Teacher perceptions of factors influencing the quality of assessment on English communication courses in Japanese universities

Paul Anthony Marshall*

Matsuyama University, 4-2 Bunkyocho, Matsuyama, Ehime, 790-8578, Japan

Abstract

This study measures the perceptions of English language teachers at Japanese universities. It combines the results of online questionnaires on assessment practices and another one on teacher autonomy. Results suggest that according to teachers' self-reports of assessment practices, certain measures of assessment quality are being affected by the high levels of teacher autonomy present in the Japanese university system. The results imply that increased top-down coordination may increase the quality of assessment according to the measures outlined in the study.

1. Introduction

1.1. What Influences the Quality of Assessment on English Communication Courses in Japanese Universities?

Certain conditions are essential for quality assessment to occur. Assessment lacking in validity, reliability, practicality, continuity, or consistency may be problematic depending on the situation. An interesting *influence* on the quality of assessment is the relationship between top-down coordination and teacher autonomy. Benson (2010) conceptualises *student* autonomy as lying between three points rather than being a simple linear scale. The three points on Benson's scale are student control, no control, and other control. If adapted to measure teacher autonomy, then the three points would be teacher control, no control, and other control which would usually mean control by the government, faculty,

* Paul Anthony Marshall
Email: paulanthony.marshall@gmail.com

or department. No control would mean that neither the institution, nor the teacher are in control, which is possible in cases where teachers are unable or unwilling to control, for example, their own curriculum, class discipline, or assessment quality. It is clear from Benson's explanation that the low level of top-down coordination discovered by Pritchard and Moore (2016) in Japanese universities must by default mean high levels of teacher control and/or no control. The study by Pritchard & Moore, (2016) supported the hypothesis that universities in Japan have a low level of top-down coordination. In a university English language context, a lack of top-down coordination can mean a lack of curricular alignment and accountability. All of these factors can then lead to a lack of the aforementioned measures of quality in assessment; reliability, continuity and consistency, which will be further discussed in the following sections.

1.2. What is Quality in Assessment?

There has been much debate about the best definition of 'assessment'. Lynch (2001:358) considers assessment to be 'the systematic gathering of information for the purposes of making decisions or judgements about individuals'. He also considers both *measurement* and *testing* to be aspects of *assessment*. Taras (2005:467) classifies educational assessment as 'a judgement which can be justified according to specific weighted set goals, yielding either comparative or numerical ratings,… justifying the judgement against the stated goals and criteria.' Furthermore, Berry & Adamson (2011:3) believe assessment to be 'one of the most emotive words in the education lexicon' and 'the subject of intense debate around the world'. For the purposes of this paper, Taras' definition is sufficient due to its inclusion of both *goals* and *criteria*, both benchmarks of quality assessment. It is vital that teachers, courses, institutions, and governments understand how much positive or negative impact assessment can have on both student motivation and learning outcomes.

The concept of quality is a complicated one; as an abstract concept it may be unquantifiable and near impossible to articulate (as deliberated over in great philosophical depth in the novel Zen and the Art of Motorcycle Maintenance, Pirsig, 1974). Firstly, different people or organisations might believe different criteria to be good measures of quality for a certain situation. Also, they might have different ideas about what standard constitutes high or low quality (Harvey & Green, 1993). Varying measures or standards of quality might also be set by different organisations; for example, by universities or by governments. Decisions about these measures and standards are obviously important and affect the quality that is aspired to and may influence many peoples' lives. The measures of quality that follow can be said to be generally accepted measures of quality in the field of assessment but are, of course, under continual debate and review.

1.2.1. Validity

Academic debate over assessment validity has been continual for almost a century (see for example, Buckingham et al., 1921; Ruch, 1924). While some have gone so far as to suggest that the term 'validity' has outlived its usefulness and discussed scrapping the term (Newton, 2013), it has also been described as the foremost aspect of test quality (Koretz, 2008: 215). Validity can be said to be 'the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests' (AERA, APA, NCME 2014: n.p.), although this is still a hotly contested topic. Validity is also often closely connected to using test scores to make decisions about the test-takers (Newton & Shaw, 2014). Assessments can be carried out using either direct tasks, which involve students carrying out the exact skill they are being assessed on, or using indirect tasks. Indirect tasks are popular because they are easier to administer but they have been criticized for not measuring what they claim to be measuring (Weir, 1990: 47); being able to choose the correct pronoun from a list on a paper test, for example, is not the same as correctly using pronouns while speaking or writing. In order to fulfil validity claims that tests assess students` communicative competence, direct tasks must be used; tasks where students are required to communicate (Talandis, 2017). In other words, in order to accurately know how well a student can speak English, it is essential that the test involves actually speaking English.

1.2.2. Reliability

Reliability is a characteristic of assessment which means that `the degree to which the measure of a construct is consistent or dependable` (Bhattacherjee, 2012: 56). For example, IELTS continually conducts testing sessions, each one with different questions. For organisations such as providers of standardised tests, which promise a high level of reliability to test-takers, considerable effort and expense have to be invested in order to establish and maintain it. In fact, this quality and the difficulty in achieving it, is one of the great strengths of exams such as TOEFL and TOEIC. In order to do this, inter-rater reliability must be established and maintained through standardisation training and monitoring procedures. Due to the high reliability of the IELTS exam, candidates can expect to have their English ability measured fairly and objectively whether they join the session in May in Tokyo, or in November in Buenos Aires. It does not matter where and when a candidate takes the test; the same performance should yield the same score. Reliability is obviously a very important test characteristic because it is necessary for the measurement to be accurate and related to criteria or descriptors in order for it to have any meaning at all (Bachman & Palmer, 1996).

1.2.3. Practicality

Despite the aforementioned methods of ensuring quality in language assessment, many teachers use indirect testing methods due to time constraints or reasons of *logistical ease*. For example, in a class of 28 students it is time-consuming and possibly inconvenient in terms of sound-levels and space to hear each student speak. As a result, validity may be compromised, and the quality of assessment can suffer.

1.2.4. Continuity

When students take *consecutive courses* such as English communication 1 and English communication 2, it follows that the courses *ought to be assessed in the same way* so that students' progress can be monitored and compared. Objectives can be set, and standards can be monitored; for both teaching and learning. The effectiveness of teaching materials and approaches can be gauged.

1.2.5. Consistency

When students take *two courses that purport to be the same*, such as two English communication 1 courses at the same institution, the experience should be similar. Materials, teaching approaches, classroom activities and *assessment methods should be coordinated*. This means that students, or potential customers, know what to expect in terms of standards and outcomes.

1.3. Hypotheses and Research Questions

It is hypothesised that the high levels of teacher autonomy on English Communication courses in Japanese universities negatively affects the quality of assessment. The resulting research questions are as follows:

RQ: To what extent are teachers on English Communication courses in Japanese universities assessing students according to some sort of coordinated system?

RQ: To what extent are teachers on English Communication courses in Japanese universities assessing students according to the measures of quality defined in this paper?

2. Method

This section will describe the participants, data collection methods, and data collection instruments used in this study.

2.1. Participants & Data Collection Methods

This study was carried out by combining the results of two online questionnaires. The first questionnaire, on assessment practices, was completed by 11 university teachers of English Communication in Japan. The second online questionnaire, on teacher autonomy, was completed by 18 university teachers of English Communication in Japan. Many of the respondents were the same for both surveys. The teachers are mostly, but not all, native-English speakers, and are based in universities both in large urban areas and in small cities in more rural areas in Japan. They teach at a mix of public and private universities. The autonomy respondents all stated that they work at universities in Japan. The job titles disclosed were 5 lecturers, 5 assistant professors, 5 associate professors, and 2 professors. One respondent did not submit a job title. Of these teachers, eight are on permanent contracts, while the other ten are on full-time, temporary contracts.

2.2. Data Collection Instruments

The assessment questionnaire was written using the surveymonkey website. The autonomy survey was written using Google Forms. Both of these were distributed separately by email. The assessment questionnaire aimed to probe the assessment practices of teachers of English Communication courses, and consisted of the questions displayed in Appendix 1.

The autonomy survey (Appendix 2) was taken from Pearson and Hall (1993). They developed and trialled the Teacher Autonomy Scale (TAS), which consists of 20 items. The items are divided into two constructs; general autonomy and curricular autonomy. Teachers were asked to rate their reaction to each statement using a 4-point Likert scale. For this study, three questions regarding respondents, and an 'additional comments' section was added. The four options and the adapted version of the TAS is depicted below.

3. Results

Teachers were asked to state which tasks they use to assess students' speaking. Multiple answers were allowed. The most popular task was group discussions (9 out of 11 teachers), followed by presentations (8). Interviews (5) and reading dialogues (2) were also popular. The following additional tasks were also mentioned in the 'Other. Please Specify' section: "poster presentations, Voice Thread, small groups, student-centred work rather than presentations, which are passive except for the presenters." And "a general participation score that I just guess at".

Next, teachers were asked to state which tasks they use to assess students' writing. Again, multiple answers were allowed. The most popular task was essays (11), but timed tests (5), emails (4), and journals (3) were also popular. The following additional tasks were also mentioned in the 'Other. Please Specify' section: "Bibliography, Essay plan/outline, Thesis statements, Paragraph reports every 3-4 weeks, Model United Nations research portfolios and papers; speeches, Edmodo".

After that, teachers were asked who decides which tasks are used to assess students. The faculty does not decide in any of the contexts where these teachers work, what tasks they should use to assess students. In fact, in most cases (7), the teachers decide by themselves, and a much smaller number (3) stated that a combination of teacher and faculty decide. The following comment was submitted: "we have common online homework, decided by all English teachers, and their completion of these assignments affects their grade in all general English courses."

The next question was: "If specific criteria are used to assess students, who decides these?" All eleven respondents said that some form of criteria are used but that these are not decided by the faculty. Eight teachers said that the teacher decides the assessment criteria, while two said that a combination of teacher and faculty makes the decision. Finally, one teacher commented: "we use an inhouse textbook for communication. For the units in the textbook, there are some tests provided for teachers. They are recommended but not required to use them."

The last question in the assessment survey asked respondents to specify the method of grade distribution used in their classes. Four teachers grade according to a bell-curve (also called normal distribution), one teacher distributes an even number of As, Bs, and Cs, one teacher grades students to an external, unchangeable standard. Five teachers chose to make comments: "often end up as a bell curve", "as earned according to set criteria", "Not evenly, as that seems artificial, but if a student gets a certain percentage, that = A, if they get another percentage that = C, according to the university's

guidelines", "The university has a system where most advanced students get As or S. This is unfair to students who try hard at 'lower' levels.", and "evaluation on an absolute scale".

Moving on to the teacher autonomy scale (TAS) survey, the results from the 12 questions concerning general autonomy, and the 8 questions concerning curricular autonomy overwhelmingly showed that teachers of English Communication courses in Japanese universities perceive that they have a high level of autonomy. For general autonomy, this was made clear in terms of freedom to be creative, selection of activities, selection of teaching methods and strategies, control of the use of classroom space, teacher discretion, and control of the use of classroom time. Furthermore, the majority of teachers responded that it is `More or less false` or `Definitely false` that they seldom use alternative procedures, do not have limited latitude in dealing with problems, evaluation and assessment activities are not decided by other people, teaching approaches are not specified by the school and instructional planning is not dictated by district needs. Perceived high levels of curricular autonomy were made clear in terms of what is taught in class, when and how topics are taught, the content and skills taught, the materials, the standards, the guidelines and procedures, and goals and objectives are determined by themselves.

4. Discussion & Conclusions

Having summarised the results, it is now necessary to reflect back on the research questions and decide to what extent they can be answered. The first question was; to what extent are teachers on English Communication courses in Japanese universities assessing students according to some sort of coordinated system? The findings shed light on the fact that in most cases teachers are independently making decisions about the tasks and criteria used to assess students. This suggests that coordinated systems of assessment are limited in some or many of these contexts.

The second research question was; to what extent are teachers on English Communication courses in Japanese universities assessing students according to the measures of quality defined in this paper? Teachers' self-reports about their assessment practices, assessment criteria, grading system, and decision-making authorities on assessment for English Communication courses in universities in Japan, shed light on both positive and negative aspects of the quality of assessment. On the plus side, it would appear that teachers are mostly using direct testing methods and therefore some level of *validity* might be assumed. Additionally, teachers are free to test students using methods

according to *practicality* in their classroom context. However, it is also apparent from the findings that teachers are assessing students using different tasks, according to different criteria, and using different grading systems. This, in itself, suggests that *reliability, continuity,* and *consistency* may be lacking.

While findings should only be interpreted with a full understanding of contextual factors, it is often the case that where assessment is not coordinated then control or monitoring of reliability, continuity, and consistency cannot be carried out between courses, departments, or universities.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

Bhattacherjee, A. (2012). *Social science research: Principles, methods, and practices*. Textbooks Collection. 3.

Benson, P. (2010). 'Measuring autonomy: Should we put our ability to the test'. In Paran, A., & Sercu, L., *Testing the untestable in language education*, 77-97. Bristol, UK: Multilingual Matters.

Berry, R., & Adamson, B. (Eds.). (2011). *Assessment reform in education: policy and practice* (Vol. 14). Springer Science & Business Media.

Buckingham, B. R., McCall, W. A., Otis, A. S., Rugg, H. O., Trabue, M. R., & Courtis, S. A. (1921). 'Report of the standardization committee'. *Journal of Educational Research*, *4*(1), 78-80.

Harvey, L. & Green, D. (1993). 'Defining Quality'. *Assessment & Evaluation in Higher Education*, 18:1, 9-34.

Koretz, D. (2008). A measured approach. *American Educator*, *32*(2), 18-39.

Lynch, B. K. (2001). 'Rethinking assessment from a critical perspective'. *Language Testing*, *18*(4), 351-372.

Newton, P. (2013). *Does it matter what "validity" means*. Lecture given at University of Oxford, Department of Education.

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. Sage.

Pearson, L. C., & Hall, B. W. (1993). 'Initial construct validation of the teaching autonomy scale'. *The Journal of Educational Research, 86(3)*, 172-178.

Pirsig, R. M. (1999). *Zen and the art of motorcycle maintenance: An inquiry into values*. Random House.

Prichard, C., & Moore, J. E. (2016). 'The Balance of Teacher Autonomy and Top-Down Coordination in ESOL Programs'. *TESOL Quarterly, 50(1)*, 190-201.

Ruch, G. M. (1924). *The improvement of the written examination*. Chicago, IL: Scott, Foresman.

Talandis, Jr., J. (2017). *How to test speaking skills in Japan: A quick-start guide*. Kyoto: Alma Publishing.

Taras, M. (2005). 'Assessment–summative and formative–some theoretical reflections'. *British journal of educational studies*, *53*(4), 466-478.

Weir, C. J. (1990). *Communicative language testing*. Hemel Hempstead: Prentice Hall.

Appendix 1: Assessment Questionnaire: Answers were multiple choice but with an additional option of 'Other. Please Specify.'

What is your current position?
What kinds of tasks do you use to assess students' speaking?

What kinds of tasks do you use to assess students' writing?
Who decides what tasks are used to assess students?
If specific criteria are used to assess students, who decides these?
How are grades distributed in your classes?

Appendix 2: Pearson and Hall`s 1993 Teacher Autonomy Scale (TAS)

1 = Definitely true

2 = More or less true

3 = More or less false

4 = Definitely false

General Autonomy	Likert Scale:
I am free to be creative in my teaching approach.	1 2 3 4
The selection of student-learning activities in my class is under my control.	1 2 3 4
My teaching primarily follows approaches that are specified by the school.	1 2 3 4
I seldom use alternative procedures in my teaching.	1 2 3 4
My instructional planning is dictated by district needs.	1 2 3 4
My job does not allow for much discretion on my part.	1 2 3 4
The scheduling of use of time in my classroom is under my control.	1 2 3 4
In my situation, I have only limited latitude in how major problems are resolved.	1 2 3 4
In my class, I have little control over how classroom space is used.	1 2 3 4
The evaluation and assessment activities used in my class are selected by people other than myself.	1 2 3 4
I select the teaching methods and strategies I use with my students.	1 2 3 4

I have little say over the scheduling of use of time in my classroom.	1 2 3 4
Curriculum Autonomy	
What I teach in my class is determined for the most part by myself.	1 2 3 4
The content and skills taught in my class are those I select.	1 2 3 4
My teaching focuses on those goals and objectives I select myself.	1 2 3 4
The materials I use in my class are chosen for the most part by myself.	1 2 3 4
In my teaching, I use my own guidelines and procedures.	1 2 3 4
In my situation, I have little say over the content and skills that are selected for teaching.	1 2 3 4
Standards for my classroom are set primarily by myself.	1 2 3 4
I follow my own dictates as to when and how topics are taught.	1 2 3 4