LEARN Journal: Language Education and Acquisition Research Network
(ISSN: 2630-0672 (Print) | ISSN: 2672-9431 (Online)
Volume: 15, No:1, January – June 2022
Language Institute, Thammasat University
https://so04.tci-thaijo.org/index.php/LEARN/index

# Reactions of Teachers and Students Towards the Implementation of Performance-based Language Assessment: A Washback Study in Hokkaido, Japan

**Bordin Chinda[a,*], Matthew Cotter[b], Matthew Ebrey[c], Don Hinkelman[d], Peter Lambert[e], Annie Miller[f]**

[a]bordin.chinda@cmu.ac.th, English Department, Faculty of Humanities, Chiang Mai University, Thailand
[b] m-cotter@hokusei.ac.jp, English Department, Hokusei Gakuen University Junior College, Japan
[c] mattebrey@outlook.com, Faculty of Humanities, Massey University, New Zealand
[d] donhinkelman@gmail.com, English Department, Faculty of Humanities, Sapporo Gakuin University, Japan
[e] pbhlambert@gmail.com, English department, Kiyota High School, Japan
[f] annie@xa.ejnet.ne.jp, English Department and C.E.P. Program, Hokusei Gakuen University, Japan
[*] Corresponding author, bordin.chinda@cmu.ac.th

A.P.A. Citation:
Chinda, B., Cotter, M., Ebrey, M., Hinkelman, D., Lambert, P., & Miller, A. (2022). Reactions of teachers and students towards the implementation of performance-based language assessment: A washback study in Hokkaido, Japan. *LEARN Journal: Language Education and Acquisition Research Network,* 15*(1),* 524-547.

Received 05/09/2021

Received in the revised form 14/12/2021

Accepted 23/12/2021

Keywords
washback; performance-

## Abstract

This qualitative study investigated the washback of performance-based assessment used by three English language teachers in Hokkaido, Japan, each of whom implemented their own course-specific assessments. The study employed qualitative research of an in-depth interview with 15 students and a self-reflective method from 3 teachers. Teachers demonstrated different approaches in implementing performance-based language

| | |
|---|---|
| based assessment; classroom-based assessment | assessment. However, they agreed that this type of assessment could enhance students' communicative skills and rubrics were crucial in the assessment process. In terms of washback on students, they found that performance-based assessment, with detailed feedback, provided learners with a comfortable and challenging learning environment, leading to self-developed English performances, feelings of accomplishment, and better retention of English used in the presentations. However, some students found performance-based assessment demanding, causing anxiety. The findings suggest that when implementing performance-based assessment teachers should consider different aspects of the assessment to create positive washback. |

## Introduction

Assessment in a classroom context has been one of the central interests in language education. Teachers have to teach and assess students, often in the ESL/EFL context through performance-based assessments. In addition, with the arrival of communicative language teaching, language testing and assessment have also shifted to focus more on students' performances. However, it is well accepted that "assessments come in all shapes and sizes, ranging from international monitoring exercises to working with individual pupils in the classroom. These assessments each have their purposes and *consequences*" (Stobart, 2003, p. 139, emphasis added). Language educators know the consequences or effects of assessment as "washback." For example, Brown and Hudson (1998) recognize both negative and positive washback effects of assessments on the curriculum.

Hamp-Lyons (1997), however, stresses that alternative assessment, including performance-based language assessment, cannot be assumed to have beneficial washback into teaching and learning (p. 300). Despite the fact that teacher assessment practice in a classroom is a complex phenomenon, there are few empirical studies investigating assessment in a classroom context, compared to considerable literature on large-scale proficiency tests. McNamara (2001) expresses the concern that "too much language testing research is about high-stakes proficiency tests, ignoring classroom contexts, and focusing on the use of technically sophisticated quantitative methods to improve the quality of tests at the

expense of methods more accessible to non-experts" (p. 329). Tsagari and Cheng (2017) suggest that more research is necessary to examine the impact of tests on students and their learning. In other words, a lack of focus in previous studies has been the direct influence of testing or assessment on learners, their learning processes, and learning outcomes.

Therefore, this study aimed to investigate the positive and negative washback of performance-based language assessment on both teachers and students in an EFL classroom context from three university level classes in Hokkaido, Japan, by employing qualitative research methodology.

## Literature Review

### Performance-based Language Assessment

Performance-based assessment, one kind of "alternative assessment," is based on "an investigation of developmental sequences in student learning, a sampling of genuine performances that reveal the underlying thinking processes, and the provision of an opportunity for further learning" (Lynch, 2001, p. 228). Davies et al. (1999) define performance-based assessment as a "test in which the ability of candidates to perform particular tasks ... is assessed" (p. 144) and McNamara (1996) states that an important aspect of performance testing is that "the assessment of the actual performances of relevant tasks is required of candidates, rather than the more abstract demonstration of knowledge, often by means of paper-and-pencil tests" (p. 6). Learners' productive language skills are assessed through performance tasks in the second language that demonstrate their abilities to use skills required in realistic situations. (Wigglesworth, 2008, p. 111). In addition, Wigglesworth (2008) notes three factors that distinguish performance tests from traditional tests of a second language: (1) there is a performance by the candidate; (2) the performance is judged using an agreed set of criteria; and (3) there is a degree of authenticity of the assessment tasks (p. 113).

### Washback

To first understand washback, it is necessary to mention the academic rationale for using testing and assessment within education. Bachman and Palmer (1996) acknowledged that tests have an impact "on society and educational systems and upon the individuals within those systems" (p. 29). Shohamy (2007) went further by proposing that assessment has been viewed as a powerful tool used by “authorities” to create change. Washback refers to how those results can in turn influence teaching pedagogies, styles, learning environments and the learning that takes place in them. Wall (1997) points out that washback is sometimes used interchangeably with "impact," but the term "washback" is "more frequently used to refer to the *effects* of tests on teaching and learning" (p. 291, emphasis added). These effects are usually perceived as negative because teachers could be forced to do what they "do not necessarily wish to do" (Alderson & Banerjee, 2001).

Alderson and Banerjee also report that researchers have argued that "tests are potentially also 'levers for change' in language education ... [i.e.] good tests should or could have positive washback" (p. 214). In other words, the effects of tests could be either positive or negative. Alderson and Wall (1993) comment on this by being of the opinion that when conducting studies in washback, researchers need to consider both negative and positive effects as washback entails that the effects of tests can be either intended or unintended, and directly or indirectly. Bailey (1996) proposes a simple washback model in which she argues that not only did tests have an impact on participants (including students, teachers, materials writers, and curriculum designers, and researchers), and the products (including learning, teaching, new materials, and new curricula, and research results), but also that participants may in turn have an impact on the tests. Based on the above review, therefore, the present paper suggests that any assessment task has the potential to affect the teaching and learning in a given context, in a positive or negative way.

In terms of washback research, Wall and Alderson (1993) were among the first researchers to call for the conducting of more studies in this area. They examined the effects of the new O-level examination on English teaching in secondary schools in Sri Lanka, in which they found that there was evidence of positive and negative washback on the content of teaching. There was no evidence of washback on teaching

methodology, but there was evidence of positive and negative washback on the way teachers and local education offices designed tests. In other words, the introduction of the new examination impacted "*what* teachers teach but not on *how* they teach" (p. 68, emphasis in original). In another context, Watanabe (1996) investigated whether the use of grammar-translation in classrooms was in fact due to grammar-translation used in university entrance examinations in Japan. The findings revealed that the presence of translation questions did not affect the two teachers in the same way. Translation-oriented entrance exams had washback effects on some teachers but not on others depending on personal beliefs, educational background, and past learning experience.

In addition, Alderson and Hamp-Lyons (1996) examined the washback effects of TOEFL (Test of English as a Foreign Language) on preparation classrooms at a language institute in the United States. Different from Wall and Alderson's (1993) study in Sri Lanka, Alderson and Hamp-Lyons found that "the TOEFL affects both *what* and *how* teachers teach, but the effect is not the same in degree or kind from teacher to teacher" and, "the simple difference of TOEFL versus non-TOEFL teaching does not explain *why* they teach the way they do" (p. 295, emphases in original). Wall and Horák conducted another washback study of TOEFL (2006; 2008) investigating the impact of the changes of the TOEFL test (to the Internet-based test, iBT) on teaching and learning in preparing students in Central and Eastern Europe to take the test. The qualitative data revealed that at the beginning of the study, teachers' awareness of the changes in the TOEFL was quite low but grew during the study. They also found that the teachers had a positive attitude toward introducing the speaking test and the integrated writing task. The teachers also expressed that the changes in the test would result in changes in their classroom.

In a different context, Cheng (2005) investigated the washback of the Hong Kong Certificate of Educational Examination in English (HKCEE), a high-stakes public exam in secondary schools. Similar to Wall & Alderson (1993), Cheng found that introducing the new examination affected what teachers teach, but not how. In other words, the change of the examination could change teachers' classroom activities, but it did not change teachers' beliefs and attitudes about teaching and learning, the roles of teachers and students, and how teaching and learning should be carried out. Wall (2005) revisited the Sri Lanka impact study (Wall &

Alderson, 1993) and concluded that the impact of exams is complex, which "should not be seen as a natural or inevitable consequence[s] of introducing a new examination into an educational setting." However, "the design of the examination will always have some effect on the way that teachers react to it" (p. 279).

Since Wall and Horák's 2008 study, the effects of washback on performance-based studies in the EFL/ESL context have seen a minimal share of research limelight when compared to large-scale assessment research. While performance-based assessment is regarded as an emerging, empowering form of assessment, classroom-based studies are needed to verify this claim. The present study, thus, aims to investigate, "What are the perceived positive and negative effects of performance-based assessment by both teachers and students in an EFL/ESL classroom context?"

## Research Methodology

### Research Design

This qualitative research employed semi-structured interviews and self-reflection (i.e., self-reports) as the data collection methods. Three teachers, also the authors of the present study, participated in the research. Each teacher made a self-report, following the same guidelines, on their thoughts about and practices in performance-based assessment by reflecting at the end of the semester after implementing the performance-based language assessment tasks (see Appendix A for the guidelines). The recordings of the reflections were transcribed and analyzed. Fifteen students, who were recruited on a voluntary basis, from the three teachers were interviewed (see Appendix B for the interview schedules). The interviews were conducted in Japanese, the first language of the students, in order to avoid any language barriers, and the interviewer was a Japanese research assistant. The interviews were recorded, and the recordings were transcribed and translated to English for the analysis. It should be noted that the self-report guidelines and the interview schedules were developed by the researchers based on the washback literature. The drafts were trialed and revised according to the comments.

## Data Analysis

For the analysis of the teachers' reflections, to create greater outsider positionality, one of the authors who was not part of the interviews was responsible for analyzing the washback on the three teachers. To analyze the students' interviews, each of the three teachers analyzed a different teacher's transcripts. To ensure the trustworthiness of the analysis, the analyses of the teachers were verified by another author. In the analysis process, if there were any disagreements, the adjustments were made by the consensus of all authors. Furthermore, the analysis of the study followed the grounded theory guidelines of doing qualitative content analysis, that is, open coding and axial coding (Corbin & Strauss, 2008). Moreover, the names and genders of the teachers were anonymized for confidentiality purposes; that is, Teacher 1, Teacher 2, and Teacher 3, and the pronoun "he" was used for all teachers.

## Research context

Since the present study was conducted during the Covid19 pandemic (December 2020 to March 2021), the researchers were restricted to interviews conducted online via the Zoom application. Moreover, classes under investigation were conducted online via either Zoom or Moodle live conferencing. Performance-based assessment, in the form of presentations or roleplays, was also carried out online. Moreover, the data were collected from three different universities in Hokkaido, Japan. Though the medium of instruction in these universities is Japanese, the teachers used English as they are native speakers of English.

In terms of participants, Teacher 1 began teaching English subjects in Japan 20 years ago. The course under investigation was a compulsory English communication course for a group of English majors. The students were first and second-year students. The assessments done were presentations on various subjects, and the students worked in pairs to prepare and present a skit related to one or more of the topics from the textbook.

Teacher 2 has been teaching EFL in university in Japan for 35 years and previously was an adult training instructor for an NGO in

different countries. The course under investigation was a compulsory second-year English for academic presentations course for second-year economics majors. The students had to do a 5-minute, 10-slide PowerPoint presentation.

Teacher 3 has had experience teaching in New Zealand and England, but for the last 20 years has been teaching in Japan young learners through to tertiary. He was teaching a content-driven elective course about international cultures. Assessment entailed active participation in class, participation in an international virtual exchange, online flipped and revision quizzes, and a group presentation. The students had to work and present in small groups with the zoom format using PowerPoint.

## Research Findings

### Teachers' Reactions Towards the Implementation of Performance-Based Language Assessment

The findings were drawn from the teachers' reflections which pointed out that the teachers had different ways to implement performance-based assessment and had different views towards the assessment.

#### *Teacher 1*

In his reflections, Teacher 1 stated how a textbook was used to teach his course, and the students were instructed to discuss the topics in breakout rooms (a function in the application Zoom where a teacher could put students into separate rooms). Although he could not monitor all students at the same time, he believed the students used English throughout their group work as he said, "I think they do a lot of speaking in groups, though I think they do keep it in English because they want to improve their skills." Therefore, Teacher 1 usually encouraged students to do repeated, lengthy pair talking and group work conversation with each other. As he put it, "I normally could tell if each group or partners were continuing to speak in English as soon as I let them go, but I think they do much speaking in groups... I think they do keep it in English because they want to improve their skills." In order to have successful group or pair work, Teacher 1 had his students work and present in

groups with members they chose themselves, which he believed could encourage natural communication. In other words, this is positive as it helps prepare them for interacting with others in employment in later life.

Moreover, Teacher 1 instructed the students to create dialogues, based on the requirements in the textbook, for the final performance task. The assessment of the dialogues was focused on grammar points and key vocabulary (as specified in the rubric). Therefore, when giving instructions, Teacher 1 made sure his students included specific grammar structures and vocabulary (i.e., what the teacher taught). Finally, in the final presentation, Teacher 1 used a comprehensive rubric to rate students' performances.

### *Teacher 2*

Teacher 2 started his class using an online text chat following a question-and-answer format. Then, he demonstrated the assignments, often showing a rubric and assessing an exemplary student's performance step by step using the rubric. For him, the rubric clearly demonstrated what was required to communicate effectively, and showing the exemplary student was an actual model of successful communication. Although Teacher 2 used a textbook, he was not happy with it because it was not interactive. Unlike Teacher 1, Teacher 2 did not pay much attention to grammar and vocabulary because he decided to teach the course with performance-based (oral presentation and communication skills) objectives.

Furthermore, Teacher 2 tried to encourage greater student responsibility by making presentations interactive. He believed that his students' self-assessment could improve their learning; consequently, he taught them how to do self-assessment. He emphasized: "I would have them make their own grading sheet on paper and take a photograph of that and send it to me because I think the teacher does too much of the assessment and they need to do more self-assessment and self-grading." Teacher 2 required the students to do three presentations, and the second was a chance to improve on the first. The advantage of this approach is that it promotes self-reflection and enables implementing better strategies upon reflection.

Finally, Teacher 2 relied heavily on his rubric and went through it with specific examples from a former student's exemplary presentation. In addition, given that the teacher and his students had gone over the rubric's elements in great detail, these high standards were meant to deliver an interactive exchange between a well-prepared presenter and an engaged audience. "I would demonstrate the assignments and walk them through, often showing a rubric of another student's work or assessing the student's performance using a rubric with a shared screen to the whole class."

### *Teacher 3*

Teacher 3 decided not to use a textbook but did Zoom sessions for each week's topic with teacher-created PowerPoint slides. Teacher 3 used 50% of the class period in breakout rooms of student pairs, doing intermittent supervision similar to Teacher 1 and Teacher 2. Teacher 3 collected student essays on Moodle and spoke of a possible discussion forum for issues related to the final presentation and to make sure students were on the right track. Moreover, Teacher 3 required his students to participate in the International Virtual Exchange and submit written essay responses to questions about the weekly lecture topics, which were not directly associated with the final performance. Unlike Teachers 1 and 2, Teacher 3 had no textbook, preferring to rely on material collected over the five years he had taught the course. Like Teacher 1, Teacher 3 had his students work and present in groups with their chosen partners. Teacher 1 emphasized the use of grammar and vocabulary studied, while Teacher 3, similar to what Teacher 2 taught, mentioned that accurate grammar and keywords were not a priority since he was teaching a content-based course.

Teacher 3 also believed that feedback could contribute to students' learning. He lamented that the amount of feedback for the students' essay writing was inadequate, saying, "I would try to give more feedback, but there is so much to do feedback-wise on assessments that, even with only the mini-essays, just reading them is hard enough. It takes much time if feedback is being given to each student. Giving them more comments would be ideal." This lack of feedback can be seen as a challenge for teachers implementing this approach.

Finally, similar to Teachers 1 and 2, Teacher 3 used a rubric for the presentation (reiterating that 10% of the grade was given for simply using the zoom format), whereas for the students' essays and blogging, no rubric was used to guide and grade students on the tasks.

## Students' Reactions Towards the Implementation of Performance-Based Language Assessment

Student interview data were transcribed from Japanese, and qualitative coding found four themes: teacher role, learning environment, assessment task, and feedback.

### *Teacher's Role*

From the interviews with 15 students, the data revealed that both the teacher's role and efforts in creating a safe, comfortable, and challenging learning atmosphere in the assessment process appear to be crucial in creating positive washback. The students of all three teachers expressed a desire for less stress in the class. Teacher 1's students felt comfortable or relaxed with their teacher, as one student said: "I don't get nervous so much anymore." Though Teacher 2's students were nervous about their English ability, their English teacher's commitment comforted them. Similarly, Teacher 3's students had similar comments, for example, "I like him because he is energetic in class and he is very responsive.", and "he is a very kind person, so I tried to create a good atmosphere in the class by mixing in my own jokes." These responses show how teachers' efforts to bring comfort to students can overcome challenging problems and tasks. The positive aspects of such safe learning environments create better learning.

Furthermore, understanding the teacher's use of the second language was a vital part of the feeling of comfort. One of Teacher 1's students found him to be "easy to understand" when explaining things suggesting "he makes sure we understand each other, so it is very easy to understand," while Teacher 3 had reservations about whether several of the students understood what was wrong with their presentations. Finally, in the assessment process, an essential part of the teacher's role was to be accessible, not just for the provision of knowledge but also in a support role. Teacher 1 "made himself available for questions and explained in a way that helped the students." One of Teacher 2's students commented that the teachers "were not doing teaching roles such as giving

lectures and tests," instead "being more of a coach and a manager." Teacher 3 being "energetic and responsive" also reinforces the teacher being in a support role as necessary.

Once establishing the safety and comfort of the class was attended to, a fun and lively teacher became important to further the learning. Teacher 1 was found by his students to be "kind and cheerful," while a large part of Teacher 3's delivery moved one student to comment that "I have gotten a sense that he is trying to keep it fun and upbeat." The findings support the idea of checking for understanding, which expands the teacher's role beyond that of a simple tutor, and into an additional supporting role, especially when assessment is concerned.

### *Learning Environment*

The interviews also found that creating a comfortable and challenging learning space inspired more significant efforts by students in the assessments, creating motivation both in and out of the class. In other words, it helped create positive washback on the students. Teacher 1's students appeared to display high levels of self-efficacy in their work by being allowed to choose their topic for the final assessment. This freedom of choice had a positive effect on their motivation. Regarding the real-life English used in the assessment situation, one student said, "In roleplaying, I could think of situations that existed in English, and I could have fun playing or acting them out." One of Teacher 2's students "felt comfortable speaking in English," which may have resulted from the authentic presentation tasks they were given in pair-work and group work. These positive reactions from students lend credence to the real-life speaking assessment task in which they took part. One of Teacher 1's students said he had developed an interest in using English outside of the classroom. Another of Teacher 1's student gave value to the exercise, stating that "we had to present what we had learned rather than take a test. But I like it because I think that if you make your own slides, you can retain the content." This improvement in language retention speaks to the effectiveness of the exercise. What can be taken from these comments is that real-life communication environments are conducive to better learning.

Moreover, the assessment task derived much more positivity from the peer work component, with the use of English in the planning and production processes and the final assessment presentation. Teacher 3's student said, "I learned the skills of presenting and discussing in a group using English." The use of English as a lingua franca for group work offers an excellent opportunity for

authentic learning. Teacher 1 also had students who enjoyed the collaboration in a group assessment project, saying, "It was really fun to make the slides and think about what we were going to say." One of teacher 2's students referred to the enjoyment of communication, saying, "I wanted to talk to other people and decide what I wanted to do, and I wanted to make a plan, so I was very happy with that."

Finally, the assessment left students with an all-around feeling of accomplishment. Teachers 1 and 2 had students who felt they had improved. One of Teacher 1's students reflected, "I was able to get more than what I learned in the class, which was a great feeling of accomplishment," suggesting more holistic learning and self-reflection. Another said that "it gave me a chance to reflect on what I had learned through the assignment," which shows improvement in another critical skill of learning. Teacher 2's students derived satisfaction from the sense of autonomy they felt through the task, explaining that, for instance, the "self-chosen topic was fun," "I had freedom to create a performance," "it was good to create a task from scratch, and "I like that I could create my task topic and skit theme." In these tasks, the content was chosen freely. Another of these teachers' students "enjoyed the challenge of performance and felt a sense of accomplishment that is different from receiving a passing score on a test." Teacher 3's students expressed satisfaction through "continuing to talk without having to stop too much" and "being told my presentation was the best in the class." The accomplishment felt by the students seems to be centered around the creation of real dialogue autonomously in companion with peers; especially with the use of English as a medium of communication, where peer work and communication have become critical aspects of their learning.

### *Assessment Task*

Anxiety, which could be perceived as negative washback on students' learning, about the assessment tasks (in performance-based assessment) also produced challenges for students. Stepping away from a traditional paper-pencil test that students are familiar with could leave some students with apprehension about how to do well in the performance-based assessment task. From the interviews, students found anxiety from the length of presentations, with a student of Teacher 2's saying, "The presentation time was so long that it was challenging." Apart from the assessment task itself, the final area of difficulty for the students was their grammar and ability to express themselves, as required

by the task. One of Teacher 3's students was concerned that the anxiety over the correct use of English might have carried over into the actual performance. However, performance-based assessment could reduce anxiety in some students, with one of Teacher 1's students remarking that "It is not a test (traditional paper-pencil test), it is just writing down what you have learned, so I did not prepare anything for that."

### *Feedback*

Finally, there was a discovery of students seeing feedback as a significant factor that affected their work. For example, the "attitude, feedback and responses" by Teacher 1 greatly affected the students' work as one of his students said that he felt invested in the course as he felt a similar investment from the teacher. He was attempting to mirror his instructor's positive outlook. Given the amount of assessing required and the feedback to compose, one teacher would select a video of a single student and comment on its positive aspects over Zoom. One student was concerned that the teacher often picked one student out of many and explained that student's work, so there were many students that he did not touch. “So, I think that one of the students who was picked up at that time understood what was wrong with him and how to fix it. I do not think the rest of them really understood what was wrong with their presentations." The students also expressed issues regarding the level of communication with the teacher and the amount and clarity of feedback they received. The timing of feedback, the specificity of feedback, the ability of students to understand the feedback, the opportunity to redo and improve the task after feedback, and the role of feedback in grading are all critical aspects of this theme.

### *Summary of Student Reactions*

In summary, student reactions were overwhelmingly positive. Students perceived the teacher’s role in a performance-based approach as a coach and a supporter that they could relate to and enjoy more than when the teacher played a role as a conveyor of information. The learning environment emphasized accomplishments that students “owned.” When students created their own performances, such as the creation of real dialogue autonomously in cooperation with peers, and used English as a medium of communication, the satisfaction gave intrinsic motivation. However, performance-based assessment

could reduce anxiety in some students, as they could control the content of the task and prepare themselves better than with a paper test requiring large amounts of memorized information. A negative washback was the anxiety of performance in front of peers, which if managed properly, was actually an indication that the student was being challenged to exceed their previous capacity and ability. Finally, students responded positively to feedback. The timing, specificity, comprehensibility, the opportunity to redo and improve the task after feedback, and the role of feedback in grading contributed to positive reactions from students.

## Discussion

Sasaki (2008) pointed out the significance of conducting studies on introducing and executing governmental policies related to language assessment and the significant impact that these policies have, particularly in Japan. According to the current policy of Japan's Ministry of Education, Culture, Sports, Science, and Technology (MEXT), starting from 2020, the English university entrance examinations will expand from two skills (listening and reading) to include four skills (speaking, writing, listening, and reading) (Bacquet, 2020). In other words, this policy change is an opportunity for writing or speaking skills to be valued and implemented in performance-based language assessment. This could create a washback effect and therefore be aligned with the Japanese government's education reform. This study, therefore, attempted to investigate teachers' experience implementing performance assessment in EFL classrooms in Japanese universities. In order to shed light on the classroom use of performance-based assessment in Japan, a washback approach was adopted.

According to Bailey's (1996) washback model, assessment and tests have an impact on participants, and the products. However, participants may also have an impact on the tests. The present study's findings indicate that the teachers' performance-based assessment in their classrooms influenced the participants (teachers and students), and the products (the ways teachers teach, and the materials they use.) In addition, the teachers' personal beliefs influenced the assessment they adopted.

Regarding washback on teachers, the findings concerning "how teachers teach" indicated that all three teachers employed rubrics

(assessment criteria or rating scales) in their assessment process, though each teacher approached the rubrics differently. In performance-based assessment, the rater needs to use a rubric in rating performance to arrive at a score for said performance. In marking any performance-based assessment tasks, whether in the classroom context or large-scale proficiency tests, there is a requirement for classroom markers/raters/teachers to make more complicated judgments than simple right-wrong decisions. Multiple-choice, true/false, error-recognition, and other item types where the candidate's responses can be marked as either "correct" or "incorrect" are examples of right-wrong decisions.

In this type of marking, sometimes referred to as subjective marking, Alderson, Clapham and Wall (1995) stress that the examiners' or teachers' job is to assess "how well a candidate completes a given task," for which they need a "rating scale" (pp. 106 - 107), also known as a rubric. However, at present, alternative approaches are discouraged, with many factors working against them. The Confucian philosophical background of the Japanese culture (Sasaki, 2008), which promotes traditional, paper-and-pencil testing, sees the teacher as a sovereign figure in the classroom, with the students but vessels for the teacher to fill. Students are experiencing little to no ownership of their learning in class sizes that often reach forty students. At the same time, many teachers are overworked and unable to commit to experimenting with new forms of assessment. Rubrics can reduce teachers' workloads without sacrificing the quality of feedback and create uniformity in assessment through different teachers at one institution (Allen & Tanner, 2017).

In the present study, students consistently expressed that they had improved or felt accomplished in the performance tasks. In constructing a rubric for tasks, Mertler (2001) identified opportunities for teachers to re-examine objectives for the task, characterizing and brainstorming how to demonstrate desired attributes in work. With rubrics constructed and definitions of success criteria accessible to all, students can create better output, as stated by Soles (2001, as cited in Turgut & Kayaoglu, 2015). Black and Wiliam (1998) support this idea proposing that knowing learning goals is fundamental to success in reaching learning objectives. As the assessments in the present study were mainly oral, rubrics could provide formative information for

students and teachers. Huang and Gui (2015) support that in spoken dialogue, the use of rubrics helps to increase discourse length, improve the level of organization, and vary the vocabulary used. Students can produce better work when rubrics provide defined performance targets and clearer paths to success.

Parallel to the student reactions, the positive washback found in the teachers was increased motivation to assess performances rather than to assess knowledge of the students. In contrast, the negative washback effects were the additional workload that teachers experienced. By preserving the rubrics and reusing materials developed for performance assessment, this time's investment is expected to decrease in subsequent iterations of the annual class cycles and eventually reduce overloaded assessment burdens.

Finally, this leads to the question of future research in the impact of spoken performance assessments on the teaching and learning. Nation-wide, researchers may investigate the distribution of impact of speaking assessments for both entrance testing and class testing. Designing entrance examinations to include positive washback is advantageous in that students can create more effective study practices. Small scale classroom studies to increase ways to teach self- and peer-assessment to students are also needed. A shift of assessment form top-down summative to self- and peer- formative is positive in regard to preparing students for assessment in later employment environments. However, it should be noted that in the analysis of the students' interviews, there could be errors due to the translation process. The interviews were conducted in Japanese and then translated to English for the analysis. Though native speakers of Japanese and English verified the translated documents, any translation could carry some error. Though three very different classes and three different assessment tasks, and different rubrics and styles of assessment could be another limitation of the study, it could yield different perspectives regarding washback effects.

## Conclusion and Implications

This qualitative study examined how performance-based language assessment impacted teachers, teaching, and students from three EFL classrooms in Hokkaido, Japan. The teachers' self-reports

revealed that Teachers 2 and 3 required short semi-formal academic type presentations, and Teacher 1 chose to use skits or roleplay types of performance. Teachers 1 and 3 seemed to be happy with their students' performance and would continue with their assessment style. Teacher 2, unhappy with the textbook, made some changes to emphasize his students' presentation performances. Teacher 2 would also like the students to take a more active role in assessing their own performances, and as audience members, he wanted those watching to be more ready with their questions.

Consequently, he introduced self-assessment rubrics and a question/answer exercise to speed up both partners' processes of asking and answering. Regarding feedback, it seems that a larger class size (as in Teacher 3's second class) limits the teacher's comments to a few quick, "that was great" or "try for better eye contact next time," and the like. For Teacher 2, the small class size allowed him to give detailed feedback, and in some cases, require a student to improve and re-record their presentation and get a new, presumably higher scored assessment. Finally, to prepare students for the final performance-based tasks, rubrics were presented by all teachers, and they devoted at least a week and a half to explaining and demonstrating, and allowing students to practice their performances.

In terms of students' interactions, the conclusion is that the positive impact derived from this study vastly outweighed the few negative impacts. Teachers provided students with a comfortable and challenging learning environment that fostered thinking. Also, expression in an English delivered performance-based assessment and using functional English in the planning and production phase of the assignments proved beneficial to the students. The peer work involved creating self-chosen topics, which led to self-developed English performances, feelings of accomplishment, and better retention of English used in the presentations. The use of technology presented challenges, and some students found the dynamic of a performance-based assessment compared to a traditional summative written exam a little out of their comfort zone. Skills learned outside English language usage were working cooperatively with other students, reflecting upon skill development, and dealing with anxiety. Therefore, teachers should adopt performance assessment and employ it for the majority of individual student evaluations. As for the negative washback, teachers

should be aware of and try to ward off unnecessary anxiety students might feel when doing the assessment, though some anxiety is needed for students to make an effort to prepare for the performance assessment.

Concerning the implications for teachers, each EFL teacher should consider the role of speaking and writing skills in the curriculum and the syllabus of the class they are teaching. When they create lessons, teachers could consider how performances are events requiring student demonstration of speaking and writing skills (i.e., writing scripts) and what types of assessment to which learners respond most positively. Teachers could also consider what conditions help students enjoy the challenge of assessment and where students feel confident because challenges were appropriate to their ability level.

In conclusion, by investigating three very different classes, with three different assessment tasks, that used different rubrics and styles of assessment, this study provides evidence that performance-based assessment has significant benefits to both students and teachers, and therefore questions the over-emphasized role that knowledge-based language testing is playing in the field of second language learning.

## Acknowledgements

This publication resulted in part from research supported by Faculty of Humanities, Chiang Mai University, Thailand, the Japan Association for Language Teaching (JALT) Hokkaido Chapter, and the School of Humanities, Sapporo Gakuin University, Japan.

## About the Authors

**Bordin Chinda:** An Assistant Professor at Chiang Mai University and a visiting professor at Sapporo Gakuin University (while conducting this study). His research interests include performance-based assessment, washback studies, and teacher education.

**Matthew Cotter:** A lecture at Hokusei Gakuen University Junior College. His research interests include indigenous cultures, intercultural communication, and assessment for learning.

**Matthew Ebrey:** A teacher at all levels of pre-tertiary education in Hokkaido and presently completing his Master of Education at Massey University, New Zealand.

**Don Hinkelman:** A professor at Sapporo Gakuin University, teaching intercultural communication skills. He has researched performance assessment and co-developed open-source video assessment modules for blended learning, emphasizing self and peer assessment.

**Peter Lambert:** A teacher at high schools in Sapporo. His research interests include cross-curricular studies, optimizing student goal setting, and teaching empathic negotiation strategies.

**Annie Miller:** A teacher at Hokusei Gakuen University. She has been teaching Oral English communication in universities in Hokkaido for the past twenty years.

## References

Alderson, J., & Banerjee, J. (2001). Language testing and assessment (Part I). *Language Teaching, 34*(4), 213-236. doi:10.1017/S0261444800014464

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing, 13*(3), 280–297. https://doi.org/10.1177/026553229601300304

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115-129. https://doi.org/10.1093/applin/14.2.115

Allen, D., & Tanner, K. (2017). Rubrics: Tools for making learning goals and evaluation criteria explicit for both teachers and learners. *Life Sciences Education*, *5*(3), 197-295. doi.org/10.1187/cbe.06-06-0168

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford University Press.

Bacquet, J. (2020). Implications of summative and formative assessment in Japan – A review of the current literature. *International Journal of Education and Literacy Studies, 8*(2), 28-35. http://dx.doi.org/10.7575/aiac.ijels.v.8n.2p.28

Bailey, K. M. (1996). Working for washback: A review of the washback. *Language Testing, 13*(3), 257-279. https://doi.org/10.1177/026553229601300303

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-74. https://doi.org/10.1080/0969595980050102

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly, 32*(4), 653-675. https://doi.org/10.2307/3587999

Cheng, L. (2005). *Changing language teaching through language testing: A washback study*. Cambridge University Press.

Corbin, J., & Strauss, A. (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (3rd ed.). Sage.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing* (Vol. 7). Cambridge University Press.

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000 - writing: Composition, community, and assessment*. Princeton, NJ: Educational Testing Service.

Huang, Y., & Gui, M. (2015). Articulating teachers' expectations afore: Impact of rubrics on Chinese EFL Learners' self-assessment and speaking ability. *Journal of Education and Training Studies, 3*(3), 126-132. https://eric.ed.gov/?id=EJ1060996

Lynch, B. K. (2001). The ethical potential of alternative language assessment. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Language testing in honour of Alan Davies (Vol. 11, pp. 228-239)*. Cambridge University Press.

McNamara, T. (1996). *Measuring second language performance*. Addison Wesley Longman Ltd.

McNamara, T. (2001). Rethinking alternative assessment. *Language Testing, 18*(4), 329–332. https://journals.sagepub.com/doi/pdf/10.1177/026553220101800401

Mertler, A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation, 7*(25), 1-9. https://doi.org/10.7275/gcy8-0w24

Sasaki, M. (2008). The 150-year history of English language assessment in Japanese education. *Language Testing, 25*(1), 63–83. https://doi.org/10.1177/0265532207083745

Shohamy, E. (2007). Language tests as policy tools. *Assessment in Education, 14*(1), 117-130. https://doi.org/10.1080/09695940701272948

Stobart, G. (2003). Editorial: The impact of assessment: Intended and unintended consequences. *Assessment in Education, 10*(2), 139-140. https://doi.org/10.1080/0969594032000121243

Tsagari D., & Cheng L. (2017) Washback, impact, and consequences revisited. In E. Shohamy E., I. Or, & S. May. (Eds.), *Language testing and assessment. Encyclopedia of language and education (3rd ed.)*. Springer.

Turgut, F., & Kayaoglu, N. M. (2015). Using rubrics as an instructional tool in EFL writing courses*. Journal of Language and Linguistic Studies, 11*(1), 47-58. https://dergipark.org.tr/en/pub/jlls/issue/36119/405585

Wall, D. (1997). Impact and washback in language testing. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education (1st ed., Vol. 7, pp. 291-302).* Kluwer Academic Publishers.

Wall, D. (2005). *The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation theory*. Cambridge University Press.

Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lanka impact study. *Language Testing, 10*(1), 41-69. https://doi.org/10.1177/026553229301000103

Wall, D., & Horák, T. (2006). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe. Phase I: The baseline study (TOEFL Monograph No. MS-34)*. Educational Testing Service.

Wall, D., & Horák, T. (2008). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe. Phase 2: Coping with change (TOEFLiBT-05)*. Educational Testing Service.

Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing, 13*(3), 318-333. https://doi.org/10.1177/026553229601300306

Wigglesworth, G. (2008). Task and performance-based assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., Vol. 7, pp. 111-122). Springer Science+Business Media.

## A: Guidelines for Teachers' Reflections

Please describe the following as detailed as possible.

1. The course under investigation:
    a. title
    b. main objectives
    c. number of students in the class
    d. the status of the course
    e. what do other teachers think about it?
    f. your impression about the course
2. How you generally conducted the class
3. How you would have conducted the class differently
4. All performance-based language assessment you used in your class this semester, especially the final project
5. How you will do the assessment for the course again
6. How you prepared the students for the final project
7. How you would have prepared the students differently

## B: Students' Interview Schedules (English version)

1. Can you describe or explain your typical class with the teacher (what usually happened in class)?
2. Do you like this class? Why? Why not?
3. How many times did your teacher test you last semester?
   - 3.1. Which one did you like the best and why?
   - 3.2. What one did you like the least (or hated) and why?
4. Can you describe the most recent performance test (for example, speaking and writing tests) your teacher used?
   - 4.1. How did you prepare for it?
   - 4.2. Did you ever feel anxious about it? Please explain.
   - 4.3. Are you satisfied with your performance?
   - 4.4. What did you like the most about it? Why? Anything you did not like? Why?
5. How did your teacher prepare you for the test?
6. What did you learn from taking that test? Please explain.
7. Do you think the assessment could reflect what you learned?
8. How well did you prepare for the assessment?
9. Do you have any suggestions for your teacher about the test and how he/she prepared you for it?