox, C. (2007). *Applying the Rasch model*. 2nd edition. London, U.K.: Lawrence Erlbaum Associates.

Burns, A. (2011). *Doing action research in English language teaching: A guide for practitioners.* Beijing: Foreign Language Teaching and Research Press.

Central Education Council. (1999). Improvement of admissions with an emphasis on the connection with the higher education and elementary and secondary education. Retrieved from <http://www.mext.go.jp/b_menu/shingi/old_chukyo/old_chukyo_index/toushin/attach/1309753.htm >.

Cohen, D. K. & Spillane, J. P. (1992*).* Policy and practice: The relations between governance and instruction*. Review of Research in Education*, *18*, 3–49.

Corlett, R. H. (2012). Overcoming communication anxiety: Observations of Japanese students during intensive communication and culture studies programmes in New Zealand. *Language Education: 江戸川大学語学教育研究所紀要*, *10*, 23-30.

Doyon, P. (2000). Shyness in the Japanese EFL class: Why it is a problem, what it is, what causes it, and what to do about it. *The Language Teacher*, *24*(1), 11-16.

Eggins, S. (2004). *An introduction to systemic functional linguistics*. London, U.K.: Pinter.

Eggins, S. & Slade, D. (1997). *Analysing casual conversation*. London, U.K.: Cassell.

Engelhard, G., Jr. (2009). Using item response theory and model-data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement, 69*(4), 585–602.

Kemmis, S. & McTaggart, R. (1988). *The action research planner*. Waurn Ponds, Vic: Deakin University Press.

Kitano, K. (2001). Anxiety in the college Japanese language classroom. *The Modern Language Journal*, *85*(4), 549-566.

Linacre, J. M. (2007). *Winsteps*. Retrieved from <http://www.winsteps.com/>.

McDonough, J. & Shaw, C. (1993). The impact of the communicative approach, in *Materials and methods in ELT: A teacher's guide.* Hoboken, NJ: Blackwell.

Marchand, T. (2012). Reticence in the classroom: examples, causes, and suggestions for improvement. *Obirin Today, 12*, 159-179.

Ministry of Education, Culture, Sports, Science and Technology, Japan (1947). *Gakushu shidou yoryu eigohen* [Study of course guideline for English education]. Retrieved from <www.nier.go.jp/guideline>.
Ministry of Education, Culture, Sports, Science and Technology, Japan (1998). *Junior high schools' courses of study.* Tokyo: National Printing Bureau.
Ministry of Education, Culture, Sports, Science and Technology, Japan (1999). *Senior high schools' courses of study.* Tokyo: National Printing Bureau.
Ministry of Education, Culture, Sports, Science and Technology (2011). *Kokusai kyotsugo toshite no eigoryoku kojo no tameno itsutsu no teigen togutaiteki shisaku* [Five Proposals to Improve the Proficiency of English as Lingua Franca]. Retrieved from <www.mext.go.jp/b_menu/houdou/23/07/1308888.htm>.
Ministry of Education, Culture, Sports, Science and Technology (2014). *Report on the future improvement and enhancement of English education: Five recommendations on the English education reform plan responding to the rapid globalization*. Retrieved from <http://www.mext.go.jp/en/news/topics/detail/1372625.htm>.
Mulvey, B. (2010). University accreditation in Japan: Problems and possibilities for reforming EFL education. *The Language Teacher, 34*(1) 15–24.
Nakane, I. (2006). Silence and politeness in intercultural communication in university seminars. *Journal of Pragmatics*, *38*(11), 1811-1835.
Ohata, K. (2005). Potential sources of anxiety for Japanese learners of English: Preliminary case interviews with five Japanese college students in the US. *TESL-EJ*, *9*(3), n3.
Rapley. (2008). *Policy and reality: The teaching of oral communication by Japanese teachers of English in public junior high schools in Kurashiki City, Japan.* Retrieved from <http://www.asian-efl-journal.com/Thesis-D-Rapley.pdf>.
Scrivener, J. (1994). *Learning teaching*. Oxford, U.K.: Heinemann.
Slade, D. & Widin, J. (2004). Teaching spoken English in the secondary school classroom. *The Zeneiren Magazine, 42*, 2–13.
Takayama, K. (2008). The politics of international league tables: PISA in Japan's achievement crisis debate. *Comparative Education, 44*, 387–407.
Turner, J. M., & Hiraga, M. K. (1996). Elaborating elaboration in academic tutorials: Changing cultural assumptions. In H. Coleman & L. Cameron (Eds.), *Change and language* (pp. 131–140). Clevedon, England: Multilingual Matters.
Watanabe, Y. (2004). Teacher factors mediating washback. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and models,* 129–146. Mahwah, NJ: Erlbaum.
Wilkins. S. G. (2018). *A genre-based approach to speaking in EFL* (Doctoral dissertation). Aston University, U.K.
Young, D. J. (1991). Creating a low‐anxiety classroom environment: What does language anxiety research suggest?. *The Modern Language Journal*, *75*(4), 426-437.