Koizumi Kaken 1st Research Meeting Jointly organized by Japan Language Testing Association (JLTA): 51st JLTA Research Meeting

November 22 (Sunday)

2020 (10:00-17:00, Japan Time)

Theme: Improving scoring methods in English speaking assessment (Part 1)

Supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant-in-Aid for Scientific Research (C), Grant Number 20K00894

Administrative information

- **During the lectures**
  - Please mute sound and turn off video camera.
- **Recording and slides**
  - All of the sessions will be recorded. We will make them accessible on YouTube (unlisted) to those who registered in advance. They will be available for one month. If you prefer not to appear in the video, turn your video function off.
  - Some slides will be available before or after the meeting on the JLTA website:
    - http://jlta2016.sakura.ne.jp/?page_id=21

Administrative information

- **Break**
  - – Feel free to take a break from time to time yourself.
- **Questions and comments**
  - – Please post your questions and comments below:
    - https://forms.gle/iaaH3T6GTTf1evhZ7
  - – You can post them in the chat box on Zoom, but priority is given to the website questions and comments.
- **Questionnaire**
  - – Please post your comments below:
    - https://forms.gle/8EtNzdtDoXdu6toy9

Kaken Project

Supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant-in-Aid for Scientific Research (C), Grant Number 20K00894

Project: Development of **scoring guidelines** for ensuring rater reliability in second language (L2) English classroom speaking assessment (SA) at senior high schools in Japan

Members

Rie Koizumi (Chair: Juntendo University), **Makoto Fukazawa** (University of the Ryukyus), **Yuichiro Yokouchi** (Hirosaki University), **Chihiro Inoue** (Centre for Research in English Language Learning and Assessment [CRELLA], University of Bedfordshire, U.K.)

Assumptions

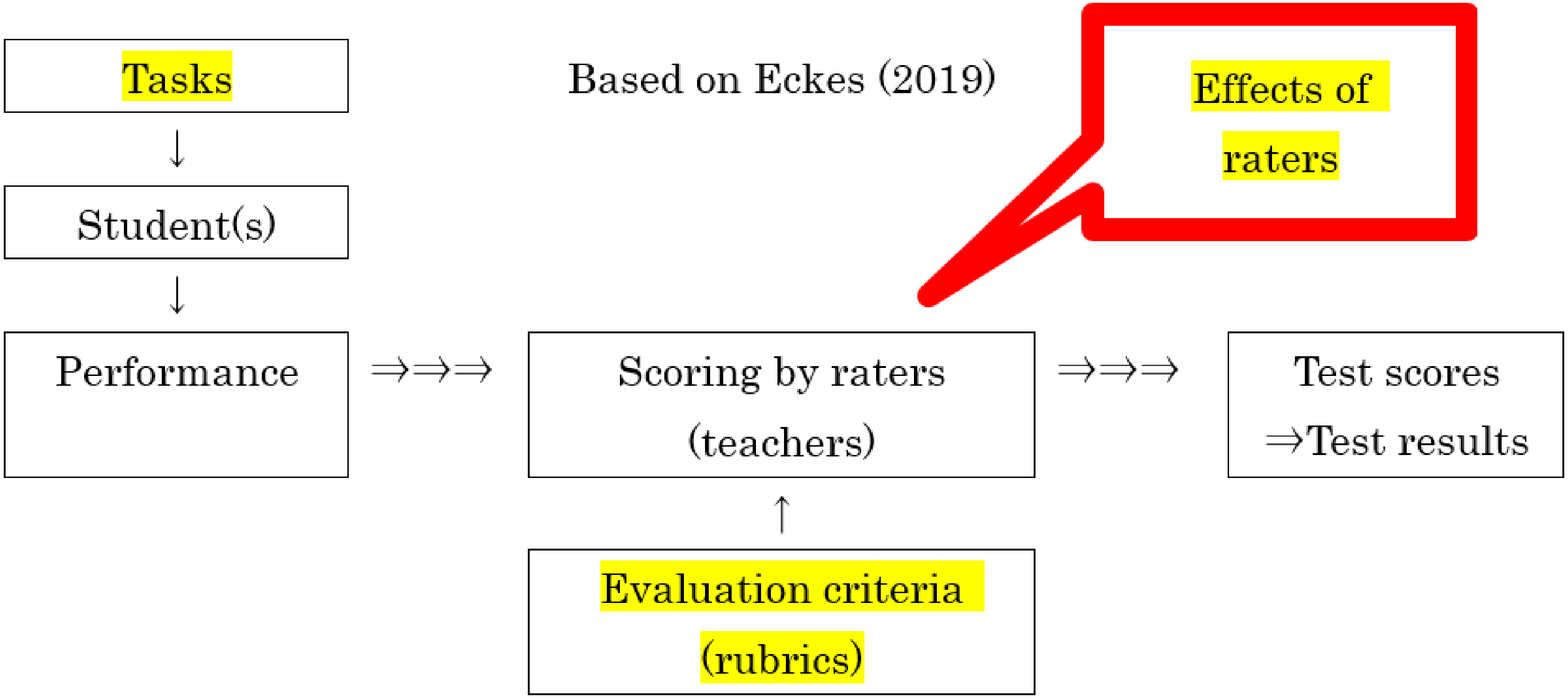
- Kaken project focuses on classroom SA involving teacher-raters.
- SA in classroom is an essential part of education (e.g., Poehner & Inbar-Lourie, 2020).
- Formative functions of L2 classroom assessment
  - -- Assessment for learning (AfL; Black & Wiliam, 2009), learning-oriented assessment (Turner & Purpura, 2016), and dynamic assessment (Leung, 2007)
- Summative use of L2 classroom assessment
- Classroom SA is not high-stakes.
- Typically used for formative and summative purposes, with more focus on the summative purpose (Bacquet, 2020).

Kaken Project

- **1. Learn assessment theories and practices in the world** regarding the improvement of scoring methods.
  - Current research meeting
  - Another meeting scheduled for February 14 (in Japanese)
- **2. Investigate the current practices in Japan**
  - SA is not conducted frequently in classroom in Japanese secondary school classrooms.
  - Tasks and rubrics are made at each school.
  - Rater training sessions are not typically conducted. Rater reliability is often not checked.
- 3. **Create** scoring guidelines and a website for ensuring rater reliability in senior high schools in Japan
- 4. **Examine** the effectiveness of the guidelines and the website

Assessment theories and practices in Japan and the world

- Aspects to consider when assessing L2 speaking



Tasks

↓

Student(s)

↓

Performance ⇒⇒⇒ Scoring by raters (teachers) ⇒⇒⇒ Test scores ⇒Test results

↑

Evaluation criteria (rubrics)

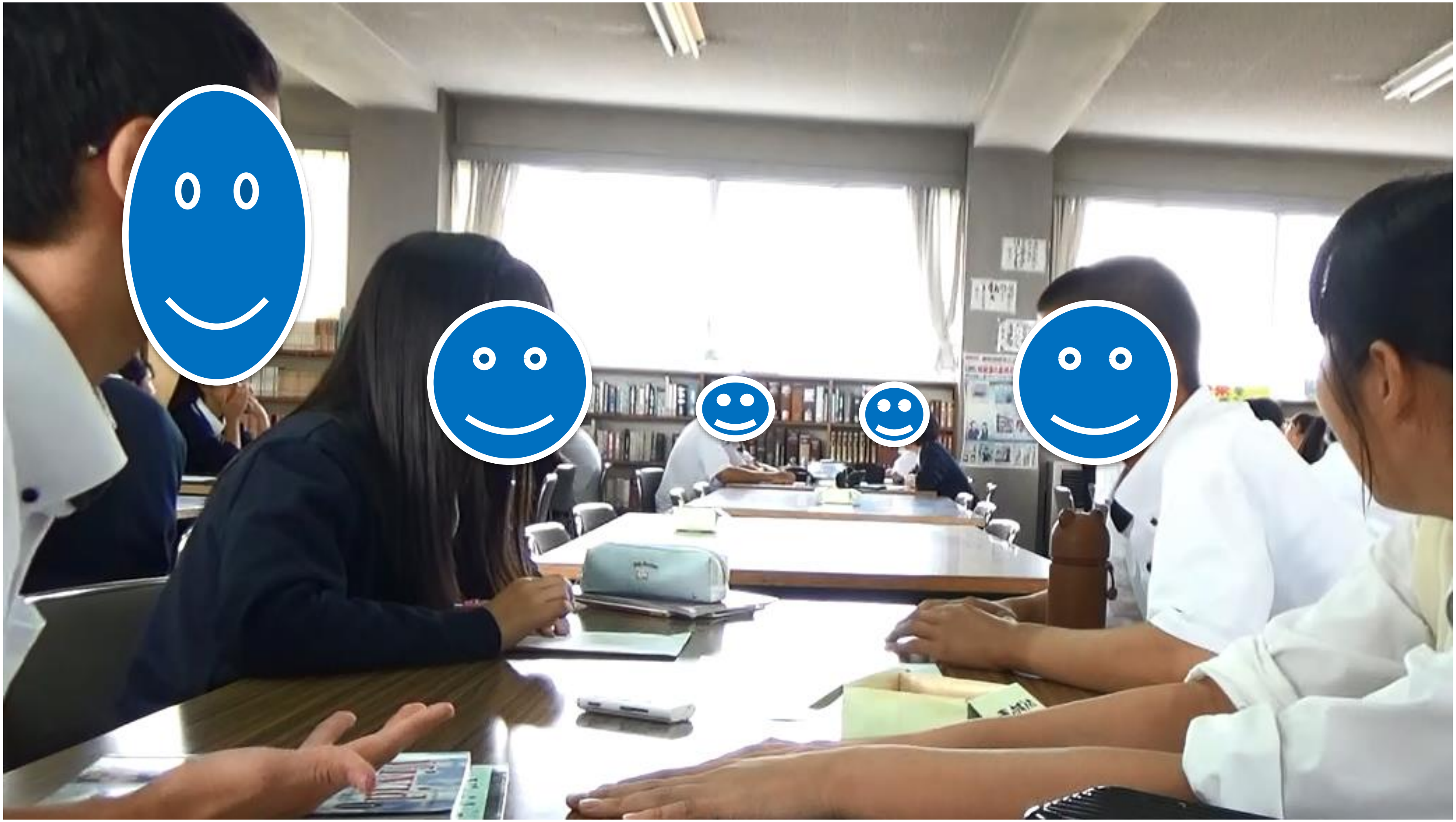
Based on Eckes (2019)

Effects of raters

Task:
Presentation & Interview

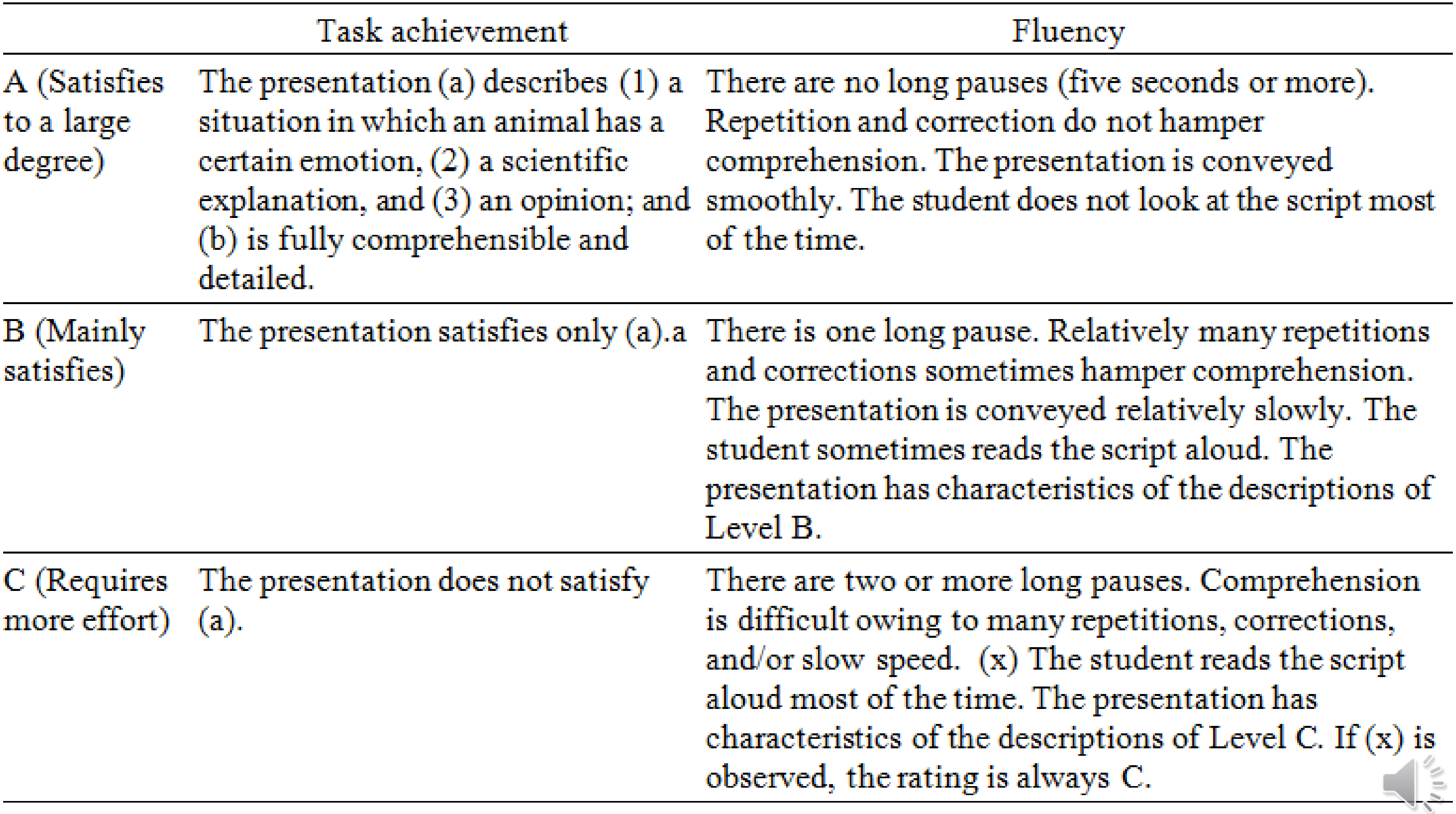
Pair talk

Group talk (e.g., by 4 students)



Rubric:

Analytic rubric (Koizumi & Yano, 2019)

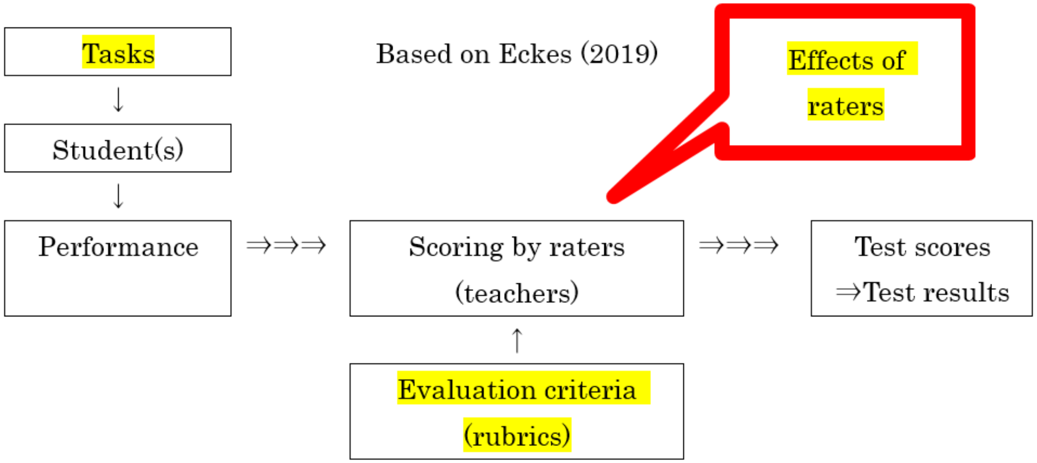


	Task achievement	Fluency
A (Satisfies to a large degree)	The presentation (a) describes (1) a situation in which an animal has a certain emotion, (2) a scientific explanation, and (3) an opinion; and (b) is fully comprehensible and detailed.	There are no long pauses (five seconds or more). Repetition and correction do not hamper comprehension. The presentation is conveyed smoothly. The student does not look at the script most of the time.
B (Mainly satisfies)	The presentation satisfies only (a).a	There is one long pause. Relatively many repetitions and corrections sometimes hamper comprehension. The presentation is conveyed relatively slowly. The student sometimes reads the script aloud. The presentation has characteristics of the descriptions of Level B.
C (Requires more effort)	The presentation does not satisfy (a).	There are two or more long pauses. Comprehension is difficult owing to many repetitions, corrections, and/or slow speed. (x) The student reads the script aloud most of the time. The presentation has characteristics of the descriptions of Level C. If (x) is observed, the rating is always C.

Tasks and rubrics (Muñoz & Álvarez, 2010)

- 1. Match **test tasks** with teaching objectives and tasks. Create authentic tasks that cover a wide range of communicative abilities.
- 2. Specify **instructions and procedures** to elicit targeted abilities.
- 3. Create **rubrics (evaluation criteria)** that clearly show the tested ability and are easy for students and teachers to understand.
- They should have positive effects on learning and teaching.
- Related to validity
  - To what degree is test interpretation and use appropriate?

Effects of raters



Based on Eckes (2019)

Tasks ↓ Student(s) ↓ Performance ⇒⇒⇒ Scoring by raters (teachers) ⇒⇒⇒ Test scores ⇒Test results

Evaluation criteria (rubrics) ↑

Effects of raters

- Rater reliability
- Includes **consistency across raters & within raters**
- Assumed, and not examined much in practices
- Teacher-raters typically do not have rater training.
- They score alone.
- Inconsistency among teachers may lead to a lack of score comparability.
- Classroom SA does not need to have a high reliability, but it should have a **moderate** reliability.

Typical rater training procedures

- Prior discussion
  - Check the rubric and performance examples. Use videos or recordings to mark separately. Discuss ratings to reduce discrepancies.
- During-test practice
  - Double rating (fully or partially).
- Post-test discussion
  - Discuss the discrepancy and identify reasons. Modify the rubric. Decide the final ratings.

Rater training procedures and practical constraints in Japan

- Prior discussion
  - Check the rubric and performance examples. Use videos or recordings to mark separately. Discuss ratings to reduce discrepancies.
  - **In Japan, teachers cannot usually hold a long training meeting.**

- During-test practice
  - Double rating (fully or partially).
  - **In Japan, a double rating may not be possible.**

- Post-test discussion
  - Discuss the discrepancy and identify reasons. Modify the rubric. Decide the final ratings.
  - **In Japan, this may not be possible.**

Koizumi & Watanabe (2020)

- Public high school SA: 116 students, 2 to 9 raters
- **To what extent can rater reliability be maintained using a simple rubric without detailed rater training?**
- Tasks: Individual presentation, paired role play, and two group discussions
- Rubric: 3 criteria with 3 levels No intensive rater training
- Data analyzed using many-facet Rasch measurement and generalizability theory
- In general, raters scored similarly and consistently.
  - Agreement: 49.5% to 80.7%
- The number of raters required to maintain sufficient reliability ($\Phi = .70$): one to four raters
- Group discussion tasks required more raters.

Koizumi, Hatsuzawa, Isobe, & Matsuoka (2020)

- Public high school SA: 232 students, 3 raters
- **To what extent can rater reliability be maintained using a simple rubric without detailed rater training?**
- Tasks: Group discussion and debate
- Rubric: 3 criteria with 3 levels
- No intensive rater training
- Data analyzed using many-facet Rasch measurement and generalizability theory
- In general, raters scored similarly and consistently.
  - Agreement: 72.9% to 81.6%
- The number of raters required to maintain sufficient reliability (Φ = .70): one rater

Rater training at Japanese schools

- With clear and simple rubrics, intensive training may not be necessary in some contexts.
- Still, during-test practice may be necessary.
  - Double rating (fully or partially).
  - Discuss the ratings for the first few performances and adjust the criteria during the discussion.

- More research is needed to identify contexts that require simple or intensive training.
- Effective practices outside Japan can provide further clues.

This research meeting

- 10:45-12:00 Scoring spoken performance in **large-scale** language testing programs **in China**
  - Lecturers: Jason Fan (University of Melbourne, Australia), Jin Yan (Shanghai Jiao Tong University, China)
- 13:00-14:15 Implementing and rating a new **peer-to-peer assessment** of speaking skills **in New Zealand**
  - Lecturer: Martin East (University of Auckland, New Zealand)
- 14:30-15:45 **Teacher 'rater' training in Hong Kong and Australia**: Different contexts, same problems?
  - Lecturer: Chris Davison (University of New South Wales, Australia)
- 16:00-17:00 Discussion (breakout session + Q&A)

References

- Bacquet, J. N. (2020), Implications of summative and formative assessment in Japan—A review of the current literature. *International Journal of Education & Literacy Studies*, *8*(2), 28–35. http://dx.doi.org/10.7575/aiac.ijels.v.8n.2p.28
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102
- Eckes, T. (2019). Many-facet Rasch measurement: Implications for rater-mediated language assessment. In V. Aryadoust & M. Raquel (Eds.), Quantitative data analysis for language assessment (Vol. I: Fundamental techniques; pp. 153–175). New York, NY: Routledge.
- Koizumi, R., Hatsuzawa, S., Isobe, R., & Matsuoka, K. (2020). *Rater reliability in classroom group discussion and debate in a Japanese senior high school*. Unpublished manuscript.

References

- Koizumi, R., & Watanabe, A. (2020). *Rater reliability in classroom speaking assessment in a Japanese senior high school*. Unpublished manuscript.
- Koizumi, R., & Yano, Y. (2019). Assessing students’ English presentation skills using a textbook–based task and rubric at a Japanese senior high school. SHIKEN, 23(1), 1–33. http://teval.jalt.org/node/87
- Leung, C. (2007). Dynamic assessment: Assessment for and as teaching? *Language Assessment Quarterly*, *4*(3), 257–278. https://doi.org/10.1080/15434300701481127
- Munóz, A. P., & Álvarez, M. E. (2010). Washback of an oral assessment system in the EFL classroom. *Language Testing, 27,* 33–49. doi: 10.1177/0265532209347148
- Turner, C. E., & Purpura, J. E. (2016). Learning–oriented assessment in second and foreign language classrooms. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 255–273). De Gruyter.