ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 9, No. 6, pp. 1138-1146, November 2018
DOI: http://dx.doi.org/10.17507/jltr.0906.03

Motivating EFL Students with Conversation Data

Austin Gardiner
Seiwa Gakuen, Sendai, Japan

***Abstract*—Motivating learners of English as a Foreign Language (EFL) to improve their speaking fluency is challenging in environments where institutions emphasize reading and listening test performance. The focus tends to shift to strategic reading and listening first in order to attain acceptable test results, often at the expense of communicative competence. Computer Assisted Language Learning (CALL) is well positioned to assess and develop communicative competence for EFL learners, and to motivate them to speak. This article introduces the Objective Subjective (OS) Scoring system, a CALL system which sets clear immediate goals on the path to better communicative competence with data from videoed conversation sessions. It motivates learners to improve on their data in every consecutive conversation session, whereby an environment is created which facilitates conversation practice as well as individual error correction.**

***Index Terms*—communication systems and technology, foreign languages, higher education, language teaching**

I. Introduction

Setting a goal involves creating a psychological discrepancy between current ability and ideal ability (Locke and Latham, 2002). Transforming current ability into ideal ability requires time, effort and motivation. When goals are set too high, learners become demotivated, and when they are too low, learners lose interest. Scoring well on an international speaking proficiency test is one example of a distal goal that represents a major discrepancy in what learners from many parts of the world are generally capable of, and what they would eventually like to be able to do communicatively. To keep them motivated, more attainable goals with periodical moments of success and rewards are needed. To this end, the Objective Subjective (OS) Scoring system sets clear immediate learning goals for every individual, which form stepping stones on the path to speaking proficiency, but concentrates on fluency, i.e. amount of language produced. It involves students rating themselves with data compiled after transcribing videoed conversation sessions. The system will be elaborated in a later section. The next section discusses some issues concerning EFL speaking proficiency rating with specific attention to motivation, which in general is a neglected attribute in proficiency rating systems.

II. EFL Speaking Skill Rating Principles

Speaking and writing skills are more complicated to assess than reading and listening skills. Reading and listening can be objectively assessed with a cloze test or multiple choice questions (Ellis, 1985; Fulcher, 2003), whereas speaking and writing need a relatively subjective perspective i.e. a human element on the assessing end. It is this human factor that modern speaking skill raters aim to emulate. Many modern speaking tests are done with CALL systems. Computer based speech raters such as the speech rater® system patented by Educational Testing Services (ETS) feature the ability to use voice recognition software in order to measure a spoken response against a corpus of previous responses with tagged data, and then predict how human raters would assess the same response. This step up is often referred to as intelligent CALL, or iCALL (Gamper, Knapp, 2002). Zechner, Chen, Davis, Evanini, Lee, Leong, Wang, and Yoon, (2015, p.1), at the time researchers at ETS, describe the steps in a typical modern automated scoring system as follows:

"...adapt a non-native English speech rater (trained on TOEFL Practice Online data) to transcribed THT (Test with Heterogeneous Tasks) task responses, then compute a set of relevant speech features to predict trained human rater scores."

In the example above, the typical challenges inherent in the process of rating a spoken response come to light: (1) Non-native English responses need to be assessed differently from native responses; (2) a criterion, usually a corpus of prior responses and a database of previous scores, is needed against which speakers are measured; (3) a set of focal speech features need to be decided and; (4) a scoring model comprising valid analytic scales and a composite score need to be put in place in collaboration with human raters. This system mirrors and improves the way human raters would assess a response against all of their previous experiences as raters. Such a system seems to be sufficiently objective; however, (2) and (3) above tend to demotivate test takers in Japan. The problem is that the corpus used by the program takes the place of previous human experience. In other words, mirroring the tagged data and focal speech features in the corpus becomes the ultimate goal, and the test taker is forced to master the particular corpus language features in order to attain a higher score. This practice often has negative effects on motivation as the goals to attain in order to appear proficient are unclear to many test takers unless they are committed to spending time and money researching the specific testing products on sale. The designers of the corpus and the language features are subjectively in charge, which in the case of high stakes assessment is a method not completely ethically sound in a world with so many first languages and variations of English. Computer based speaking tests are valuable because of their superior accuracy; however, the majority of the CALL systems and tests are not motivational. Bodnar, Cucciarini, Strik, and van Hout (2016) outlined various issues concerning CALL systems with special emphasis on motivation. They agree that not enough research has gone into making CALL systems inherently motivational. This article aims to add to the body of research concerning this issue. It calls for the localization of speech raters, especially in Japan, which would add beneficial elements of objectivity and validity and enhance motivation toward producing more language, regardless of whether the language produced is exactly in line with an Anglo-American dominant corpus.

A speaking assessment must be valid as well as sufficiently objective; however, it should also be a fair, calculated measurement of proficiency that takes into account variables such as linguistic interference and educational background. Many CALL systems address this through student modeling, but few of these systems take motivation into account (Bodnar et al., 2016). As an additional overall guide to test validity in general, Bachman (1990, p.300) states that an assessment should *"capture or recreate…the essence of language use."* The speaking tasks, in other words, need to reflect real world usage and a speaker who scores well on the assessment should also be able to use English communicatively outside of the testing environment. In the experience of the researcher, this is rarely the case with internationally sold computer based tests in Japan. There is as of yet no test able to assess and predict proficiency for Japanese students conversing with each other. The types of expressions Japanese students use naturally in conversations are not paid much attention in the literature. Instead, students are generally expected to model their expressions around what is dictated by international testing companies. As a result, many students start to focus on beating the international test, and regard being able to converse in English with their peers as less important.

Motivation requires constant *"feed-forward control in addition to feedback."* (Locke et al., 2002, p.708). Without attainable intervals that feed forward through periodical successes, most people become demotivated regardless of the nature of the final goal. EFL learners in Japan are no exception. To make matters worse, some stop improving communicative ability altogether when they have received an acceptable score on an international speaking test. Learners want to improve their ability, in other words close the gap between current ability and future ability. Speaking proficiency rating has become a very lucrative business which feeds off an ultimately elusive, idealized final goal.

III. Proficiency Goals and Attainable Fluency

A clearly defined set of goals to attain proficiency often eludes students in parts of the world such as Japan where English is seldom used outside of pedagogical situations. The International English Language Testing System (IELTS) designed and sold by the British Council Inc., for example, rates speaking proficiency on nine bands, according to four factors: fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation (IELTS, 2015). In North America, the American Council on the Teaching of Foreign Languages (ACTFL) suggests four factors, namely function, content, context, and accuracy as indicative of speaking proficiency. In fact, most regions have their own innate concept of proficiency and these overlap somewhat to form an intuitive rather than a concretely quantifiable concept of proficiency internationally. For EFL learners in general, staying motivated throughout the long path toward proficiency as prescribed by these and other testing institutions is no small task.

The OS scoring system considers these developments, and is designed in such a way that the students become the creators of their own corpus. They are ultimately measured against their former selves—not against a foreign corpus as they do not reside in an English-dominant region. Atkinson (1958) has shown that motivation decreases in the face of a task or goal deemed to be distal and out of reach. The OS Scoring system aims to shift the focus of the learner from proficiency as perceived by independent international testing companies, to fluency more immediately attainable by improving on prior data. It aims to motivate speakers in conversations by lowering their affective filter, by giving them a sense of control over their assessment, and by allowing them to analyze their own data to see where they need to make progress in their expressions.

IV. OS Scoring and Conventional Speech Rating

The OS Scoring system represents the culmination of fourteen years of trial and error using video to assess speaking skills. Peter Wanner (2002) described in detail the first attempts at the Kyoto Institute of Technology to assess group discussions using video and the subjects themselves as raters. Since then, technology has improved, and the process has been streamlined to deliver more accurate qualitative as well as quantitative data. This article serves to bear on these developments using the most recent speech rating research and to introduce OS Scoring as an effective motivator and assessment tool.

OS Scoring differs from conventional rating in the high entropy tasks it requires in each conversation session. High entropy tasks include more unpredictable activities such as expanding on the opinions of others, not only question responses with predictable answers. Unlike the human to computer interaction, which is the norm in conventional rating, OS scoring involves human to human interaction which necessitates real time responses to natural, unpredictable conversation. Table 1 below outlines the five main differences in the processes of conventional speech rating and OS Scoring:

TABLE 1.
THE FIVE MAIN DIFFERENCES BETWEEN CONVENTIONAL SPEECH RATING AND OS SCORING.

Conventional Speech Rating	Objective Subjective Scoring
(i) Subjects are generally assessed one at a time by a computer.	Subjects assess themselves in a group with the aid of a computer.
(ii) Subjects are compared to tagged data.	Subjects are compared to their last session.
(iii) Scores are predicted by a computer and verified by human raters afterwards.	Scores are exact analyses of utterances, done by the subjects themselves.
(iv) The test result is a once off analysis of a comparison with tagged data.	The subjects assess themselves a number of times per semester and are able to document development.
(v) Tasks are planned and corpus specific.	Tasks are topic specific with natural unplanned interaction.

As can be inferred from Table 1 above, OS Scoring makes the subjects responsible for their own scores. This eliminates obstacles common in speech rating such as rater bias (Caban, 2003) and accent comprehension (Major, Fitzmaurice, and Balasubramanian, 2002). The natural unplanned interaction allows subjects to interact by asking for clarification, as well as perform *"confirmation checks"* and *"comprehension checks"* as nonnative speakers are reported to do more often than native speakers (Zhang, 2009, p.92). OS Scoring therefore enables a more personalized evaluation of nonnative speakers than a once off computer based test.

The system sets itself apart from conventional speech rating by keeping the subjects involved in the rating process and introducing them to the basic measuring methodology for phrases such as mean length of turn (MLT) and total TIME spent speaking English. These will be elaborated later. Total TIME spent speaking English is considered by the researcher here to be a valuable indicator of motivation, especially in Japan where students can be reluctant to speak in front of their peers.

V. OS SCORING PRINCIPLES

Thus far, some of the literature concerning motivation as an important factor in assessment systems as well as what sets OS Scoring apart from conventional systems have been discussed. This section moves on to introduce OS Scoring as a motivational CALL tool that can be utilized in university English courses.

Setting up the system is easy. Similarly proficient students are divided into conversation groups of six to eight members. Each member of the group is assigned a chair, labeled for the camera as A to H. These labels are used in the transcription stage to identify which subjects spoke at which turns. Conversation topics can be chosen from the English curriculum for every session. In this case, topics included a *Self-introduction* as warm-up, followed the next week by *Conflict*, then *Population*, and this investigation concluded with *The Environment* as the last topic although there were six more topics following the investigation.

At the beginning of each session, the camera is switched on and students have twenty minutes to converse. After twenty minutes, the session is concluded and a copy of the video file is sent to each student in the group. Students then use the CLAN program (MacWhinney, 1995) to transcribe their utterances with the assistance of a researcher as they watch themselves speak, adhering to the Codes for Human Analysis of Transcripts (CHAT) protocol. Students also transcribe what their friends say before they speak in order to pay attention to how dialog flows from one student to the next. At the time of this research, instructions for this protocol were available online in both English and Japanese, free for students to download from the CHILDES website (MacWhinney, 1995), which also supports the transcribing process and helps students familiarize themselves with language transcription in general. The program allows readouts of frequency, mean length of utterance, mean length of turn, and other data which the students compile in a spreadsheet. This data is submitted and the researcher then records it into a master spreadsheet as a collection of analytic scales that will finally make out a percentage of the composite score for each student.

Students are given their data readouts each week. In this way, they are constantly aware of the time they speak individually during every session, their ability to form and pronounce utterances that are lexically and syntactically coherent, and the pauses they take between utterances. This self-awareness serves to motivate them to challenge the data, to speak more, to form better coherent utterances, and to pause less during each following conversation session.

VI. OS SCORING VALIDITY

For the OS Scoring system to be a valid speaking proficiency assessment tool with the benefit of improving subject fluency in conversations, there must be proof, both that the system delivers results indicative of communicative speaking skills, and that it improves speaking fluency. The first of these two attributes to be proven has to do with traditional concepts of validity: concurrent validity, construct validity, content validity, face validity, and predictive validity (Bachman & Palmer, 1996; Li, 2011). Since OS Scoring is a relatively new concept, concurrent validity and face validity for OS Scoring as a system indicative of speaking proficiency have not been established, and much needs to be done in the future to develop its validity on these fronts; however, the fact that the subjects rate themselves in real interactions attest in some degree to construct validity and predictive validity. As for content validity, at this point it is perhaps adequate to choose relevant topics for conversation, and to make sure that sessions include various language tasks such as arguing opinions and expanding on the opinions of others, or other content which may be specific to course curricula. The second attribute concerning validity, a measurable effect on fluency, will be discussed in light of statistical evidence in the following sections.

VII. AIM AND RESEARCH QUESTION

The aim of the research is to find out whether there is an increase in the overall fluency of 45 subjects after four sessions of OS Scoring. The research question is formulated as follows: Is there a significant difference in the overall mean length of speaking turn (MLT) and total time spent speaking English measured during the *Conflict* session, and the overall (MLT) and total time spent speaking English measured during *The Environment* session? In other words, is there proof of an increase in the willingness to converse in English?

VIII. HYPOTHESES

In order to establish empirical grounds for the claim that OS Scoring improves speaking fluency, two variables, MLT and total time spent speaking English (TIME), in three data sets, TOPIC1, 2, and 3, were subjected to a one way analysis of variance (ANOVA) in order to test the following two hypotheses:

(1) OS Scoring administered over three sessions increases MLT.

(2) OS Scoring administered over three sessions increases TIME.

In both instances, the null hypothesis holds if there is an increase not significant enough, or which results in $p>0.01$, i.e. $\alpha=0.01$, a cumulative probability of less than 99%.

IX. METHODOLOGY

The subjects in this study were second-year science course undergraduates in one of the top-rated public universities in Japan. They were on average 20 years of age, 50 students in total, and English was a compulsory subject. Five students out of fifty unfortunately submitted incomplete or corrupted data and were disregarded, thus N=45. The students exchanged opinions in conversations on the following topics: *Conflict* (TOPIC1) during session two; *Population* (TOPIC2) during session three; and *The Environment* (TOPIC3) during session four. The first session, *Self-introduction*, was not included in the analysis because students were still learning how to use the CLAN program at this stage and needed some time to settle into the assessment environment. The first session is therefore excluded in order to minimize confounding variables and is treated as an introduction to OS Scoring in general.

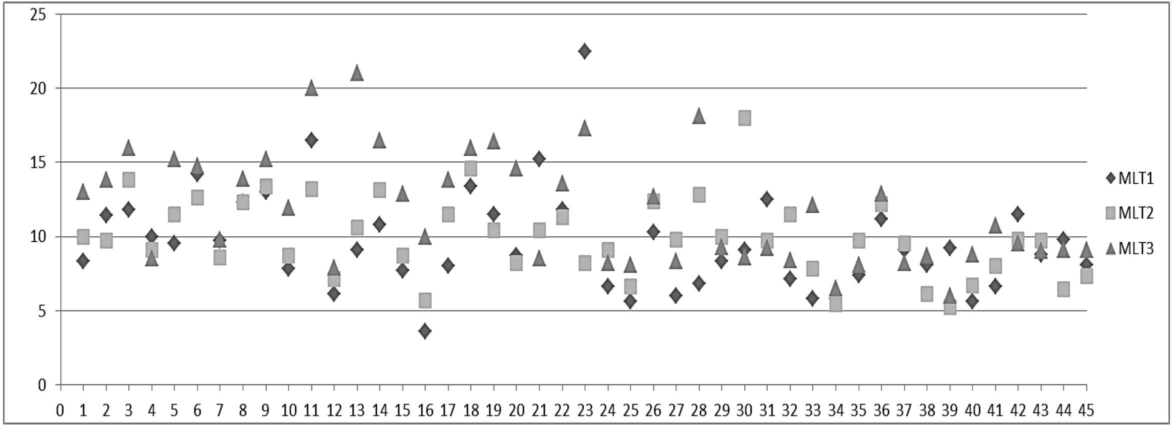
After each session, every student was given a video file of their group talk and tasked with analyzing their individual data. They created three files for each session in the CHAT program and submitted these to the researcher. This procedure is similar to the one discussed in detail in Wanner (2002). Each of the three CHAT files contains readouts of various language related data, but only two variables are of interest in this investigation: MLT measured in words and total TIME measured in seconds.

Analyzing data is a specific and difficult task, and many students were reluctant at first; however, it soon became clear that, as Locke and Latham (2002, p.706) have also found, *"specific, difficult goals consistently led to higher performance."* Many students found the transcription tasks tedious and daunting at first, but all the students seemed to like crunching the data in the end; they enjoyed seeing their speech reflected in the data, and like most language learners, jumped at every opportunity for corrective feedback.

X. DESCRIPTIVE STATISTICS

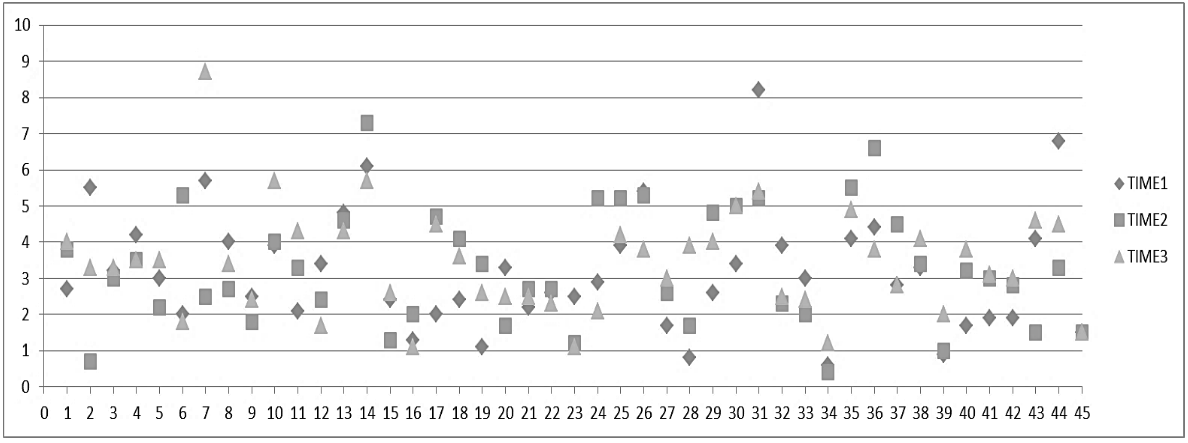
A one way ANOVA was conducted to look for variation in the sum of squares (SS). The SSW (within) was subtracted in both cases from the SST (total) to arrive at the SSB (between). The results were used with the factor degrees of freedom (k-1=2), and error degrees of freedom (135-k=132) to arrive at a Fisher ratio F(2,132)=(SSB/(k-1=2))/(SSW/(135-k=132)). Effect sizes were calculated by finding eta (η2), which was arrived at by η2=(SSB/SST). The one way ANOVA was performed using Microsoft Excel 2013, and all calculations were confirmed with the scientific calculator of an Apple iPhone 7. A significant increase in MLT is illustrated in the plots for MLT 1, 2, and 3 below in Table 2.1:

TABLE 2.1.
MLT1 ◆, MLT2 ■, AND MLT3 ▲ PLOT FOR EVERY SUBJECT.



In Table 2.1 it can be seen that 29 out of 45 subjects increased their MLT from TOPIC1 to TOPIC3, illustrated by the MLT3 ▲ higher than MLT1 ◆ in each of the 29 subject plots. The second variable, TIME measured in minutes, showed a surprising tendency: most subjects moved closer to the mean as they progressed through the sessions, illustrated in Table 2.2 below.

TABLE 2.2.
TIME1 ◆, TIME2 ■, AND TIME3 ▲ PLOT FOR EVERY SUBJECT.



The plot for TIME1, 2, and 3 shows that only 16 out of 45 subjects spoke longer during TIME3 ▲; however, note how TIME3 ▲ moves closer to the mean for most subjects, and how there are many instances of TIME3 ▲ lying between TIME1 ◆ and TIME2 ■ (19 out of 45 cases). This can be interpreted as subjects tuning in to one another or maturing in the test environment as those who initially spoke a lot started giving others time to speak, and those who did not speak much during TIME1 ◆ and TIME2 ■ spoke longer in order to increase their score for TIME3 ▲. As an example, subject 34, noticeably the most reluctant throughout the study, increased speaking time by 30sec. during TIME3 ▲ in order to beat the data and get a passing score.

XI. RESULTS

MLT describes the average length of every speaking turn, and this study found that there was a considerable difference between MLT on average measured after TOPIC1 *Conflict* , 9.6 words per turn, and measured after TOPIC3 *The Environment*, 11.78 words per turn (a total increase of 2.17 words per turn), and in this case the increase proved significant (p=0.004, η2=0.079). MLT showed a definite increase—and for the purposes of this investigation—a positive effect on motivation. As can be seen in Table 3 below, MLT increased significantly.

TABLE 3.
SIGNIFICANT INCREASE IN MLT.

Mean MLT		INCREASE	**yes/no**	Total
MLT1	9.604444			Increase
MLT2	9.92	0.315556	**yes**	1st to 3rd
MLT3	11.78222	1.862222	**yes**	2.177778 p=0.004, η2=0.079

MLT increased from 9.6 to 11.78. It has to be mentioned that the researchers constantly motivated subjects beforehand to use their 20 minutes as effectively as possible in order to increase their chances of getting a higher score by beating the data. Therefore, a type 1 error is present in this one way ANOVA; the same participants were measured over time without a control group. This renders the statistics of little use when it comes to isolating the cause of the increases. Nevertheless, as the increases were significant, the fact that the system is motivational cannot be ruled out. It is not designed to be an exact measurement of the source of the increases; rather, it is designed to help students keep track of progress. The system requires only that they beat their previous data with a statistical significance. Later sections in this article discuss excerpts with data that show specific instances of increases in MLT and TIME.

TIME showed an increase in averages as well, from 3.1 min. after TOPIC1 *Conflict* to 3.4 min. after TOPIC3 *The Environment* (a total increase of 15 sec.), but did not reach the significance threshold (p=0.74, η2= 0.004). However, OS Scoring showed a surprising tendency: Here, the one way ANOVA proved extremely useful. Subjects tended to tune in to one another as time spent speaking English for each subject became less varied. In other words, those who may have spoken overbearingly for the 20 min. during TOPIC1, tended to speak less during TOPIC2 and then less or more during TOPIC3, and those who spoke little at first tended to speak more towards the end. This notion was arrived at by looking at the shrinking differences in within-group variances of TIME after each session; after TOPIC1 *Conflict*: VAR=2.59, after TOPIC2 *Population*: VAR=2.53, and after TOPIC3 *The Environment*: VAR=2.03. Table 4 highlights the shrinking variances that show how students matured or became more familiar in the test environment:

TABLE 4.
RAW MEANS FOR TIME WITH SHRINKING VARIANCES.

Mean TIME		INCREASE	**yes/no**	VARIANCE	Total
TIME1	3.171111			2.59	Increase
TIME2	3.264444	0.093333	**yes**	2.53	1st to 3rd
TIME3	3.422222	0.157778	**yes**	2.03	0.251111
				Shrinking Variances ⬆	p=0.74, η2= 0.004

TIME increased by 0.25 minutes for every subject on average, in other words, an increase of 15 seconds on average during TOPIC 3.

XII. CONVERSATION DATA AS MOTIVATOR AND INDICATOR

Concerning MLT, a definite increase in overall mean length of speaking turn indicated that OS Scoring administered over three sessions has a high probability of increasing sentence length as well as amount of sentences uttered per speaking turn. The null hypothesis is rejected. This bodes well for data as an energizing goal (Locke, E.A., Latham, G.P. 2002). The subjects also significantly increased MLT throughout the course, which can be interpreted as a significant increase in willingness to converse. The statistics show that data can motivate subjects to speak in longer utterances. However, they do not guarantee motivational impact in this study, because of the type 1 error inherent in a one way ANOVA done over a span of time. Nevertheless, a closer look at some random samples of the data shows that the subjects seemed motivated to increase their MLT and TIME. They show signs of challenging themselves to speak with varied vocabulary, and to elicit more conversation so that their partners have a better chance at beating previous data. Note the spontaneous question by Student A (STA), below as well as the repetition of *"Generally."* to elicit conversation from Student B (STB) (see APPENDIX A for full excerpts).

EXCERPT 1. AN EXCERPT FROM A GROUP CONVERSING ON THE FIRST TOPIC, *CONFLICT*.

**STE:I wrote my essay just like you. I thought, when I'm tired, very tired and so sleepy, I have a report to hand in tomorrow, so there is the inner conflict. I wrote it already essay so I understand it. It is very vrey difficult but I must win my desire.*

**STA:What is your conflict that you wrote?*

**STB:Generally.*

**STA:Generally.*

**STB:How prevent from conflict so very difficult to think.*

**STA:How do you think to prevent conflicts? It's that you can't prevent my inner conflict?*

STA prompts STB with the question *"What is your conflict that you wrote?"* This question is answered with one word, which STA repeats in order to keep the conversation going. Throughout the course, many strategies like this one emerged as students did their best to increase their MLT and TIME in order to beat the data from previous sessions.

Students also had a chance to focus on innate Japanese-English errors. In the last utterance in Excerpt 1 above, the pronoun *my* was wrongly used for the possessive pronoun *your*, a common error among Japanese subjects which occurs because of the ambiguity of *自分の jibun no*, which translates as *one's own* in their first language. This type of interference is well documented. The transcripts doubled as a speaking assessment and an exercise in lexis, orthography and syntax where the subjects were able to receive instruction afterwards on the mistakes made in their transcriptions. Adams (1980) noted that vocabulary and grammar are the two factors that distinguish proficiency most dramatically. OS Scoring proved useful for indicating irregularities, and facilitated subsequent formative instruction in vocabulary and grammar, with the added benefit of being able to focus on innate Japanese-English errors. This proved to be highly motivational as subjects looked forward to receiving constructive feedback on their transcripts.

Students are motivated when they feel like they are making progress, and the system became a useful CALL tool for pinpointing where progress is most necessary. The group of students in Excerpt 1, especially STA, showed improvement in consecutive sessions. Excerpt 2 below shows progress in the form of longer utterances and better coherence, although there is much room for improvement in spelling.

EXCERPT 2. AN EXCERPT FROM THE SAME GROUP CONVERSING ON A LATER TOPIC, *POVERTY*.

**STA:Speaking of poverty, we likely to think that the problem is only in developing countries but, certainly, there is a problem about poverty in developed coutries. Then, what is like the problem about poverty in developed countries? We can pick up, for instance, working poor problem. Workong poor is the people who can't get enough money to live even though they get regular jobs. The trouble is, they can't get much money to live. but they have a certain job so they can't use public welfare system, in Japanese, Seikatsuhogo. So they are suffering from serious poverty.* [sic]

Excerpt 1 and 2 above are samples from the files of the same student (STA). Notice that the differences between Excerpt 1 and 2 are significant when it comes to MLT and TIME (longer, more coherent utterances as well as more language produced in the allocated time). Both variables increased. STA seemed adamant to generate conversation in the first excerpt already by asking questions and repeating responses. This initial positivity seems to bloom in Excerpt 2 where there are almost monologue length utterances and a discernable freedom in the flow of expression.

Almost all of the submitted CHAT data files, of which Excerpt 1 and 2 above are examples, show improvement in coherence and increases in amount of language produced. These excepts serve as examples of how beating the data in subsequent sessions can become a positive influence as a goal, and ultimately a valuable motivational strategy.

XIII. CONCLUSION

Although there was a significant increase in the overall mean length of speaking turn (MLT), this research looked at speaking data through an ANOVA over a span of time. Other influences may have contributed to the increase, for example reminders by researchers to beat the data of each previous session. A follow-up study is in the works which will include a control group where students will not analyze their data. This group will then be compared to a group subjected to OS Scoring with analyses as discussed in this article. Thus, the efficacy of OS Scoring and its motivational impact can be isolated and measured with more accuracy. Nevertheless, since the focus is on measuring an increase in speaking time and length of utterances regardless of the cause, i.e. to make sure that an awareness of the data is at least a factor to some degree, the current results seem to be proof enough that conversation data can be a motivator as well as an illuminating indicator of changes in student speaking behavior in an authentic conversational setting.

The variances in TIME became smaller after each session. The researcher interprets it as students tuning into one another over time. As they matured in the testing environment, less communicative speakers were coerced into speaking more in order to beat their data from previous sessions. At the same time, overbearing speakers started to hold back in order to give their reluctant friends more time.

Because OS Scoring involved subjects in the speech rating process, it facilitated formative instruction in morpho-syntactic errors specific to Japanese as first language learners, and thereby doubled as a self-motivational learning tool. This is perhaps its most valuable aspect, and it is sincerely hoped that other researchers in the field will recreate and improve this study by trying the system in their courses, collecting data of conversations, and by finding ways in which it can be used to motivate students to have more confidence in conversations in English and to scaffold skills in preparation for international computer based tests.

APPENDIX. TRANSCRIPTS OF CHAT DATA

EXCERPT 1. AN EXCERPT FROM A GROUP CONVERSING ON THE FIRST TOPIC, *CONFLICT*.

**STE: I wrote my essay just like you.*
**STE: I thought, when I'm tired, very tired and so sleepy, I have a report to hand in tomorrow.*
**STE: so there is the inner conflict.*
**STE: I wrote it already essay so I understand it.*
@Time Duration: 05:47-06:08
**STA: It is very vrey difficult but I must win my desire.*
@Time Duration: 08:10-08:13
**STA: what is your conflict that you wrote?*
@Time Duration: 08:22-08:23
**STB: generally.*
@Time Duration: 08:23-08:24
**STA: generally.*
@Time Duration: 08:30-08:40
**STB: how prevent from conflict so very difficult to think.*
@Time Duration: 09:23-09:27

**STA: how do you think to prevent conflicts?*
@Time Duration: 10:30-10:35
**STA: it's that you can't prevent my inner conflict?*
@Time Duration: 10:38-10:39
**STA: no no no no sorry sorry.*

EXCERPT 2. AN EXCERPT FROM THE SAME GROUP CONVERSING ON A LATER TOPIC, *POVERTY*.

**STA: speaking of poverty, we likely to think that the problem is only in developing countries but, certainly, there is a problem about poverty in developed coutries.*
@Time Duration: 0:36-0:45
**STA: then, what is like the problem about poverty in developed countries?*
@Time Duration: 0:47-0:55
**STA: we can pick up, for instance, working poor problem.*
@Time Duration:0:56-1:14
**STA: workong poor is the people who can't get enough money to live even though they get regular jobs.*
@Time Duration:1:16-1:28
**STA: the trouble is, they can't get much money to live.*
@Time Duration:1:29-1:47
**STA: but they have a certain job so they can't use public welfare system, in Japanese, Seikatsuhogo.*
@Time Duration:1:48-2:00
**STA: so they are suffering from serious poverty.*
@Time Duration:2:07-2:14
**STA: why they are suffering from poverty?*
@Time Duration:2:15-3:25
**STA: most of them engaged in non-regular job or daytime job and they are so hard to make a livng the day so they don't afford improve job skill to increase their income and it is too difficult to get regular job.*

REFERENCES

[1] ACTFL Proficiency Guidelines. (1999). Hasting-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

[2] Adams, M. L. (1980). Five co-occurring factors in speaking proficiency. In J. Firth (ed.), *Measuring Spoken Proficiency*, 1-6.Washington, DC: Georgetown University Press.

[3] Atkinson, J. (1958). Towards experimental analysis of human motivation in terms of motives, expectancies and incentives. In J. Atkinson (ed.), *Motives in fantasy, action and society*, 288-305. Princeton, NJ: Van Nostrand.

[4] Bachman, L. F. (1990). Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

[5] Bachman, L. F. & Palmer, A.S. (1996). Language Testing in Practice: Designing and Developing Useful Language Tests. Oxford: Oxford University Press.

[6] Bhat, S., Su-Yoon, Y. (2015). Automatic Assessment of Syntactic Complexity for Spontaneous Speech Scoring. *Speech Communication*, 67, 42-57.

[7] Bodnar, S., Cucciarini, C., Strik, H., van Hout, R. (2016). Evaluating the Motivational Impact of CALL Systems: Current Practices and Future Directions. *Computer Assisted Language Learning*, 29, 1, 186-212.

[8] Caban, H. (2003). Rater Group Bias in the Speaking Assessment of four L1 Japanese ESL Students. *Second Language Studies*, 21, 2, 1-44. University of Hawaii.

[9] Cheng, L. (2004). ESL/EFL Instructors' Classroom Assessment Practices: Purposes, methods, and Procedures. *Language Testing*, 21, 3, 360-389.

[10] Cumming, A. (2003). A Teacher-Verification Study of Speaking and Writing Prototype Tasks for a New TOEFL. *TOEFL Research Report*, 1-41.

[11] Ellis, R. (1985). Understanding Second Language Acquisition. Oxford: Pergamon Institute of English.

[12] Flesch, R. (1948). A New Readability Yardstick. *Journal of Applied Psychology*, 32, 3, 99-112.

[13] Fulcher, G. (2003). Testing Second Language Speaking. Harlow. Pearson Longman.

[14] Gamper, J., Knapp, J. (2002). A Review of Intelligent CALL Systems. *Computer Assisted Language Learning*, 15, 4, 329-342.

[15] Gardner, R. C. (1985). Social Psychology and Second Language Learning: The Role of Attitudes and Motivation. London, GB: Edward Arnold.

[16] IELTS. (2015). SPEAKING: Band Descriptors (public version). Retrieved from http://takeielts.britishcouncil.org/find-out-about-results/ielts-assessment-criteria (accessed 12/06/2018).

[17] Iwashita, N., Brown, A., McNamara, T., O'Hagan, S. (2008). Assessed Levels of Second Language Speaking Proficiency: How Distinct? *Applied Linguistics*, 29, 1, 24-49.

[18] Levine, T. R., Hullett, C. R. (2002). Eta Squared, Partial Eta Squared, and Misreporting of Effect Size in Communication Research, *Human Communication Research*, 28, 4, 612-625.

[19] Li, W. (2011). Validity Considerations in Designing an Oral Test. *Journal of Language Teaching and Research*, 2, 1, 267-269.

[20] Locke, E.A., Latham, G.P. (2002). Building a Practically Useful Theory of Goal Setting and Task Motivation. *American Psychologist*, 57, 9, 705-717.

[21] MacWhinney, B. (1995). The CHILDES project: Tools for analyzing talk. Hillsdale, NJ: Erlbaum.

[22] Major, R. C., Fitzmaurice, S.F., Bunta, F., Balasubramanian, C. (2002). The Effects of Nonnative Accents on Listening Comprehension: Implications for ESL Assessment, *TESOL Quarterly*, 36, 2, 173-191.

[23] Sawaki, Y. (2007). Construct Validation of Analytic Rating Scales in a Speaking Assessment: Reporting a Score Profile and a Composite. *Language Testing*, 24, 3, 355-390.

[24] Ushioda, E. (2012). Motivation: L2 Learning as a Special Case? In Mercer, S. et al. (eds.), *Psychology for Language Learning: Insights from Research, Theory and Practice*, 58-73: Palgrave Macmillan.

[25] Wanner, P. J. (2002). Student Self-Evaluation Skills Using CHILDES Programs. In Lewis, P. (ed.), *The Changing Face of CALL: A Japanese Perspective*, 138-154. Netherlands: Swets & Zeitlinger.

[26] Zechner, K., Chen, L., Davis, L., Evanini, K., Lee, C.M., Leong, C.W., Wang, X., Yoon, S.Y. (2015). Automated Scoring of Speaking Items in an Assessment for Teachers of English as a Foreign Language. *ETS Research Report RR*, 18, 15-31.

[27] Zhang, S. (2009). The Role of Input, Interaction and Output in the Development of Oral Fluency. *English Language Testing*, 2, 4, 91-100.

Austin Gardiner holds an MA in Applied Linguistics from The University of Birmingham, UK, and is currently finishing his PhD in Linguistics in Japan. His interests include student-centered approaches, statistics in language research, and computational approaches to language acquisition.