The Language Teacher

jalt journal

The research journal of the Japan Association for Language Teaching

Volume 46 • No. 1 • May 2024

JALT
全国語学教育学会
Japan Association for Language Teaching
¥1,900 ISSN 0287-2420

JALT Journal

Volume 46 • No. 1

May 2024

Contents

May 2024
Volume 46 • No. 1

Japan Association for Language Teaching

A Nonprofit Organization

The Japan Association for Language Teaching (JALT) is a nonprofit, professional organization dedicated to the improvement of language teaching and learning in Japan. It provides a forum for the exchange of new ideas and techniques and offers a means of keeping informed about developments in the rapidly changing field of second and foreign language education. Established in 1976, JALT serves an international membership of approximately 3,000 language teachers. JALT has 32 JALT chapters and 32 special interest groups (SIGs) and is a founder of PAC (Pan-Asian Consortium), which is an association of language teacher organizations in Pacific Asia. PAC holds annual regional conferences and exchanges information among its member organizations. JALT is the Japan affiliate of International TESOL (Teachers of English to Speakers of Other Languages) and a branch of IATEFL (International Association of Teachers of English as a Foreign Language). JALT is also affiliated with many other international and domestic organizations.

JALT publishes *JALT Journal*, a semiannual research journal; *The Language Teacher*, a bimonthly periodical containing articles, teaching activities, reviews, and announcements about professional concerns; and the annual *JALT Postconference Publication*.

The JALT International Conference on Language Teaching and Learning and Educational Materials Exposition attracts some 2,000 participants annually and offers more than 600 papers, workshops, colloquia, and poster sessions. Each JALT chapter holds local meetings. and JALT's SIGs provide information and newsletters on specific areas of interest. JALT also sponsors special events such as workshops and conferences on specific themes and awards annual grants for research projects related to language teaching and learning.

Membership is open to those interested in language education and includes copies of JALT publications, free or discounted admission to JALT-sponsored events, and optional membership in one chapter and one SIG. For an annual fee of ¥2,000 per SIG, JALT members can join as many additional SIGs as they desire. For information about JALT membership, contact the JALT Central Office or visit the JALT website.

JALT National Officers (2024)

President: . . . . . *Clare Kaneko*
Vice President: . . . . . *Kenn Gale*
Auditor: . . . . . *Robert Chartrand*
Director of Treasury: . . . . . *Michael Mielke*
Director of Records: . . . . . *Samantha Kawakami*
Director of Program: . . . . . *Chelanna White*
Director of Conference: . . . . . *Wayne Malcolm*
Director of Membership: . . . . . *Julie Kimura*
Director of Public Relations: . . . . . *William Pellowe*

Chapters

Akita, East Shikoku, Fukuoka, Gifu, Gunma, Hiroshima, Hokkaido, Ibaraki, Iwate-Aomori, Kitakyushu, Kobe, Kyoto, Matsuyama, Nagano, Nagoya, Nankyu, Nara, Niigata, Oita, Okayama, Okinawa, Osaka, Saitama, Sendai, Shizuoka, Tochigi, Tokyo, Tottori, Toyohashi, West Tokyo, Yamagata, Yokohama

Special Interest Groups

Accessibility in Language Learning; Art, Research, and Teaching; Bilingualism; Business Communication; CEFR and Language Portfolio; College and University Educators; Computer Assisted Language Learning; Critical Thinking; Extensive Reading; Gender Awareness in Language Education; Global Englishes; Global Issues in Language Education; Intercultural Communication in Language Education; Japanese as a Second Language; Learner Development; Lifelong Language Learning; Listening; Literature in Language Teaching; Materials Writers; Mind, Brain, and Education; Mixed, Augmented, and Virtual Realities; Other Language Educators; Performance in Education; Pragmatics; School Owners; Study Abroad; Task-Based Learning; Teacher Development; Teachers Helping Teachers; Teaching Younger Learners; Testing and Evaluation; Vocabulary

JALT Central Office

Level 20, Marunouchi Trust Tower–Main,
1-8-3 Marunouchi, Chiyoda-ku, Tokyo 100-0005 Japan
Tel.: (+81) 3-5288-5443; Email: jco@jalt.org;
Website: <https://jalt.org>

In This Issue

Articles

In the first full-length research article, **Thomas Stones** discusses the assessment of interactive speaking skills, specifically how rubrics used for that purpose can be considerably refined and improved through the use of many-facet Rasch measurement. The author also recommends strategies for creating rubrics of relevance across assessment contexts. The second article by **Clay Williams** and **Naeko Naganuma** looks at L2 vocabulary acquisition among young learners. Through analysis of empirical evidence gathered at seven elementary schools in northern Japan, they note heightened acquisition of L2 vocabulary units among learner-participants through the use of pictures. Their analysis further questions the applicability of the revised hierarchical model, which posits that beginning learners must rely on translation for L2 knowledge development. In the third article, **Yoko Suganuma Oi** examines the effects of student feedback, to include self-assessment and peer assessment, on students' writing and perceptions of the writing task. The research is based on survey data collected from nearly 300 students along with 14 follow-up interviews with both students and teachers. The study is of relevance to language teachers who are considering involving their students in the assessment process, while being concerned about the reliability and effectiveness of such assessments.

The *Expositions* article is by **Sowyma Vajjala**. In this article, Vajjala provides a broad and informative overview of Artificial Intelligence (AI) in the field of applied linguistics. Vajjala brings not only an accessible explanation of how generative AI works, but also shares how applied linguists and language teachers might utilize generative AI in their work, and discusses the moral and ethical implications of AI in applied linguistics. Also of note, is that the author equips the interested reader with a cornucopia of references to delve deeper into the discussion of generative AI.

Reviews

This issue contains the first three of *JALT Journal*'s new style of book reviews. Based on the exposition written by Melodie Cook in the November 2023 *JALT Journal* issue, we have expanded the scope and modified the style of our reviews. In this issue, we present three reviews following our new guidelines. **Julia Christmas** provides a review of Hiroe Kobayashi and Carol Rinnert's, *Developing Multilingual Writing—Agency, Audience, Identity.* Particularly notable are her comments on how novice and experienced readers may need to familiarize themselves with recent research concepts including multicompetence, complex systems theory, adaptive transfer, multilingual motivation, and translanguaging. In the second review, **Mary Hillis** and **Anastasia Khawaja** offer a uniquely structured discussion of *English Language Teacher Education in Changing Times: Perspectives, Strategies, and New Ways*

of Teaching and Learning, edited by Liz England, Lía D. Kamhi-Stein, and Georgios Kormpas. The book focuses on the impact of COVID-19 on teaching, and the review authors discuss their individual and shared perspectives on the concepts raised in the book, such as student wellbeing, teacher training, and teacher resilience, among others. The third review by **Mariana Oana Senda** and **Karmen Siew** looks at Naoki Fujimoto-Adamson's *Globalisation and its Effects on Team-Teaching*. Both reviewers are former team-teachers, and review the book from their unique perspectives. They highlight various points of interest, ranging from team-teaching around the world to the team-teaching situation in Japan. We hope that readers will enjoy these, and will be inspired to write book reviews for us in the future!

From the Editors

This *JALT Journal* issue hones in on a pivotal moment in language learning, emphasizing the transformative impact of generative AI in education. The past year has showcased AI's pervasive influence across educational disciplines, prompting a re-evaluation of the relationship between target knowledge and knower, and more broadly speaking, the production, consumption, learning, and assessment of knowledge. Navigating generative pre-trained transformers (GPT) discourse elicits a mix of fear and excitement, urging a nuanced understanding of how such AI might be applied in learning and research contexts. Discussions of GPT platforms in language learning and research go beyond generic calls for responsible usage; they demand active and ethical engagement by educators and learners with AI-related technologies, so as to fulfill the transformatory and emancipatory purposes of language education.

JALT Journal's commitment to rigorous research can be seen in how it facilitates critiques and discussions in applied linguistics. Amidst ethical considerations, it is crucial to remember that human control over AI, and the necessary role of human ethics in shaping AI's integration into education, are imperative. Offering an introductory view of AI in language education, this issue's *Expositions* article underscores the powerful constraining and enabling influences of generative AI. In the process, it highlights the need for more research on human agency and ethics in AI-influenced language education. Indeed, *JALT Journal* encourages readers to refine their knowledge of generative AI as a new and increasingly powerful technology, and in the process, reflect on the centrality of human agency in the larger project of maintaining the integrity and quality of language education.

— Dennis Koyama, Editor

— Jeremie Bouchard, Associate Editor

— Joe Geluso, Assistant Editor

Articles

Developing a Rubric for Interactional Competence Using Many-Facet Rasch Measurement

Thomas Stones
Kwansei Gakuin University

The teaching and assessment of interactive speaking skills is a key aim in many English-language programs, and the right assessment rubric is a key component of any effective course. There is no one-size-fits-all approach and rubrics need to be validated, feedback collected, and revisions made. This paper reports on a small-scale, exploratory study undertaken to develop and improve a rubric for assessing interactive discussion skills. Utilizing many-facet Rasch measurement (MFRM), the paper reports on an analysis of the rubric originally used on the course. Based on these findings, the rubric was revised and subject to a second round of analysis. The findings indicate that the revisions led to considerably improved rubric function, due primarily to a reduced number of scale points and more clearly defined rubric categories. Finally, the paper suggests a number of recommendations for rubric creation that can be applied to a range of assessment contexts.

概要インタラクティブなスピーキングスキルの指導と評価は、多くの英語学習プログラムにおいて重要な目標であり、適切な評価ルーブリックは効果的な授業の重要な要素である。万能なアプローチは存在しない為、フィードバックを収集し、ルーブリックは検証、改善される必要がある。本稿では、対話型ディスカッションのスキルを評価するためのルーブリックを開発・改善するために行われた、小規模で探索的な研究について考察する。多相ラッシュ測定（MFRM）を活用し、当初コースで使用されていたルーブリックの分析について報告する。その結果に基づき、ルーブリックが改訂され2回目の分析が行われたところ、ルーブリックの機能に大幅な改善が見られた。その主な理由は、尺度点数を減らし、ルーブリックのカテゴリーをより明確に定義したことに

https://doi.org/10.37546/JALTJJ46.1-1

JALT Journal, Vol. 46, No. 1, May 2024

あることがわかった。最後に、本論文は様々な評価の文脈に適用可能なルーブリック作成に関する多くの提言を提示している。

Keywords: interactional competence; many-facet Rasch measurement; rubric development; speaking assessment; test validation

The effective development of speaking skills has become a core component of many English-language programs, and the range of approaches to assessing speaking has grown concurrently. Speaking assessments themselves can range in orientation from individual, paired and group and include tasks that test global speaking proficiency, interaction in specific scenarios or the use of pre-defined language forms or discourse functions (Luoma, 2004). In most cases, the speaking test represents a performance assessment (Johnson et al., 2009) where learners demonstrate the spoken language skills in authentic or semi-authentic ways. Central to the assessment of speaking is the use of a rubric to base judgements on participant performance. Any rubric should reflect the performances of the test participants and ability with the target skill as accurately as possible (Green, 2013). However, in many contexts various pressures mean that it is not often possible to investigate the effective functioning of rubrics or scoring systems, with teachers frequently left to trust in their own professional judgement as to how well these documents are functioning. Indeed, Janssen et al., (2015) note that in-depth studies of rubric development are relatively few and far between. Thus, in small part, this study aims to address this by exploring the validity of a rubric for a paired speaking test, primarily using many-facet Rasch measurement (MFRM).

Paired Speaking Tests and Interactional Competence

In recent years, the use of paired or group discussion tasks in universities has increased (e.g., Bonk & Ockey, 2003; Leaper & Brawn, 2019; Nitta & Nakatsuhara, 2014), as well as in a number of the higher-level Cambridge assessments (e.g., Cambridge 2008). The use of group discussions necessarily requires the incorporation of rubric categories that deal with interactional competence, which is the ability to effectively co-construct an interaction with an interlocutor within a specific context (Kramsch, 1986). High-profile examples are the interaction component in the CEFR (Council of Europe, 2001), which formed the basis of the similarly named ‘interactive communication’ category in the Cambridge exams (Galaczi et al., 2011). The inclusion of interactive competence measures is necessary to represent the

co-constructed nature of dialogic speech in group speaking tasks (Nitta & Nakatsuhara, 2014). Indeed, its inclusion in speaking tests has also been persuasively argued for by Roever and Kasper (2018) who note that it can lead to a far richer range of information on participants' ability to engage in the type of interactive, group-based talk that is common to many academic and professional contexts than potentially unstable predictions from assessments focused primarily on monologic production. Galaczi (2014) highlights several key areas of interaction that are central to the maintenance of an effective, co-constructed discussion which are topic development, listener support & turn-taking management, with greater topic development across turns and speakers particularly noticeable at higher CEFR levels. Similarly, Leaper and Brawn (2019) analysed the progression of learners over a two-year period focusing on four main areas of interaction: initiating, responding, developing and collaborating. Therefore, including interactional components into a rubric can help educators promote essential interactive skills for use beyond the classroom, as well as provide valuable reference information that enhances the validity and transparency of the awarding of scores (Jeong, 2015).

Many-Facet Rasch Measurement & Rating Scales

Once a suitable rubric has been created it is important to investigate how well it is functioning, which is where MFRM is of great use. MFRM is a statistical technique devised by Linacre (1994) that provides a 'rich set of highly efficient tools' (Eckes, 2015, p. 19) to examine how various facets of assessments interact to contribute to the assignment of scores. MFRM, therefore, can contribute to test development and administration due to the range of facets that can be compared and analysed (McNamara, 1996) and can inform revisions to rating scales for more meaningful and accurate scoring (Bond & Fox, 2015). MFRM has been gradually adopted in a variety of fields but has also become increasingly influential within applied linguistics and language teaching over the last 20 – 30 years (McNamara & Knoch, 2012). It has featured in a range of journals and has been used to investigate a variety of assessment types (Aryadoust et al., 2021) and was in fact foundational in the formation of the 6 CEFR levels (Council of Europe, 2001). More specifically, MFRM can be used to detect a variety of rater effects, such as leniency/severity, central tendency, randomness, halo, and restriction of range (Myford & Wolfe, 2003) as well as other demographic factors including gender, age, or attractiveness (Murphy & DeShon, 2000), format of test delivery (Nakatsuhara et al., 2020) or difficulty of assessment topics (Engel-

hard, 1992). In terms of research on scale function, Chen and Liu (2016) found that a 5-point rather than a 10-point scale functioned more effectively when evaluating written discourse completion for an email task. Janssen et al. (2015) similarly found that reducing the number of points on scales of a variety of sizes, some up to 20 points per rubric section, led to far more reliable scoring. Further, McDonald (2018) was able to considerably improve the functioning of a 9-point rubric to assess speaking skills by adopting a 5-point scale. Bonk and Ockey (2003) also utilised MFRM to explore the functioning of their group oral assessment in a Japanese University. They found that raters varied considerably in terms of the severity of scoring by as many as 2 points on a 9-point scale, despite training and practice on rubric use. Thus, MFRM is a highly flexible tool that can bring focus to areas of rubric and assessment performance that are difficult to obtain through other methods.

Rubrics & Raters

In addition to the rubric, raters can also introduce a large amount of unwanted variability to any score. The assessment of any spoken performance necessitates a subjective judgement on the part of the rater (McNamara, 1996), and human raters are inevitably fallible and may imperfectly represent any given performance (Eckes, 2015). This can add levels of construct-irrelevant variance, known generally as 'rater effects', which are a consequence of rater and not candidate performance (Scullen et al., 2000). There are myriad ways in which raters can differ in their application and interpretation of scoring rubrics and learner performances as well as potentially exhibiting other biases based on length of experience, pedagogical preferences, and educational background (Eckes, 2015). Furthermore, teachers can incorporate external additional factors when assigning grades, such as effort and behaviour throughout the course (Randall & Engelhard, 2009), or including factors such as body language and gaze despite them not being part of the scale (Orr, 2002). Rater training does help improve accuracy and eliminate extreme scoring phenomenon (Davis, 2016; Yan & Chuang, 2022), but despite even substantial and sustained attempts at rater training, some errors and biases can persist (McNamara, 1996; Myford & Wolfe, 2003). Therefore, it is essential to take remedial action where appropriate, as such running a MFRM analysis and providing the results to the teachers themselves (Myford & Wolfe, 2003). Validation of rater performance is also central to rubric development as it can lead to unreliable scoring or can indicate that rubric wording is not providing sufficient clarity to raters.

Research Context

This research took place at a Japanese university in the Kansai region. The speaking test used is part of the seminar-skills course, which is, in turn, part of a two-year (four semester), compulsory English-language program aimed at developing basic EAP skills. Student proficiency levels vary widely from A2 up to B2 and in some rarer cases C1. This seminar-skills course is the level 3 (of 6) course. The speaking test is a 5-minute paired discussion on a topic that participants had been studying for the previous 3 weeks. The weeks prior to the tests also introduced various discussion skills to be used in the test. As such, the test represented a summative assessment of content covered. At this level, the discussions skills are introductory and are closer to more general interactive competence than a full academic discussion. The original rubric included three categories: Discussion Skills, Discussion Questions, and Delivery and Effectiveness and could be described as a hybrid checklist-rating scale model whereby two of the rubric sections specify language to be used and checked off, but those sections were scored on a scale. The third section is a more typical rating scale with only the relevant constructs listed within it. The three rubric categories have scales of 10, 20 and 20 points respectively (see Appendix A). The scoring system was initially this single sheet but based on feedback received that it was difficult to discriminate between points on this scale, a more detailed explanation of the different bands was added to the primary rubric (Appendix B). Thus, the teachers had a 'live' rubric (Appendix A) for use while scoring participants as well as a 'detailed rubric' (Appendix B) that was also used as a reference to offer more guidance on the requirements for each category. In addition, the university-wide scoring policy sets 60% as a pass and states that the average score should be around 70 – 75%. The rubric was pre-existing within the program and had undergone various minor adjustments over a number of years and frequently received a range of feedback, from very positive to very negative. This feedback acted as the trigger for this research which intends to more deeply explore where the strengths and weaknesses in the rubric lie and find ways to improve it with the use of MFRM.

Research Aims

This research aims to explore the effective functioning of a rubric used for a paired-speaking test on a discussion skills course at a Japanese university. The aims of the research are as follows:

1. Does a many-facet Rasch measurement analysis reveal any issues in rubric functioning in the original rubric?

2. Based on question 1, what revisions should be made to the rubric?
3. Does a many-facet Rasch measurement analysis reveal any improvements in rubric functioning in the revised rubric?

Methods

For the examination of the original rubric, three speaking tests consisting of two participants each (total 6 participants) were video recorded and graded by 11 raters. The discussion topic for this test was 'Accommodation Options for University Students' and the problems and benefits associated with the various options. The participants were drawn from classes that covered the range of levels represented in the course with participants from the lowest, highest classes and mid-level classes selected. Students were originally assigned to classes based on TOEFL ITP scores taken at the program entry point. All raters were teachers that have had some experience on the course, work at the featured institution and held at least master's level-qualifications in TESOL or a related field and/or TESOL teaching certificates. All raters rated all speaking tests, meaning that the data was fully crossed. After teachers graded the speaking tests, they were asked to respond to a short questionnaire that focused on the validity and usability of the rubrics.

Thus, this is a small-scale, exploratory study taking an investigative approach to instrument development, looking to explore the functioning of the rubric and the teachers views thereof without aiming to prove or disprove a predetermined hypothesis (Singh, 2007). It is part of a broader study in which the primary area of data-collection is quantitative, with qualitative data used to supplement the quantitative findings (Morgan, 1998). The qualitative data would serve to add completeness to the picture gained from the Rasch analysis as well as provide methodological triangulation and facilitate instrument development (Bryman, 2006). This aims to align with the view of teachers and assessors as a community of professional practitioners who have the responsibility to uphold standards and contribute to a dialogue of continual, iterative improvement (Fulcher & Davidson, 2007) that should be mutually developed by key stakeholders in the local context (Hamp-Lyons, 1991; Ockey et al., 2013). This mixed methods approach is also in line with the Common European Framework (2001) recommendations on rubric development which suggests the use of intuitive, qualitative, and quantitative methods, where intuitive and qualitative elements can include informed, experience-based contributions with opportunities for feedback and review.

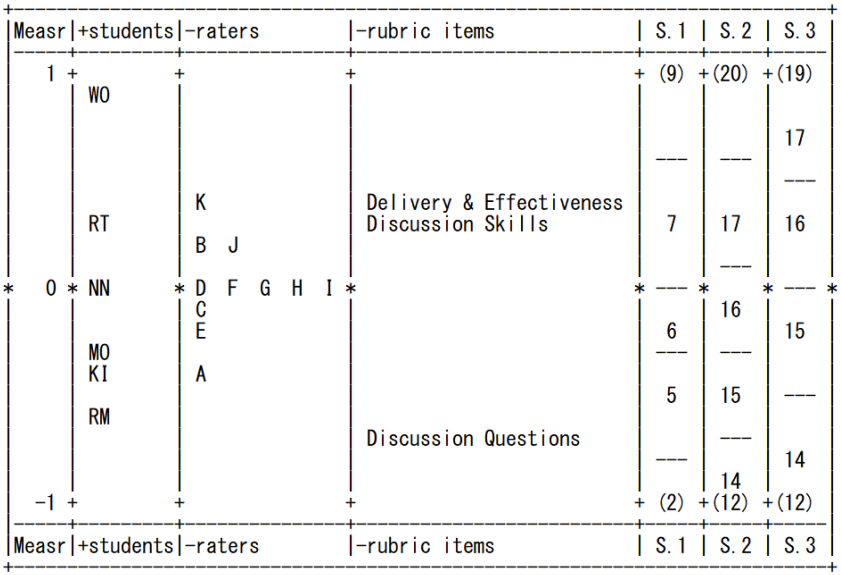
However, although the data collected from the questionnaire provided valuable insights, space restrictions preclude a detailed analysis of the responses. Also, despite the small sample size, it should be noted that MFRM does not necessarily need a large data set if the data fits the Rasch model. Indeed, Linacre's foundational work on MFRM (1994) reanalysed Guilford's (1954) data that featured only three raters and seven participants.

Findings from the Many-Facet Rasch Measurement Analysis

The raters' scores were input to the *Facets* (Linacre**,** 2001) program for performing MFRM. The Partial Credit Model was used to better compare the functioning of the three rubric categories. The Partial Credit Model analyses the rubric categories separately, so allows for more precision compared to analysing the rubric as a whole (Bond & Fox, 2015). This is also helpful where the rubric categories have different length scales as is the case here.

Figure 1 shows the Wright Map for the MFRM analysis. The Wright Map displays all facets ordered along a vertical logit scale (the leftmost column). The first column, students, orders the participants' ability from higher ability at the top to lower ability at the bottom. The next column gives the rater severity/leniency information with more severe raters placed at the top. 'Rubric items' orders the difficulty of the three rubric categories from easier at the top to harder at the bottom. The final three columns compare the relative difficulty of individual scale points. An initial analysis yields several interesting findings. Firstly, there is a narrow range of ability indicated on the logit scale, covering only 1.53 logits. The Rubric Items column shows the Discussion Questions section is the hardest, with the remaining two categories, Discussion Skills and Delivery and Effectiveness, exhibiting similar levels of challenge to each other. Rubric categories varying in difficulty is not necessarily an indicator of a misfunctioning rubric and can be desirable as different subskills may pose differing levels of challenge. In fact, the relative difficulty of the Questions section is likely a factor of poor assessment design. To score well on this section, most of the pre-determined discussion questions need to be asked. Within a short discussion, it is virtually impossible for both participants to ask all 4 questions, especially as some become redundant after one person has used them. Furthermore, the three columns on the far right raise some concern as they are considerably misaligned. Ideally, a score of 16, for example, on one component should align with a score of 16 on another if they are of similar difficulty. However, there is considerable misalignment, which will be further explored below.

Figure 1
Wright Map for Student, Rater & Rubric items



Statistical Findings from the Rasch Analysis

Tables 1–3 provide more detailed statistical information on rubric function. Examination of the student facet in Table 1 shows a separation index of 3.77 with a high reliability coefficient (0.93). This figure indicates how many different levels of ability were found among the learners, and a figure of 3.77 suggests that there were just under 4 ability levels across the cohort. The rater facet had a separation index of 0.76 and a reliability coefficient of 0.36. A separation index closer to 1.00 is ideal as it indicates that the raters are 'functioning as one' and producing similar scores for similar performances (Eckes, 2015). A higher separation index would indicate large differences in scores awarded for the same performance and would thus be undesirable. A reliability coefficient for raters closer to 0 rather than 1 is preferable (Eckes, 2015), so 0.36 is relatively good. These are encouraging findings in terms of rater reliability and suggest that a high level of consistency among raters. However, this could be partly due to the university's grading policy, which can have a narrowing effect on scoring. The rubric facet has a separation statistic of 5.19 with a high reliability coefficient (0.96). This is somewhat

problematic as it shows the rubric can only distinguish 5 ability levels. In and of itself, that is not a problem, but given that two of the category scales have 20 points available, it implies that only 25% of the scale points are being used, leading to a significant amount of redundancy.

Table 1

Separation Statistics for the Three Facets

	Root-mean Square Error	Separation Index	Reliability Coefficient	x^2
Student facet	0.13	3.77	0.93	0.00
Rater facet	0.18	0.76	0.36	0.07
Rubric items	0.09	5.19	0.96	0.00

Table 2

Measures and Fit Statistics for Raters and Rubric Categories

	Measure	SE	Infit MNSQ	Outfit MNSQ
Rater Facet				
A	-0.44	0.19	1.11	1.00
B	0.20	0.17	1.34	1.28
C	-0.08	0.18	0.93	0.98
D	-0.05	0.18	1.42	1.45
E	-0.24	0.18	1.28	1.20
F	-0.01	0.18	1.12	0.97
G	-0.05	0.18	0.56	0.57
H	0.05	0.18	0.32	0.32
I	-0.05	0.18	1.40	1.39
J	0.23	0.17	0.70	0.86
K	0.43	0.17	0.72	0.71
Rubric Facet				
Discussion Questions	-0.70	0.10	1.12	1.06
Discussion Skills	0.29	0.08	1.02	1.04
Delivery and Effectiveness	0.41	0.10	0.82	0.82

Table 2 gives more details on the raters' performances, with Infit MNSQ statistics largely falling within acceptable ranges. Infit MNSQ square statistics detail the extent to which the data matches the Rasch model and can serve to highlight various phenomenon among individual raters, such as erratic or conservative scoring (Myford & Wolfe, 2003). The acceptable range for this statistic varies depending on the purposes of the instrument with tighter ranges, for example between 0.8 and 1.2, preferred for higher-stakes situations and 0.7 – 1.3 for 'run of the mill' situations (Wright & Linacre, 1994), although often 0.5 – 1.5 is used, especially as small samples can widen the range of fit statistics (Wu & Adams, 2013). The Infit MNSQ statistics of the data analysed for this study generally fall between 0.7 – 1.3, suggesting good model fit and no erratic scoring, with no raters above 1.5. Two raters fell below the lower threshold, with Rater H at 0.32 and Rater G at 0.56. These indicate 'overfit' meaning that the raters more conservatively stuck to a narrow range of scores.

Overall, these figures would generally suggest fairly good rubric functioning and good model fit, with raters scoring in a fairly consistent manner. However, the narrow range of difficulties the rubric can discriminate and the misalignment of the scoring thresholds warrant further investigation.

Table 3

Rubric Functioning

Scale Point	Number of Observations	Av. Measure	Outfit MNSQ	Rasch-Andrich Threshold	Standard Error
Discussion Questions					
2	1	0.14	1.0		
3	0	N/A	N/A	N/A	N/A
4	2	0.16	0.8	-0.59	1.02
5	5	0.22	1.0	-0.73*	0.61
6	4	0.49	1.5	0.52	0.41
7	16	0.62	1.2	-0.95*	0.35
8	25	0.68	1.2	0.19*	0.27
9	13	1.06	1.0	1.56	0.33

Scale Point	Number of Observations	Av. Measure	Outfit MNSQ	Rasch-Andrich Threshold	Standard Error
Discussion Skills					
12	5	-0.76	1.0		
13	8	-0.83*	0.6	-1.24	0.49
14	8	-0.41	1.9	-0.67	0.34
15	9	-0.62*	1.4	-0.67	0.31
16	20	-0.28	1.1	-1.18*	0.29
17	2	0.22	0.5	2.13	0.35
18	12	0.17*	1.0	-1.71*	0.36
19	1	0.42	0.9	2.82	0.75
20	1	0.68	0.9	0.53*	1.04
Delivery and Effectiveness					
12	6	-0.87	1.0		
13	7	-0.98*	0.5	-0.99	0.45
14	23	-0.61	0.9	-1.89*	0.33
15	10	-0.50	1.2	0.3	0.29
16	11	-0.11	0.7	-0.40*	0.32
17	4	0.27	0.7	0.97	0.42
18	4	0.38	0.8	0.21*	0.52
19	1	0.56	0.8	1.79	1.03

* indicates where scale points do not advance in a linear fashion.

A Closer Look at Rubric Functioning

It is in the analysis of how individual points on the rating scales were awarded that the issues implied from the misalignment in the Wright Map and narrow separation index of 5.19 are fully explained. We can see that for all rubric categories a profound clustering of scores around particular scale points occurred. For example, the Number of Observations column shows that the majority of scores for Discussion Questions fall at scale points 7, 8, and 9 and at 14, 15, and 16 for Delivery and Effectiveness. Discussion Skills showed a greater spread, but a number of scale points were seldom selected,

most notably 17, 19, and 20. Furthermore, even though Discussion Skills and Delivery and Effectiveness are 20-point scales, less than half of these scale-points were actually used, with no scores awarded below 12 for either Discussion Skills or Delivery and Effectiveness. Even when considering that the university's scoring policy, and the small sample are likely having a narrowing effect, these results still suggest there are more scale points than there are levels of ability. Linacre's (2002) recommendation is that at least 10 observations per scale-point is needed for reliable analysis. With a total of only 66 ratings collected for this study (from 11 raters scoring 6 students across each of the 3 criteria), this is mathematically impossible with a 20-point scale, but the trends that are visible here are likely to be repeated with a greater number of raters and test takers, although a larger sample would be needed to confirm this.

Further issues can be seen in the average measures and Rasch-Andrich Threshold scores. In both cases, these should increase in line with the increase in the rating-scale points to suggest that a higher score on the rubric represents a higher-level of ability on the latent variable. Disordered categories, where the average measure and Rasch-Andrich threshold for a higher scale-point are below that of a lower scale point, reveal instances when the thresholds do not advance in a step-by-step manner and indicate that the rubric scale points are overlapping in the minds of the raters and, therefore, do not represent a distinct level of ability on the latent variable (Linacre, 2020). These points are marked with an asterisk in Table 3. The recommended distance between scale points is 1.4 – 5 logits (Linacre, 2002). Again, with a total range of 1.5 logits, this is clearly impossible in this data set and is a function of the extremely narrow range of scores awarded relative to the far wider span of the rubric. Another problem with some of the Rasch-Andrich thresholds is the large standard error figures associated with some of them. This is caused by the very low number of observations for several scale points, thus reducing their precision.

A visual representation of the trends indicated in Table 3 can be seen in Figures 2 – 4. These graphs display the probability of a particular score on the scale being awarded as difficulty increases. With a well-functioning rubric, the graph should appear as several distinct curves, similar in appearance to bell-curves, with the peak of each clearly separate from its neighbour, thus indicating that at each point on the latent variable, that score is the most likely. No lines should be subsumed by others, and curves should cross around their mid-points. Figures 2, 3, and 4 are obviously some distance from such a pattern.

Figure 2

Category Probability Scores: Discussion Questions

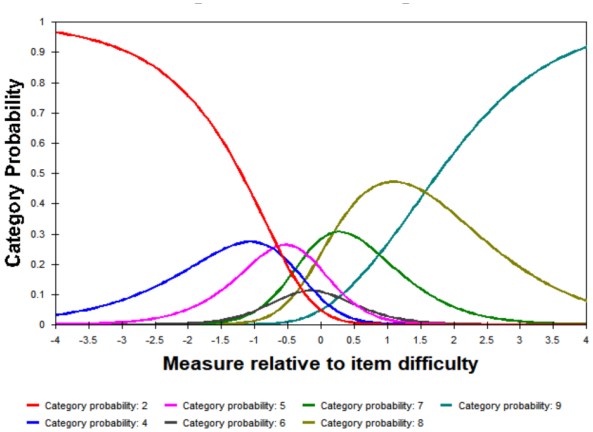


Figure 3

Category Probability Scores: Discussion Questions

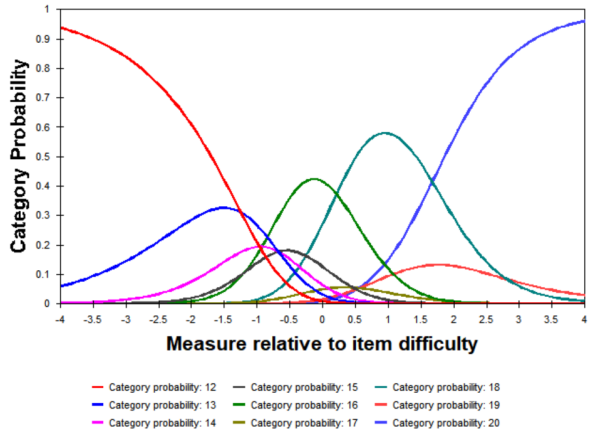
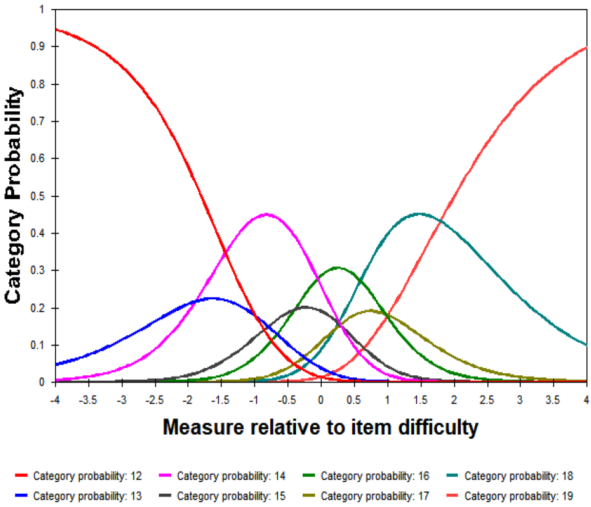


Figure 4

Category Probability Scores: Delivery and Effectiveness



Each of the three graphs show an amount of chaos in their alignment, with very few peaks distinct from the next and with a number subsumed by others. This suggests that raters do not have a clear idea of what level of performance is reflected by each point on the scale and indicates inconsistency in how points are awarded. The typical recommendation in such cases is to collapse the scale-points (Bond & Fox 2015; Eckes, 2015; Linacre, 2002).

Rubric Development Process

In light of the Rasch findings and feedback from teachers, a rubric revision process was undertaken that involved extensive discussions and multiple stages of drafting and redrafting. The major changes are summarized below.

Reduced number of scale points. In line with other studies where fewer scale-points improved functioning (Bonk & Ockey, 2003; Janssen et al., 2015; McDonald, 2018) recommendations for interpreting the output of an analysis using MFRM (Bond & Fox, 2015; Eckes, 2015; Linacre, 2002) and the results of the statistical analysis that the rubric distinguishes five levels of ability, the number of rubric categories was reduced. The revised scale goes from 1 – 5, with half scores at 3.5 and 4.5, so ultimately contains seven points. This is also the suggested maximum of seven that human raters can

deal with in short-term memory (Miller, 1956). As there were virtually no failing scores in the analysis of the original rubric, only two were awarded for Discussion Questions, it was thought that there needed to be some options for poor performances not fully represented in the original sample. Likewise, in creating the descriptors, it was felt that teachers would want more than three options for passing scores, especially as a maximum score is rarely awarded.

Move to a more general assessment of interactional goals. The original rubrics required learners to produce specific language, but this created several issues, so the descriptors will focus on interaction in general, rather than specific phrase production. Specific language should be taught in the course, but not mandated to be used within the rubric itself. Wiliam (2011) suggests including course language in the rubric itself to provide a connection to the course content, but its use should be subordinate to the achievement of interactional goals and avoid construct reductionalism (Green, 2013). Therefore, the Discussion Questions and Skills will be merged into a general 'Interaction' category, with the descriptors drawn from the interactional competence rubric developed by May et al. (2020) and the findings of Galaczi (2014).

Separate and reduce the constructs in the Delivery and Effectiveness category. This section was divided into two categories: Fluency and Language Use. The fluency category is based on that used by Nitta & Nakatsuhara, (2014), Iwashita et al. (2001), and later incorporated by McDonald (2018), as well as the criteria for the IELTS Speaking Test (IELTS, n.d.). Similarly, the Language Use category aims to incorporate the constructs of complexity and fluency and drew heavily on the IELTS criteria (IELTS, n.d.). This replaced the 'unit language' section as it was felt that the load placed on raters to reliably track the usage of 15 or so words that were included in each unit added to the already heavy cognitive burden that is often characteristic of scoring a performance with multiple traits (Hamp-Lyons & Henning, 1991).

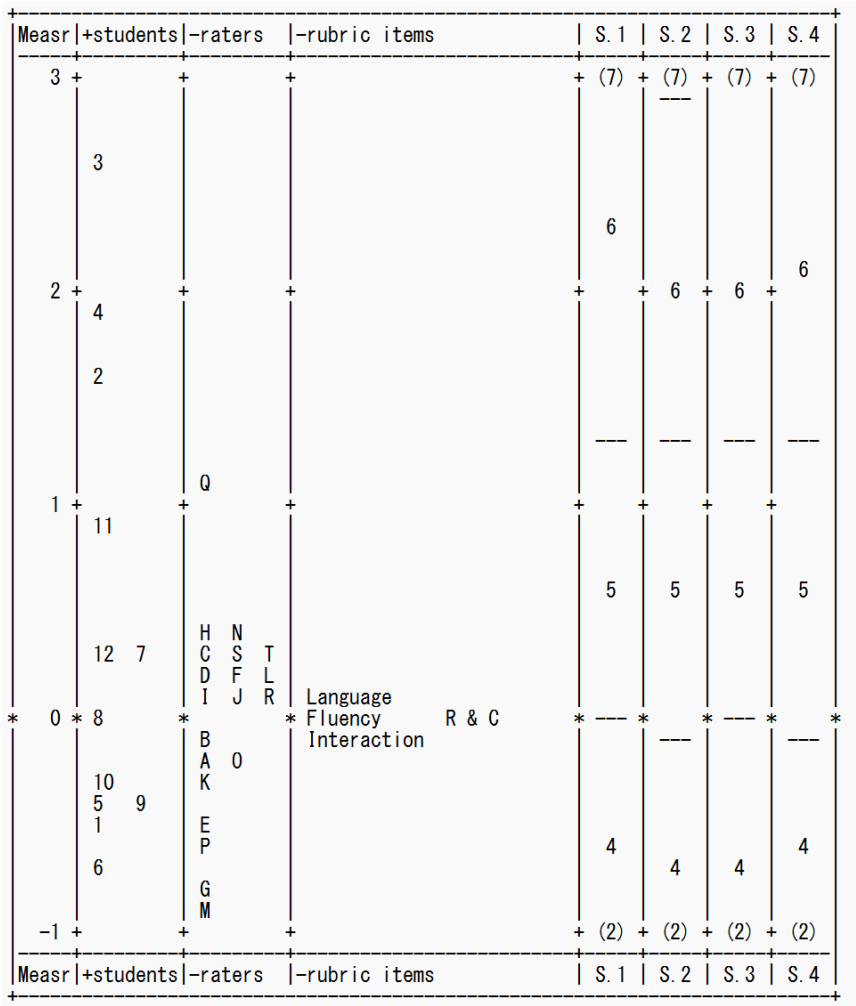
Add a Relevance & Content Category. This category was added as one intended outcome of the course is that the students engage with articles on the unit topics. Thus, this category was added to provide performance-based evidence that this goal has been achieved and thus the assessments align with the intended learning outcomes (Wiggins & McTighe, 2005). It was partly based on the descriptors in the Discourse Management Category for the Cambridge First Certificate (Cambridge, 2008).

Analysis of Revised Rubric

The redesigned rubric (see Appendix C) was tested with an expanded group of raters. In total, 20 raters scored the videos, all of whom had master's degrees in TESOL and/or extensive experience of teaching language. Some raters were from recruited from outside of the program; this was deemed important as the program coordinators suggested that teachers within the program tended to award most scores of around 70%, regardless of the wording of the rubrics. Using raters without such preconceptions should prove a better test of the validity of the descriptors. No mention was made in the instructions that a passing score was 60% and that an average of around 70 – 75% is expected by the institution. Additional speaking test videos were also added to provide a better spread of performances. All videos from the original rating session were included, plus three more, making a total of 12 performances. These additions now mean that this is not now a direct A to B comparison, but it was felt that expanding the sample was more important to explore rubric function more fully. In fact, it would have been ideal to have a greater number of performances to be rated, but logistical issues limited this. Further, standardisation training was not conducted before scoring. This is clearly less than ideal but does reflect the reality of the program, where standardization cannot always occur, and, therefore, provides a robust test of rubric performance in context-realistic conditions. Additionally, it was necessary to recode the scale as *Facets* cannot take decimals. Therefore, for the purposes of the Rasch analysis the scale points were recoded as follows: 3 = 3, 3.5 = 4, 4 = 5, 4.5= 6, 5 = 7.

Overall, the Wright Map (Figure 5) and Tables 4 – 6 shows several interesting findings. Regarding the spread of student abilities, the rubric identified a wider range of abilities as the separation statistics stood at 8.08, as shown in Table 4, slightly wider than the 7-point scale. Also, it is clear that the sample is skewing positively as no learners displayed abilities below -1 logits, but three above +1 logits. This is an artifact of the university policy of setting a passing grade at 60%, so the scale points 0 to 2 are less extensively used, as a pass should be achievable for most. The wording of the rubric was deliberately chosen such that the majority of scores would fall above this threshold. Also, given that a number of raters were unaware of this, it suggests the descriptor wording is targeting a suitable difficulty level for this cohort and the resulting skew in fact aligns the assessment institutional expectations, while still maintaining the ability to distinguish differing ability levels.

Figure 5
Wright Map for Revised Rubric



The spread of rater severity, column 2, now ranges from -0.94 to 1.10, a total range of 2.04 logits, an increase from 0.87 from the original rubric. Also, the separation statistic of 2.58 and reliability at 0.87 now suggest at least two statistically significant different levels of severity and a lower likelihood

of repeatability. These figures have increased from the original separation of 0.76 and reliability of 0.36. This is clearly worse than the original rubric and could be due to the novelty of the rubric, which was new to all raters. Rater training, ideally over a period of time, should bring scores closer into alignment. Indeed, it has been found that experienced teacher-raters can provide more-or-less reliable scores using their background and experience, as is likely the case here, but specific training with a given rubric can lead to considerable improvement in reliability and reduced severity ranges (Yan & Chuang, 2022).

The data on the rubric categories in column three now show the rubric categories as bunching very tightly together, with a separation statistic of 0.60, suggesting similar difficulty levels. The low reliability coefficient (0.26) supports this and demonstrates that the different rubric categories are similarly difficult. This may or may not be an improvement, as it could be indicative of halo effects (Myford and Wolfe, 2004).

Table 5 shows the fit statistics for the rubric, as all fall very close to 1 and within the narrower range of 0.7 – 1.3 (Wright & Linacre, 1994) suggesting good fit to the Rasch model.

Table 4
Separation Statistics for Revised Rubric

	Root-mean Square Error	Separation Index	Reliability Coefficient	x^2
Student facet	0.13	8.08	0.98	0.00
Rater facet	0.17	2.58	0.87	0.00
Rubric items	0.08	0.60	0.26	0.26

Table 5
Rubric Categories in Fit Order

Category	Measure	SE	Infit MNSQ	Outfit MNSQ
Interaction	-0.11	0.08	1.26	1.30
R & C	0.01	0.07	0.96	0.96
Language	0.10	0.08	0.94	0.94
Fluency	-0.01	0.08	0.80	0.79

Table 6 gives details on the raters, and overall, the raters fit the model well. Almost all fall between 0.5 – 1.5, with three underfitting with Infit MNSQ between 1.5 and 2.0. Two raters, R and F, are relatively close to the 1.5 threshold; however, rater N is somewhat higher. Only two raters, C and L, exhibited overfit, but overfit rarely causes any validity issues for measurement, especially when rater agreement is encouraged (Linacre, 2020). However, it could be indicative of halo effects where examiners show less variance than expected and assign identical scores across categories despite differing performances within each category (Myford & Wolfe, 2004). One simple method for investigating this suggested by Myford and Wolfe is to calculate the percentage of grades awarded by each rater that are identical across categories. This is shown in the rightmost column, and there appears not numerous incidences of halo effect. The two most overfitting raters, perhaps unsurprisingly, had the highest percentages of identical scores, but the 3rd lowest had none. However, further training would likely be beneficial (Linacre, 2012), especially for the three that underfit. The underfit exhibited here does not appear large enough to invalidate the measures, and so for the purposes of this study, it is not necessary to remove these ratings. Overall, without any formal training on the use of the rubric, these figures are encouraging and would improve with a standardisation session. Further encouraging statistical support is the close match of exact agreements, the Rasch Model expect this to be 31.1%, and the data yields a score of 31.2%.

Table 6
Raters in Fit Order

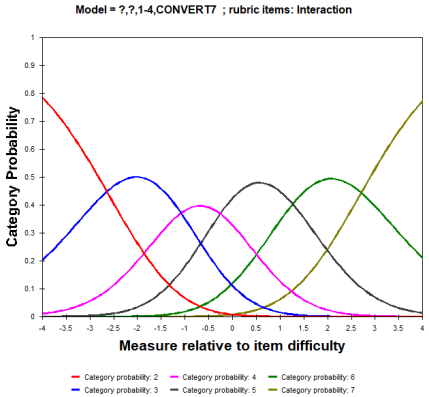
Rater	Measure	SE	Infit MNSQ	Outfit MNSQ	% Identical
N	0.40	0.16	1.87	1.91	8.33
R	0.10	0.17	1.65	1.63	16.67
F	0.15	0.17	1.59	1.58	8.33
J	0.10	0.17	1.28	1.39	16.67
Q	1.10	0.17	1.31	1.33	8.33
K	-0.30	0.17	1.27	1.27	33.33
G	-0.78	0.18	1.08	1.15	0.00
P	-0.56	0.17	1.06	1.07	8.33
O	-0.24	0.17	1.05	0.98	8.33

Rater	Measure	SE	Infit MNSQ	Outfit MNSQ	% Identical
T	0.26	0.17	0.99	1.02	8.33
M	-0.94	0.18	0.89	0.86	16.67
S	0.31	0.17	0.82	0.81	0.00
A	-0.21	0.17	0.77	0.77	0.00
B	-0.07	0.17	0.73	0.76	16.67
D	0.21	0.17	0.71	0.70	16.67
H	0.37	0.16	0.68	0.69	8.33
E	-0.47	0.17	0.61	0.62	8.33
I	0.07	0.17	0.59	0.57	0.00
C	0.31	0.17	0.45	0.46	25.00
L	0.21	0.17	0.39	0.38	33.33

Figure 6 gives the combined Category Probability Curves for the revised rubric overall. In general, the results here are very positive as each scale point is relatively distinct from its neighbour, the peaks are even and are not overlapping, and the peaks are not subsumed by others. In general, this points to a well-functioning rubric and is largely what could be hoped for in this context.

Figure 6

Overall Category Probability Curves



In addition to the overall rubric performance, it is also important to look at the individual category response curves (Andrich, 1996), as shown in Figures 7 – 10. Similarly positive results to the overall category curves are evident; however, some areas where further progress could be made. On the positive side, most peaks occupy their own space along the latent variable, but there are also clear exceptions to this, especially scale-point 4 in the Interaction and R and C categories, and to a lesser extent point 6 for fluency, where peaks are subsumed. Despite this, the improvement from version 1 is clear and substantial.

Figure 7

Category Probability Curves: Interaction

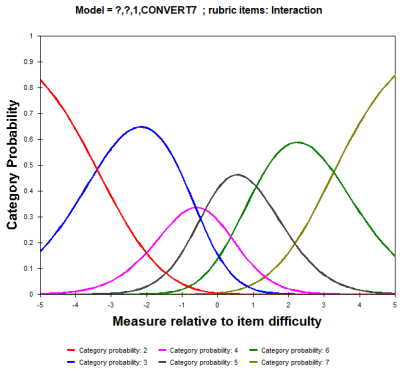


Figure 8

Category Probability Curves: Fluency

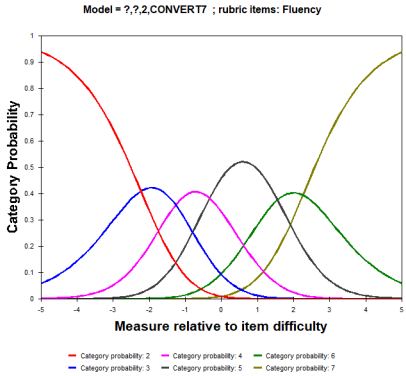


Figure 9

Category Probability Curves: Language

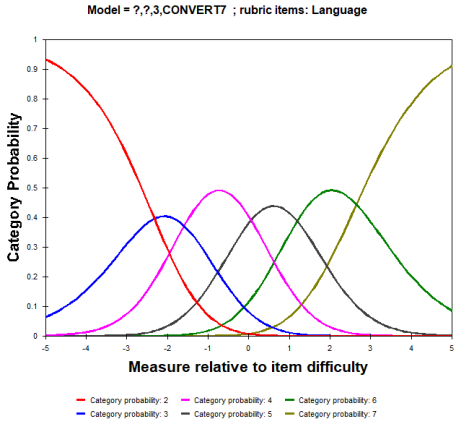


Figure 10

Category Probability Curves: Relevance and Content

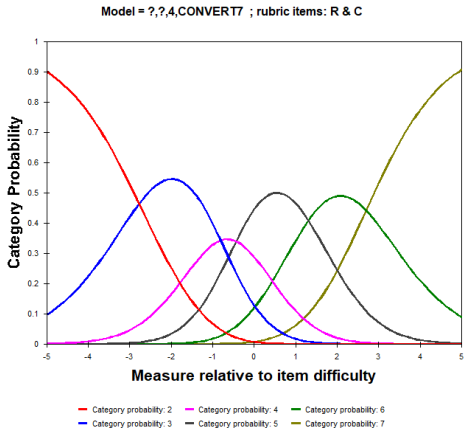


Table 7 gives specific statistical information for the four rubric categories. For all categories, the step calibrations advance monotonically, as per Linacre's recommendation (2002), a clear contrast from the first rubric iteration. Also, all passing grades, scores 3 to 7, have more than 10 observations, and so similarly meet Linacre's (2002) minimum requirement for stability.

However, Linacre (2002) also recommends that there be a minimum of 1.4 logits between the category thresholds, as can be seen in the Rasch-Andrich Threshold column, but this is not the case as a number of places where the spacing is below these recommendations can be seen. Instances of where the distance is below 1.4 logits are shown in bold in Table 7. This implies these scale points do not represent a suitably distinct level of ability on the latent variable; however, all scales do advance monotonically, which represents significant progress.

Table 7
Revised Rubric Step Calibrations

Scale Point	Number of Observations	Av. Measure	Rasch-Andrich Threshold	Distance to Next Category	Standard Error
			Interaction		
1	0	N/A	N/A	N/A	N/A
2	2	-1.12	N/A	N/A	N/A
3	34	-0.40	-3.39	2.77	0.72
4	48	0.28	-0.62	0.27	0.20
5	78	0.39	-0.35	1.44	0.16
6	63	1.08	1.09	2.18	0.17
7	15	2.48	3.27	N/A	0.30
			Fluency		
1	0	N/A	N/A	N/A	N/A
2	5	-1.23	N/A	N/A	N/A
3	24	-0.47	-2.23	0.96	0.47
4	59	-0.25	-1.27	0.95	0.22
5	87	0.27	-0.32	1.88	0.16
6	41	1.52	1.56	0.70	0.19
7	24	2.27	2.26	N/A	0.26

Scale Point	Number of Observations	Av. Measure	Rasch-Andrich Threshold	Distance to Next Category	Standard Error
			Language		
1	0	N/A	N/A	N/A	N/A
2	5	-1.29	N/A	N/A	N/A
3	24	-0.61	-2.32	0.73	0.47
4	75	-0.26	-1.59	1.65	0.22
5	71	0.37	0.06	1.13	0.16
6	47	1.20	1.19	1.46	0.19
7	18	2.20	2.65	N/A	0.28
			R & C		
1	0	N/A	N/A	N/A	N/A
2	4	-0.95	N/A	N/A	N/A
3	33	-0.51	-2.77	1.96	0.52
4	51	-0.30	-0.81	0.38	0.20
5	83	0.40	-0.43	1.74	0.16
6	50	1.42	1.31	1.39	0.18
7	19	1.89	2.70	N/A	0.28

Conclusion & Reflections

Overall, the use of a Rasch analysis has led to considerable improvements in the rubric functioning, with scale points and categories far more clearly delimitated, leading to far more reliable scoring. However, more work needs to be done in terms of validation as the small sample of test takers mean there could be more clarity in terms of the number of levels of ability the rubric can identify. Also, several scale points still have relatively narrow logit distances between them, so closer attention to the wording of the descriptors or a merging of some scale points could be areas that would improve functioning still further. Indeed, it has been argued that adhering to a consistent number of scale points across categories, although the norm and appearing 'neat' on the surface, may come with validity issues (Humphry & Heldsinger, 2014) as unnecessary scale points may be added for the sake of appearances.

Furthermore, although the categories now appear to be better matched in terms of overall difficulty, this can in fact provide less information on the sub-skills that make up the assessment, making it potentially less valuable. In our case, it appears that the apparent lack of halo effect means that the categories are of a similar level of difficulty, but care needs to be taken when interpreting such trends.

Through the process of developing this rubric there emerged some general principles that could be generally applied to rubric development, namely:

- **Less is more regarding scale points**. Frequently, a small number of scale points have been found to perform better than a larger number (Bonk & Ockey, 2003; Janssen et al., 2015; McDonald, 2018). This increases clarity as to what a particular score means and therefore allows for better feedback and clearer performance expectations. Although it may be tempting to allow a large range of points to be awarded for greater flexibility; in reality, this can lead to inconsistent scoring across raters and so should be avoided.
- **Separate constructs into clear categories.** In the original version of the rubric, there was some confusion arising from indistinctly defined constructs. By separating these into categories with clearly defined boundaries, raters and test takers alike will have clearer expectations as to what any rubric category is trying to target. This also helps to add to the granularity of the assessment as specific information can be provided about sub-dimensions of an overarching skill. This can reveal information on which aspects of performance pose differing levels of challenge to learners and action can be taken accordingly. Of course, this is assuming raters are scoring each category distinctly from the others and that halo effects are not evident. This is important as categories that align well on the Wright Map may look tidy but could indicate other issues.
- **Look to the bigger picture, avoid a check box approach.** The original rubric included individual phrases that were checked when used. Such an approach can be appropriate in some cases, but it has been argued that it can be reductionalist (Green, 2013) as it ignores certain aspects of performance. Some teachers commented, for example, that it is unclear if any phrases need to be pronounced perfectly or with 100% grammatical accuracy for points to be awarded. As such, seemingly simple checkbox approaches can in fact add complexity and reduce reliability if expectations are not clearly set.

- **Carefully word the descriptors based on the performance expectations of the cohort.** If an institution, as was the case here, has guidelines in terms of the passing score, then descriptors need to be written such that the minimum expected performance is 'set' to this benchmark. Knowledge of cohort ability and the general levels of performance they are capable of is essential here, as is teacher and assessor input.
- **Involve colleagues in the process of rubric development.** Despite teacher comments not featuring in this paper, they did play a significant role in the development process and provided valuable insights into teacher perceptions of rubric function and its usability. Adding a learner perspective in any future studies would strengthen any future research findings and involve more key stakeholders, as suggested for the development of any well-rounded testing instrument (Fulcher & Davidson, 2007; Hamp-Lyons, 1991; Ockey et al., 2013).

These recommendations need to be caveated with the proviso that the needs of all stakeholders in the local context need to be considered in the design of assessment instruments, but MFRM would likely be a useful tool where rater-mediated assessment is employed, regardless of the form of the rubric.

Thomas Stones has been working in language teaching for more than 15 years and currently works at the Department of Economics at Kwansei Gakuin University. He has a range of research interests including developing skills in interactional competence, assessment validation using Rasch-based methods, the teaching and assessment of listening skills as well as developing skills in self-directed learning. He has presented and published on all of these topics.

Appendices

All appendices are available from the online version of this article at https://jalt-publications.org/jj.

References

Andrich, D. A. (1996). Measurement criteria for choosing among models for graded responses. In A. von Eye & C. C. Clogg (Eds.), *Analysis of categorical variables in developmental research* (pp. 3–35). Academic Press. https://doi.org/10.1016/B978-012724965-0/50004-3

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 38*(1), 1–35. https://doi.org/10.1177/0265532220927487

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89–110. https://doi.org/10.1191/0265532203lt245oa

Bryman, A. (2006). Integrating quantitative and qualitative research: How is it done? *Qualitative Research, 6*(1), 97–113. https://doi.org/10.1177/1468794106058877

Cambridge. (2008). Assessing speaking performance – level B2. https://www.cambridgeenglish.org/images/168619-assessing-speaking-performance-at-level-b2.pdf

Chen, Y., & Liu, J. (2016). Constructing a scale to assess L2 written speech act performance: WDCT and e-mail tasks. *Language Assessment Quarterly, 13*(3), 231–250. https://doi.org/10.1080/15434303.2016.1213844

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* https://rm.coe.int/16802fc1bf

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117–135. https://doi.org/10.1177/0265532215582282

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement.* Peter Lang.

Engelhard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5*(3), 171–191. https://doi.org/10.1207/s15324818ame0503_1

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge. https://doi.org/10.4324/9780203449066

Galaczi, E. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics*, *35*(5), 553–574. https://doi.org/10.1093/applin/amt017

Galaczi, E. D., ffrench, A., Hubbard, C., & Green, A. (2011). Developing assessment scales for large-scale speaking tests: A multiple-method approach. *Assessment in Education: Principles, Policy & Practice, 18*(3), 217–237. https://doi.org/10.1080/0969594X.2011.574605

Green, A. (2013). *Exploring language assessment and testing: Language in action*. Routledge. https://doi.org/10.4324/9781315889627

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). McGraw-Hill.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241–276). Ablex.

Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning, 41*(3), 337–373. https://doi.org/10.1111/j.1467-1770.1991.tb00610.x

Humphry, S. M., & Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher, 43*(5), 253–263. https://doi.org/10.3102/0013189X14542154

IELTS (n.d.). *Speaking: Band descriptors* (public version). https://www.ielts.org/-/media/pdfs/speaking-band-descriptors.ashx

Iwashita, N., Elder, C., & McNamara, T. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information processing approach to task design. *Language Learning, 51*(3), 401–436. https://doi.org/10.1111/0023-8333.00160

Janssen, G., Meier, V., & Trace, J. (2015). Building a better rubric: Mixed methods rubric revision. *Assessing Writing*, *26*, 51–66. https://doi.org/10.1016/j.asw.2015.07.002

Jeong, H. (2015). Rubrics in the classroom: Do teachers really follow them? *Language Testing in Asia*, *5*(6), 1–14. https://doi.org/10.1186/s40468-015-0013-5

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks.* Guilford.

Kramsch, C. (1986). From language proficiency to interactional competence, *The Modern Language Journal*, *70*(4), 366–372. https://doi.org/10.2307/326815

Leaper, D. A., & Brawn, J. R. (2019). Detecting development of speaking proficiency with a group oral test: A quantitative analysis. *Language Testing*, *36*(2) 181–206. https://doi.org/10.1177/0265532218779626

Linacre, J. M. (1994). *Many-Facet Rasch Measurement*. (2nd ed.). MESA Press.

Linacre, J. M. (2001). *FACETS* [Computer program, version 3.36.2]. MESA Press.

Linacre, J. M. (2002). Optimal rating scale category effectiveness. *Journal of Applied Measurement*, *3*, 85–106.

Linacre, J. M. (2012). *Many-Facet Rasch Measurement: Facets tutorial*. https://www.winsteps.com/a/ftutorial2.pdf

Linacre, J. M. (2020). *A user's guide to Facets*. https://www.winsteps.com/manuals.htm

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press. https://doi.org/10.1017/CBO9780511733017

May, L., Nakatsuhara, F., Lam, D., & Galaczi, E. (2020). Developing tools for learning oriented assessment of interactional competence: Bridging theory and practice. *Language Testing, 37*(2), 165–188. https://doi.org/10.1177/0265532219879044

McDonald, K. (2018). Post hoc evaluation of analytic rating scales for improved functioning in the assessment of interactive L2 speaking ability. *Language Testing in Asia*, *8*(19), 1–23. https://doi.org/10.1186/s40468-018-0074-3

McNamara**,** T. F. (1996). *Measuring second language performance.* Longman.

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555–576. https://doi.org/10.1177/0265532211430367

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2), 81–97. https://doi.org/10.1037/h0043158

Morgan, D. L. (1998). Practical strategies for combining qualitative and quantitative methods: Applications for health research, *Qualitative Health Research*, *8*(3), 362–376. https://doi.org/10.1177/104973239800800307

Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*(4), 873–900. https://doi.org/10.1111/j.1744-6570.2000.tb02421.x

Myford, C. M., & Wolfe, E.W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.

Myford, C. M., & Wolfe, E.W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189–227.

Nakatsuhara, F., Inoue, C., & Taylor, L. (2020). Comparing rating modes: Analysing live, audio, and video ratings of IELTS speaking test performances. *Language Assessment Quarterly*, *18*(2), 83–106. https://doi.org/10.1080/15434303.2020.1799222

Nitta, R., & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-task planning effects on paired oral task performance. *Language Testing*, *31*(2), 147–175. https://doi.org/10.1177/0265532213514401

Ockey, G. J., Koyama, D., & Setoguchi, E. (2013). Stakeholder input and test design: A case study on changing the interlocutor familiarity facet of the group oral discussion test. *Language Assessment Quarterly, 10*(3), 292–308. https://doi.org/10.1080/15434303.2013.769547

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores, *System 30*(2), 143–154. https://doi.org/10.1016/S0346-251X(02)00002-7

Randall, J., & Engelhard, G. (2009). Examining teacher grades using Rasch measurement theory. *Journal of Educational Measurement, 46*(1)*,* 1–18. https://doi.org/10.1111/j.1745-3984.2009.01066.x

Roever, C., & Kasper, G, (2018). Speaking in turns and sequences: Interactional competence as a target construct in testing speaking. *Language Testing*, *35*(3), 331–355. https://doi.org/10.1177/0265532218758128

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, *85*(6), 956–970. https://doi.org/10.1037/0021-9010.85.6.956

Singh, K. (2007). *Quantitative social research methods*. Sage. https://doi.org/10.4135/9789351507741

Wiggins, G., & McTighe, J. (2005). *Understanding by design* (2nd ed.). ASCD.

Wiliam, D. (2011). *Embedded formative assessment.* Solution Tree Press.

Wright, B., & Linacre, M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Wu, M., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement, 14*, 339–355.

Yan, X., & Chuang, P. L. (2022). How do raters learn to rate? Many-facet Rasch modelling of rater performance over the course of a rater certification program. *Language Testing*, *40*(1), 153–179. https://doi.org/10.1177/02655322221074913

Young Learner L2 Vocabulary Acquisition: Does the Revised Hierarchical Model Apply to Child Learners?

Clay Williams
Naeko Naganuma
Akita International University

Vocabulary learning is a process requiring the connection of mental concepts to new word-labels. The Revised Hierarchical Model claims that beginning learners recognize the meaning of L2 words via a process of translation, needing considerable time and effort to forge direct connections between L2 words and mental concepts. However, might young children, as they are still rapidly acquiring L1 vocabulary, be able to bypass the L2-to-L1 translation required by adult L2 learners, and instead link new L2 words directly to pre-existing mental concepts? This study tested over 1,000 4th-6th graders in Japanese elementary schools on their ability to match newly learned L2 words with corresponding pictures or L1 translations. The results demonstrate that students connect L2 vocabulary to pictures more quickly, and this effect becomes more robust when students are taught via pictures, which suggests that young learners are indeed capable of accessing concepts without translating from their L1.

語彙学習とは学習者の脳内に存在する概念を新しい語彙に連結するという過程である。改訂階層モデルによると、初級者は翻訳という過程を経ることで第二言語(L2)での意味を認識する。そのため、L2語彙と脳内概念への直接連結するにはかなりの時間と努力を要する。しかし、まだ急速に第一言語(L1)語彙習得過程にある子どもたちはどうだろうか。成人学習者が必要とするL2からL1への翻訳を介さずに、脳内概念とL2語彙へ直接連結することが可能なのではない

https://doi.org/10.37546/JALTJJ46.1-2

JALT Journal, Vol. 46, No. 1, May 2024

だろうか。本研究では四年生から六年生の日本人小学生千人以上を対象に、新出L2語彙を対応する絵または日本語へ照合する能力について検証した。結果として、L2語彙に対応する日本語への照合よりもイメージへの照合がより早く行われ、またこの効果が絵を用いて教わった場合により強く見られた。このことから、年齢が低い学習者はL1への翻訳することなしに脳内に存在する概念にアクセス可能であることが示唆できる。

Keywords: conceptual access; Japanese learners of English; RHM; second language education; vocabulary learning

The importance of vocabulary acquisition in learning languages has been widely researched, across a variety of perspectives, and it is safe to claim that its significance is no longer a matter of debate. While vocabulary learning is one of the most basic aspects of language learning, and indeed, one of the basic units by which we can measure such learning, our understanding of the psychological processes, undergirding and driving the acquisition of words, is still only in its early stages. Especially when we cross-analyze first and second language acquisition dynamics, a number of interesting—and, as of yet unanswered—questions raise themselves immediately, such as the relative degree of difference and sameness in process. One issue that is often debated is the role of creation of direct links between vocabulary and mental concepts. While it is often taken for granted that word acquisition and conceptual access occur simultaneously in an L1 context, learning in an L2 context (wherein the conceptual links between L1 words and concepts already exist) opens up other possibilities. Might these fundamental psychological processes for learning vocabulary be different between L1 and L2 learners, could this be affected by learner variables such as age, and how would this affect learning?

Literature Review

Research on Conceptual Access

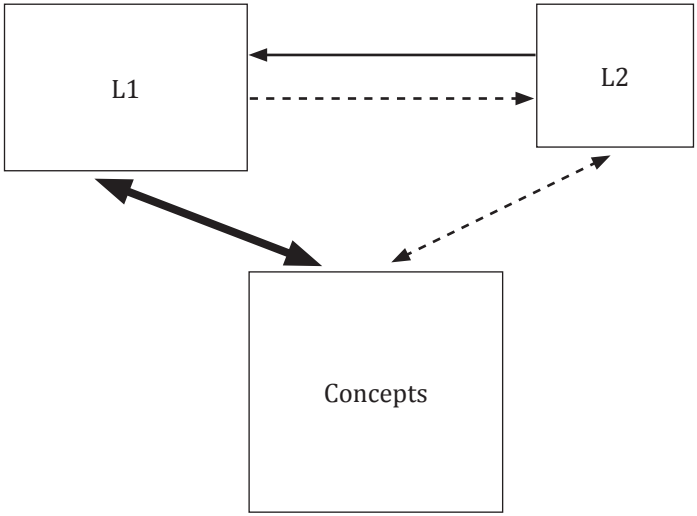
It is believed that most of the information in the brain is stored conceptually as images and/or ideas rather than aurally (or orally), and L1 users can make a link between those images without interference. However, applying this capability to L2 acquisition is still controversial. Potter et al. (1984)'s research was one of the first to investigate conceptual access. They proposed two hypotheses on how words in two languages are associated with each other. According to the *word association* hypothesis, mental concepts can only be accessed via words in the L1, and thus, L2 words have to be translated

to L1 before meaning can be unlocked. By contrast, the *concept mediation* hypothesis suggests that concepts are directly linked to both L1 and L2 vocabulary equally. These two conflicting theories were tested by a number of researchers. The *word association* hypothesis could be proven if translation from L1 to L2 were faster than naming the image in L2. This would verify that L2 word retrieval was conducted through L1 (i.e., the concept triggering the L1 label, which in turn triggers the L2 label), not via direct conceptual access. On the other hand, if the *concept mediation* hypothesis were proven true, recall of the names of objects in pictures (i.e., picture naming tasks) in L2 should take the same amount of time as word translation from L1 to L2. The findings by Potter et al. (1984) showed that it took the same amount of time for advanced L2 learners to translate from L1 to L2 as picture naming, while for lower level L2 learners, it took much less time to translate from L1 to L2 than picture naming. Therefore, it was suggested that both of the models required revisions in order to explain the subsequent results. Note that, while mental image and mental concepts are not synonymous, there is broad overlap, at least in terms of concrete, easily-visualizable vocabulary. Given the past methodology's reliance on picture naming as a means of testing conceptual access, this paper will use the terms *image* and *concept* fairly interchangeably. This is merely reflective of the concentration on concrete vocabulary (which both past research and the current study will concentrate upon), and is not meant to suggest a broad equivalence between mental image and concepts beyond the sorts of vocabulary items dealt with herein.

The Revised Hierarchical Model (hereafter the RHM) was devised by Kroll and Stewart (1994) to account for these differences in basic word processing by L2 students of different proficiency levels (refer to Figure 1). They claimed that L2 learners would develop the ability to directly link concepts to L2 target words over time. Their study found that there were differences in the speed of translation from L1 to L2 and picture naming, and the asymmetry could be found in the direction of translation: translation from L1 to L2 being slower than L2 to L1. They also found another asymmetry in terms of categorical interference. Tests devised to elicit semantic interference effects by grouping vocabulary found significant interference effects when translating from L1 to L2, but not the other way around. The other asymmetry found in the study by Sholl et al. (1995) was difference in priming facilitation occurring only when participants translated from L1 to L2 after a picture naming task, irrelevant to the language used in the task.

Figure 1

Revised Heretical Model (RHM)



Criticism on the Revised Hierarchical Model

In fact, the RHM has been criticized because the studies which followed Kroll and Stewart (1994) found contradictions in the proposed theory (Brysbaert & Duyck, 2010; de Groot et al., 1994; van Hell & de Groot, 1998). In particular, Brysbaert and Duyck suggested "leaving behind" (p. 359) the RHM based on their rather comprehensive review of research which was published after the theory was originally published. They concluded that since there have been previous models made for investigating bilingual language processing, more research should have been implemented to check how to adapt the existing models, including the Bilingual Interactive Activation model proposed by Dijkstra and van Heuven (2002). Another revision suggested for the RHM was on the interpretation of L1 and L2. When learners are in situations where they are immersed in the L2, making L2 the dominant language, reversed results were found (Heredia, 1997). Therefore, some researchers proposed that the RHM should be referred to from the perspectives of dominant vs. additional languages, rather than L1 vs. L2 (Heredia, 1997; Linck et al., 2009). Additionally, the study by Sunderman and Kroll (2006) found no appreciable differences in the degree of semantic sensitivity between low- and high-proficiency L2 learners. Williams (2017, 2018), in a series of studies on semantic priming sensitivity, found that orthographic properties of Japanese

and Chinese script may impair the development of certain types of semantic sensitivity in L2 learners of English from Japanese/Chinese L1 backgrounds.

Furthermore, Kroll et al. (2010) suggested acknowledging the possible weakness in the RHM that there is a bidirectional weak link between L2 and concept. The study found that the results of production tasks (i.e. naming) and receptive tasks (i.e. comprehension) showed significant gaps, indicating the link between L2 and concept is actually asymmetrical. To pursue possible reasons for an asymmetrical result in Kroll and Stewart (1994)'s study, the category facilitation effect in the L2-to-concept and L1-to-concept directions was further investigated by Wu and Juffs (2019). They tested whether they could find a category facilitation effect in both the L2-to-concept and L1-to-concept directions by providing the categorized list conditions and the randomized list conditions. Their results indicated a significant category facilitation effect in both L2-to-concept for young Chinese adults and L1-to-concept for young English adults when the number of trials was increased. They argue that the general L2-to-L1 null category effect discovered by Kroll and Stewart (1994) could not be used to disprove concept mediation in backward translation.

Responding to the general criticism on the RHM, Kroll et al. (2010) argued against the idea of "leaving behind the RHM" (Brysbaert & Duyck, 2010, p. 359), claiming that in principle "models provide a means to approach problem solving and to refine our thinking" (Kroll et al., 2010, p. 381) instead of being tested and rejected. While various issues have been evident in the model itself, the RHM has remained one of the most dominant in explaining concept mediation. It suggests that L2 access to concepts in the brain can only be achieved for lower-level L2 learners through use of the L1; however, learners would develop the ability to link concepts directly to L2 words as they develop proficiency. This, in turn, raises other questions, such as whether or not younger L2 learners are similarly restricted from connecting L2 vocabulary directly to mental concepts, and whether they could develop such conceptual links differently from adults.

Research of the RHM on Japanese Learners of English

As the study which will be presented in this paper is focused on Japanese learners of English, it would be useful to review prior research investigating the RHM as pertaining to Japanese learners of English. In one of the earliest studies, Kawakami (1994) tested how three groups of Japanese learners of English with different proficiencies (English-major university students, high school students, and junior high school students) would perform in learning new English vocabulary words. In her study, the group of higher proficiency

(English-major university students) performed similarly both in the Japanese and English priming tasks, which led her to conclude that both English and Japanese vocabulary were accessible to the higher proficiency group at the similar levels due to more direct links between meaning and concept than less-proficient learners. She found that the data generally corroborated RHM predictions because the group of higher proficiency evinced more direct links between meaning and concept, while the patterns that less-proficient learners produced matched more with the *word association* model as Potter et al. (1984) suggested.

Similarly, the study by Anezaki (2006) attempted to investigate if there would be any difference between first-year students who had learned English for three months and third-year students who had a longer experience in learning English for two years and three months in a formal school setting at junior high school. With a two-choice reaction time task, he discovered that it took longer for L2 learners at a very early stage to engage in backward translation than those learners with more experience in learning L2 at school. Another finding was that this asymmetry disappeared among the second group of learners. Therefore, Anezaki concluded, "the results of this study are congruent with the prediction of the Revised Hierarchical Model" (p. 128).

Some studies on Japanese learners of English focused on the concreteness of target vocabulary words to examine the RHM. Habuchi (2003) investigated how words would be processed in translating between English and Japanese. According to her study, advanced-level Japanese learners of English seemed to go through the process suggested by the RHM when they dealt with concrete words (e.g., fox and fish), whereas the results showed that the participants were processing words in accordance with the *word association* hypothesis (Potter et al., 1984) when dealing with abstract vocabulary words in L2. The study by Nakagawa (2009) explored interrelatedness among L1 and L2 lexicons and concepts through her experiment on Japanese first-year university students, finding that more concrete and higher frequency target words were found processed via concept mediation, while abstract words seemed to be processed via word association. In addition, Nakamura (2007) found that translation of concrete words from L2 to L1 would be processed as suggested in the *word association* hypothesis, while the L1 to L2 translation would be done through the process according to the *concept mediation* hypothesis (Potter et al., 1984). In addition, he explored differences between the direction of the translation, and the results supported the RHM, L1 to L2 translation of concrete words requiring concept mediation while L2 to L1 translation of concrete words done via word association.

Conceptual Access of Young Learners

The RHM has received tremendous attention and has been a target to be tested on bilingual adult learners, but less attention has been on young L2 learners. Young learners, in fact, exhibit distinctive traits from adult L2 learners. It is estimated that 6-year-old to 8-year-old children learn 6 to 7 new words per day, and this rate increases to 12 words per day at the age of 8 to 12 (Bloom & Markson, 1998). While young children are expanding their L1 vocabulary, creating links between new words and concepts, it can be assumed that it might not be as difficult to create direct links between L2 words and respective concepts. This was found to be the case in the study by Comesaña et al. (2009). After one vocabulary session, Spanish-L1 elementary school students showed semantic interference effects. It turned out that it was hard for them to reject incorrect translations that were semantically related when trying to acquire target L2 words (in this case Euskera, Basque language). Comesaña et al. (2012) replicated the study by Comesaña et al. (2009) with Portuguese L1 speaking children who are learning Euskera. The researchers found that the participants displayed similar results of semantic interference effects. In addition, it was found that the degrees of semantic interference increased when target words were instructed via pictures, and also the delayed posttest conducted one week later revealed that the semantic interference effect increased regardless of the different teaching methodologies being used.

Another study by Poarch et al. (2015) investigated how Dutch L2 fifth graders after receiving English instruction for eight months connected new vocabulary words to mental concepts. Their results generally corroborated what Comesaña et al. (2009) and Comesaña et al. (2012) found. Young Dutch learners of English at early stages of their L2 learning were found to be able to actively exploit conceptual links when they translated from English to Dutch. The study by Sheng et al. (2013) examined if Spanish-English bilingual children would prove to be influenced by their age and previous learning experience of L2 in semantic development. The study concluded that their results were "consistent with predictions of the Revised Hierarchical Model of bilingual lexical organization" (p. 1023).

The Present Study

The current study aimed to investigate whether young L2 learners can create conceptual links to L2 vocabulary which they learn as new words, and the researchers also tried to identify at what age the ability to build direct conceptual links to L2 words might cease. Additionally, the study looked at

whether pedagogical methods would influence the degree of the L2 conceptual connection strength or not.

Study Participants

During the data collection phase of the present study, 1,260 elementary-aged children, ranging from 4th to 6th graders, participated. The breakdown according to age/grade was: 4th graders = 437; 5th graders = 346; and 6th graders = 477. They were all monolingual native speakers of Japanese. Classroom teachers were consulted to identify any students who had multilingual backgrounds (e.g., students who had spent significant time abroad, who lived in households where languages other than Japanese were spoken, or who engaged in English study in private educational centers). Students with such backgrounds were still permitted to participate in the study, but their test results were not included in the analysis.

Participants were recruited from seven elementary schools in one city in northern Japan. The research team visited those seven elementary schools during the periods from July 2017 to November 2018 for data collection. The Ministry of Education, Culture, Sports, Science, and Technology (hereafter, MEXT) stipulates the school curriculum and releases new versions of course of study every 10 years. MEXT (2017) announced the new course of study in July 2017, suggesting to begin preparatory measures to ensure a smooth transition from April 2018 and to complete the transition before April 2020. When the current study was conducted, the new course of study had been released; however, the actual implementation of the new curriculum had not been in progress. Therefore, it was considered to be safe to assume that students below the 5th grade had no or minimal exposure to English in a formal classroom setting, and 5th and 6th graders had undergone one 45-minute English lesson per week since the beginning of the 5th grade when the study was executed in 2017-2018.

Materials

The vocabulary items used in the current study were decided in conjunction with the teachers at one of the elementary schools used for the pilot study (wherein all instruments and materials were calibrated). To maximize the probability of the participants never having heard the target words before the study, the researchers chose the target vocabulary through a careful discussion with a group of teachers beforehand. Additional efforts were made to avoid selecting English words which have been used as *katakana-eigo*,

borrowed foreign words which had been already integrated into the Japanese language, including many food names, e.g., *soup* and *broccoli*. Since public elementary schools in the same district use the same textbook and follow the same curriculum, it was presupposed that those 45 items to be used in the study would be English words that elementary-aged participants had not heard or acquired yet through the English lessons at school. A full list of the 45 items is available in Appendix A. Those 45 target vocabulary items were printed and made into laminated cards, one set of cards with an image to represent the target word, and the other set with the Japanese translation in *kana*. Card examples are available in Appendix B. The other research material included computers with the DMDX software (Forster & Forster, 2003) installed for the participants to take a computer-mediated test of vocabulary recognition with images and sound files of vocabulary items. The test itself was written with the DMDX software by one of the researchers.

Procedures

The study was conducted in two days for each participant group. Each class was randomly divided into two groups of equal size by their homeroom teacher beforehand. On the first day of the study, the two groups would go to separate classrooms where they would participate in an English lesson for 45 minutes taught by a graduate school student from the English Language Teaching Practices program at the university to which the researchers are affiliated. The graduate student instructors were from Japan, China, and Vietnam. They were scheduled to teach lessons according to their class schedules, avoiding any time conflicts with their academic activities on campus. All of them have experienced teaching demo lessons in classes but had not completed their teaching practicums yet. All the graduate instructors were required to participate in an explanatory session by one of the researchers beforehand. In addition, before each lesson, the graduate instructors reviewed the pronunciations of all the target words with one of the researchers to consistently present similar oral production to each other in terms of stress and pronunciation.

The graduate student instructors used the first 20 to 25 minutes to teach the 45 target vocabulary items. The words presented were the same in both groups, and all the vocabulary cards which instructors used for instruction were laminated. However, the teaching method and the information on the vocabulary cards differed.

In one group, the instructor utilized vocabulary cards with only images; this group will hereafter be referred to as the "Picture Group." On one side of

each vocabulary card for the Picture Group is a picture which can clearly represent the image of the target word, and the other side shows the target word itself for instructors to refer to when they teach the target vocabulary words. To illustrate, a vocabulary card would include an illustration of a dustpan on one side of the card, and the other side had an English word "dustpan" for reference. The Picture Group instructor was prohibited from using Japanese translations during the lesson, instead presenting the English vocabulary orally, and allowing the visual aid to convey meaning to the children. Therefore, when instructors were asked for the Japanese translation, they were told never to respond in Japanese; instead, they pointed to the target picture card.

The other group was taught using Japanese translation, hereafter referred to as the "L1 Group." The set of the vocabulary cards used for this group includes the Japanese translation in *hiragana* or *katakana* on one side with the target English word on the other side, without any image to represent target words. One example card contains the target English word "dustpan," and the other side shows the Japanese translation "ちりとり" in *kana*. The instructor used the Japanese translation of the vocabulary word along with the Japanese word cards during the lesson for the L1 Group.

The instructors were provided specific directions to follow for their lessons. Approximately the first 20 to 25 minutes of each lesson was devoted to vocabulary instruction of the target words. First, they had students repeat after the instructor while students were looking at each vocabulary card for three rounds. Then, they chose those target words that students seemed to have difficulty and repeated the "repeat-after-me" practice. Afterwards, the instructors randomly chose vocabulary cards to quiz students through asking them to say the English word aloud quickly as soon as they flipped the vocabulary cards. Each lesson was 45 minutes long, so instructors played some fun games for the rest of the lesson time, including playing musical chairs, drawing, crafting, and so on. One important point that the instructors were told to avoid including in the fun activities was not to have students exposed to those 45 target words in any part of the activities.

On the second day of the study (i.e., the day immediately following when the vocabulary was presented), the computer test was administered. Another English lesson was offered to the entire class (i.e., they were not separated into two groups this time), and during the lesson, students were invited in groups of six to go to a separate room to take a computer-mediated test of the vocabulary items they had learned on the first day. In the test, the students would hear (via a headset) vocabulary words from the list of 45 target words presented one at a time. Immediately following the presentation of

the word, a pair of pictures or a pair of words (written in Japanese *kana*) would be displayed on the screen. The pictures used in the test were of the same vocabulary items taught the previous day, but not the same pictures that were presented to the group taught with picture cards. Participants were asked to select which picture or word would best correspond to the vocabulary word they had heard by pressing either the RIGHT or LEFT SHIFT key (corresponding to the choices on the left and right sides of the screen, respectively). Research assistants encouraged participants to answer as quickly as possible, and feedback on accuracy and response time was displayed after each response (therefore encouraging test takers to try to make a game of it, and answer as quickly as possible). Each set of the choices included two *kana* words or two pictures, and the image and *kana* only presentations alternated for counterbalancing the total number of test items. The order of the test item presentations was randomized, and the reaction times were recorded for analysis.

Data Analysis

The reaction times were analyzed via multi-factor ANOVA. Each grade was analyzed to determine time latencies in matching the spoken target word to the picture vs. to the Japanese translation. Further analysis was conducted comparing groups within grade-levels to determine whether the teaching condition affected response times. An error cut-off rate of more than 20% resulted in participants' exclusion from analysis. Given the 20 minutes of actual vocabulary study and a one-day gap before testing, in addition to the participants being young learners, the error rate was high, which necessitated a rather large subject pool in order to gather enough reliable data.

Results

Comparison of reaction times revealed that every single group, whether taught via pictures or taught via L1 translation, was significantly faster at matching L2 words to pictures than they were at matching L2 words to L1 translation equivalents (F1, i.e., analysis of all groups: $p < 0.01$). The individual reaction times can be seen in Table 1.

Table 1
Reaction Times across Grades (Measured in Milliseconds)

Grade	n	Teaching condition (translation vs pictures)	Reaction time for matching L2 word with Picture	Reaction time for matching L2 word with L1 word
6th	199	L1 Translation	1330	1498
6th	190	Picture	1251	1482
5th	120	L1 Translation	1390	1576
5th	124	Picture	1326	1537
4th	130	L1 Translation	1540	1652
4th	121	Picture	1422	1629

Sub-analyses of interactions found a significant item effect (F2, i.e., analysis of the difference between reaction times in item types) whereby reaction times for matching the L2 words to pictures was significantly faster than that for matching them to L1 translations, but only in the Picture Groups: 4th Graders, $F2\ (1,8) = 5.37$, $p < 0.049$ ($\eta_p^2 = 0.40$, small effect); 5th Graders, $F2\ (1,8) = 6.10$, $p < 0.039$ ($\eta_p^2 = 0.43$, small effect); 6th Graders, $F2\ (1,8) = 6.88$, $p < 0.031$ ($\eta_p^2 = 0.46$, small effect). None of the L1 Groups demonstrated any significant item effects (i.e., all $p < 0.08$). While direct comparisons between the teaching conditions within grade levels revealed no significant differences, among 4th graders, the faster mean times for the Picture Group vs. the L1 Group nears significance: $F1\ (1,126) = 3.79$, $p < 0.0539$. Nonetheless, this near-effect fades in 5th grade: $F1\ (1,124) = 2.63$, $p < 0.107$, and it disappears entirely by 6th grade; however, in the 6th grade, there was also an item effect between the two teaching conditions, favoring the Picture Groups: $F2\ (1,8) = 6.98$, $p < 0.0297$ ($\eta_p^2 = 0.47$, small effect).

Discussion

The results do suggest the possibility that elementary-aged students can forge direct cognitive links between L2 labels and mental concepts. The faster speed of picture-matching across the board is highly suggestive of such. If the students were required to connect L2 words to the concept via L1 translation, we would see a slow-down in picture-matching, much as early testing on the *word association* model and the *concept mediation* model found

in picture naming among low-level students (Potter et al., 1984). The truly potentially surprising aspect of this study is that such robust acceleration in picture-matching vs. L1-matching was found only one day after the L2 label entering into the students' receptive vocabularies. It is possible, albeit entirely speculative at the moment, that if conditions allowed longitudinal instruction, such effects would likely have been even larger. It is important to note that some of the slow-down in L1 translation could possibly be due to latencies in reading speeds. Studies with young learners always face certain limitations due to their individual cognitive development levels; however, the fact that the latencies between picture-matching and L1-matching remained significant through 6th grade is suggestive that reading speed was not a critical factor (and early calibration efforts of the testing materials abandoned testing with students under 4th grade for precisely this reason – reading speeds were so slow with some learners as to make the comparison between categories invalid).

In addition to showing that elementary school-aged learners have the ability to bypass the RHM constraints by accessing mental concepts via L2 labels immediately after vocabulary acquisition, the study results also seem to support our earlier hypothesis that teaching methodology may have impacted the degree of latency between picture-matching and L1-matching. The significant item effects across all grades demonstrate that those taught with pictures are significantly faster at connecting pictures to L2 labels, compared with those taught via L1-translation. The reverse was not shown to be the case (i.e., those taught via translations were not significantly faster at matching L1-translations), so this is not simply a teaching effect, but instead, instruction using visual illustrations seems to better reinforce the conceptual links created during vocabulary acquisition, thus permitting faster recall, and may even be helping with translation between the L1 and L2 (as evidenced by the lack of significant advantage for L1-translation by the group taught via explicit use of L1).

From pedagogical points of view, the current study can suggest some important implications. One of such is benefits of utilizing images in teaching vocabulary to young learners. As the data in the present study suggested, the participants who were taught with images were significantly faster when they were matching images with L2 translation when compared with those who were taught the same set of vocabulary items by way of Japanese translations. It became evident that teaching with visuals appeared to have facilitated the conceptual access, which was demonstrated by the higher speed of recall in the dataset. Therefore, it could be suggested that using pictures to teach new

L2 terms, rather than L2 translation, could produce better results of learning in teaching L2 to young learners because using visuals seems to have facilitated establishment of direct conceptual access in learners.

Another pedagogical implication of the present study might be related to age and developmental factors. The younger the participants were, the more robust effects were found in the study. This could generally suggest that it might be useful for elementary school teachers in charge of third- and fourth-graders to make more use of pictures when teaching vocabulary words with concrete meaning (e.g., a dustpan), rather than those with abstract (less concrete) meaning (e.g., love). The results of the present study discovered the significant effects of use of images in teaching vocabulary; however, the target vocabulary words taught through the instructions were chosen carefully to avoid any misunderstanding of the target word meanings, specifically selecting rather concrete words which are found easy to be understood only with an image. Thus, future research could look into the effects of image usage in teaching more abstract vocabularies to confirm effectiveness of use of visuals over L1 translation in teaching L2 vocabulary in general.

While the study did not find a "cut-off" developmental period for conceptual access, the gradual decline in significance of teaching condition comparisons as students age very well may be indicative of an approaching point where conceptual access can no longer be achieved in the short-term. Finding an absolute point where conceptual access is no longer a factor and the RHM is in full effect will likely require extension of study in junior high school (or even high school) groups.

Conclusion

Japan made a significant change in the age when school children begin to learn English as a second language as part of the official school curriculum. At present, 3rd- and 4th-graders take a 45-minute lesson per week (in total 35 hours per grade year), and for 5th and 6th graders, English is one of the official school subjects in elementary schools (70 hours per grade year in total), as MEXT (2017) stipulates. This shift in the elementary school curriculum has resulted in the dramatic change in the junior and high school English curricula especially with the significant increase in the vocabulary size to be acquired before starting the 10th grade. With all these drastic changes being made, it is imperative that the English language educators develop effective pedagogical methods in terms of overall language instructions as well as vocabulary teaching strategies to young learners of English. Understanding how the young learner's brain functions in acquiring new words is an important step

forward to finding best methodologies of teaching L2. One of the possible ways for such investigation could be to test the RHM. Since not many studies have investigated the RHM with Japanese learners of English, more research could be conducted to determine most appropriate methods of teaching English vocabulary to Japanese learners with various proficiency.

This study aimed to provide evidence of elementary-aged students creating conceptual links to new L2 vocabulary, though the exact developmental period when the capability of utilizing direct links from L2 words to concept may cease (thus introducing the RHM dynamic whereby beginning adult learners are incapable of linking concepts to L2 vocabulary) was not clarified. Thus, more research is required to ascertain the present study findings. It was found that young learners of English seem to be capable of forming direct links between L2 vocabulary and concepts, but this ability disappears at a certain stage later in life and it is still not clear exactly at what age the ability ceases. Figuring out the exact age would be invaluable toward enabling both young learners and their teachers to determine maximally appropriate methods for teaching and learning new vocabulary words in elementary school contexts (and possibly even in junior high).

Another point of suggestion could be related to tools of testing concept mediation. Accuracy or error rate was not accounted into in the current study, and there might be some impact on the results, though elementary-aged children would have difficulty avoiding guessing answers or making mistakes in choosing answers since they are not used to using the computer keyboard in general. Therefore, it would be beneficial to create a testing tool which would make the testing of L1, L2, and concept linkage feasible for and more easily accessible to young participants.

In researching vocabulary acquisition, the "distance" between L1 and L2 can be considered one important factor to be accounted for since L2 being distant from L1 (e.g., Japanese L1 speakers learning English) has been found more challenging and thus present more difficulty in acquiring L2 than L2 being close to L1 (e.g., Spanish L1 speakers learning English). Crystal (1987) defined interlingual distance to be "[t]he structural closeness of languages to each other" (p. 371). Since then, it has been well established that L2 being distant from L1 (e.g., Japanese L1 speakers learning English) might be found more challenging and thus present more difficulty in acquiring L2 than L2 being close to L1. Burrows (2012) suggested that the language distance between English and Japanese might have been one of the crucial factors for Japanese learners of English having difficulties in learning English. Similarly, it might be interesting to investigate into how learners in English as a Foreign Language

(EFL) situations and those in English as a Second Language (ESL) situations might differ in terms of building vocabulary conceptual access. Consequently, similar studies to the present study can examine such contrasts among those with different L1 backgrounds and also in the situation where English is the learner's dominant language (e.g., in international school settings) so that most appropriate vocabulary teaching methodologies and strategies can be discovered, developed, and utilized in schools for maximized benefit.

In conclusion, the present study succeeded in garnering evidence that young learners of English could forge direct conceptual access to L2. However, the question of when the child ability to do so "switches off" remains unanswered, as the RHM model shows that adult learners are unable to directly connect mental concepts to L2 vocabulary at beginning stages of learning. This study can suggest future research directions in order to further knowledge of how young learners acquire L2 vocabulary.

Acknowledgements

The research project described in this paper was funded by a grant from the JSPS Grant-in-Aid for Scientific Research (C) (17K03013).

Clay Williams is a professor in the English Language Teaching Practices program of the Graduate School of Global Communication and Language at Akita International University. His research interests include cross-script word recognition, literacy acquisition, and online and virtual reality technology integration into L2 acquisition.

Naeko Naganuma is an associate professor of the English for Academic Purposes (EAP) program at Akita International University. Her research interests include integration of well-being education into curricula, teaching reading and vocabulary with technology, and use of self-reflection in classrooms.

Appendices

All appendices are available from the online version of this article at https://jalt-publications.org/jj.

References

Anezaki, T. (2006). Exploring the developmental shift of translation recognition: Japanese junior high school students. *ARELE: Annual Review of English Language Education in Japan, 17*, 121–130. https://doi.org/10.20581/arele.17.0_121

Bloom, P., & Markson, L. (1998). Capacities underlying word learning. *Trends in Cognitive Sciences, 2*(2), 67–73. https://doi.org/10.1016/S1364-6613(98)01121-8

Brysbaert, M., & Duyck, W. (2010). Is it time to leave behind the revised hierarchical model of bilingual language processing after fifteen years of service? *Bilingualism: Language and Cognition, 13*(3), 359–371. https://doi.org/10.1017/S1366728909990344

Burrows, C. (2012). English linguistics: Difficulties in English for Japanese EFL Learners. *Bulletin of International Pacific University, 6*, 73–76. https://doi.org/10.24767/00000331

Comesaña, M., Perea, M., Piñeiro, A., & Fraga, I. (2009). Vocabulary teaching strategies and conceptual representations of words in L2 in children: Evidence with novice learners. *Journal of Experimental Child Psychology, 104*(1), 22–33. https://doi.org/10.1016/j.jecp.2008.10.004

Comesaña, M., Soares, A. P., Sánchez-Casas, R., & Lima, C. (2012). Lexical and semantic representations of L2 cognate and noncognate words acquisition in children: Evidence from two learning methods. *British Journal of Psychology, 103*(3), 378–392. https://doi.org/10.1111/j.2044-8295.2011.02080.x

Crystal, D. (1987). *The Cambridge encyclopedia of language.* Cambridge University Press.

de Groot. A. M. B., Dannenburg, L., & van Hell, J. G. (1994). Forward and backward word translation by bilinguals. *Journal of Memory and Language, 33*(5), 600–629. https://doi.org/10.1006/jmla.1994.1029

Dijkstra, A., & van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition, 5*(3), 175–197. https://doi.org/10.1017/S1366728902003012

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers, 35,* 116–124. https://doi.org/10.3758/BF03195503

Habuchi, Y. (2003). Word processing in cross-language translation between Japanese and English by advanced second-language learners: A test of the revised hierarchical model. *Japanese Journal of Educational Psychology, 51*(1), 65–75. https://doi.org/10.5926/jjep1953.51.1_65

Heredia, R. R. (1997). Bilingual memory and hierarchical models: A case for language dominance. *Current Directions in Psychological Science, 6*(2), 34–39. https://doi.org/10.1111/14678721.ep11512617

Kawakami, A. (1994). The effect of proficiency in a second language on lexical-conceptual representation. *Japanese Journal of Psychology, 64,* 426–433. https://doi.org/10.4992/jjpsy.64.426

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual and memory representations. *Journal of Memory and Language, 33*(2), 149–174. https://doi.org/10.1006/jmla.1994.1008

Kroll, J. F., van Hell, J. G., Tokowicz, N., & Green, D. W. (2010). The revised hierarchical model: A critical review and assessment. *Bilingualism: Language and Cognition, 13*(3), 373–381. https://doi.org/10.1017/S136672891000009X

Linck, J. A., Kroll, J. F., & Sunderman, G. (2009). Losing access to the native language while immersed in a second language evidence for the role of inhibition in second-language learning. *Psychological Science, 20*(12), 1507–1515. https://doi.org/10.1111/j.1467-9280.2009.02480.x

Ministry of Education, Culture, Sports, Science and Technology. (2017). *Shōgakkō gakushū shidō yōryō (Heisei 29-nen kokuji) kaisetsu gaikokugokatsudō gaikokugo-hen* [Foreign language activities and foreign language edition of the courses of study for elementary schools]. https://www.mext.go.jp/content/20220614-mxt_kyoiku02-100002607_11.pdf

Nakagawa, C. (2009). Examination of the developmental hypothesis on the revised hierarchical model. *ARELE: Annual Review of English Language Education in Japan, 20*, 121–130. https://doi.org/10.20581/arele.20.0_121

Nakamura, T. (2007). *Tango no ninchi ni okeru gainen shohyo* [Conceptual representation in word recognition]. *STEP Bulletin, 19,* 23–40.

Poarch, G. J., van Hell, J. G., & Kroll, J. F. (2015). Accessing word meaning in beginning second language learners: Lexical or conceptual meditation? *Bilingualism: Language and Cognition 18*(3), 357–371. https://doi.org/10.1017/S1366728914000558

Potter, M. C., So, K.-F., von Eckardt, B., & Feldman, L. B. (1984). Lexical and conceptual representation in beginning and more proficient bilinguals. *Journal of Verbal Learning and Verbal Behavior, 23*(1), 23–38. https://doi.org/10.1016/S0022-5371(84)90489-4

Sheng, L., Bedore, L. M., Peña, E. D., & Fiestas, C. (2013). Semantic development in Spanish-English bilingual children: Effects of age and language experience. *Child Development, 84*(3), 1034–1045. https://doi.org/10.1111/cdev.12015

Sholl, A., Sankaranarayanan, A., & Kroll, J. F. (1995). Transfer between picture naming and translation: A test of asymmetries in bilingual memory. *Psychological Science, 6*(1), 45–49. https://doi.org/10.1111/j.1467-9280.1995.tb00303.x

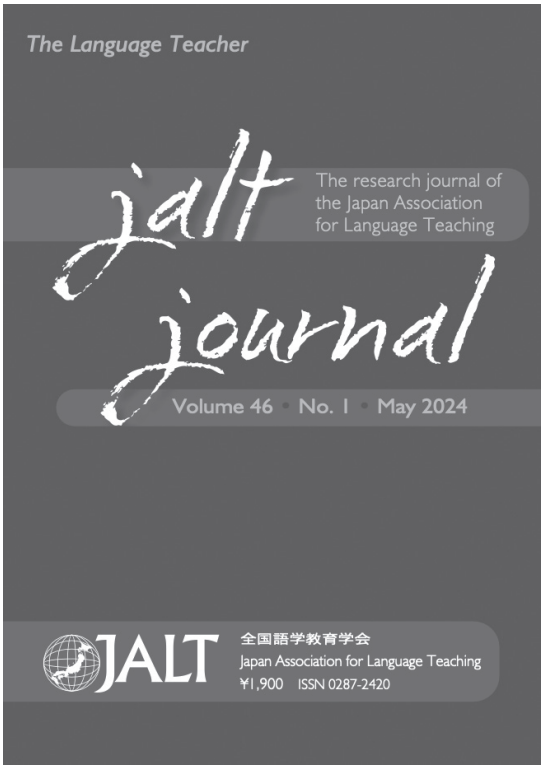
Sunderman, G., & Kroll, J. F. (2006). First language activation during second language lexical processing: An investigation of lexical form, meaning, and grammatical class. *Studies in Second Language Acquisition, 28*(3), 387–422. https://doi.org/10.1017/S0272263106060177

van Hell, J. G., & de Groot, A. M. B. (1998). Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. *Bilingualism: Language and Cognition, 1*(3), 193–211. https://doi.org/10.1017/S1366728998000352

Williams, C. (2017). Word-level decoding mechanisms for Japanese and Chinese L2 English learners: L1 interference effects. *Studies in English Language and Literature, 39,* 51–76.

Williams, C. (2018). Word recognition and semantic processing by Japanese English learners. In R. Ruegg & C. Williams (Eds.), *Teaching English for academic purposes in Japan: Studies from an English medium university* (pp. 59–75). Springer. https://doi.org/10.1007/978-981-10-8264-1_4

Wu, Z., & Juffs, A. (2019). Revisiting the revised hierarchical model: Evidence for concept mediation in backward translation. *Bilingualism: Language and Cognition, 22*(2), 285–299. https://doi.org/10.1017/S1366728917000748



JALT Journal Online

https://jalt-publications.org/jj

46 Volumes
85 Issues
Hundreds of articles!
1979~2024

Expositions

Generative Artificial Intelligence and Applied Linguistics

Sowmya Vajjala
National Research Council, Canada

Since the advent of ChatGPT in November 2022, there has been a growing interest and widespread speculation on how Artificial Intelligence (AI), more specifically generative AI, has the potential to revolutionize research and applications across disciplines, with real-world implications. Applied linguistics researchers and practitioners have long adapted to the use of technology in language learning and teaching, where AI already plays a role in the form of Natural Language Processing (NLP), Machine Learning (ML), and other related technologies. This *Expositions* article introduces generative AI, explains how it works and what distinguishes it from other AI technologies, and discusses its growing influence in the applications relevant to applied linguists. The article concludes with some guidance on how to navigate the generative AI space as an applied linguist while acknowledging the current limitations, including how to use generative AI in research and practice.

2022年11月のChatGPTの登場以来、人工知能（AI）、より具体的には生成AIが、現実の世界にどのような影響を及ぼし、分野横断的な研究や応用に革命をもたらす可能性があるのか、関心が高まり、さまざまな憶測が現在広がっている。応用言語学の研究者や実践者は、自然言語処理（NLP）、機械学習（ML）、その他の関連技術の形でAIがすでに役割を果たしている言語学習や教育における技術の使用に、長い間適応してきた。このExpositionsの論文では、生成AIを紹介し、その仕組みと他のAI技術との違いを説明し、応用言語学にどのように影響をもたらすのかについて論じている。最後には、応用言語学者として、現在の限界を認識しながら、生成AIをどのように研究・実践に活用するかなど、生成AIの空間をどのようにうまく使っていけるかを紹介している。

https://doi.org/10.37546/JALTJJ46.1-3

JALT Journal, Vol. 46, No. 1, May 2024

Keywords: educational technologies; generative AI; natural language processing; speech recognition

Artificial Intelligence technologies in the form of Natural Language Processing (NLP) and Machine Learning (ML) have been used in Intelligent Computer Assisted Language Learning (ICALL), particularly in the development of support tools for reading, writing, speaking, or listening, for intelligent tutoring systems and automated assessment (Heift, 2012). Software applications targeting teachers as well as students that rely on such technologies have been developed and researched for almost two decades now. The potential of AI technologies in building support tools for teachers to select course materials (Brown & Eskenazi, 2004; Sheehan et al., 2014), automated grading of written and spoken language (Burstein et al., 2013; Chen et al., 2018), and automated creation of questions and assessment items (Chinkina & Meurers, 2017) have all been well explored in the past. These technologies have also been employed to build language learner support tools for writing (Madnani et al., 2018) and speaking (Kheir et al., 2023). The availability of language learning mobile apps such as Duolingo (duolingo.com), general purpose writing assistants such as Grammarly (grammarly.com), pronunciation and speaking apps such as elsaspeak (elsaspeak.com) are examples of how this strand of research evolved into practical everyday tools. The use of AI in language learning and teaching can thus be considered an active and established area of research and practice. Vajjala (2018) and Meurers (2021) give an overview of some of the research on the role of machine learning and natural language processing respectively in language learning and teaching.

ChatGPT (https://chat.openai.com/) was released as an open-access web tool in November 2022 and has since played an important role in the discussion around the applications of artificial intelligence, more specifically, generative artificial intelligence in various areas. Within the realm of education, it has been utilized across a range of subject areas such as medical education (Kung et al., 2023; Tsang, 2023), computing education (Denny et al., 2023), and science education (Cooper, 2023). Although there was an initial wave of skepticism and a call to ban the use of such tools in education contexts, there are now calls to think about ways to incorporate them into education policies. The New York City public schools AI policy lab is an example of such an initiative (Klein, 2023). Khan Academy (https://www.khanacademy.org/), an online education provider, announced Khanmigo, an AI-powered teaching assistant, earlier this year, leveraging GPT-4, ChatGPT's successor

(Khan Academy, 2023). In the language learning space, applications such as Duolingo (Duolingo, 2023) and Grammarly (Grammarly, 2023) quickly moved towards incorporating generative AI into their existing software applications. The rapid adaptation of this new technology into these various areas of educational technologies indicates its importance for educational technologies in general, and language learning technologies in particular.

Considering that AI-based technologies have already been used in language learning and teaching for some time now, what new things does generative AI bring into the picture? Does it just do existing things better, or does it enable new possibilities? This *Expositions* article explores the role of generative AI in language learning and teaching technologies and addresses the following questions:

1. What is generative AI and how does it work?
2. How does it impact the technologies related to language learning and teaching?
3. How should one work with generative AI, as an applied linguist?
4. What are some limitations of generative AI, and caveats to working with it?

The target audience is expected to be primarily applied linguists familiar with the use of language technologies and artificial intelligence in the context of language learning and teaching, and interested in knowing more about how the recent developments in generative AI are useful for research and practice in this area. The next four sections address the four questions listed above, respectively.

Here is a quick note on the terminology before diving in: While discussing what AI systems can and cannot do, it is common to use words such as "learning", "understanding", "reasoning" etc. Such words are only used in a metaphorical sense, for easier comprehension, and there are no parallels with human learning/understanding/reasoning abilities. Readers are advised to not conflate machine processes with human processes as they explore this article further.

Generative AI—An Overview

The ultimate goal of any AI system is to achieve a semblance of human-like intelligence in the tasks it is expected to perform. There are many ways of achieving this goal, from using hard-coded rule-based reasoning to learning to perform different tasks in a data-driven manner, from a large volume of exam-

ples, without explicit specification of rules. *Generative AI* refers to the form of AI that is capable of processing and generating new content for a range of input/output forms (e.g., text, image, audio, video, a combination of these et cetera). Generative AI models today are responsible for creating human-like texts, realistic images and videos, and natural-sounding audio. *Deep learning*, a form of data-driven learning based on artificial neural networks, is the force behind all the recent developments in generative AI. There are many different forms of generative models for processing different forms of data, and some of these models can also learn multimodally i.e., working with different forms of input or output at the same time. In this article, we will focus on one type of generative AI that is more relevant to our context - *Large Language Models* (LLMs).

Language models learn to assign probabilities to a sequence of words (Jurafsky & Martin, 2023; Chapter 3). They learn the probabilities of word sequences by using the frequency information from large amounts of textual data. Readers familiar with the use of word concordance models in corpus linguistics may be familiar with this approach. Language models go further and use that knowledge to predict probabilities for future sequences, which can then be used to perform a range of language processing tasks, from text classification to machine translation. *Neural language models*, based on artificial neural networks, use massive amounts of textual data to learn these probabilities. Such massive data is available in many languages in the form of web texts, Wikipedia dumps, and other such sources. This process of learning the probabilities is known as "pre-training". Performing pre-training on increasing amounts of textual data from various sources resulted in more and more powerful language models over the past five years, since the arrival of the BERT language model a few years ago (Devlin et al., 2018). A pre-trained language model that passed through this process with massive amounts of generic text data can be further "fine-tuned" with smaller amounts of task-specific data to perform specific tasks (e.g., question answering, machine translation et cetera), by a process known as "transfer learning" (Howard & Ruder, 2018).

Autoregressive language models are a form of neural language models that undergo pre-training by repeatedly predicting the next token given a sequence of tokens. A token can be understood as a machine equivalent of a word. Note that what humans understand as one word is considered to be composed of multiple tokens by a neural language model. For example, consider this sentence - "Sara vociferously denied to comment". A traditional NLP system may split it linguistically and identify six tokens [sara, vociferously, denied, to, comment] in this sentence, like a human would perhaps do. However, GPT-4 splits it into 9 tokens instead, as ['S', 'ara', ' voc', 'ifer', 'ously', ' denied', ' to',

‘ comment’, ‘.’]. The tokens are not necessarily morphologically meaningful, and this tokenization is machine-learned by processing word patterns in the data, to create a finite vocabulary for the language model.

The task of next token prediction may seem like a simple task from a layperson’s perspective. Yet, it forms the foundation for all the modern-day LLMs, as many NLP tasks can be framed as text completion tasks, laying the foundations for a generative language model. For example, if one gives an input “What is the capital of Canada?”, a pre-trained LLM can respond with “Ottawa” as an answer. As the amount of pre-training data increased, the models became capable of learning to perform a task based on a description, with very few or no examples, without requiring any explicit further fine-tuning. *GPT-3* (Brown et al., 2020), an LLM developed by OpenAI and trained on half a trillion tokens, is an example of such a general-purpose LLM. Today’s LLMs (such as ChatGPT) follow this autoregressive approach to text generation and show some ability to process human input and generate an appropriate output for a given input from a human user, in a human language. The current generation of LLMs can also generate natural-sounding text following human instructions. Development of new techniques to improve over what a language model “learns” during pre-training resulted in the latest generative large language models we see today, such as ChatGPT, GPT-4 (OpenAI, 2023), Gemini (Gemini Team, 2023) and Claude (Anthropic, 2023). In addition to such commercial LLMs, a wide range of non-commercial, open-source alternatives, such as Zephyr (Tunstall et al., 2023), Falcon (Almazourei et al., 2023), and LLaMa2 (Touvron et al., 2023), to name a few, are other alternatives. There is also a growing body of work on developing small, focused language models (e.g., Li et al., 2023; Zhang et al., 2024) that are good at reasoning from data and performing tasks that require some form of natural language understanding. The generative LLMs mentioned here are only a few examples, and the readers are suggested to refer to Zhao et al. (2023) for a detailed listing of LLMs.

Two key ideas that made large language models go from models such as BERT to systems like ChatGPT are *Supervised Fine-Tuning* (*SFT*) and *Reinforcement Learning with Human Feedback* (*RLHF*), both of which involve a large number of human annotators. In SFT, the LLM is taught to follow instructions for different use cases (e.g., machine translation, text classification, chat, writing a short story, et cetera), by providing task descriptions along with example items and soliciting responses from humans for a large data sample. This data is then used to fine-tune and optimize the original pre-trained LLM to perform diverse tasks. For a given prompt, many outcomes are possible from a language model, considering that the output generation

process is probabilistic. Which is the most preferred by human users? If a human user ranks a set of responses by an LLM for a given prompt in terms of how good they are, can a model learn to generate "good" responses? RLHF is the technique that addresses this question by learning a "reward model" and optimizing an LLM to generate responses that align with human preferences. The data to learn such a reward model is again collected on a large scale by setting up an annotation task where humans choose a preferred output from the given machine responses. InstructGPT (Ouyang et al., 2022), a generative language model from OpenAI which is a predecessor of ChatGPT, and GPT4, was among the first to describe this approach, which soon became a standard procedure for building large generative AI models.

Any computer system built for a specific purpose can be evaluated on how it performs on specific tasks that achieve that purpose, and machine-learned systems are no exception. However, how should we evaluate Generative AI systems, more specifically, LLMs such as ChatGPT? This is an ongoing and active area of research, and the current practices include evaluating LLMs on popular benchmarks that cover multiple tasks and languages as well as other aspects such as toxicity and harmfulness. Note that there are several *LLM evaluation benchmarks*, and there is no single LLM that performs the best on all the benchmarks. A *public leaderboard* offers a quick lookup of how different LLMs compare against each other on various benchmarks (Huggingface.co, 2023). Liu et al. (2023) and Guo et al. (2023) present comprehensive surveys on the evaluation of large language models. Note that the performance on such standard evaluation benchmarks should not be equated to real-world performance in a given application scenario and it is possible for an LLM to do well on such benchmarks but not be useful for a given real-world task.

There is much more to LLMs and generative AI than what was presented so far, and this only aimed to provide a short overview of what generative AI is, how it differs from other forms of AI, and how generative LLMs such as ChatGPT are built, trained, and evaluated. For a more comprehensive discussion about the topic, refer to Jurafsky and Martin (2023). For a contemporary introduction to the artificial neural network models that power modern generative AI, refer to Prince (2023). With this introduction to what generative AI is, let us now turn to how it is impacting the language learning and technology space.

Impact of Generative AI on Language Learning Technology

The past year witnessed the impact of generative AI in a range of disciplines that were not already adapted to AI in general. Hence, it is natural that edu-

cational technologies, that have already adapted AI across many applications, were impacted by generative AI. Some applications such as providing reading/writing/speaking support for learners or teaching support in the form of grading and creating assessment items have improved, and others that were previously considered too specific, such as providing personalized, explicit feedback, are now enabled by these new advances. There is also a huge potential for previously under-explored use cases for AI such as helping teachers with lesson planning or for multimodal content generation. Recent research on the use of generative AI, more specifically large language models, in language learning technologies can perhaps be grouped into three categories: content and test generation, assessment, and assistive tool development. Let us take a closer look at each of them below:

Test item generation: Generation of diverse, high-quality questions from a given content, adhering to a given criteria, can reduce the teachers' workload while increasing content quality. It is also useful in the development of intelligent tutoring systems. NLP techniques have been used for various forms of automated question generation in the past, ranging from fill-in-the-blank and multiple-choice questions to generating open-ended questions. Recent research discussed the utility of large language models for question item generation for English and Swedish texts (Elkins et al., 2023; Goran & Abed Bariche, 2023). Other research also showed how ChatGPT can be useful in generating questions for assessing English reading comprehension (Lee et al., 2023; Shin & Lee, 2023). Human validation studies were conducted in all these studies to verify the usefulness of machine-generated questions. Going a step further, Xiao et al. (2023) demonstrate the usage of ChatGPT for both reading text generation as well as exercise generation for English reading comprehension. They also report an evaluation study with Chinese middle school teachers who concluded the generated texts and exercises to be appropriate for their students.

Assessment: Assessment is another area in which the important application of Natural Language Processing and AI for language learning and technology has been investigated. Automated scoring of essays for language proficiency or short answers for content accuracy has been well-studied in the literature. Over the past year, some work in the NLP community has explored the usefulness of generative AI models for this purpose. Naismith et al. (2023) show the use of GPT4 in automatic writing evaluation for discourse coherence. Their research showed that GPT4's ratings correlate well with human evaluations, and GPT4 performance is better than a linguistic feature-based model baseline for the dataset under consideration. Further, the GPT4 response can

be accompanied by rationales for the evaluations, if necessary. Note that the "rationales" are generated by the model, and need not necessarily align with a human evaluator's rationales.

In contrast to Naismith et al. (2023), another recent work evaluating the ability of GPT3.5 and GPT4's ability to rate short essays on the CEFR scale (Yancey et al., 2023) showed that although GPT4 performs on par with existing approaches when calibration examples are provided in the prompt, agreement with human ratings vary depending on the test taker's first language. Another recent work by Mizumoto and Eguchi (2023) shows that a GPT-based LLM model combined with linguistic feature information performs better than just using an LLM by itself. One major concern with using some inherently opaque large and complex models is the lack of interpretability and explainability of their predictions. Fiacco et al. (2023) developed a method to extract and understand the implicit rubrics of such neural network models when used as essay scorers. Even though this discussion is not exhaustive, it clearly shows the adaptation of generative AI and LLMs into automated language assessment research, and we could expect more practical utilities in the coming years.

Support tools for language learners: Davis et al. (2024) present a comprehensive evaluation of both open-source and proprietary LLMs for (English) Grammatical Error Correction tasks and show that they do not always outperform custom-built machine learning models for the task when used as-is. However, the quick adaptation to generative AI by language learning and writing support software such as Duolingo and Grammarly, which was discussed earlier, clearly points to the value these technologies bring to language learners when customized to the task. Beyond a language learner context, Speakerly (Kumar et al., 2023), a new language learning platform by Grammarly, shows how large language models and speech recognition can be integrated to build a voice-based writing assistant. Raheja et al. (2023) explored instruction tuning, which was described in Section 2, to build a text editing system for writing assistance. Expanding the horizons beyond the commonly seen applications of NLP in the development of such support tools, emerging research has begun to investigate using generative large language models for grammatical error correction beyond English (Kwon et al., 2023). Duolingo (2023) discusses the use of LLMs for generating personalized feedback for learners. Kew et al (2023)'s recent work on benchmarking large language models for automatic text simplification shows that such generative AI models can assist in making texts easier to read for learners, by producing rephrased versions of the input text with simpler vocabulary and syntactic

structure.

The use of AI in most of the above-mentioned areas is an existing practice, which underwent considerable improvement with the new generative AI methods. Language technologies such as machine translation and chatbots too have been studied in the context of language learning and teaching in the past for quite some time (Freyer et al., 2020; Hellmich et al., 2023). However, the limitations of the technologies themselves resulted in their use being limited to research studies. Recent advances in neural network techniques improved the generative capability of NLP systems. Hence, we may see more research into the usefulness of such technologies in language learning research in the future (Huang et al., 2022; Tyen et al., 2022; Zhou et al., 2023).

New developments in generative AI can potentially enable new use cases too. Several recent studies (Kasneci et al., 2023; Yan et al., 2023; Yu & Guo, 2023) provide a broader overview of the potential applications and challenges of using generative AI technologies in various aspects of education (not specifically language education). Caines et al. (2023) take the specific case of language teaching and assessment technologies and discuss how generative AI technologies such as large language models can be used in novel ways for content generation, providing feedback, open-ended chatting at the level of a learner, providing document level assessment and feedback, and supporting "plurilingual" learning. Aryadoust et al. (2024) studied the use of LLMs for developing listening assessments targeting test takers at different proficiency levels and concluded that LLMs can be adopted at different stages of listening test development and validation.

Considering pronunciation training in particular, Kheir et al. (2023) predict that the advances in conversational capabilities of generative AI models, coupled with other developments in low-resource and end-to-end speech processing, may lead to the development of more sophisticated and personalized virtual tutors, and support multilingual applications for spoken language learning resources such as pronunciation tutors, which have been primarily English-focused so far (e.g., Ding et al., 2019; Thompson, 2012; Yonesaka, 2017). Asthana et al. (2023) describe an initiative to incorporate generative AI into a higher education course and study how automated generation of course metadata could support broader instructional goals. Matelsky et al. (2023) explore how large language models can be used to provide rapid personalized feedback to students for open-ended questions. The discussion in this article revolved around written or spoken texts, but we have to remember that language learning involves interaction between learners and a range of semiotic modes beyond printed or spoken texts. Future developments may

lead to the maturing of multimodal learning environments with text, images, audio, and other media integrated into the learning process using generative AI technologies.

Most of the developments discussed in this section so far show a high degree of interest in utilizing generative AI in the language learning and technology space. This interest and the push towards adopting generative AI into applied linguistics research and practice necessitates a discussion around the ethics of using generative AI in this context, particularly on how to use the technology appropriately and responsibly. How do applied linguists working in the language teaching and assessment context see the rise of these technologies so far?

There has been some discourse in this regard, particularly in language testing research. Summarizing the debate on allowing the use of assistive technologies including generative AI by test takers for language assessment, Voss et al. (2023) suggest that language teachers must have sufficient expertise to understand and integrate such technologies into their language instruction and assessment practice and recommend collaboration between test creators and AI developers for ensuring appropriate usage of assistive technologies. Taking a holistic perspective on the role of AI methods in the language testing and assessment process, Bolender et al. (2023) also recommend a collaboration between AI scientists, psychometricians, and subject matter experts to address issues around reliability, validity, and fairness in language test development. Another recent article by Xi (2023) echoes this strand of thought, emphasizing developing best practices for the ethical and responsible use of generative AI technologies specifically in the context of language testing. It is not surprising that the discussion around the responsible use of generative AI in this area started with language testing, as that can be considered as a high-stakes application scenario for generative AI compared to others such as the development of teaching and learning support tools.

Working with Generative AI

We've seen how recent advances in generative AI, especially with large language models, have improved upon existing use cases within the realm of language learning and technology, and how they opened pathways for potential new use cases that were not possible before. Kohnke et al. (2023) in a recent study on generative AI preparedness among university language instructors pointed to the need for tailored support for teachers to develop AI-related competencies. Some research recommends training both the faculty and the students about the effective use of these new technologies (Fuchs, 2023; Huallpa et al., 2023). With widespread speculation around how ubiquitous

generative AI would be in our personal and professional lives, how should applied linguists learn to work with generative AI? There are two ways:

Prompting: The most common means of interacting with such systems is through *prompting*. A prompt is similar to a "query" given to a search engine and can be understood as the input (including any instructions) to the AI describing the expected outcome. While having a natural language interface to generative AI systems is tempting to get started right away, creating proper prompts is more of an art than a science, and it would be useful to know some basics to get started. Saravia (2022) provides a comprehensive, constantly updated, collection of resources on prompting large language models. Understanding efficient and effective prompting methodologies could lead to applied linguists exploring the use of generative AI to pursue some of the prospective directions mentioned earlier, as well as add another tool to their research methods basket. Vee et al. (2023) compiled exercises to incorporate generative AI into the practice of teaching writing, which could serve as a useful resource for applied linguistics interested in pursuing this direction.

AI Coding Assistant: Another interesting possibility to work with generative AI as an applied linguist is by using it as a software coding assistant. Applied linguists, especially those who work on topics such as corpus linguistics or CALL have been learning to write software programs across universities. However, available teaching and learning material is not often geared towards students coming from a language teaching background, making learning challenging. The advent of generative AI-based assistants to write code over the past few years has shown promising results in its use in introductory programming classrooms (Porter & Zingaro, 2024; Puryear, 2022).

As for how generative AI is useful in applied linguistics research, the applications discussed in the previous section hopefully provide useful pointers in that direction. Most such research has been traditionally conducted on English language resources, considering the amount of available datasets and software support. The advent of large language models that have some form of knowledge about various languages provides an opportunity to explore them for other languages (e.g., in the Japanese as a Second Language context). The same applications (content generation, question generation, content assessment, learner support tools, etc.) can all be explored and the capabilities and limitations of current generative AI methods in a broader language learning and teaching technology context can be evaluated for other languages as well.

Let us turn to the question: What can applied linguistics contribute to the discourse around generative AI itself? With the widespread increase in both interest and adoption of generative AI technologies in various application

domains, there is also a lot of emerging discourse around the responsible usage of the technologies to ensure reliability and integrity. Note that this discussion is field-specific. For example, a discussion around the ethics of AI system development typically focuses on issues such as fairness and bias in the models, privacy concerns, explainability, and accountability. But when it comes to actually using such AI systems in, say, education, there are other (or additional) concerns such as the question of what is appropriate usage for a student who is learning a topic, or taking part in an assessment to evaluate their understanding. This is where the applied linguistics community can contribute to the general discourse around the ethics of generative AI usage.

A guideline on the ethical usage of generative AI in language teaching, learning, and testing (and more broadly, encompassing other areas of applied linguistics) is needed considering the growing interest in the community on the topic. Yan et al. (2023) discuss ethical concerns around the use of AI broadly in the context of education, and Mohammad (2022) suggests an "ethics sheets" approach for different AI applications, listing the specific questions that need to be addressed, which can have different answers depending on the task at hand. Both these references are useful in thinking about developing guidelines for applied linguistics. The call for developing best practices in using generative AI for language testing (Xi, 2023) can be considered a starting point in this direction.

The annual state of AI reports published by the Montreal AI Ethics Institute (Gupta et al., 2023) are useful to give a broader perspective on various topics around AI ethics, for readers interested in exploring this aspect further. The EU AI Act (European Union, 2024) which proposes to regulate the development, deployment, and use of AI in the European Union region is another example of a broader discussion around addressing the ethical issues around AI and ensuring responsible development of technology.

Limitations and Caveats

Generative AI and its implications and applications are speculated upon and adopted widely, but this is not immune to challenges. Over the past year, researchers have widely discussed the technological as well as behavioral limitations of generative AI systems (see Kaddour et al., 2023 for a comprehensive discussion). Here are some limitations one needs to be aware of while using generative AI systems:

- Brittleness of the prompt-based querying process: Small changes in the prompts given to generative AI systems can sometimes result in

drastic changes in output, which pose problems in terms of reliability and reproducibility of the process.
- Hallucinations: Generative AI systems such as large language models can produce potentially inaccurate, and at times, completely false information, which may be hard to detect, as the text itself is highly fluent.

Although the above-mentioned limitations arise from the working of the systems themselves, the abilities of these systems, along with their ubiquity now, pose two other problems:

- Distinguishing between machine and human-generated output is sometimes difficult owing to the fluency and human-like text patterns. However, there is some ongoing research into watermarking AI-generated output, which can potentially help address such issues in the future.
- Access to such AI systems could potentially compromise the integrity of computer-based testing scenarios, as some recent research showed (de Winter, 2023). Research into alternative formats of assessment may help overcome the challenges that arise out of this issue.

With existing limitations and the potential problems that may arise from the use of these technologies, should we prohibit their use until some solutions have been found? Current discussion in the research community instead suggests acknowledging the ubiquity of generative AI today, and adapting the teaching and evaluation approaches accordingly (Yu, 2023). Finally, it has to be noted that these limitations and caveats reflect the current state-of-the-art, and the mitigation of such issues is currently an active area of research. Thus, we could expect future research to develop new systems that can overcome such challenges, as far as the technology itself is concerned. However, responsible and ethical use of any technology needs to be separately addressed for a given application scenario, irrespective of how good or advanced the technology is. The guidelines on the responsible use of generative AI should be field-specific, and application-specific, and developing more specific guidance on the use of generative AI covering different topics in applied linguistics would be a worthwhile direction to pursue, as the adoption of these technologies increases.

Summary

In this *Expositions* article, I aimed to give a broader overview of generative AI and its implications for applied linguistics researchers and practitioners. In doing so, I attempted to summarize recent research on generative AI in areas related to applied linguistics from the Natural Language Processing community, as well as the perspectives from applied linguistics research and practice. Some guidelines were provided for applied linguists who are interested in getting started with generative AI technologies, and potentially new research directions were identified. Generative AI was described as an active research area with a blurring divide between research and practice today. Hence, it is important to be aware of its current limitations and potential issues that may arise, and I have provided some guidance in that direction. The capabilities of current generative AI methods open up new avenues for applied linguists, and I hope this article serves as a starting point for a deeper exploration of these technologies and their relevance to the field.

Sowmya Vajjala works as a Natural Language Processing (NLP) researcher at National Research Council, Canada's largest federal research and development organization. Her research interests lie in information extraction from text, multilingual modeling, and studying the relevance of NLP in other disciplines. She co-authored a book: "Practical Natural Language Processing: A Comprehensive Guide to Building Real World NLP Systems", published by O'Reilly Media, which was also translated into Japanese as "実践 自然言語処理 ―実世界NLPアプリケーション開発のベストプラクティス"in 2023.

References

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Hesslow, D., Launay, J., Malartic, Q., Mazzotta, D., Noune, B., Pannier, B., & Penedo, G. (2023). *The Falcon series of open language models* [Preprint]. arXiv. https://doi.org/10.48550/arxiv.2311.16867

Anthropic. (2023, March 14). Introducing Claude. *Anthropic.* https://www.anthropic.com/index/introducing-claude

Aryadoust, V., Zakaria, A., & Yichen, J. (2024). Investigating the affordances of OpenAI's large language model in developing listening assessments. *Computers and Education: Artificial Intelligence, 6*, Article 100204. https://doi.org/10.1016/j.caeai.2024.100204

Asthana, S., Arif, T., & Thompson, K. C. (2023, December 15). *Field experiences and reflections on using LLMs to generate comprehensive lecture metadata* [Workshop]. Generative AI for Education (GAIED): Advances, Opportunities, and Challenges, San Diego, CA, USA. https://gaied.org/neurips2023/files/31/31_paper.pdf

Bolender, B., Foster, C., & Vispoel, S. (2023). The criticality of implementing principled design when using AI technologies in test development. *Language Assessment Quarterly, 20*(4–5), 512–519. https://doi.org/10.1080/15434303.2023.2288266

Brown, J., & Eskenazi, M. (2004). Retrieval of authentic documents for reader-specific lexical practice. In R. Delmonte, P. Delcloque, & S. Tonelli (Eds.), *Proceedings of InSTIL/ICALL2004 – NLP and speech technologies in advanced language learning systems.* Unipress.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33,* 1877–1901.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 55–67). Routledge/Taylor & Francis.

Caines, A., Benedetto, L., Taslimipoor, S., Davis, C., Gao, Y., Andersen, Ø., Yuan, Z., Eilliott, M., Moore, R., Bryant, C., Rei, M., Yannakoudakis, H., Mullooly, A., Nicholls, D., & Buttery, P. (2023). *On the application of large language models for language teaching and assessment technology* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2307.08393

Chen, L., Zechner, K., Yoon, S. Y., Evanini, K., Wang, X., Loukina, A., Tao, J., Davis, L., Lee, C. M., Ma, M., Mundkowsky, R., Lu, C., Leon, C. W., & Gyawali, B. (2018). Automated scoring of nonnative speech using the speechrater sm v. 5.0 engine. *ETS Research Report Series, 2018*(1), 1-31. https://doi.org/10.1002/ets2.12198

Chinkina, M., & Meurers, D. (2017). Question generation for language learning: From ensuring texts are read to supporting learning. In J. Tetreault, J. Burstein, C. Leacock, & H. Yannakoudakis (Eds.), *Proceedings of the 12th workshop on innovative use of NLP for building educational applications* (pp. 334–344). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-5038

Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology, 32*(3), 444–452. https://doi.org/10.1007/s10956-023-10039-y

Davis, C., Caines, A., Andersen, Ø., Taslimipoor, S., Yannakoudakis, H., Yuan, Z., Byrant, C., Rei, M., & Buttery, P. (2024). *Prompting open-source and commercial language models for grammatical error correction of English learner text* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2401.07702

de Winter, J. C. F. (2023). Can ChatGPT pass high school exams on English language comprehension? *International Journal of Artificial Intelligence in Education.* https://doi.org/10.1007/s40593-023-00372-z

Denny, P., Prather, J., Becker, B. A., Finnie-Ansley, J., Hellas, A., Leinonen, J., Luxton-Reilly, A., Reeves, B. N., Santos, E. A., & Sarsa, S. (2023). *Computing education in the era of generative AI* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2306.02608

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Ding, S., Liberatore, C., Sonsaat, S., Lučić, I., Silpachai, A., Zhao, G., Chukharev-Hudilainen, E., Levis, J., & Gutierrez-Osuna, R. (2019). Golden speaker builder–An interactive tool for pronunciation training. *Speech Communication, 115,* 51–66. https://doi.org/10.1016/j.specom.2019.10.005

Duolingo. (2023, March 14). Introducing Duolingo Max, a learning experience powered by GPT-4. *Duolingo Blog*. https://blog.duolingo.com/duolingo-max/

Elkins, S., Kochmar, E., Serban, I., & Cheung, J. C. K. (2023). How useful are educational questions generated by large language models? In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Home artificial intelligence in education. Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky* (pp. 536–542). Springer. https://doi.org/10.1007/978-3-031-36336-8_83

European Union. (2024). *Artificial intelligence act.* https://artificialintelligenceact.com/

Fawzi, F., Amini, S., & Bulathwelaa, S. (2023). *Small generative language models for educational question generation* [Workshop]. Generative AI for Education (GAIED): Advances, Opportunities, and Challenges, San Diego, CA, USA. https://gaied.org/neurips2023/files/18/18_paper.pdf

Fiacco, J., Adamson, D., & Ros, C. (2023). Towards extracting and understanding the implicit rubrics of transformer based automatic essay scoring models. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)* (pp. 232–241). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.20

Fryer, L. K., Coniam, D., Carpenter, R., & Lăpușneanu, D. (2020). Bots for language learning now: Current and future directions. *Language Learning & Technology, 24*(2), 8–22. https://doi.org/10125/44719

Fuchs, K. (2023). Exploring the opportunities and challenges of NLP models in higher education: Is Chat GPT a blessing or a curse? *Frontiers in Education, 8*, Article 1166682. https://doi.org/10.3389/feduc.2023.1166682

Gemini Team (2023). *Gemini: A family of highly capable multimodal models*. DeepMind. https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf

Godwin-Jones, R. (2023). Emerging spaces for language learning: AI bots, ambient intelligence, and the metaverse. *Language Learning & Technology, 27*(2), 6–27. https://doi.org//10125/73501

Goran, R., & Abed Bariche, D. (2023). *Leveraging GPT-3 as a question generator in Swedish for high school teachers* [Bachelor's thesis, KTH Royal Institute of Technology]. Digital Scientific Archive. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1789082&dswid=6855

Grammarly. (2023, April 25). Grammarly brings personalized generative AI to your writing process*. Grammarly Blog*. https://www.grammarly.com/blog/grammarlygo-personalized-ai-writing/

Guo, Z., Jin, R., Liu, C., Huang, Y., Shi, D., Yu, L., ... & Xiong, D. (2023). *Evaluating large language models: A comprehensive survey* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2310.19736

Gupta, A., Wright, C., Bergamaschi Ganapini, M., Sweidan, M., & Butalid, R. (2022). *State of AI ethics report (Volume 6, February 2022)* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2202.07435

Heift, T. (2012). Intelligent computer-assisted language learning. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics.* Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal0548

Hellmich, E. A., & Vinall, K. (2023). Student use and instructor beliefs: Machine translation in language education. *Language Learning & Technology, 27*(1), 1–27. https://doi.org/10125/73525

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 328–339). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1031

Huallpa, J. J., Arocutipa, J. P. F., Panduro, W. D., Huete, L. C., Limo, F. A. F., Herrera, E. E., Callacna, R. A. A., Flores, V. A. A., Romero, M. Á. M., Quispe, I. M., & Hernández Hernández, F. A. (2023). Exploring the ethical considerations of using Chat GPT in university education. *Periodicals of Engineering and Natural Sciences, 11*(4), 105–115.

Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning, 38*(1), 237–257. https://doi.org/10.1111/jcal.12610

Huggingface.co. (2023). *Open LLM leaderboard*. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd ed). https://web.stanford.edu/~jurafsky/slp3/

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). *Challenges and applications of large language models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2307.10169

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103,* Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kew, T., Chi, A., Vásquez-Rodríguez, L., Agrawal, S., Aumiller, D., Alva-Manchego, F., & Shardlow, M. (2023). BLESS: Benchmarking large language models on sentence simplification. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on empirical methods in natural language processing* (pp. 13291–13309). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.821

Khan Academy. (2023, March 14). Harnessing GPT-4 so that all students benefit. A nonprofit approach for equal access. *Khan Academy Blog.* https://blog.khanacademy.org/harnessing-ai-so-that-all-students-benefit-a-nonprofit-approach-for-equal-access/

Kheir, Y., Ali, A., & Chowdhury, S. (2023). Automatic pronunciation assessment: A review. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 8304–8324). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.557

Klein, A. (2023, October 5). 180 degree turn: NYC district goes from banning ChatGPT to exploring AI's potential. *EducationWeek*. https://www.edweek.org/technology/180-degree-turn-nyc-schools-goes-from-banning-chatgpt-to-exploring-ais-potential/2023/10

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). Exploring generative artificial intelligence preparedness among university language instructors: A case study. *Computers and Education: Artificial Intelligence, 5*, Article 100156. https://doi.org/10.1016/j.caeai.2023.100156

Kumar, D., Raheja, V., Kaiser-Schatzlein, A., Perry, R., Joshi, A., Hugues-Nuger, J., Lou, S., & Chowdhury, N. (2023). Speakerly: A voice-based writing assistant for text composition. In M. Wang & I. Zitouni (Eds.), *Proceedings of the 2023 Conference on empirical methods in natural language processing: Industry track* (pp. 396–407). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-industry.38

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digital Health, 2*(2), Article e0000198. https://doi.org/10.1371/journal.pdig.0000198

Kwon, S., Bhatia, G., & Abdul-Mageed, M. (2023). Beyond English: Evaluating LLMs for Arabic grammatical error correction. In H. Sawaf, S. El-Beltagy, W. Zaghouani, W. Magdy, A. Abdelali, N. Tomeh, I. A. Farha, N. Habash, S. Khalifa, A. Keleg, H. Haddad, I. Zitouni, K. Mrini, R. Almatham (Eds.), *Proceedings of ArabicNLP 2023* (pp. 101–119). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.arabicnlp-1.9

Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2023). Few-shot is enough: Exploring ChatGPT prompt engineering method for automatic question generation in English education. *Education and Information Technologies*, 1–33. https://doi.org/10.1007/s10639-023-12249-8

Li, Y., Bubeck, S., Eldan, R., Del Giorno, A., Gunasekar, S., & Lee, Y. T. (2023). *Textbooks are all you need II: phi-1.5 technical report* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2309.05463

Liu, Y., Yao, Y., Ton, J. F., Zhang, X., Cheng, R. G. H., Klochkov, Y., Taufiq, M. H., & Li, H. (2023). Trustworthy LLMs: *A survey and guideline for evaluating large language models' alignment* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2308.05374

Madnani, N., Burstein, J., Elliot, N., Klebanov, B. B., Napolitano, D., Andreyev, S., & Schwartz, M. (2018). Writing Mentor: Self-regulated writing feedback for struggling writers. In D. Zhao (Ed.), *Proceedings of the 27th international conference on computational linguistics: System demonstrations* (pp. 113–117). Association for Computational Linguistics. https://aclanthology.org/C18-2025

Matelsky, J. K., Parodi, F., Liu, T., Lange, R. D., & Kording, K. P. (2023). *A large language model-assisted education tool to provide feedback on open-ended responses* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2308.02439

Meurers, D. (2021). Natural language processing and language learning. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics.* Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal0858.pub2

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heinz, I., & Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys, 56*(2), 1–40. https://doi.org/10.1145/3605943

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), Article 100050. https://doi.org/10.1016/j.rmal.2023.100050

Mohammad, S. (2022). Ethics sheets for AI tasks. In S. Muresan, P. Nakov, A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 8368–8379). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.573

Naismith, B., Mulcaire, P., & Burstein, J. (2023). Automated evaluation of written discourse coherence using GPT-4. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)* (pp. 394–403). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.32

OpenAI. (2023). GPT-4 technical report. arXiv:2303.08774

OpenAI. (2023, December 20). *Prompt engineering.* Openai.com. https://platform.openai.com/docs/guides/prompt-engineering

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, Z., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Siemens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730–27744.

Porter, L., & Zingaro, D. (2024). *Learn AI-assisted Python programming: With Github Copilot and ChatGPT*. Manning Publications.

Prince, S. J. (2023). *Understanding deep learning*. MIT Press.

Puryear, B., & Sprint, G. (2022). Github copilot in the classroom: learning to code with AI assistance. *Journal of Computing Sciences in Colleges, 38*(1), 37–47.

Raheja, V., Kumar, D., Koo, R., & Kang, D. (2023). COEDIT: Text Editing by task-specific instruction tuning. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 5274–5291). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.350

Saravia, E. (2022). *Prompt engineering guide*. https://github.com/dair-ai/prompt-engineering-guide

Sheehan, K. M., Kostin, I., Napolitano, D., & Flor, M. (2014). The TextEvaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *The Elementary School Journal, 115*(2), 184–209. https://doi.org/10.1086/678294

Shin, D., & Lee, J. H. (2023). Can ChatGPT make reading comprehension testing items on par with human experts? *Language Learning & Technology, 27*(3), 27–40. https://doi.org/10125/73530

Thomson, R. I. (2012). *English accent coach* (Version 2). [Computer program]. https://www.englishaccentcoach.com/

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlylov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Bleacher, L., Canton Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … & Scialom, T. (2023). *Llama 2: Open foundation and fine-tuned chat models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2307.09288

Tsang, R. (2023). Practical applications of ChatGPT in undergraduate medical education. *Journal of Medical Education and Curricular Development, 10*. https://doi.org/10.1177/2382120523117844

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., H, Shengyi., von Werra, L., Fourrier, C., Habib, N., Sarrazin, N., Sanseviero, O., Rush, A. M., & Wolf, T. (2023). *Zephyr: Direct distillation of LM alignment* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2310.16944

Tyen, G., Brenchley, M., Caines, A., & Buttery, P. (2022). Towards an open-domain chatbot for language practice. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 17th workshop on innovative use of NLP for building educational applications (BEA 2022)* (pp. 234–249). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.bea-1.28

Vajjala, S. (2018). Machine learning in applied linguistics. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics.* Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal1486

Vee, A., Laquintano, T., & Schnitzler, C. (Eds.) (2023). *TextGenEd: Teaching with text generation technologies.* The WAC Clearinghouse. https://doi.org/10.37514/TWR-J.2023.1.1.02

Voss, E., Cushing, S. T., Ockey, G. J., & Yan, X. (2023). The use of assistive technologies including generative AI by test takers in language assessment: A debate of theory and practice. *Language Assessment Quarterly, 20*(4–5), 520–532. https://doi.org/10.1080/15434303.2023.2288256

Xiao, C., Xu, S. X., Zhang, K., Wang, Y., & Xia, L. (2023). Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)* (pp. 610–625). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.52

Xiaoming, X. (2023) Advancing language assessment with AI and ML–Leaning into AI is inevitable, but can theory keep up? *Language Assessment Quarterly, 20*(4–5), 357–376. https://doi.org/10.1080/15434303.2023.2291488

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2023). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology, 55*(1), 90–112. https://doi.org/10.1111/bjet.13370

Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (2023). Rating short L2 essays on the CEFR scale with GPT-4. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)* (pp. 576–584). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.49

Yonesaka, S. M. (2017). Learner perceptions of online peer pronunciation feedback through P-Check. *JALT CALL Journal, 13*(1), 29–51. https://doi.org/10.29140/jaltcall.v13n1.210

Yu, H. (2023). Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching. *Frontiers in Psychology, 14,* Article 1181712. https://doi.org/10.3389/fpsyg.2023.1181712

Yu, H., & Guo, Y. (2023). Generative artificial intelligence empowers educational reform: Current status, issues, and prospects. *Frontiers in Education, 8*. https://doi.org/10.3389/feduc.2023.1183162

Zhang, P., Zeng, G., Wang, T., & Lu, W. (2024). *TinyLlama: An open-source small language model* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2401.02385

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … & Wen, J. R. (2023). *A survey of large language models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2303.18223

Zhou, W. (2023). *Chat GPT integrated with voice assistant as learning oral chat-based constructive communication to improve communicative competence for EFL learners* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2311.00718

Japanese-Language Article

「学習を促す評価 (Learning-Oriented Assessment (LOA)」としての自己評価とピア評価：日本の高校生のライティング授業を対象として

Self-Assessment and Peer Assessment as Learning-Oriented Assessment: A Focus on the Writing Classes of Japanese Senior High School Students

大井洋子　（おおい　ようこ）
Yoko Suganuma Oi
清泉女子大学
Seisen University

自己評価やピア評価のような学習者による評価は英語教育の現場で活用されてきたが、それぞれの評価タイプと「学習を促す評価(learning-oriented assessment (LOA)」との関係は論じられてこなかった。また、高校生を対象としたライティングの授業は、他の三技能であるリーディング、スピーキング、リスニングに比して、授業の頻度が少なく、その内容も文法等の学習に偏る傾向があり (Mulvey, 2016; 行森, 2018)、ライティング授業の改善が待たれている。本研究はライティング授業を活性化させる方法の一つとして学習者による評価を提案し、その有効性と学習者情意への影響を考察することを目的とする。研究には293人の15〜18歳の日本人高校生が参加した。生徒は自己評価グループ(147名)とピア評価グループ(146名)に分かれ、それぞれのタイプの評価訓練を受けた後に10日間のあいだに5

https://doi.org/10.37546/JALTJJ46.1-4

回英作文を書き、その直後にそれぞれの評価活動を行った。この集中的なライティング活動の後に、研究に参加した12人の生徒と研究を観察した二人の英語教師が半構造的面接に参加した。自己評価とピア評価は似ているが異なる効果を英語ライティング力の向上や情意にもたらし、「学習を促す評価(LOA)」を効果的にするためには、教師が適切に介入することが必要であることがわかった。

This study explored how student assessments, that is, self-assessment and peer assessment, were related to learning-oriented assessment (LOA) by exploring the reliability and the effects of each assessment type on students' writing performance and learner affect. A total of 293 students aged 15 to 18 years participated in the survey. They were divided into two groups, self-assessment group (147 students) and peer assessment group (146 students). Each group had a rater training and then had five consecutive writing sessions for ten days. Just after writing English compositions, each assessment group evaluated their own writing or a peer's writing. After the sessions, 12 students and two teachers were interviewed. It was found that each assessment type had similar but different effects on writing performance and learner affect. It was also found that teachers were expected to intervene in self- and peer assessment to make learning-oriented assessment more successful.

キーワード: 自己評価;ピア評価;英語教育;学習を促す評価

Keywords: English education; learning-oriented assessment; peer assessment; self-assessment

コミュニケーション力育成を目指した英語教育の改革が叫ばれ、スピーキングやライティングなどの相手あるいは読み手に意図を伝達する「産出的(Productive)　技能」の強化の必要性が強調されている。だが、スピーキングに比してライティングを扱う授業の頻度が他の技能の授業頻度に比べて少ないことが指摘されている(文部科学省、2018)。また、ライティングの授業内容が文法や語彙学習に大きく時間を割く傾向があることや(文部科学省、2011)、中高では学年の上昇につれ、ライティング活動がさらに減少し、訳読によるリーディングや語彙・文法指導に時間が多くとられている(Mulvey, 2016;　行森, 2018)。こうした現状の中でライティング授業を充実させていくために、学習者による評価を積極的に取り上げることを提案したい。なぜならば、学習者による評価は、「学習を促す評価」(Learning-oriented assessment (LOA): Turner & Purpura、2016)として、学習者を学びの中心に据えることが可能なので、ライティング授業の活性化につながると考えられるからだ。学習者による評価には、自分の学習成果を評価する自己評価(Boud、1992)と学習者同士による評価であるピア評価(相互評価・他己評価・他者評価)がある(Topping, 1998)。こうした学習者による評価は、形成的評価 (formative assessment) の一つでもある (Black他, 2003;Wiliam他, 2000)。なぜならば、形成的評価は、学習目標に照らし合わせて学習者がどの程度まで学習目標に到達しているかを評価し、次の学習段階へ導き、学習を促すことを目標にする評価であるからだ (Leahy & Wiliam, 2012)。つまり、学習者が立場を変えて評価者になることによって、学習目標と今の学習レベルとの開きを意識できることが形成的評価の目的である (Leahy & Wiliam, 2012)。教授者が自己評価とピア評価というそれぞれの特質を理解すれば、自己評価とピア評価を効果

的に授業に活用できると考えられる。しかしながら、自己評価とピア評価を比較研究したものは少なく、研究対象の多くは大学生で、特に日本の高校生を対象にした研究は少ない (Oi、2021)。また、これら二つのタイプの評価法が「学習を促す評価 (LOA)」としてどのような意味があるのかは探究がされていない。なぜならば、自己評価やピア評価の先行研究は、その信頼性に着目して教師評価の代替評価として可能かどうかを検証したものが多く(Oi、2021)、学習を促す評価としての役割を追求したものは少ないからだ。学習を促す評価では、教師と学習者は評価の責任を共有すべきであるとされている。生徒が評価活動に取り組むことによって、学習を促す評価を前進させる可能性があることからその探求は喫緊の課題である。また、生徒評価を授業に導入することによって、学習者の英語力の進展にどのような影響を与えるかについても、教授者にとっては欠くことのできない命題だと言える。したがって、本研究は、自己評価とピア評価を比較することによって、それぞれの評価法の特徴を信頼性とライティング能力、そして情意に与える影響から分析し、学習を促す評価として期待できることは何かを考察する。

先行研究

先行研究の傾向として、自己評価とピア評価をそれぞれ単独で分析するものが多く、この二つを直接比較して、それぞれの評価方法の特徴を明確にした研究は少ない(Oi, 2021)。また、先行研究のメタ・アナリシスによると、1980～2018年までに行われた学習者による評価に関する研究の約7割はヤング・アダルト(young adult)と呼ばれる大学生を対象としている(Oi, 2021)。学習者評価の信頼性は、学習者の教育的背景に影響を受けるという先行研究結果もあるので(Boud & Falchikov, 1989)、本研究は分析が進んでいないヤング・ラーナー(young learner)である高校生対象に研究を行った。

学習を促す評価 (Learning-oriented assessment: LOA)

教室評価に対する理論的枠組みとして、「学習を促す評価(LOA)」には様々な定義がある。その中でも意義深いのは、教育における評価の役割は学習を促すのが目的だとして、評価の意味を再定義したことにある(Carless, 2007)。つまり、評価を行う目的はあくまでも学習の促進であり、評価は学習者の成長のための足場かけ(scaffolding)として機能すべきというものだ。「学習を促す評価 (LOA)」においては、評価は、学習の根本的な要素であり、教授内容や指導方法を決定する拠り所となるものなのである(Mok、2012; Ploegh, 2009)。言い換えれば、総括的評価と形成的評価を分離させずに協同的に学習者の学びを促進させ、公式的な評価と教室評価を協調的に機能させようという意図が土台となっている(Carless, 2007; Jones & Saville, 2016)。「学習を促す評価(LOA)」には、理論的枠組みにとどまらず、この枠組みを基本にして様々なEFLのコンテクスト合わせて言語教育に応用していく流れがある。例えば、言語技術の向上の観点からLOAが論じられたり(Green, 2017; Hamp-Lyons, 2017)、実用性(Alsowat, 2022)や教師のLOAに対する認知度(Derakhshan & Ghiasvand, 2022)等も研究の対象となっている。Hamp-Lyon (2017) は、「学習を促す評価(LOA)」を教室評価に取り入れるモデルを構成する要素として、第一に、学習に焦点を当てた課題、第二に、学習者による評価、そして第三に、学習者に焦点を当て

たフィードバックの3つを掲げている。このモデルでは、学習者と教授者の協同的関係が求められているのが特徴である。このように「学習を促す評価(LOA)」は、比較的新しい理論的枠組みで、それと実践的試みの関係を探究する研究がまだ少ない。よって、本研究では、Hamp-Lyon によって提唱をされた「学習を促す評価(LOA)」を教室評価に採用するモデルの一つである「学習者による評価」に着目をして、その有効性について分析をすることとした。

学習者による評価の信頼性

学習者による評価の信頼性に関する研究は、教師評価と比較する形式で行われてきた。先行研究によれば、自己評価の教師評価に比した信頼性は、ある程度高い傾向があるとされている。例えば、Andre他(2010)の研究によれば、教師評価との相関関係はある程度高く、特に文法評価においては高い傾向があるという結果が示されている。しかし、Runnels (2014)は、大学生を対象にした研究において、自己評価の信頼性は評価項目の数に左右されることを報告している。また、学習者の英語力の高さや評価の訓練度、年齢や文化的背景が信頼性に影響を与えるとしている (Dieten, 1989; Peirce他、1993)。一方、ピア評価は、教師評価と正の相関関係があるとする研究結果が多いが(Weaver 他, 2011; Orsmond他, 2000)、自己評価と同様に評価項目の理解度や評価項目の数、文化的背景にその信頼性が影響をされる(Falchikov & Goldfinch, 2000; Matsuno, 2009)。また、ピア評価を行う学習者同士が、人間関係を円滑にしようと正直な評価をしないこともある(Oi, 2021)。両方の評価方法ともに、その信頼性には、評価項目に対する理解や訓練、文化や教育的背景が影響を与えることは共通であるが、信頼性を一貫性と厳格度から分析したものは少ないので本研究はこの点について論じることにした。

学習者による評価のライティングへの学習効果

1980～2018年の間に行われた学習者による評価を対象としたメタ・アナリシスによると、学習効果があったとするのは、自己評価で約2割、ピア評価で約3割とされている(Oi、2018)。学習者によるライティング能力への学習効果については、主に文法や語彙の習得に対するものが多い(McDonald & Boud, 2003; Lee, 2017; Sadler & Good, 2006)。特に、ピア評価に関する研究では、文法や校正力の向上に寄与したという研究があるが(Oi, 2021)、自己評価を対象とした研究では、これらに着目した研究は少ない。また、二つの評価タイプを比較した場合、ピア評価の方が自己評価よりもライティング能力の向上に貢献をしたという結果を示した研究や(Matsuno, 2009)、友達の評価の方が深刻に受け止めるので教師評価よりも効果的であるとする先行研究もある(Zarei & Mahdavi, 2014; Tsui & Ng, 2000)。自己評価とピア評価の学習者に及ぼす影響の共通点としては、学習者の評価項目への理解が深まり、それが学習効果につながることである(Orsmond他, 2000)。しかし、自己評価とピア評価を直接比較して、どちらがライティング能力の改善により効果的かを示した研究は少ないが、教師評価と同様な使用効果をもっているとする研究(Farrokhi他, 2012)や、自己評価とピア評価を行う学習者グループの間には評価姿勢に違いが見られると論じる研究もある(Aslanoglu他, 2020)。また、信頼性を対象にした先行研究に比して、ライティン

グ能力の向上に着目した研究が少ないために、どちらの評価タイプがより効果的かを判断するのは難しいように思われるので、この二つのタイプの評価法を直接比較し、それぞれのライティング能力向上への影響を分析するのは意義があると考えられる。さらに、自己評価の文法や校正力の伸長に与える影響に関する研究は少ないので、研究の必要があると言える。

学習者による評価の学習者情意に与える影響

自己評価は、学習目的を学習者に明確に意識をさせるので、自己達成感や学習意欲、そして自信の向上に効果がある (Butler & Lee, 2010; Zarei & Usefli, 2015)。Black他 (2004)は、学習目的を意識化させることによって、内省する習慣を学習者につけることが可能と述べている。一方、ピア評価については、相互評価による協同作業を通して、他者との交流が生まれ、分析力が身につき、評価をする技術が身に付くという先行研究がある(van Gennip他, 2010; Saito & Fujita, 2009; Cho他, 2006)。また、自己効力感を高め、緊張感を減少させ精神的安心感を与える (Cho他, 2006)。Matsuno (2009)によれば、自己評価とピア評価の共通点としては、両評価タイプともに学習に対する責任感や自律性に良い影響を与えると報告している。しかし、両者の違いを論じた研究は少なく、効果的に授業に取り入れていくには、二つの評価タイプの相違点にも着目していくべきだと考える。

目的と研究課題

本研究は、学習者による評価法である自己評価とピア評価をその信頼性・ライティング能力に与える影響の観点から比べ、それぞれの特徴を明らかにするのが目的である。また、それぞれの評価タイプが「学習を促す評価(LOA)」としてどのように機能できるかを論じるものとする。なお、ここでの信頼性は評価者の一致度や安定性の他に評価者間の厳格度や評価者内の一貫性を含む。したがって、以下の研究課題を掲げる。

1. 教師評価に比して自己評価とピア評価の信頼性はあるのか。
2. 自己評価とピア評価はどのようにライティング能力に影響を及ぼすのか。
3.「学習を促す評価」として自己評価とピア評価の役割は何か。

分析方法

研究参加者

研究には、15～18歳の293名の日本の高校生(表1)と4名の英語教員(日本人2名、英語母語話者2名)が参加をした(表2)。参加生徒の中には英語母語話者はいなかった。多相ラッシュモデル (many-facet Rasch measurement; MFRM; Eckes, 2015; Linacre, 1994) の分析は、英作文評価の総合得点と分析的評価の得点について別々に行った。また、英作文の評価(教師・自己・ピア評価)は、成績には関与しないことを事前に参加者には伝えた。ピア評価については、英作文を授業内で交換する形式で行い、評価表は書き手に直接返却をしてもらい、書き手が読み終わった後に英語教員が評価表を回収した。

表1.
実験参加生徒のグループ構成

	自己評価グループ				ピア評価グループ			
	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8
学年	1	1	2	3	1	1	2	3
科目名	I	I	II	III	I	I	II	III
人数	36	36	36	39	36	36	36	38
合計人数	147				146			

注. I = English Communication I; II = English Communication II; III = English Communication III

表2.
実験参加教師の英語教育経験

	JET A	JET B	NET C	NET D
年齢	40	55	38	50
性別	Male	Female	Female	Male
英語教育経験年数	15	30	7	20

注. JET = Japanese English teacher; NET = native English teacher.

研究手順

表1のように、既存のクラスを自己評価グループとピア評価グループの二つに分け、生徒たちは教師と同じ評価表を使い(付録A)、10日間のライティング活動の直前に評価訓練をした。評価訓練1回とライティング・テストを含むライティング活動5回の時に、15分の間に辞書なしで60〜80語の英作文を書いてもらった。英作文を書いた直後に、自己評価かピア評価のどちらかの評価活動を行った。ピア評価では、1名の生徒のライティングを、別な生徒1名が行った。評価表は、高校教師を対象にした高校生の英作文を評価する際に最も重視をする項目の調査結果をもとに作成した「課題達成度」、「構成と内容の一貫性」、「語彙の適切な使用」、「正確な文法の使用」の4観点の4段階(レベル1〜4)、合計16点満点を利用した(Oi, 2019)。評価活動には10分をとり、英作文課題は、高校生を対象とした英語表現の教科書で最も頻度が高いトピックを選んだ(付録B)。なお、英作文課題の難易度は、英検3級の英作文問題と同程度である(Oi、2018)。

研究方法

量的研究(研究課題1〜2)においては、多相ラッシュモデル(many-facet Rasch measurement; MFRM; Eckes, 2015; Linacre, 1994)を使い自己評価グループの生徒とピア評価グループの生徒の英作文の総合得点と英作文の分析的評価の得点に関する信頼性を分析した。また、同じく多相ラッシュモデルにより英作文能力に対するそ

れぞれの評価者タイプの影響を分析した。ソフトウエアはMinifac (Linacre, 2021)を用いた。

ラッシュモデルは、人と項目の両方を潜在変数（構成概念）上に位置づけ、受験者の能力と項目の難易度を一元的尺度で分析をする項目反応の確立論的モデルである。受験者能力、項目難易度に加えて、第3の相(facet) として評価の安定度を入れることができ、ロジット (logits) を単位とした同じ尺度で受験者能力、項目難易度、評価者の厳しさを推定することができる。ロジットの値が小さいほど、受験者能力は低く、項目難易度は低く、評価者は甘いことを示す。具体的には、受験者能力と他をそれぞれ独立に並べて推定することが可能である。

分析には、4人の教師と293人の生徒、293人の生徒によって書かれた586の英作文(事前と事後のテストの2回分)なので、2930のデータを使った。しかし、一部の生徒が欠席等のために参加ができなかったので、有効となるデータ数は243人の受検者、6人の評価者、事前・事後の2回のテストから得られた2430のデータとなった。内、自己評価グループは244、ピア評価グループは242、そして教師の評価データ数は1944であった(表3)。本研究では、教師評価、自己評価グループの生徒、そしてピア評価グループの生徒をラッシュモデルを使って比較したが、これら3つの評価タイプを同時に比較した先行研究は乏しい（Farrokih他, 2012）。また、ラッシュモデルを使用時に、評価者グループまたは評価者個人として扱うかは、研究参加者の数や教育環境にも作用される（Aslanoglu他, 2020; Farrokih他, 2012）。本研究では、評価者一人一人の評価結果をラッシュモデルを使って分析を行ったが、紙幅の関係で評価グループとして報告をしている。また、本研究のoccasion相は、pre-testとpost-test が異なっていることから“Occasion by Task”相とした。Pre-testのmeasureの値は .03、一方、post-testのmeasureの値は- .03で、値の差は .06で、先行研究の結果に基づき（Oi、2018）、二つのタスクの困難度はほぼ同程度とみなした。また、タスクは高校生を対象とした教科書の頻出度上位のタスクから選んだ（Oi, 2018）。

表3.
多相ラッシュモデル分析に使われた有効なデータ数

	参加者数		評価数	欠席者合計	有効な評価数		評価数合計
自己評価グループ	147	147 students x 2 tests	294	25	122	122 students x 2 tests	244
ピア評価グループ	146	146 students x 2 tests	292	25	121	121 students x 2 tests	242
Teachers	4	4 teachers x 2 tests x 293 students	2,344	0	4	4 teachers x 2 tests x 243 students	1,944
テスト回数	2				2		
合計		2,930				2,430	

質的研究(研究課題1〜3)においては、12名の生徒(自己評価とピア評価の各グループから英作文の評価結果をもとに6名ずつ抽出、表4)と2名の教員(日本人1名JETB、英語母語話者1名NETC、表2)に対して半構造的面接データを行った。なお、実験前後の英作文の点数は、表2に記された4名による英語教師の評価平均である。生徒には自分の考えを語りやすい日本語で、英語母語話者を含むことから教師には英語で、学習者による評価の信頼性やライティング能力の向上に関する効果についての面接を実施し録音をした(表5)。

表4.
インタビューを受けた生徒一覧

	全体平均	SHSS 1	SHSS 2	SMSS 1	SMSS 2	SLSS 1	SLSS 2	PHSS 1	PHSS 2	PMSS 1	PMSS 2	PLSS 1	PLSS 2
実験前の英作文の点数	12.22	14	15	12	12	10	9	15	15	12	12	9	8
実験後の英作文の点数	12.43	15	14	12	12	10	9	14	14	12	12	11	11
年齢	16.3	15	16	18	16	17	16	18	15	16	17	16	16
性別		男子	女子	男子	女子	男子	女子	男子	女子	男子	女子	男子	女子

注. SHSS = 自己評価グループで高得点生徒(high-scoring students in the self-assessment group); SMSS = 自己評価グループで中程度得点生徒(middle-scoring students in the self-assessment group); SLSS = 自己評価グループで低得点生徒(low-scoring students in the self-assessment group;) PHSS = ピア評価グループで高得点生徒(high-scoring students in the peer assessment group); PMSS = ピア評価グループで中程度得点生徒(middle-scoring students in the peer assessment group); PLSS = ピア評価グループで低得点生徒(low-scoring students in the peer assessment group)

表 5.
半構造面接質問項目

Interviewees	
Students	Teachers
• あなたの意見では、生徒評価(自己評価またはピア評価)は英語のライティング能力を向上させるのに効果的ですか。 • もしそうならば、どんな点で重要なのですか。 • 生徒評価(自己評価またはピア評価)は自分の気持ちや学習態度に何か影響を与えましたか。 • 生徒評価(自己評価またはピア評価)の良い点、または悪い点は何ですか。	• In your opinion、is student assessment important for developing writing ability? • If so、in what ways? • Did you observe that learner affect such as anxiety and autonomy did students change by using self-assessment/peer assessment? • How can you utilize self-assessment/peer assessment in writing classes? • What are the effects and challenges of student assessment?

録音は筆者が書き起こしを行い、グラウンデッド・セオリー(Glaser & Strauss, 1967)を用いて分析をした。これは、他の質的分析に比べ、グラウンデッド・セオリーは、人間の語りや行為の意味を分析、コード化し、それら上位の概念的カテゴリーを考察するのに確立された手法だからである。グラウンデッド・セオリーは、データから概念を抽出し、概念同士を関連付けようとする方法(戈木、2018、p.2)であり、抽象度の異なる4種類の概念、プロパティ(特性)とディメンション(次元)、ラベル、カテゴリーが核となっている。その内、抽象度が低いプロパティとディメンションは分析を通して使う分析の中心的なもので、それには5つの役割があるとされている(戈木、2018)。第一に、データから概念を抽出する土台となること、第二に概念を理解するヒントを提供すること、第三に概念同士につながりを見出すこと、第四に現象の中にある変化のパターンを提示すること、そして最後に分析者がデータの解釈をする際に解釈の理由を説明できることである。

本研究では、抽出のレベルを増すために半構造インタビューのデータを5段階に分けて分析を行った:(1)データを内容ごとに切片に分け、(2)オープン・コーディング、及びアキシャル・コーディング(軸足コーディング)でカテゴリーを関連付け、(3)カテゴリーを現象ごとにセオレティカル・コードに分類、(4)セオレティカル・コードのカテゴリー関連図の作成、(5)各カテゴリー関連統合図の中心となっているカテゴリーを、プロパティとディメンションで関係づけた。コーディングには、著者の他、日本人英語教師(JETA)がコーダーとして加わった。著者のモデル・コーディングの後に、例を付したガイドラインを作成し、JETAがコーディングを見て矛盾点がないかについてのチェックを行った。分析に不一致があった場合は、その都度話し合いを行いコーディングの修正を行った。

分析の最終段階で、生徒評価と学習を促す評価(LOA)の関係を考察するために探索的に質的・量的研究分野を統合した。

量的研究分析結果

1. 教師評価に比して自己評価とピア評価の信頼性はあるのか。

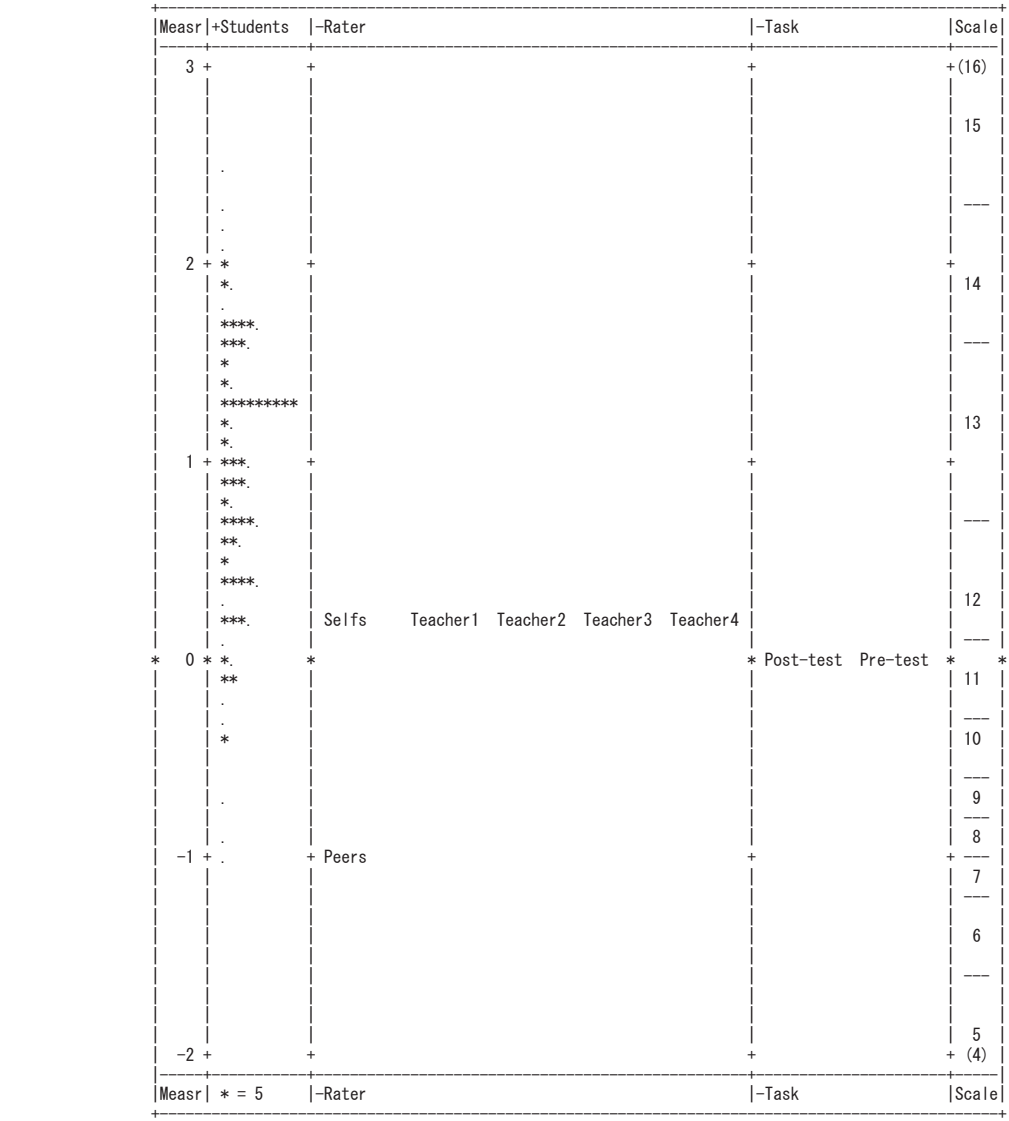
総合得点に関する信頼性

教員評価に比した生徒評価の信頼性に関する結果を提示する前に、MFRMの前提であるモデルの全体的適合度についてライティングの総合得点について計算をした。予想外の回答(Unexpected responses) における標準化残差を用い、±2を超えた標準化残差が約5％以内、±3を超えた標準化残差が約1％以内であれば、データがラッシュモデルに全体的に適合したと考えた (Linacre、2021、p. 178)。総合得点については、データポイント数2、920でそれぞれが3.4％、1.4％であったので、全体としてこの基準は満たされていた。

図1の変数マップ (Wright マップ) は、縦軸Measr はロジットの尺度、Students は本研究に参加をした生徒、Rating methodsのSelfは自己評価、Peers はピア評価、Teacher 1〜4は4人の教師評価それぞれを指す。変数マップの上位に位置するほど、評価が厳しいことを示し、下位に位置するほど評価が甘いことを意味する。また、Occasion はPre-testとPost-test は実験前後の英作文テストを意味し、変数マップの上に位置するほどテストが難しいことを意味し、下方にいくほど困難度が低いことを意味する。Scaleは評価表の合計点を示している。この結果によると、実験前後に行われた英作文テストの難易度はほぼ同じであり、生徒を意味するアステリスクがロジットの尺度0のやや上に集中することから比較的易しいテストだったことがわかる。また、4人の教師評価と自己評価がほぼ同じ所に位置することから、自己評価グループの生徒の評価の厳しさは教師評価と同程度だったと推定される。一方、ピア評価のグループの生徒は最も下方に位置するので、教師評価と自己評価グループの生徒に比べ評価が甘かったと言える。

次に評価の一貫性についてだが、インフィット及びアウトフィット統計量が評価者のラッシュモデルに対する適合度を表すので、こちらを参考に考える。表6が示すように、4人の教師全員のInfitとOutfitは、0.60〜0.61に分布している。一般的には、0.7〜1.3が適合を示すフィット統計量の基準だが(Bond & Fox, 2007)、本研究のテストで重要な判断を行わないことにより、0.5〜1.5の時に、評価がモデルの予測値に適合していることを示すとした(平井 他、2018)ので、教師評価は4人が全員ともに評価の一致度は適合度を示していた。教師評価、自己評価、ピア評価をそれぞれ分析した所、自己評価グループの生徒とピア評価グループの生徒についてはインフィットとアウトフィットが1.4以上を示すので、両評価タイプ共に評価は予測不可能で不一致度が高いと考えられる。つまり、受験者能力の低い受験者に高い点数を与えたり、受験能力の高い受験者に低いスコアを与えている可能性がある。なお、表は、評価者の厳格度の順に並んでいて、一番厳しい評価者は自己評価のグループの生徒で、一番評価が甘いのは、ピア評価のグループの生徒ということになる。また、教師評価者の人数が生徒評価グループの参加者よりも少ないことがこの結果に影響を与えたとも考えられる。

図1.
ライティング・テスト総合得点の変数マップ



注. Pre-testとpost-testはpre-task とpost-taskを指す。

表6.
総合得点に関する評価者特性

Raters	Observed Average	Measure logit	Model *SE*	Infit *M* Sq	*Z* Std	Outfit *M* Sq	*Z* Std
Self	12.42	.21	.05	2.91	9.0	2.94	9.0
Teacher 3	12.32	.20	.03	.60	-6.4	.61	-6.5
Teacher l	12.33	.20	.03	.60	-6.4	.61	-6.4
Teacher 2	12.33	.20	.03	.60	-6.4	.61	-6.4
Teacher 4	12.33	.20	.03	.60	-6.3	.61	-6.4
Peers	13.97	-1.01	.05	2.44	9.0	2.57	9.0
Mean	12.76	.00	.04	1.29	-1.3	1.33	-1.3
SD (pop.)	.61	.45	.01	.99	7.3	1.02	7.3
SD Sample	.66	.50	.01	1.08	8.0	1.11	8.0

注. 下記はそれぞれの評価者タイプによるラッシュ分析の要約統計による。

全ての評価者(4人の教師、自己評価グループの生徒、ピア評価グループの生徒):Separation = 12.70; Strata = 17.26; Reliability = .99; χ 2: 504.5; Significance = .00.

4人の教師:Separation = .00; Strata = .33; Reliability = .00; χ 2; .00; Significance = 1.00.

自己評価グループの生徒: Separation = 1.85; Strata = 2.80; Reliability = .00 χ 2: 1429.7; Significance = .00.

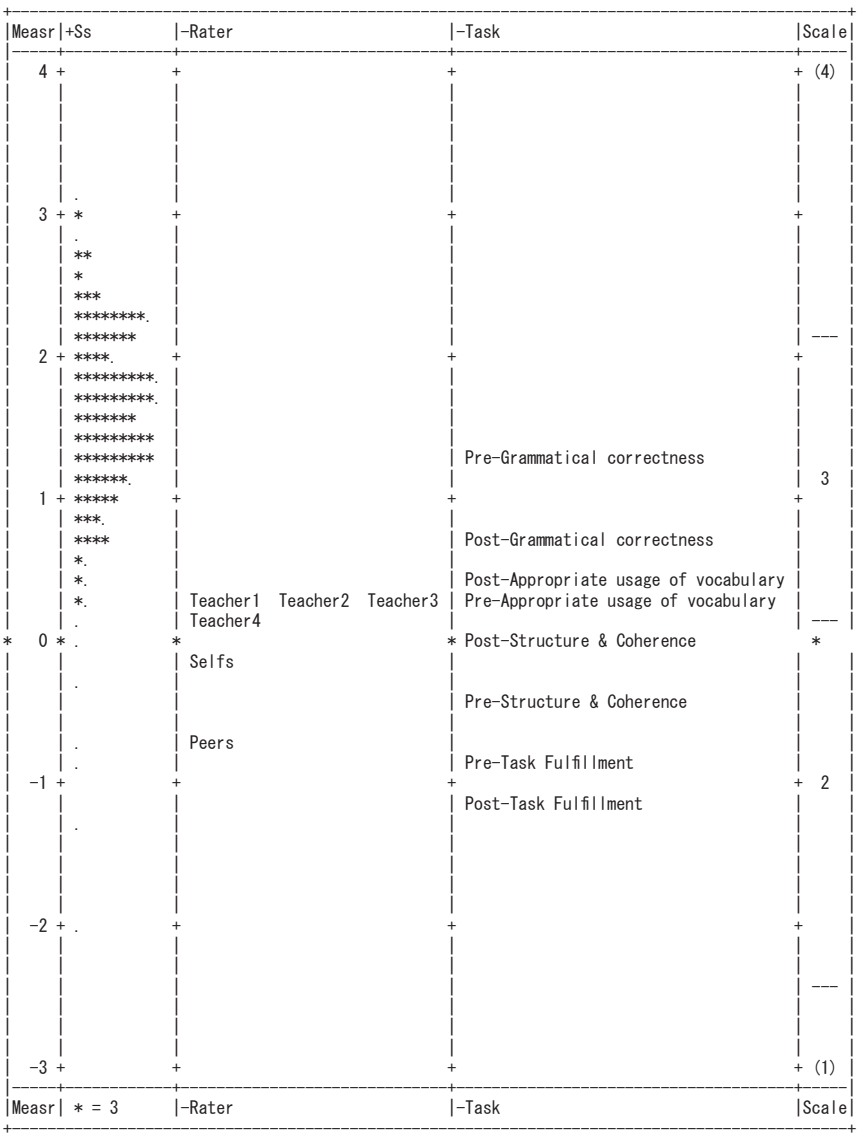
ピア評価グループの生徒:Separation = 2.06; Strata = 3.08; Reliability = .81; χ 2 :679.3; Significance = .00.

分析的評価に関する信頼性

4つの項目からなる英作文の分析的評価の結果について、モデルの全体的適合度を調べたところ、予想外の回答(Unexpected responses) における±3を超えた標準化残差のデータポイント数が14、016の内、7.13％だった。よって、対数尤度カイ二乗検定(log-likelihood chi-square)を調べた所、対数尤度カイ二乗値が26、237.32(approximate model df = 13、709; p > .001)だったので、MFRMのモデルの全体適合度は満たされていると判断した。

図2は、実験前後の英作文テストの分析的評価の変数マップで、図の構図は、図1とほぼ同じであるが、四列目のOccasion * Rating scale（Occasion by Rating scale） は本研究実験前後の4つの観点に関する評価を表す点が異なっている。これは、Occasionによる実験前後それぞれの分析的評価のタスクの困難度を示す。第2列のテスト受験者である生徒の能力値は3.1から-2 logitを指しているが、大半は0より上なので難易度は低かったと考えられる。第3列の各観点の困難度の順番は、実験前の「正確な文法の使用」が一番高く、実験後の「課題達成度」が一番低かった。また、実験後の「正確な文法の使用」は他3つの実験後の各観点よりも困難度は一番高かった。要するに、4つの評価項目の中で一番困難度が高いのが「正確な文法の使用」で「課題達成度」は一番易しい項目であり、これは実験前後でも変わらなかった。左から第3列目の

図2.
ライティング・テスト分析的評価の変数マップ



評価の厳しさによれば、最も評価が厳しかったのはTeacher 1、次にTeacher 3、そしてTeacher 4 と続く。生徒評価については、自己評価グループの生徒の厳格度が教師評価のそれに続き、最も評価が甘いのはピア評価グループの生徒の評価だった。この結果表示においては、図1と評価者の厳格度において教師評価と自己評価グループにおいてずれが見られるが、分析的評価では、Occasion ではなく、Occasion * Rating scale としたことが影響を与えたと考えられる。

次に分析的評価の一貫性についてだが、表7が示すように、4人の教師の4観点からなる分析的評価平均はいずれも3.06　もしくは　3.07で、分離信頼性(reliability statistics)が .00 を示しているので、評価の一貫性があり、おしなべて同程度の評価の厳格さがあると推定できる。対照的に、自己評価グループは、高い分離信頼性である0.75 を示し、ピア評価のグループも 0.82を示している。評価の厳格度の一貫性をみるには、低い信頼性が好ましいので両評価タイプともに分析的評価の厳格度の一貫性はかなり低いと推定できる。さらに、両評価タイプともに、評価のばらつき(variability of severity)が自己評価グループでは1.75であり、ピア評価のグループでは2.16だったので、評価の厳格度にもばらつきがあると考えられる。

表7.

英作文の分析的評価に関する評価者特性

Raters	Observed Average	Measure logit	Model SE	Infit M Sq	Z Std	Outfit M Sq	Z Std	Separation ratio (G	Separation (strata) index (H)	Separation reliability (R)
Teacher 2	3.06	.25	.03	.87	-5.2	.86	-5.1			
Teacher 3	3.06	.23	.03	.86	-5.5	.85	-5.4	.00	.33	.00
Teacher 1	3.07	.22	.03	.85	-7.4	.83	-7.7			
Teacher 4	3.07	.21	.03	.84	-6.4	.83	-6.3			
Self	3.23	-.15	.04	1.50	9.0	1.48	9.0	1.75	2.67	.75
Peers	3.45	-.77	.04	1.57	9.0	1.62	9.0	2.16	3.21	.82
Mean	3.21	.00	.03	1.08	-1.1	1.08	-1.1			
SD (population)	.17	.37	.01	.32	7.2	.34	7.2			

注. 下記はそれぞれの評価者タイプによるラッシュ分析の要約統計による。

全ての評価者(4人の教師、自己評価グループの生徒、ピア評価グループの生徒): Separation = 12.19; Strata = 16.58; Reliability = .99; χ 2 : 541.1; Significance = .00

4人の教師: Separation = .00; Strata = .33; Reliability = .00; χ 2 :1.7; Significance = .65

自己評価グループの生徒: Separation = 1.76; Strata = 2.67; Reliability = .76 χ 2 :1748.9; Significance = .00

ピア評価グループの生徒: Separation = 2.17; Strata = 3.22; Reliability = .82; χ 2 :1764.2; Significance = .00

次に、評価者間の得点の一致率だが、教師評価が97.6％(17093/17520)を示す一方、自己評価グループは76.7％(20256/26421)、ピア評価グループは79.0％(17714/24966)だった。また、教師評価のフィット統計値は、0.84-0.87の範囲におさまるので、予測可能で一致度が高いと考えられるが、両生徒評価タイプともフィット統計値は1.5よりも高かった。

2. 自己評価とピア評価はどのようにライティング能力に影響を及ぼすのか。

英作文の総合得点に与える影響

研究課題2は、研究課題1で使ったラッシュアナリシスのデータを使う形で二つの評価法のライティング能力向上に対するそれぞれの影響について分析をしていく。表8は、実験前後の両評価タイプの総合得点の記述統計を表す。自己評価グループの生徒の平均点はピア評価のグループの生徒の平均点よりも実験前後ともにやや高いが近似している。標準偏差については、実験前後双方において、ピア評価のグループの方が自己評価グループよりも高い値を示している。

表8.
自己評価グループとピア評価グループの複合得点の記述統計

		人数	最低点	最高点	平均	標準偏差
事前英作文テスト	自己評価	147	4	16	12.27	1.64
	ピア評価	146	4	16	12.16	1.87
	合計	293	4	16	12.21	1.76
事後英作文テスト	自己評価	147	6	16	12.52	1.44
	ピア評価	146	6	16	12.33	1.62
	合計	293	6	16	12.43	1.54

しかし、表9の総合得点の課題測定レポート(task measurement report)によれば、自己評価の事前テストのmeasure logitは .33 であるが、ピア評価のグループの事前テストのmeasure logit は -.33 なので、事前テストは自己評価グループにとっての方がより難しかったといえる。事後テストについても同様の結果で、自己評価グループのmeasure logit は .05 であるが、ピア評価のグループの measure logit は -.05 だった。また、生徒全員のmeasure logit は事前テストが .12 で事後テストは -.12なので、両タイプともに生徒全体のライティング能力の向上が見られたと言えるが、自己評価グループの方がピア評価のグループよりもライティング能力が向上したと考えられる。

表9.
総合得点の課題測定レポート

Statistic	SA		PA		All students	
	Pretest	Post-test	Pretest	Post-test	Pretest	Post-test
Observed Average	12.27	12.17	12.71	12.25	12.23	12.43
Measure logit	.33	.05	-.33	-.05	.12	-.12
Model SE	.05	.04	.06	.04	.03	.03
Infit M Sq	.95	1.03	1.04	.93	1.00	.97
Z Std	-.8	.5	.6	-1.2	.0	-.7
Outfit M Sq	.92	1.03	1.04	.93	.99	.96
Z Std	-1.3	.5	.6	-1.3	-.1	-.8
Estm. Discrm	1.06	1.00	.95	1.02	1.02	.99

注. SA＝自己評価グループの生徒; PA ＝ピア評価グループの生徒

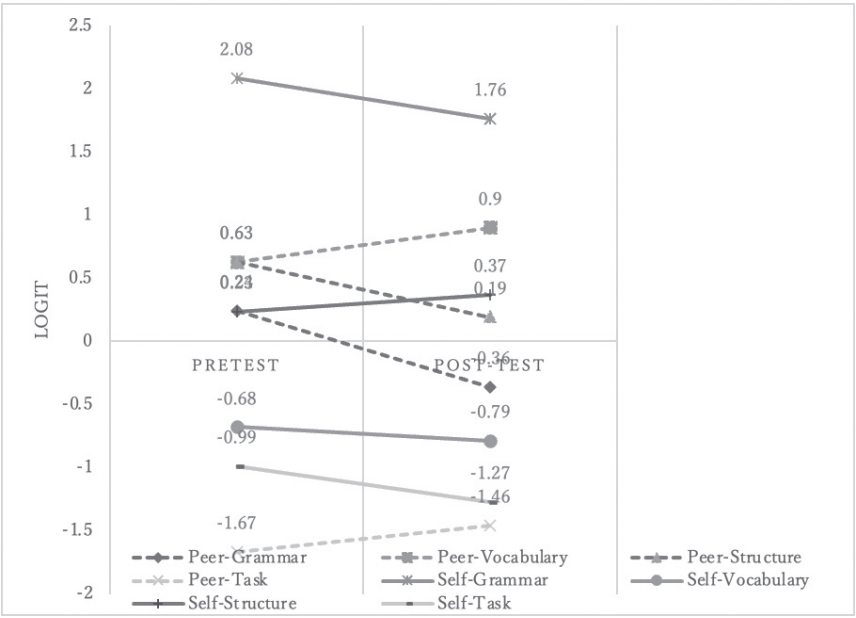
英作文の分析的評価の得点に与える影響

次に4つの観点における自己評価とピア評価の影響について考える。これについてもラッシュアナリシスのモデルの全体的適合度を調べたところ、一次元性(unidimensionality)がラッシュ測定によって説明された変数の20％以上を占めたので十分に適合していた(Engelhard、2013)。モデルの全体的適合度もデータポイント9、344の内、11.5％が標準化残差の絶対値が3以上に関連していたので、対数尤度カイ二乗検定(log-likelihood　chi-square)を調べた所、対数尤度カイ二乗検定　(log-likelihood chi-square) の16,898 (approximate model df = 16,864.73; p > .001) が統計的に有意ではなかったのでラッシュアナリシスに適合すると考えた。課題測定レポート(task measurement report)に基づいて、事前事後テストの自己評価グループとピア評価グループの違いを分析するために図3を作成した（Appendix G参照）。

自己評価とピア評価グループ双方において課題の困難度から見た4つの観点は、事前事後テスト間で「構成と内容の一貫性」の観点以外の順位は変わらなかったが、事後テストは、事前テストに比べ、課題の困難度はやや変化をした。最も困難度が高かった観点は自己評価グループの「正確な文法の使用」だった。一方、最も困難度が低かった観点は、ピア評価のグループの「課題達成度」だった。図が示すように、自己評価とピア評価グループ両方において、「正確な文法の使用」のロジットが低下しているので、実験前に比べて「正確な文法の使用」の課題が両グループにとって簡単になったことを意味する。他の観点は、両グループとも事前・事後テストでロジットの増減が見られなかった。

表10は、自己評価とピア評価グループに関する課題測定のmeasure logitsとWald Statistics の結果を表す。この結果をもとに (1) タスク測定のmeasure logits; (2) Wald Statisticsの二つの観点からライティング能力の向上についてそれぞれの評価タイプ

図3.
自己・ピア評価グループの4観点からなる分析的評価の事前事後テストの推移



注. Pre-testとpost-testはpre-task とpost-taskを指す。

のライティングの能力の変化についての分析をする。なお、タスク測定に関する報告は、付録Gを参照。

まず、measure logits の結果についてだが、表10によると、自己評価グループのmeasure logits は事前テストの「正確な文法の使用」では2.08であり、事後テストでは1.76だった。同様に、ピア評価グループのmeasure logits は事前テストの「正確な文法の使用」では0.24だが、事後テストでは0.32だった。自己評価グループの事前事後テストの差は0.32で、ピア評価のグループの事前事後テストの差は0.60だったので、ピア評価のグループの方が差が大きかったと言える。

「課題達成度」のmeasure logits は自己評価とピア評価グループで違いを見せた。自己評価グループのlogit estimate は-0.99が事前の「課題達成度」の値であるが、事後テストの「課題達成度」は-1.27だった。一方、ピア評価のグループは、事前「課題達成度」は-1.67で、事後のそれは-1.46だった。端的に言えば、「文法の正確さ」と「課題達成度」において両グループ共に生徒評価の影響を受けたと考えられる。

次にWald Statistics を行い、それぞれの評価タイプにおける事前・事後テストの4つの観点それぞれの評価について統計的な有意差があるかを調べた。Wald Statistics は、有意水準0.05で臨界値1.96の時の帰無仮説を検定するものである。表10によれ

ば、自己評価グループでは、事前事後テストにおいて、総合得点と「正確な文法の使用」の観点で統計上の有意差があった。総合得点については、課題困難度の観点から見ると、1.44logits の差があった。Wald Statistics はこれを統計上有意だと確認した:t pre-composite、post-composite (283) = 5.63、p < .05、d = .50。よって、帰無仮説は棄却された。また、事前事後の自己評価グループの「正確な文法の使用」もまた帰無仮説は棄却された:t pre-grammar、post-grammar (113) = -2.74、p < .05、d = .60。つまり、自己評価グループの生徒は自己評価活動の後に総合得点と「正確な文法の使用」において課題困難度に違いを見せたということで、事前テストよりも課題困難度は低くなったので、能力の向上が見られたという意味になる。一方、ピア評価の生徒は、事前事後テスト間の「文法の正確さ」の観点において帰無仮説は棄却された:t pre-grammar、post-grammar (181) = -5.25、p < .05、d = .40。だが、自己評価グループの生徒とは異なり、総合得点の事前事後テスト間の帰無仮説は棄却されなかった。まとめると、自己評価グループは、ライティングの後に自己評価活動を行うとライティングの総合得点が向上し「正確な文法の使用」の観点も向上を見せた。一方、ピア評価グループの生徒は、総合得点の向上はできなかったが、「正確な文法の使用」においては向上をすることができた。

表10.

自己評価グループとピア評価グループの生徒に関するMeasure LogitsとWald Statistics

	Self-assessment group					Peer assessment group				
	Observed Average		Logit		Wald (pretest & post-test)	Observed Average		Logit		Wald (pretest & post-test)
	Pretest	Post-test	Pretest	Post-test		Pretest	Post-test	Pretest	Post-test	
Composite	12.27	12.71	.33	-.33	-5.63*	12.17	12.25	.05	-.05	-1.02
Grammar accuracy	2.33	2.64	2.08	1.76	-2.74*	2.22	2.74	.24	-.36	-5.25*
Vocabulary	2.98	3.07	-.68	-.79	-1.06	2.80	2.96	.63	.9	-1.61
Structure & Coherence	3.34	3.42	.23	.37	-.94	2.96	3.39	.63	.19	-.43
Task fulfilment	3.58	3.62	-.99	-1.27	-.47	3.52	3.68	-1.67	-1.46	-1.61

注. *p < .05

質的研究結果

生徒12名と教師二人を対象とした半構造インタビューの分析をもとに課題1と2についての分析をする。

自己評価グループ生徒のセオレティカル・コード

自己評価グループの半構造インタビューのデータは、第一段階で52のオープン・コードにまとめられ、それらは帰納的に16のアキシャル・コードに抽出された。これらは、10のカテゴリーに関連付けられ、それが4つのセオレティカル・コード(emergent themes)に分類された。この分類をもとに、カテゴリーとコード間の関係を理論化した。4つのセオレティカル・コードは、(A)評価基準と評価内容に対する意識化、(B)外部の存在、(C)英作文に対する学習態度、(D)英作文と自己評価によって促される情意、である。

(A)評価基準と評価内容に対する意識化

自己評価グループの生徒に対するインタビューの中で最も頻繁に登場したのが評価基準に対する意識づけに関する言葉だった。つまり、自己評価グループの生徒は、語彙、文法、英文構成や英文内容の一貫性に対する意識が高まったと英語力如何にかかわらず、強調をしていた。特に、語彙や文法を正確に使うように心がけたというコメントが多かった。こうした意識の高まりが英作文を校正し、書き終わった後に誤りをチェックする習慣の定着につながったようだった。例えば、英作文の得点が高かった生徒(SHSS1)は次のように答えている:

> *自分は英作文を書き終わる前に正確さをチェックする習慣が身に着きました。英作文を提出した後でさえ、自分が書いた英文の単語や文法の正確さを調べるために辞書を使って調べたり、友達に尋ねたりしました。(SHSS1)*

このように、自己評価は評価基準に気をつけることによって校正をして見直す習慣づけにつながったと言える。また、英文構成や内容の一貫性についても自己評価グループ6人の内3人が英作文を書き終わった後に、書いた内容に筋が通っているか構成が整っているかを見直すようになったと答えている。英作文が高得点だった生徒(SHSS2)は次のように述べている:

> *自己評価をすることによって、私は英文と英文のつながりが自然かどうかを考えるようになりました。授業で習った英文構成に自分の英作文構成がしたがっているかを考えるようになりました。例えば、理由の述べ方や英作文での結論文の書き方などです。自己評価表は、自分の英文がうまく構成されているかどうか振り返るのに役立ちました。でも、自分一人で振り返るのは難しかったです。(SHSS2)*

このように、SHSS2は、評価基準が英文校正や一貫性を見直す助けになったと認め、授業で学習したディスコース・マーカーを使って自分の英文を構成しようとした。以上のように、自己評価グループの生徒は、自己評価の時に使用した評価基準を語彙、文法、構成や一貫性を見直す道具として使い、その振り返りが英文を構成する意識の向上につながったと考えられる。二人の教師であるJETB と NETC も生徒たちが特に語彙や文法等の言語使用に注意を払い、それを見直すようになったと発言をしていた。また、JETBによれば、評価基準表が生徒たちに刺激を与え、生徒たちが現在知っている知識を最大限発揮しようという意欲につながったという。まとめると、自己評価グループの生徒は、自己評価の活動が評価基準の意識付け、特に語彙や文

法などの言語使用の正確さに対する意識の向上に良い影響を与えたと考えていることが分かった。

(B)外部の存在

自己評価グループでインタビューに答えた生徒の内半分が外部の存在である「読み手の存在」と「教師の評価」についての言及をした。英作文の点数がほぼ平均点だった自己評価グループの生徒(SMSS1)と英作文の点数が低かった生徒(SLSS2)は、自分が書いた英作文が読み手にとって理解可能かどうかについてわからないという不安を語った。自己評価をすることによって、生徒たちは自分の英作文に客観性があるかどうかや読者にとって理解が可能かどうかについて考えるようになった。今回の研究では、教師や友人が英作文を読むということは設定されていなかったからこそ、自己評価グループの生徒にとって外部の存在が浮かび上がってきたのかもしれない。

さらに、自己評価グループの生徒たちが外部の存在として、教師の存在の重要性や教師の指導の必要性を指摘した。これは、生徒たちが通常は教師の指導や助けを期待して英文を書いているので、それがない時の不安感が吐露されたのである。また、自己評価の信頼性について疑いの気持ちをもつ生徒もいて、教師評価がこうした不安を解消してくれると英作文が平均点くらいだった生徒(SMSS2)は次のように訴えていた。

> *英作文の評価は先生によってされるべきだと思います。なぜならば、私は自分の英語力に自信がないからです。自分の英語が正しいかどうか判断するのは自分にはとても難しかったです。(SMSS2)*

また、インタビューに答えた半分の生徒が自己評価の効果を認めながらも、読者がいないので自分の考えを共有できない不満や自己評価が自己満足で終わってしまう不安もコメントしていた。英作文の点数が低かった生徒(SLSS2)は次のように述べている。

> *私は英作文は読む人にとってわかりやすくなければいけないと思います。でも、自分の英語が他の人にとって意味が通じるものなのか自分ひとりでは、よくわかりません。ひとりでは、自分の英語が理解可能かどうかは想像できません。(SLSS2)*

本研究では、自己評価の活動に教師や友人のフィードバックを介在させなかったので、教師などの外部の存在に対する意識が生徒たちの中でより強くなったようである。特に教師は生徒たちに信頼のおけるフィードバックや新しい知識を与える存在としてその必要性が強調されていた。インタビューを受けた二人の教師も客観性や学習意欲が向上をするので、読者の存在が生徒たちに与える効果や重要性について言及をしていた。

(C)英作文に対する学習態度

英作文に対する学習者の意識は、「英文を書くことに対する態度」や「学習者の考

え方に対する効果」の観点から浮き彫りになった。インタビューに答えた6人全ての自己評価グループの生徒が自己評価は自分の英文を書くことに対する態度に良い影響を与えたとコメントをした。例えば、英作文の点数が高かった生徒(SHSS1)は、現在の自分の能力と学習目標の間にあるギャップを理解させてくれたので、自己評価が自分の英語を書く学習態度が改善されたと次のように言及をしている。

> *自己評価をすることによって、自分の弱点がわかったのでもっと英語を勉強しようとやる気が出ました。今のレベルと最終ゴールのギャップを埋めなければだめだと思いました。自己評価をすると、自分を振り返ることができます。それがもっと勉強しようという気持ちにさせるのです。(SHSS1)*

一方で、6人の内2人の生徒が学習の方向性を見失ったと報告をしている。例えば、英作文の点数が平均程度だった生徒(SMSS2)は教師の指導がなかったことや他者からの導きがなかったことによって不安を感じたと述べている。

> *自分が何をすべきなのかわからなくなってしまったので自己評価について不安を感じました。誰かの助けがあれば、もっと自信をもって英作文を書いて自己評価をすることもできたと思います。(SMSS2)*

このように、自己評価の英作文に向かう態度に良い影響をインタビューに答えた全員が認めながらも一部の生徒は英語を書くことに対する意識についての負の影響も指摘していた。インタビューに答えた教師も自己評価が生徒たちの英作文に対する意識に正と負の影響の両方を与えると述べていた。JETBは、正の影響としては、自己内省を深めることによって、自己校正力を向上させ、より適切に英文を書いていけるようになると指摘している。また、NETCは、ほとんどの生徒たちが自分の考えや経験を英語で書くことを楽しみ、英語で書くことを価値ある体験だと認識していると述べている。一方、学習者の意識に及ぼす負の影響としては、JETBとNETCの両者ともに生徒によっては自己評価が学習に向かう姿勢を消極的にさせる原因にもなったとコメントした。本研究は、10日間の短期間に集中的に教師や友達の助けや指導がない形式の実験だったので、こうした負の部分が解決策のないまま、一部の生徒の情意の中で増強されたのかもしれない。よって、自己評価が生徒に与える負の影響を防止するためには、こうした生徒のストレスを緩和することが必要だ。そのためにも教師の助けや指導が良いタイミングで必要とされると二人の教師は回答をしていた。まとめると、自己評価は生徒の意識や学習態度に概して良い影響を与えるが、欲求不満などの負の影響を与えることもあるということがわかった。また、それらを抑止するためにも教師の適切なガイダンスが期待されている。

(D)英作文と自己評価によって促される情意

インタビューに答えた6人の内4人の自己評価グループの生徒が自己評価をすることが自分に対する自信につながったと答えた。具体的には、課題を完成したときに感じる自己達成感や十二分に努力をして課題をやり遂げたことに対する充足感を指摘した生徒が多かった。英作文が平均点程度だった生徒(SMSS1)は、英作文は苦手だったが今回の活動で努力をしたことに触れ、達成感を感じたとコメントをしている。

英語は得意ではないので、英語で自分の考えを書くということはとても難しかったです。でも、自分は課題に対して頑張って取り組みました。ただ、自己評価に使ったルーブリックには、どれだけ努力したかを書き込む欄がなかったので、少し残念でした。でも、とにかく頑張って、努力をして英作文を書きました。だから、今は英作文にたいして苦手意識はありません。(SMSS1)

また、インタビューに答えた自己評価グループの6人の生徒の内4人が英語の勉強に対する意欲が高まったとコメントをしている。英作文で高い点数であった生徒(SHSS2)は自己評価活動のおかげで英語の勉強への動機付けとなり、学習方法を再考するようになったと答えている。しかし、英作文の点数が低かった生徒(SLSS2)は、自己評価活動は、自分の現在の能力と目標のレベルのギャップを埋めることはできず、かえって自己評価活動が彼女をいらだたせることになったという。つまり、自己評価活動が、自分の英語力の欠如を際立たせることになり、英作文を書くことに対して意欲が減退したとコメントしていた。よって、この生徒は自己評価活動について負の感情を抱いていると述べている。

私も英語で文章を書こうと努力をしたし、自分の英作文を見直し評価をしようとしました。でも、自分の英語のできなさは全く変わらなかった。とにかく、(自己評価をしても)英語は不得意なままであることがよくわかりました。だから、英作文に対して良い気持ちは持っていません。(SLSS2)

インタビューに答えた二人の教師も、自己評価活動と英作文を集中的に書いたことが生徒の情意に与える正と負の影響について報告をしていた。NETCによれば、自己評価は生徒によってその情意に与える影響は異なり、英語力が高い生徒は肯定的に自己評価の結果を受け入れるが、英語力が低い生徒は劣等感を強く感じる傾向があるとのことだった。JETBもこのNETCのこの観察に同意をしていて、自己評価が自分の成果と目標値との距離を自覚させるので、英語力が低い生徒は自分の置かれた厳しい現実にやる気が萎えてしまうのかもしれないと答えていた。以上のように、自己評価は英語ライティングに対して自己達成感や満足感、そして意欲を高めることもできるが、生徒によっては、学習目標と自分のレベルの開きに目覚めストレスを感じることもあることがわかった。

ピア評価グループ生徒のセオレティカル・コード

ピア評価のグループの生徒のインタビューからは、最初のコーディングから38のオープン・コードを抽出し、それを22のアキシャル・コードに分類することができた。また、これらは8つのセオレティカル・コードにまとめられた。8つのセオレティカル・コードの関係を分析するために、それぞれのカテゴリーの核となる要素と概念を明確に表現している部分をコード化し、続けて互いの関係に注意を払いながら、3つのセオレティカル・コードを決定した。8つのセオレティカル・コードは、以下の3つのセオレティカル・コードに抽出された:(E)友達からの学習効果、(F)メタ認知の発達、そして(G)英語学習に対する意識。

(E)友達からの学習効果

ピア評価のグループの生徒へのインタビューから最も頻繁に言及されたのは友達から受ける影響についてだった。インタビューを受けた6人全員の生徒が友達の英作文から何かしらを学ぶことができたとコメントをしていた。6人の内、5人の生徒が友達の英作文を言語使用、構成、課題達成、そして英文内容の独創性の観点から賛辞を送っていた。さらに、6人全員が友達の英作文を読むのは楽しく、英作文を読むことによって友人関係が深まり、友達の考えや経験を知ることができるのは喜びであると語っていた。また、友達の英作文を評価するのも楽しいと述べていた。これが、ピア評価のグループの生徒のインタビューの最も際立つ特徴だった。よって、生徒たちは、ピア評価活動を通して、友達とのインタラクションを確立することができたことを肯定的に受け止めていた。例えば、英作文の評価が低かった生徒(PLSS1)は、友達の英作文を読むことによって、友達の考えを知ることができたとして次のように述べている。

> *友達の考えや興味を発見できるので、クラスの人たちの英作文を読むのはとても楽しかったです。同じクラスでもあまり話す機会がなかった人たちと会話をするチャンスになりました。英作文の授業やピア評価活動が終わった後も、もっと話がしたいなと思いました。(PLSS1)*

また、ピア評価のグループの生徒6人の内4人の生徒が異なる意見や考え方を知ることができるので、他者の見方を共有する大切さについてコメントをしていた。英作文の評価が平均点レベルである生徒(PMSS1)は、ピア・リーディングによって友達から英語の書き方を学ぶことができるし、新しい知識を得ることもできると言及している。

> *友達の書いた英文を読んだおかげで、新しい単語や使い方を知ることができて良かったです。英文をどう構成するかを学ぶこともできました。だから、次に英作文を書くときには、友達の英文の書き方をまねたいと思っています。友達が頑張っているので、自分も頑張らなきゃという気持ちにとてもなりました。(PMSS1)*

PMSS1の言葉は、ピア評価が生徒に、友達、読者、評価者、そして書き手という複数の役割をする機会を与えるということを示唆している。言い換えれば、ピア評価をした生徒は、多くの視点で友達の英作文を読んで評価をしているのである。こうした体験が英文の書き手としての成長に肯定的な影響を与えていると考えられる。この点については、二人の教師も同意をしていた。特に、JETBは、ピア評価を授業の学習ツールとして有効であるとコメントをしている。つまり、ピア評価はライティングの授業を活発な雰囲気にさせ、生徒たちを協同的に学ばせることを可能にさせるのだ。

(F)メタ認知の発達

「メタ認知の発達」というセオレティカル・コードは、「読者の存在を意識化すること」と「読み手にとって読みやすい英文を書くことを意識化すること」の二つのコードから構成される。なお、メタ認知とは、自分の認知活動を客観的にとらえることを意味する(三宮、2018)。まず、「読者の存在を意識化すること」については、ピア評価グループのインタビューに答えた6人の生徒全員が、自分の英作文を読む友達の存在を意識

して書くようなったと答えていた。英作文の評価が平均点レベルだった生徒(PMSS1)は、友達が読むことを強く意識して、例示を工夫するなどして説得力がある英文になるように努力をしたとコメントをしていた。

> *クラスメイトが自分の英文を読むことがわかっていたので、わかりやすい英文を書くために注意を払って書こうと思いました。特に、友達に自分の考えや体験をわかってもらうために客観的に書こうと思ったので、具体的な例を出そうと思いました。ピア評価をする時には、友達の英文を読みながらも自分の英作文のことを思い出して、頭の中で(自分の英作文と友達の英作文を)比較をしました。(PMSS1)*

また、英作文の評価が高い生徒　(PHSS2)は、友達から自分の英作文をほめてもらい学習意欲が高まったと述べている。さらに、この生徒は、友達が評価をするので、英作文を書く時の意識に変化があったとも言及した。この意識の変化こそが、他者の存在の認識であり、それを意識するメタ認知の発達につながったと考えられる。

> *ピア評価はいつも自分に読む人のことを考えさせてくれました。ピア評価がなければ、自分が書いた内容は違うものになっていたと思います。友達が自分の英文を読んでくれることやピア評価が良いことかどうかはわからないが、書いている間、クラスメイトの誰かがこれを読むということはいつも意識していました。自分の英作文を読んで良い評価をくれるといいなと思っていたし、自分が評価を意識していることも自覚していました。(PHSS2)*

これついてJETB は、ピア評価活動の最中に生徒が、書き手や読み手、評価者、そして友達などの立場を自然に変化させていることがメタ認知の発達を導いたとコメントをしている。つまり、様々な役割を担っていることが、生徒のモニター的意識を目覚めさせ、ライティングの力を向上させると語っている。このように、ピア評価は、学習者にライティングや評価活動をモニタリングさせることが可能で、それがメタ認知の発達に寄与すると考えられる。

(G)英語学習への意識変化

ピア評価に参加をしてインタビューを受けた6人の生徒全員が、ピア評価が英語で書くことに与える心理的影響について語っていた。また、肯定的影響を意味するコード数の方が、否定的な影響のコード数よりも多かった。肯定的な言及としては、ピア評価が英語学習に対する意欲を高め、意欲的に英作文を書きたいという刺激を与えたということだった(PMSS2;PHSS1)。この刺激は、ライティングのみならず、リーディングや英語の各種検定試験、そして大学入試に対する意欲にもつながったとコメントをしている(PHSS1)。また、友達の英文を読みフィードバックをもらうことによってクラスに良い雰囲気が醸造されることを高く評価している生徒もいた(PMSS1)。

一方で、ピア評価が生徒の情意に与える負の影響についてコメントをしている生徒もいた。例えば、ピア評価の難しさについて、英作文の評価が平均点のレベルだった生徒(PMSS2)は以下のように述べている。

ピア評価は、いつもではないが、時々、勉強への意欲を刺激してくれるし他の人からたくさん学ぶことができるので良いと思います。でも、課題(英作文)を全く完成していない人もいるので、そういう人の英作文を読み評価をする場合は、どう評価をしてよいのかわからなくて困りました。(PMSS2)

このようにピア評価の意義について疑問を投げかけている。また、ピア評価は評価をする友人との人間関係に影響をされるので、良い評価を友達からもらってうれしいが、果たしてその評価は正直なものなのか疑問を抱くともPMSS2はコメントしている。別の英作文評価が低い生徒(PLSS2)は、ピア評価の英語ライティング向上に効果的かどうかについても疑問を呈した。なぜならば、彼女自身が自分の英語力に不安を感じていて友達の英語を読み信頼性のある評価ができる自信がないとコメントをしているからだ。

私は英語が不得意なので、他の人の英作文を判断する能力はないと思います。だから、自分のピア評価は良いものではないし正しくないと思います。だから、評価をされた人には悪いなと思っているので、ピア評価は好きではありません。自分には、他の人の英文が正しいのかどうかよくわからないし、一人では判断するのは難しかったです。(PLSS2)

このように、ピア評価は個々の生徒の英語力や英語に対する態度、そして生徒同士の関係性によって影響を受けるようである。これらの生徒によるインタビュー内容に呼応するように、二人の教師もピア評価には正の局面と負の局面があるので、授業への導入には注意が必要だと語っていた。JETBは、ピア評価を授業で扱うときには、特にクラスの人間関係に気配りをし、ピア評価の前に簡単なルールを設けてやる気をなくす生徒が出ないようにしていると述べている。

ピア評価は生徒に競争意識を持たせ、生徒の客観性ややる気を向上させることができます。でも、クラス内の人間関係に対する配慮は必要でしょう。匿名でのピア評価をするのも一つの方法でしょう。ピア評価をする前に、生徒には、他者に意見を言うときに注意をしなければならないこと、悪い点ばかりではなく、必ず良い点を見つけてあげてそれを言うようにと指導をしています。(JETB)

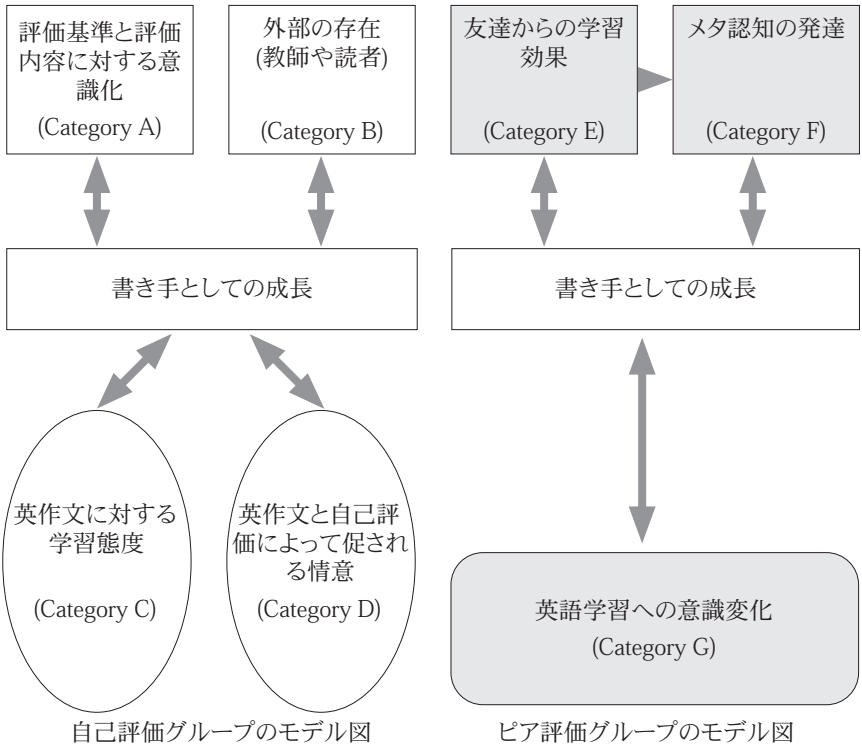
このように、教師達もピア評価をより実りあるものにするように努力をしていることがわかる。NETCも生徒は教師からの言葉や評価よりも友達の評価の方を深刻に受け止める傾向があると言及している。よって、教師たちは、ピア評価が時に生徒同士の関係に影響を与えることも考えて注意を払うべきだろう。以上のように、ピア評価がライティングへの意識に与える影響は、教師による配慮や導入前のルールで肯定的なものになるようだ。

半構造インタビューの結果から見た自己評価とピア評価の比較

Corbin と Strauss (2015)によると、コーディングの過程でコードの互いの関係性に着目をしてダイアグラムを作成することがコーディングの関係性や秘匿された重要なテ

ーマを暴くので、図4のように自己評価とピア評価がライティング活動に与える影響を二つのモデルにまとめた。

図4.
自己評価グループとピア評価グループモデル比較



自己評価とピア評価の生徒に共通するのは、学習者評価を通じて両タイプの生徒が英語の書き手としての成長を目標(コア)に置いていることである。さらに、サブ・カテゴリー(セオレティカル・コード)の矢印は全てこの目標(英語の書き手としての成長)を目指している。一方、自己評価とピア評価のモデルで違う点は、サブ・カテゴリーの「外部の存在」の意味が異なっている。これは、自己評価では、外部の存在は物理的にはないが、ピア評価では友人からの影響を直接受けることを意味している。他の相違点としては、ピア評価モデルにはメタ認知の発達が提示されているが、一方、自己評価モデルには自己評価活動が与える自己達成感などを意味する情意が提示されている。

考察

ここでは、量的研究結果と質的研究結果を統合し、まず二つの研究課題を考察し、その後に研究課題3である「学習を促す評価(LOA)」における二つの学習者評価の役割を考える。最初の研究課題である学習者評価の信頼性についてであるが、量的研究結果の通り、厳格さにおいては、自己評価は教員評価と同等の厳しさを示した。これは、質的研究結果も、自己評価活動が英作文の評価基準や評価項目に対する理解を深めるのに有効であったことを示すので、自己評価は評価の厳格さにおいては信頼性の高い評価法といえよう。だが、量的研究結果によれば、自己評価は一貫性には乏しく、質的研究結果の生徒からの発言からも、生徒たちは自己評価に対する自信が不足気味であることがわかったので、質的研究結果は量的研究結果を支持しているといえる。先行研究によれば、自己評価の信頼性は研究対象によって若干の差異はあるもののある程度高いとされているが、年齢が若い学習者の場合は評価の信頼性に揺れが生じると言われているので(Esfandiari & Myford、2013)。本研究も若い学習者を対象としたので、信頼性に年齢が影響を与えたと考えられる。また、本研究は、信頼性を厳格さと一貫性の観点から分析したので、自己評価の信頼性の新しい側面がわかったといえる。

一方、ピア評価は、量的研究結果によれば、教師評価や自己評価にくらべ、評価は甘く、評価の一貫性も見られなかった。質的研究結果の半構造的面接からは、ピア評価は生徒の英語力や生徒同士の人間関係にも影響を受けていることが観察され、ピア評価の信頼性を疑問視する意見も一部の生徒から発声された。よって、量的と質的ともに本研究におけるピア評価の信頼性については低いと考えられる。ピア評価の信頼性は高めであるとする先行研究結果が多いが、その多くは大学生を対象とした研究である。日本の高校生を対象とした分析結果は少ないので、若年学習者によるピア評価の信頼性が新たにわかったといえる。

次に、第二の研究課題である生徒評価のライティング能力への影響に関してだが、自己評価とピア評価は共に文法の正確さに良い影響を与えるが、その影響の質が異なることが考察された。上記のように、厳格さにおいて教師評価と同等の自己評価は英作文の評価基準に対する認識を高め、自己統制や内省力を深めると考えられる。質的研究結果が示すように、自己評価によって醸成されたその強い自己認識は、語彙や文法知識の不足を意識することにおいて特に強く現れていた。しかし、こうした強い自己認識や内省は、英語力に対する自信喪失にもつながることもあるが、文法の正確さを向上させるために肯定的な影響を与えた。さらに、自己評価グループにおいては、総合得点の向上も見られた。また、質的研究結果の二つの評価法を比較したモデル(図4)が示すように、自己評価グループの生徒の方が、ピア評価グループの生徒よりもより評価基準や評価内容について頻繁に言及をしていたので、評価基準をより強く意識し、その気づきが英作文の全体的向上に役立ったと考えられる。生徒評価の学習効果としてライティングにおける校正能力への向上が先行研究でも報告されている。本研究の結果でも英文を書いた後に校正する習慣に結び付いたという結果が出ているので、先行研究の結果と同じである。

一方、ピア評価においては、生徒同士で英作文を読み合い評価をすることによって、同級生との交流を楽しみ、友達が書いた英語から学ぶことが文法力向上の助けになったと思われる。ピア評価のグループの生徒たちの多くは、友達の英作文を読む楽しさを指摘し、大半の生徒が肯定的にピア評価やピア・リーディングを受けとめてい

た。自己評価グループの生徒たちよりも自然に楽しんで評価活動を行っていたことが質的研究結果から観察された。友達同士で励ましや誉め言葉を交換することが、教室内の雰囲気を和やかにし、生徒間の交流を深め、そうした協同的学習スタイルが英作文に取り組む態度に効果的な影響を与えたと考えられる。また、ピア評価を通して、書き手、評価者、読み手、友達等の多様な役割を自然に担うことが、メタ認知の発達を促したことも、友達の英作文から学ぶことを効果的に促したと考えられる。

次に、第三の研究課題である「学習を促す評価(LOA)」における二つの学習者評価の役割について考察をする。先行研究が示すように、自己評価とピア評価の双方とも、普段は学ぶ立場の学習者が立場を変えて評価をする役割に代わることによって、学習者を学びの中心に置くことができるので(Leahy & Wiliam, 2012)、ライティング授業の活性化につながり、学習者による評価は「学習を促す評価(LOA)」の目的に合致すると考えられる。また、信頼性には欠けていたが、二つの評価方法ともに学習者に評価基準や評価内容に関するメタ認知力を上昇させた。それによって、学習者にとって何が必要で何が不足しているのかを自分で考えさせることができた。これが、学習者のメタ認知の発達や自律的学習の助けとなり、学習を促していくことになることが本研究では明らかになった。また、教師にとって適切なフィードバックをするための重要な情報となりうることも本研究の結果によってわかったことである。本研究では、教師は観察者で評価活動の介在やその後のフィードバックはしなかったが、質的研究結果によれば、生徒は学習者による評価だけではその信頼性に対し不安を感じ、教師によるフィードバックを期待していることがわかった。また、知識の不足を自覚した時に、次の学習段階に進むための助け(scaffolding)も望んでいた。自己評価によって内省し次に何をすべきかを自覚することはできる。また、友達から賛辞をもらっても、具体的に次にどう前に進めばライティング能力が向上するのかを具体的に助言してくれる友達は少ないことがピア評価グループの生徒は不満だったようだ。このように、フィードバックのみならず、前に進ませる(feed-forward)指導や助言こそが、生徒が教師に求めることだろう(Hamp-Lyons, 2017)。よって、「学習を促す評価 (LOA)」をうまく機能させるためにも、学習者による評価によって、学習者を評価の中心に置きながらも、教師による適切な介在は必要だろう。これは、学習者による評価のみではその信頼性のおいて不十分であり、たとえ学習者の自立度が向上したとしても、学習者の生徒評価に対する自信や信頼性には不安が残るためである。教師は、今後の学習に必要なことを示す役割が求められていて、自己評価とピア評価の結果は、具体的なその判断を下す資料となる。これまでの「学習者による評価」に対する先行研究は、生徒評価が教師評価の代替評価として可能かということが主要な研究課題の一つだったが、学習者による評価の本質は、学習を前に推し進めるための重要な情報を教師に提供することにあると考えられる。

結論

本研究は、日本の高校生を対象に、自己評価とピア評価という二つのタイプの評価法の信頼性やライティング能力への影響を探究することを目的とした。二つの評価方法を比較することによって、両者の類似点や相違点を明らかにし、それぞれの特質を具体化し「学習を促す評価(LOA)」を進める足掛かりとすることも目的とした。生徒評価は単なる教師評価の代替ではなく、教師評価とは異なる視点や特質をもつ評価法であるので、教師はその特質を理解し授業に採用することが求められている。本研究

は、自己評価とピア評価各々の特徴を明らかにするために、できるだけ他の教育的影響を避ける目的で、短期間に集中的に実験を行った。しかしながら、長期的な視点が不足するために、「学習を促す評価(LOA)」が目指す学習の永続性に関する分析と考察が不足している点が限界点だと言える。また、それぞれの評価タイプのデータ数に違いがあるので、それが誤差の大小に影響している可能性がある点も限界点だと言える。今後の研究としては、生徒評価へのフィードバックやフィードフォワードを教師がどのようにいつすべきかについての分析や生徒評価の信頼性と英語力の関係についての分析、さらに、自己評価とピア評価を両方取り入れたハイブリッド形式の効果や適切な運用方法についてさらに進めていく必要がある。

注. 付録は https://osf.io/nuvgx/?view_only=f1699ee01b454d029ee48e94b007edfeを参照。

All appendices are available from the online version of this article at https://jalt-publications.org/jj.

すべての付録は、https://jalt-publications.org/jjのオンライン版から入手できます。

謝辞

本調査を実施するにあたり、参加者の皆様には多大なご協力をいただきました。また、原稿執筆にあたっては、査読者の皆様から貴重なご助言をいただきました。心より感謝申し上げます。本研究は一部著者の博士論文に基づいています。

著者略歴

大井洋子は、神奈川県で高校教諭として英語教育に携わり、早稲田大学大学院にて博士号を取得。現在は清泉女子大学言語教育研究所で言語教育コーディネーターとして言語教育に携わっている。

参考文献

戈木クレイグヒル滋子(2018).『グランデッド・セオリー・アプローチ 改訂版 理論を生みだすまで』新曜社.

三宮真智子(2018).『メタ認知で＜学ぶ力＞を高める:認知心理学が解き明かす効果的学習法』北大路書房.

平井明代・横内裕一郎・加藤剛史(2018).「項目応答理論:標本依存と項目依存を克服した測定を実現する」平井明代編『教育・心理・言語系研究のためのデータ分析— 研究の幅を広げる統計手法』(pp. 94-137)東京書籍.

行森　まさみ(2018)「学習指導要領における英語4技能再考」『表現学』第4号　(pp.10-17) 大正大学表現学部表現文化学科.

Alsowat, H. H. (2022). An investigation of Saudi EFL teachers' perceptions of learning-oriented language assessment. *European Journal of English Language and Literature Studies, 10*(3), 16–32. https://doi.org/10.37745/ejells.2013/vol10no3pp.16-32

Aslanoglu, A. E., Karakaya, I., & Sata, M. (2020). Evaluation of university students' rating behaviors in self and peer rating process via many facet Rasch model. *Eurasian Journal of Educational Research, 89*, 25–46. https://doi.org/10.14689/ejer.2020.89.2

Andrade, H., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education*, *17*(2), 199–214. https://doi.org/10.1080/09695941003696172

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Open University Press.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan*, *86*(1), 8–21. https://doi.org/10.1177/003172170408600105

Bond, T. & Fox, C. (2015). *Applying the Rasch model fundamental measurement in the human sciences*. Routledge.

Boud, D. (1992). The use of self-assessment schedules in negotiated learning. *Studies in Higher Education, 17*(2), 185–200. https://doi.org/10.1080/03075079012331377621

Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education*, *18*, 529–549. https://doi.org/10.1007/BF00138746

Butler, Y. G. & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing*, *27*(1), 5–31. https://doi.org/10.1177/0265532209346370

Carless, D. (2007). Learning-oriented assessment: Conceptual basis and practical implications. *Innovations in Education and Teaching International*, *44*(1), 57–66. https://doi.org/10.1080/14703290601081332

Chang, W-.C., & Chan, C. (1995). Rasch analysis for outcomes measures: Some methodological considerations. *Arch Phys Med Rehabil, 76*, 934–939. http://www.archives-pmr.org/article/S0003-9993(95)80070-0/pdf

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, *98*(4), 891–901. https://doi.org/10.1037/0022-0663.98.4.891

Corbin, J., & Strauss, A. (2015). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage.

de Wever, B., van Keer, H., Schellens, T., & Valcke, M. (2011). Assessing collaboration in a wiki: The reliability of university students' peer assessment. *The Internet and Higher Education*, *14*(4), 201–206. https://doi.org/10.1016/j.iheduc.2011.07.003

Derakhshan, A., Shakki, F., & Sarani, M. A. (2020). The effect of dynamic and non-dynamic assessment on the comprehension of Iranian intermediate EFL learners' speech acts of apology and request. *Language Related Research*, *11*(4), 605–637. https://lrr.modares.ac.ir/article-14-40648-en.html

Dieten, A. J. (1989). The development of a test of Dutch as a second language: The validity of self-assessment by inexperienced subjects. *Language Testing*, *6*(1), 30–46. https://doi.org/10.1177/026553228900600105

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Peter Lang.

Esfandiari, R., & Myford, C. M. (2013). Severity differences among self-assessors, peer-assessors, and teacher assessors rating EFL essays. *Assessing Writing*, *18*, 111–131. https://doi.org/10.1016/j.asw.2012.12.002

Engelhard, G. (2018). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, *70*(3), 287–322. https://doi.org/10.3102/00346543070003287

Farrokh, F., Esfandiari, R., & Schaefer, E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *JALT Journal, 34*(1), 79–102. https://doi.org/10.37546/JALTJJ34.1-3

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Aldine. https://doi.org/10.1097/00006199-196807000-00014

Green, A. (2017). Learning-oriented language test preparation materials: A contradiction in terms? *Language Testing and Assessment*, *6*(1), 112–132. https://doi.org/10.58379/SFUN3846

Hamp-Lyons, L. (2017). Language assessment literacy for learning-oriented language assessment. *Language Testing and Assessment*, *6*(1), 88–111. https://doi.org/10.58379/LIXL1198

Jones, N., & Saville, N. (2016). *Learning oriented assessment*. Cambridge University Press.

Leahy, S., & Wiliam, D. (2012). From teachers to schools: Scaling up professional development for formative assessment. In J. Gardner (Ed.), *Assessment and learning* (pp. 49–72). Sage. https://doi.org/10.4135/9781446250808.n4

Lee, H. (2017). The effects of university English writing classes focusing on self and peer review on learner autonomy. *The Journal of Asia TEFL*, *14*(3), 464–481. https://doi.org/10.18823/asiatefl.2017.14.3.6.464

Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). MESA Press.

Linacre, J. M. (2021). *A user's guide to FACETS Rasch-model computer programs: Program manual 3.83.5.* http://www.winsteps.com/manuals.htm

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing*, *26*(1), 75–100. https://doi.org/10.1177/0265532208097337

McDonald, B., & Boud, D. (2003). The impact of self-assessment on achievement: The effects of self-assessment training on performance in external examinations. *Assessment in Education: Principles, Policy & Practice*, *10*(2), 209–220. https://doi.org/10.1080/0969594032000121289

Ministry of Education, Culture, Sports, Science and Technology. (2011). *Wagakuni niokeru kokuren jizoku kano na kaihatsu no tameno kyoiku no 10nen jisshi keikaku* [Education for sustainable development for our country]. https://www.mext.go.jp/unesco/004/detail/1359220.htm

Ministry of Education, Culture, Sports, Science and Technology. (2018). *Heisei 29 nenndo eigokyoiku chosa kekka* [The results of fiscal 2017 survey of proficiency in English]. https://www.mext.go.jp/component/a_menu/education/detail/__icsFiles/afieldfile/2018/04/06/1403469_02.pdf

Mok, M. M. C. (2012). Assessment reform in the Asia-Pacific region: The theory and practice of self-directed learning-oriented assessment. *Examinations Research*, *4*(33), 79–89. https://doi.org/10.1007/978-94-007-4507-0_1

Mulvey, B. (2016). Writing instruction: What is being taught in Japanese high schools, why, and why it matters. *The Language Teacher, 40*(3), 3–8. https://doi.org/10.37546/JALTTLT40.3-1

Oi, S. Y. (2018). The relationship between writing tasks in textbooks and can-do lists in terms of task difficulty. J*ournal of Pan-Pacific Association of Applied Linguistics*, *22*(2), 53–70. https://doi.org/10.25256/PAAL.22.2.3

Oi, Y. (2019). Japanese high school English teachers' perspectives on classroom writing assessment criteria: A needs analysis. *The Bulletin of the Graduate School of Education of Waseda University*, *27*(1), 159–176. http://hdl.handle.net/2065/00063295

Oi, Y. (2021). *Efficacy of student assessment as part of English writing instruction for Japanese high school student* (Publication No. 6413) [Doctoral dissertation, Waseda University]. National Diet Library. https://iss.ndl.go.jp/books/R100000039-I003071522-00

Orsmond, P., Merry, S., & Reiling, K. (2000). The use of student derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education*, *25*(1), 23–38. https://doi.org/10.1080/02602930050025006

Peirce, B. N., Swain, M., & Hart, D. (1993). Self-assessment, French immersion, and locus of Control. *Applied Linguistics*, *14*(1), 25–42. https://doi.org/10.1093/applin/14.1.25

Ploegh, K. (2009). In search of quality criteria in peer assessment practices. *Studies in Educational Evaluation*, *35*(2–3), 102–109. https://doi.org/10.1016/j.stueduc.2009.05.001

Runnels, J. (2014). Japanese English learner self-assessments on the CEFR-J's A-level can-do statements using four and five-point response scales. *The Asian Journal of Applied Linguistics*, *1*(2), 167–177. https://caes.hku.hk/ajal/index.php/ajal/article/view/13

Sadler, P. R., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment*, *11*(1), 1–31. https://doi.org/10.1207/s15326977ea1101_1

Saito, H., & Fujita, T. (2009). Peer-assessing peers' contribution to EFL group presentation. *RELC Journal*, *40*(2), 149–171. https://doi.org/10.1177%2F0033688209105868

Topping, K. J. (1998). Peer assessment between students in college and university. *Review of Educational Research, 68*(3), 249–276. https://doi.org/10.3102/00346543068003249

Tsui, A., & Ng, M. (2000). Do secondary L2 writers benefit from peer comments? *Journal of Second Language Writing*, *9*(2), 147–170. https://doi.org/10.1016/S1060-3743(00)00022-9

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. 255-274). De Gruyter. https://doi.org/10.1515/9781614513827-018

van Gennip, N. A., Segers, M. S. S., & Tillema, H. H. (2010). Peer assessment as a collaborative learning activity: The role of interpersonal variables and conceptions. *Learning and Instruction*, *20*(4), 280–290. https://doi.org/10.1016/j.learninstruc.2009.08.010

Wiliam, D. (2000). *Integrating summative and formative functions of assessment* [Keynote address]. European Association for Educational Assessment, Prague, Czech Republic. https://discovery.ucl.ac.uk/id/eprint/1507176/1/Wiliam2000IntergratingAEA-E_2000_keynoteaddress.pdf

Zarei, A. A., & Mahdavi, A. (2014). The effect of peer and teacher assessment on EFL learners' grammatical and lexical writing accuracy. *Journal of Social Issues & Humanities*, *2*(9), 92–97.

Zarei, A. A., & Usefli, Z. (2015). The effect of assessment type on EFL learners' goal orientation. *Journal of Language, Linguistics and Literature*, *1*(4), 112–119. http://creativecommons.org/licenses/by-nc/4.0/

Reviews

Developing Multilingual Writing: Agency, Audience, Identity. Hiroe Kobayashi and Carol Rinnert. Springer, 2023. xv + 353 pp. ¥24,337. https://doi.org/10.1007/978-3-031-12045-9_1

Reviewed by
Julia Christmas
University of Niigata Prefecture

L1 and L2 writing and the influence of both on each other have been researched extensively (Gonca, 2016; Uzawa, 1996; van Weijen et al., 2009). Fewer observations are available in one volume of L1, L2, and L3 writing, the developmental trajectories of multilingual writers (two or more languages), the specific strategies they use, or how they use their unique identities to create meaning in text and connect with the reader. Using a style that flows from theory to quantitative data and then to deeper qualitative analysis, the authors provide comprehensive insight into the concepts of agency, audience, and author identity.

While exploring new and previously gathered data, the authors promise a study of the development of multilingual writing. As the title indicates, the book offers a focus on multilingual writers (those writing with two or more languages) and the agency (control over text features and conventions), audience (potential, imagined, future readers), and identity of the writers that is co-created by the reader and progresses along a developmental path. It argues that this research includes new and innovative approaches to theories of SLA and L1/L2 writing that add to the current body of best pedagogical and research practices.

The book is organized into ten sections that fall under three main divisions: Part I: Development of Multilingual Writing; Part II: Interconnectedness of Agency, Audience, Identity; and Part III: Synthesis and Implications. In addition to these three Parts, the About and the Introduction sections guide the reader by outlining the theoretical approaches, aims, research

questions, and methodologies, and a review of current literature. It is here that, along with an explanation of where this research falls within the current literature, the authors illustrate how this book works to fill gaps in previous research. Part I offers four text-analysis-based studies that focus on English and Japanese writing using a cross-sectional approach. Part II presents more in-depth case studies to give a deeper examination of the "interconnectedness between text, audience by individual multilingual writers in two or three languages" (p. ix) and broadens the writing focus to include academic, creative, scholarly, and artistic genres. Part III is a synthesis of the research outcomes that emerged from Parts I and II and supplies the implications relating to theory, methodology, and pedagogy.

The academic writing style, while not particularly dense, does require either scaffolding or a solid knowledge base of best practices for teaching writing with writers or students who are bilingual or multilingual. While this volume is not recommended as an introductory text, readers familiar with SLA and writing-related concepts such as writer motivation, discourse analysis, and rhetorical features will not struggle with the ideas contained therein. However, further reading may be essential for understanding newer and emerging theories about multicompetence, complex systems theory, adaptive transfer, multilingual motivation, and translanguaging.

The authors specifically note, and I agree, that the book will be practical for researchers and teachers. One of the books appeals is the potential for its use in many educational and research settings. The authors clearly present how they assembled and examined their data, making it an asset to courses focused on teaching ESL writing in Master's or Ph.D. programs. Similarly, as the subject matter is focused on the newest set of best practices in multilingual writing instruction, it will add to the curriculum used by advisors of students in EFL teacher training programs as they work to help their budding practitioners understand how writing skills are developed. In my own case, I felt that it would be beneficial for experienced instructors who are already familiar with writing theory as an update to their current teaching of writing with multilingual learners.

The authors draw extensively on past and present scholarly research in the fields of SLA and L1/L2 Writing. In the review of literature, readers can find a neat and easily understandable explanation of previous views held by the SLA community regarding concepts such as multiple competence, complex systems theory, transfer (as a creative, dynamic, and fluid process), and translingualism contrasted with developing theories about these concepts. Furthermore, the authors add to the current scholarly literature by

offering innovative L2 text analysis of multiple genres to apply in other language contexts and break new ground by extending the empirical analysis of L2 writing to the same writers' L1 and L3 texts. The volume includes a comprehensive discussion of how this research builds on previous research while it gathers and adds to it, particularly regarding the social view of writing (Hyland, 2011; Prior, 2001). Closely related to the research of writing as a social act, the volume also explores the role of co-constructed writer identity, which has not been explored in length, where it connects to the relationship between the writer and the audience and the writer's awareness of the audience.

The aims of the authors, reached through extensive examination of both quantitative and qualitative data, are "to find out how multilingual writers become able to take conscious control over their own text construction so they can respond effectively to their expected audiences and realize their full potential as multiliterate members of society" (p. 2). Additionally, while taking the view that linguistic development is an integral part of being able to write well, they primarily focus on the process of writer development and agency rather than what is produced (p. 2). Speaking further of their aims and the results that emerge from their research, Kobayashi and Rinnert delineate three concerns related to the construction of texts as a social act: writer agency, audience expectations, and co-constructed nature of writer identity by writer and reader (p. 2). To deal with these concerns, they present a wealth of research that connects good writing with the level of writer awareness toward the audience.

Additionally, while highlighting limitations of their work, they point out potential areas for future research and help fill gaps in previous research, which looks at the role of writer identity in relation to audience expectations and writer agency as well as research gaps that offer an integrated focus on all three concerns. Their approach attempts to connect rhetorical text features, composing processes, and how "composing activities are associated with specific text features for individual writers" (p. 239). This allowed them to uncover "individual writers' distinctive use of a variety of strategies at both local and global levels" (p. 239). Furthermore, looking at voice as it relates to identity, their research also challenges conventions held about its development (p. 162). Thus, we can not only understand old lenses and frameworks, but also have access to new ones that will allow further study and pedagogical support for our students.

Moreover, by using case studies, which included traditional students as well as academics, and an artist, and detailed cross-sectional examination

of essays, they fulfill their aim to supply a guide for teachers by showing the "general developmental path from novice to advanced writers" (p. v). Via their methodology and through their SLA and L1/L2 writing theories, which view writing as a social action, readers are shown how students learn to draw on their own learned writing knowledge to raise levels of sophistication in their writing and make more connections with perceived audiences. From this, readers can also picture how to focus on developing these abilities in our own students—or students of our students.

Each part of the book builds on the previous section, helping the reader gain a deeper understanding of the prior discussion. At the same time, readers learn how the progression of ideas connects and integrates with the three concepts of focus: the development of writer agency, audience, and identity. The plentiful student writing samples greatly enhanced my understanding of how to compare writers at different stages of development, and the case studies reminded me that each of my learners comes to the classroom with a unique background and set of learning experiences. There are few specific examples of classroom pedagogy; however, from a teacher perspective, the look into how one can parse student work to track their writing development and how one could use that data, paired with the to support student growth, is something that I want to work on. After reading this book, I feel I am more equipped to help my students gain more meta-awareness of how to use their L1, L2, and L3 (when applicable) as resources upon which they can draw to have better writing experiences in all languages. I believe that I am now more knowledgeable and more likely to be able to empower my students with the knowledge that they have a more extensive skills repertoire than they may already be aware of.

Further strengths of this volume are found in the arguments and their substantiation addressed in the data. As previously discussed, the authors contend that writing is a social action, and substantial evidence for the accuracy of this is found in their demonstration of how students made decisions to use the rhetorical constructions of their essays based on their own experiences and beliefs about what is right for a specific audience. Another strength is the delineation of how to empower—how to help writers move from novice to beyond and how to support their journey towards becoming independent writers who have meta-awareness of agency, audience, and their ever-changing identity. This volume gives teachers new means to assess the strengths and weaknesses of their students and do further research in this field by looking at their students' writing using the same methods this volume's authors use.

To conclude, Kobayashi and Rinnert posit that writing is a social act; writers assess the audience, set goals, choose appropriate text features, and communicate ideas to an audience—perceived, imagined, or future. In agreement with previous research, they assert that there is a positive relationship between linguistic development and writing ability development—and choose a focus on the latter. Using a considerable amount of data, they look at two not-yet-well-researched elements: the exploration of the role of multilingual writer identity in writer development, particularly in its relationship with agency (the text features writers use including "diverse or innovative ways of using text features that they had internalized" (p. 238) which are then interpreted by the audience (reader). They assert that a gap also exists in writing research, i.e., there is no existing comprehensive and integrated examination of three main theoretical and pedagogical concerns of multilingual writing: writer agency, audience expectation, and the co-constructed (by writer and audience) identity of the writer—particularly the multilingual writer. The authors build a case for their arguments and, using a hybrid style of quantitative and rich case-study-based qualitative data resulting from in-depth interviews, surveys, recorded talk-aloud writing sessions, and retrospective stimulated recall pause data, show us the advantages of being a multilingual writer. We learn that multilingual writers have a repertoire of strategies that arise from their languages and that these are accessible to some extent, even for novice L2 writers. They give us answers to questions about strategies students use and ways to nurture a "balance between writing knowledge and language proficiency in L1 and L2" (p. 170) and L3. They show us that there is a continuum of development and use their research, along with that of others, to highlight the unending potential for growth. I plan to use it to inform my teaching and firmly believe it serves a valuable purpose for anyone involved in multilingual writing.

References

Gonca, A. (2016). Do L2 writing courses affect the improvement of L1 writing skills via skills transfer from L2 to L1? *Educational Research and Reviews, 11*(10), 987–997. https://doi.org/10.5897/ERR2016.2743

Hyland, K. (2011). The presentation of self in scholarly life: Identity and marginalization in academic homepages. *English for Specific Purposes, 30*(4), 286–297. https://doi.org/10.1016/j.esp.2011.04.004

Prior, P. (2001). Voices in text, mind, and society: Sociohistoric accounts of discourse acquisition and use. *Journal of Second Language Writing, 10*, 55–81. https://doi.org/10.1016/S1060-3743(00)00037-0

Uzawa, K. (1996). Second language learners' processes of L1 writing, L2 writing, and translation from L1 into L2. *Journal of Second Language Writing, 5*(3), 271–294. https://doi.org/10.1016/S1060-3743(96)90005-3

van Weijen, D., van den Bergh, H., Rijlaarsdam, G., & Sanders, T. (2009). L1 use during L2 writing: An empirical study of a complex phenomenon. *Journal of Second Language Writing, 18*(4), 235–250. https://doi.org/10.1016/j.jslw.2009.06.003

English Language Teacher Education in Changing Times: Perspectives, Strategies, and New Ways of Teaching and Learning. Liz England, Lía D. Kamhi-Stein, and Georgios Kormpas (Eds.). Routledge, 2023. xxii + 235 pp. Approx. ¥7,024. https://doi.org/10.4324/9781003295723

Reviewed by
Mary Hillis
Ritsumeikan University
Anastasia Khawaja
University of South Florida

In *English Language Teaching Education in Changing Times: Perspectives, Strategies, and New Ways of Teaching and Learning*, editors Liz England, Lía D. Kamhi-Stein, and Georgios Kormpas explore the impact of the COVID-19 pandemic on the field of teacher education in ELT. Their focus extends beyond a retrospective analysis of the challenges but offers a forward-thinking perspective, encouraging readers to consider the lessons learned, the adaptations made, and the continuously evolving needs of our profession. This volume, marked by its inclusion of authors from numerous countries and teaching contexts, underscores the shared challenges faced by educators worldwide. While coping with physical, emotional, and economic aspects of the pandemic, teachers and leaders crafted their responses to unprecedented challenges in the workplace. Ultimately, the book encourages readers to reflect on what can be applied to their unique contexts, thereby fostering a collective understanding of the changing landscape of English language teacher education.

In the introduction, the editors write that they aim to address "... how the COVID-19 pandemic acted as a catalyst for new ways of teaching, learning, and leading" and "... how the innovative practices will continue to inform and impact the ELT field for years to come" (p. 1). The three parts of English Language Teacher Education in Changing Times each focus on one key aspect: learning, teaching, and leading and management. There are fifteen chapters in total (five in each section) written by contributors from around the world. Before describing the research, each author begins with a scenario, and many chapters also include tables summarizing the main points, both of which make the volume more accessible to readers.

Part one is titled "Learning in English Language Teacher Education in Changing Times," and its five chapters delve into teacher education in both formal education and continuous professional development contexts. The book begins with "Learning to Lead in Language Education" by Andy Curtis which describes a Leadership and Management in Language Education (LaMiLE) course with global participants who were able to further explore leadership styles by observing how leaders in their respective countries responded to the COVID-19 pandemic. The author makes comparisons between leaders in government and leaders in English language teaching although the alignment between these two concepts appears tenuous. Of note, chapter four also focuses on a course for teachers during the pandemic period. "Training Teachers in an Interdisciplinary Approach through EMI: A Case Study in Greece" by Chrysoula Lazou, Nikolaos Panagiotou, and Avgoustos Tsinakos, outlines the implementation of an online course pairing pre-service and in-service teachers to develop engaging materials for their Gen Z students. Finally, Georgios Kormpas and Christine Coombe's chapter "English Language Teacher Education and Development through Language Teacher Associations: Opportunities and Challenges" rounded out part one by reporting on how LTAs coped with the abrupt shift from face-to-face to online events, homing in on their challenges, opportunities, and future plans.

The middle section "Teaching in English Language Teacher Education in Changing Times" covers timely topics such as teacher wellbeing and teaching diverse learners. "Learning to Surf the Pandemic Wave: Interventions for Wellbeing and Inner Peace in an EFL Practicum Course" by María Matilde Olivero and María Celina Barbeito explains how teacher wellbeing activities were integrated into a virtual course for preservice teachers (PST). They utilized the PERMA model which consists of five key areas: positive emotions, engagement, relationships, meaning, and accomplishment (Seligman, 2011). Chapter 8, Leveraging Virtual Professional Development to Promote

Computer Science Education for Multilingual Students by Donna Eatinger et. al, reports on a study of participants' experiences in a professional development course for K-12 teachers in the United States preparing to teach computer science skills to multilingual students. And Chapter 10, Transforming Pre-Service Educators' Preconceived Ideas of Teaching General Education Content through Task-Based Hybrid Instruction by Kate Mastruserio Reynolds emphasizes a shift during the pandemic period to intentionally create supportive classroom communities and position herself as a mentor and coach, rather than primarily as a content expert. The chapter includes comments and reflections from PSTs enrolled in the course.

Including chapters from diverse contexts, part three focuses on the theme of "Leading and Management in English Language Teacher Education in Changing Times". Its chapters address the complexity of leading programs and carrying out partnerships during the pandemic and reflect on the growth of leaders and teachers resulting from these challenges. In chapter 11, Joan Kang Shin, Rebecca Kanak Fox, and Dildora Khakimova recount the reimagination of a training of trainer program in Uzbekistan in "Reaching Program Outcomes during Pandemic: English Language Teacher Professional Development in Uzbekistan". After their training of teacher trainers program was shifted to an online format, they found that participants benefited from communicating through online tools and developing e-portfolios. From Turkey, Bahar Gün writes of "Unravelling the Quality Conundrum: Teacher Education Program Administration in the New Normal", in chapter 12, and two additional chapters based on the process of TESOL Teacher Preparation programs shifting online based in the United States are included. In the final chapter, with members from around the globe, volunteer leaders of a large online network, Julie Lake and Liz England, share their perspectives in "Worldwide TESOL Career Path Development: We Lift Each Other Up When We Fall" which discusses how teachers supported each other through virtual activities, such as webinars, mentoring programs, and social events.

The uniform organization of the book's chapters is one of its strongest points, each one orienting the reader to the specific educational context before explaining the shift that occurred as a result of the COVID-19 pandemic. As can be seen from the chapter titles, the book's focus on teacher education did not limit its scope to formal learning (i.e., post-graduate programs, preservice teacher training), but also included accounts of teacher professional development through language teaching associations and online communities of practice. Furthermore, the chapters provide diverse perspectives on the subject although there are few specific examples from the African and

East Asian contexts, with most of the studies coming from Europe, North America, and South America. A global health emergency, the pandemic disrupted classroom environments around the world, and educators should not have difficulty relating or adapting information to their own teaching and learning contexts. Overall, the book makes an important and timely contribution to the field of English language teaching, describing teacher education during a tumultuous period.

Anastasia Khawaja (**AK):** One of the points that struck me was that many of the examples in this book circled back to the whole student. With the need of shifting online due to the pandemic, there was also a shift in a focus on student wellbeing that really has not been seen as vividly in practice until there was a literal health crisis in the world. This was exhibited explicitly in Olivero and Barebito as well as Reynolds chapters where wellbeing exercises were built into the curriculum, but also implied in most of the other chapters. This in turn affected how we as educators ran our assessments, and in turn made us reflect a lot deeper in how we teach and why we teach. Currently, we are faced with a discussion of what practices to keep, what to adapt, and what to discard now that we are more or less back to a "new normal."

Mary Hillis (MH): Conducting teacher training during the pandemic affected many areas, which all relate to the students, as you mentioned. For example, when creating materials for online environments, teachers found that replicating the face-to-face classroom experience was not actually their primary goal but rather finding ways to adapt and improve upon their materials. Concerning the program in Uzbekistan, Shin, Fox, and Khakimova wrote, "… the new program components involved transforming the teaching and learning into new spaces with a fuller array of options for participant engagement" (p. 169). Another point of interest was that teachers became more familiar with online tools through participation in online communities of practice.

AK: Community building was such a critical area, and arguably continues to be so. During the pandemic when everyone rapidly shifted to online, it opened the door for anyone to join gatherings, classes, meetings, conferences, and the like. Coombe and Kormpas highlighted the expanse of connectivity within English Language Teaching Associations (ELTAs), "ELTAs gained access to teachers (potential members) that would never have had the opportunity to attend a face-to-face conference. Underrepresented populations were able to attend, but also to present to world-renowned TESOL conferences including TESOL, IATEFL, and others" (p. 73). I really took note

where they mentioned potential members of the organization, as these are individuals who would probably not otherwise have had access to the conference, and by extension the organization itself. Even as we have returned to more face-to-face offerings, the global reach that online events have created cannot be ignored, and many organizations offer various online meetings to keep these communities going. Lake and England describe how they met language teachers' needs globally, building their professional communities through a variety of synchronous and asynchronous initiatives such as the Career Path Development's many free webinars, social events, and online resources that were all accessible for any TESOL educator around the world regardless of official organization affiliation. There are so many communities that would not have been created unless the online opportunities were available not just to organization members, but to non-members as well. Accessibility was and still is everything in global community building.

MH: Yes, community building was paramount during the pandemic, and we witnessed the professional development offerings at workplaces or through language teaching associations, with new online communities of practice springing up as needs surfaced. Teacher education is a continuous process, and the editors' focus on teachers and leaders' positive experiences during the pandemic was noteworthy. This was exemplified in chapter 11:

> … the shift to emergency remote teaching actually created an opportunity to explore virtual spaces and expand on the original pre-pandemic plan to deliver a more innovative and sustainable approach to teacher professional learning and collect and more robust set of data to inform program effectiveness and teacher growth" (p. 161).

These lessons will be useful in the post-pandemic too; the future is uncertain, and classes may still shift online due to a variety of factors, such as disease, natural disaster, or conflict.

AK: The one constant that book also has is that educators keep going. In every chapter, there was a scenario, and there was a clear issue with the pandemic. However, the subheading, "Shifting from the traditional to new ideas" that just about every chapter has conveys a sense of resilience. We saw stories from multiple parts of the world where educators discussed how they were given one to two weeks to essentially make what mainly existed as in-person only options to be completely accessible online. We were expected to make the transition seamless for our students and our faculty. Educating young students, university students, teacher educators, and teacher trainers

did not stop during the pandemic. It, as the majority of these chapters can attest to, flourished despite the obstacles. We adapt; we find ways to make education happen. I believe we are also still seeing the results of that education as through that medium, and we started noticing the digital divide. People started recognizing inequities that many may not have been aware of before, which has in turn raised an awareness in oneself. We have learned to turn our attention further towards greater societal issues. One only need look at the massive increase in engagement shown through protests of various causes around the world to see that this awareness has been augmented globally, and no one is easily coming back from that.

MH: Education did not stop, it expanded in contexts all around the world. The authors of the chapters in this book continued advancing education: they had a scenario, a context, and a shift which kept their classes and programs moving forward. As educators continue to face challenges in our ever-changing world, readers will find accounts of previous successes in England, Kamhi-Stein, and Kormpas' edited volume *English Language Teacher Education in Changing Times: Perspectives, Strategies, and New Ways of Teaching and Learning.*

References

Seligman, M. E. P. (2011). *Flourish: A visionary new understanding of happiness and well-being*. Atria.

Globalisation and its Effects on Team-Teaching. Naoki Fujimoto-Adamson. Cambridge Scholars Publishing, 2020. ix + 274 pp. ¥13,640.

Reviewed by
Mariana Oana Senda
Meiji University
Karmen Siew
Kaichi Nihonbashi Gakuen

Globalization and Its Effects on Team-Teaching is a seminal work by Naoki Fujimoto-Adamson. An esteemed associate professor at the Niigata University of International and Information Studies (NUIS), Fujimoto-Adamson

brings her extensive expertise in English Language Teaching (ELT), the historical nuances of ELT in Japan, team-teaching dynamics, and Content and Language Integrated Learning (CLIL) to this book. The central focus of the book revolves around unraveling the intricate web of connections between global issues, national education policies, and local practices related to team-teaching.

Not only does she explore CLIL and other partnership teaching schemes, but also extends her examination to other government-initiated team-teaching programs in East Asia. She discusses the Native-speaking English Teachers (NET) Scheme in Hong Kong, the Foreign English Teachers in Taiwan (FETIT) program, and the English Programs in Korea (EPIK) in South Korea, treating them as parallel case studies. Fujimoto-Adamson's book, tailored for educators, institutional leaders, educational policymakers, and other stakeholders, serves as an extensively researched record of the history of team-teaching in Japan from both educators' and students' perspectives, and provides insightful reflections on optimizing team-teaching strategies.

In a world continually reshaped by globalization, every facet of human existence, notably education, undergoes transformative changes. As educational systems evolve in response to global influences, novel policies, and innovative teaching practices, the pedagogical methodologies, particularly in Japanese classrooms, must keep pace. As the Japan Exchange and Teaching (JET) program served as the introduction to team-teaching in Japan, Fujimoto-Adamson prefaces her research by delving into the roots and goals of the program (Fujimoto-Adamson, 2020). This is followed by an investigation of team-teaching practices and analysis of pedagogic interactions at three Japanese schools. Her book takes a broader perspective by considering other team-teaching models worldwide, providing a comprehensive examination of partnership and team-teaching within the Japanese high school environment. Despite presenting her findings in book form, Fujimoto-Adamson's work follows a structured research framework, comprising essential elements such as an introduction, literature review, research methodology, findings and discussion, and conclusion. The core of "Globalization and Its Effects on Team-Teaching" is structured around five meticulously researched chapters, each shedding light on varying facets of team-teaching in Japan.

Chapter One contains an introduction, and it serves as a foundational backdrop, delineating the research objectives, scope, and key inquiries. It also furnishes readers with the historical examination of team teaching. Fujimoto-Adamson postulates that a comprehensive exploration of glo-

balization's imprint on language classrooms necessitates the inclusion of methodologies like linguistic ethnography. Such approaches are pivotal in discerning the intricate connections between classroom discourse, pedagogical interactions, and their broader social matrices.

Chapter Two is a literature review of team-teaching that is presented focusing on dissecting the extant literature around collaborative and team-teaching in contexts beyond ELT. It meticulously delves into the dynamics between native and non-native linguistics, underscored by discussions around native speaker ideologies and the burgeoning perspective of English as a lingua franca. The narrative further extends to spotlight team-teaching experiences from diverse locales, encompassing South Korea, Hong Kong, Taiwan, and both JET and non-JET programs within Japan. A critical appraisal of Japan's policy documentation is anchored around the roles of Assistant Language Teachers (ALTs), whether as pedagogical assistants or trainers, and the symmetrical partnership sought with Japanese Teachers of English (JTEs) (Fujimoto-Adamson, 2020).

Chapter Three is an explanation of research methodology, particularly highlighting the triangulation method which fuses classroom observations with semi-structured interviews across three schools located in Nagano Prefecture. Fujimoto-Adamson harnesses the potential of sociolinguistics, or linguistic ethnography, as her chosen lens. This approach demonstrates her understanding of the interplay between language and its enveloping social milieu, bringing into perspective the overarching global and political dynamics influencing localized team-teaching practices.

The final chapter, Chapter Four, contains research findings and a discussion that delves into the empirical outcomes derived from investigations at three junior high schools situated in cities in Nagano prefecture. Observational data spotlighted the nuanced interpretations of national policy directives vis-à-vis team-teaching methodologies. Furthermore, interviews shed light on a striking revelation: the absence of foreign educators recruited via the JET Program. Instead, private linguistic institutions and local educational boards were found to enlist educators with a diverse array of qualifications, professional experiences, and nationalities. This landscape rendered certain ALTs as domain experts, while relegating others to assistant roles.

Chapter five contains concluding insights, and it weaves together the pivotal insights unraveled throughout the research, casting a spotlight on the nexus between globalization and team-teaching, the contours of Japanese national educational mandates, and tangible team-teaching paradigms observed in the studied institutions. Moreover, the chapter furnishes action-

able recommendations aimed at fostering a synergistic classroom dynamic between ALTs and JTEs. Fujimoto-Adamson concludes with an imperative: the necessity for deeper dives into the evolving landscape of team-teaching within the Japanese context

The book is composed in a scholarly manner, offering a thorough examination of how global economic and political factors impact team-teaching. Despite the complexity of the topic, the author provides a clear and accessible explanation of the research process, beginning with the discussion of the topic's significance and scope, followed by the analysis of existing sources, primary data collection, evaluation of findings, and concluding remarks.

The insights presented in the research are valuable for educators, government officials, and educational institutions, even though the target audience is not explicitly specified. This aligns with previous research by McConnell (2002) and Reed (2015), which emphasized the significance of examining Japanese education within the context of political internationalization dynamics. Japan's increasing integration into the global arena is manifesting itself within the education sector. Therefore, studying Japanese education within a global context is imperative.

The relevance of team-teaching in English classrooms extends beyond Japan, finding applicability in schools worldwide. Moreover, similar collaborative team-teaching models are employed in British, Australian, and European schools with a focus on Content and Language Integrated Learning (CLIL) (Fujimoto-Adamson, 2020). Fujimoto-Adamson's book examines the unique Japanese context while considering the various factors influencing English education in Japanese secondary schools within the broader framework of team-teaching. Consequently, the book's audience may hail from any country or educational institution. While the author offers an historical perspective on team-teaching and provides valuable recommendations for enhancing future practices, the information presented accurately reflected the state of team-teaching practices in Japan up until 2020.

Due to pandemic restrictions that lasted from 2020 to 2022, a number of ALTs (both JET and non-JET) returned to their home countries and contracting organizations such as Boards of Education (BOEs) and individual schools struggled to find replacements due to tight restrictions on foreigners entering Japan, even for school and work. Following the initial surge in hires of ALTs between 2015-2020 to prepare for the Tokyo Olympics, there have been post-pandemic government initiatives to improve the level of English spoken in Japan. She noted that the number of JETs in Tokyo dropped to four in 2002, but as of 2023, Tokyo private schools currently hire 191 JET-ALTs,

and the Tokyo Metropolitan Board of Education (TBOE) currently hires 289 JET-ALTs and 1 Coordinator for International Relations (CIR); TBOE intends to increase the number in 2024, having 2 JET-ALTs assigned to each school. While Fujimoto-Adamson's observation that "the number of JET participants is decreasing" may not be accurate, her observation that overreliance on government schemes is insufficient, and that individual schools and local contexts should consider what they need is important, is extremely poignant.

The book can be viewed as a critical review of the existing literature on team-teaching practices in Japanese schools. Furthermore, the author acknowledges that the team-teaching approach is implemented in various countries, suggesting that Japan's experiences can serve as a valuable reference. This reflects the author's comprehensive understanding of not only Japan's education system but also global education dynamics. Her top-down analysis of diverse team-teaching practice perspectives, beginning with the consideration of globalization and international politics between the US and Japan and descending to Japanese domestic politics, the JET Program, personal experiences of JET teachers and local schools, as well as underscores the depth of her analysis.

The book effectively fulfills its stated objectives by providing comprehensive answers to research questions supported by a combination of primary and secondary sources. Using the triangulation method enhances the volume of available data, enabling a multifaceted understanding of various aspects of team-teaching practices and affirming research findings. The application of linguistic ethnography contributes to the deconstruction of social beliefs and provides insights into the benefits of different approaches to English teaching (Marine & Čermáková, 2021). The use of ethnography in English teaching helps explore cultural phenomena, examining the behavior of social situation participants, and understanding the interpretation of this behavior by group members. It contributes to the endeavor to commonly use this approach in teaching practices.

The book provides extended possibilities to the target audience due to the detailed discussion of parties engaged in the education system and related policy making. Therefore, it can be claimed that the book has few weaknesses as it provides readers with many advantages. The book is presented in the form of a research paper, providing a clear understanding of what is going to be achieved. Moreover, it offers a critical analysis of the current literature sources on team-teaching practices with the consideration of different perspectives. Finally, the use of the method of triangulation helped to

obtain comprehensive information, which provides a deep insight into the nature and core aspects of team-teaching.

Fujimoto-Adamson's significant contribution lies in establishing a connection between globalization and team-teaching. The recommendations provided by the author may help institutions adopt more effective educational practices that focus on challenges relevant to teaching English as a foreign language. This book can be used as a guide for school managers, policymakers, teachers, and other stakeholders as the information received from school observation represents the real state of teaching in most Japanese schools. Finally, this volume enables readers to learn about the complex power dynamics of different ministries and enhance an understanding of their role in the system of education in Japan.

Fujimoto-Adamson's book makes a significant contribution to the field of education by establishing a link between globalization and team-teaching. Her recommendations offer insights into adopting more effective educational practices, particularly in the realm of teaching English as a foreign language. Her assertion that Japanese education requires change resonates deeply, as local teaching practices have been hindered by untrained JTEs, exam-oriented classes, and teachers merely serving as "human tape recorders" (Reed, n.d.). She emphasizes the need for adapting the local educational system to align with the growing influence of political and international dynamics, which have become increasingly pronounced in the age of globalization. Additionally, this work enriches our understanding of the intricate power dynamics among various ministries and their roles in Japan's educational system, addressing the evolving challenges of contemporary education.

Fujimoto-Adamson's book offers a comprehensive exploration of team-teaching practices between ALTs and JTEs. Through three selected observations at local schools, the author explains her belief that JTEs should take a leading role in conducting lessons, while ALTs should be relegated to secondary teaching roles. Additionally, she underscores the importance of considering each individual context since every situation is different – the ALT's and JTE's backgrounds, and the individual students' strengths, weaknesses, and needs.

The author not only identifies urgent issues in English language teaching in Japan but also offers recommendations for addressing these challenges, not only in Japan but also globally. Fujimoto-Adamson's extensive teaching experience and insider perspective attest to her expertise in the field. Furthermore, her recommendations span different levels and promise benefits

for both educators and learners. The author's suggestion to enhance dialogue between teachers and government officials for policy changes via a bottom-up approach demonstrates her awareness of common communication issues within Japanese schools. As the information derived from school observations reflects the actual state of teaching in some Japanese schools, we reiterate that this book can serve as a useful resource for school administrators, policymakers, teachers, and other stakeholders in understanding and implementing effective educational strategies.

References

Fujimoto-Adamson, N. (2023). From JTE to team-teaching researcher: Autoethnographic reflections. In T. Hiratsuka (Ed.), *Team teachers in Japan: Beliefs, identities, and emotions* (pp. 32–43). Routledge. https://doi.org/10.4324/9781003288961-4

Kano, A., Sonoda, A., Schultz, D., Usukura, A., Suga, K., & Yasu, Y. (2016). Barriers to effective team teaching with ALTs. In P. Clements, A. Krause, & H. Brown (Eds.), *Focus on the learner* (pp. 74–82). JALT.

Marine, F., & Čermáková, A. (2021). Using linguistic ethnography as a tool to analyse dialogic teaching in upper primary classrooms. *Learning, Culture and Social Interaction, 29*, Article 100500. https://doi.org/10.1016/j.lcsi.2021.100500

McConnell, D. (2001). Importing diversity: Inside Japan's JET program. *Contemporary Sociology: A Journal of Reviews, 30*(6), 596–598. http://doi.org/10.2307/3089007

Reed, N. (2015). *Contemporary roles of foreign English teachers in Japanese public secondary schools: An exploratory study* [Master's thesis, University of Birmingham]. Asian EFL Journal. https://asian-efl-journal.com/wp-content/uploads/Nathaniel-David-Reed.pdf

JALT Journal Aims and Scope

JALT Journal is a bi-annual, Scopus-approved research journal of the Japan Association for Language Teaching (全国語学教育学会). JALT's larger mission is to support the research programs and professional development of JALT members, promote excellence in language learning, teaching, and research, and provide opportunities for those involved in language education. In line with this mission, *JALT Journal* publishes high-quality English- and Japanese-language, quantitative and qualitative, theoretically-informed and empirically-grounded studies of relevance to second/foreign language education in Japan. Although emphasis is placed on the Japanese context, *JALT Journal* values contributions which also transcend geographical boundaries to illuminate the complex interaction between language, language use, people, education, and society across cultural and socio-political contexts.

When possible, submissions to *JALT Journal* should aim to be both descriptive (*What is my data?*) and explanatory (*Why is my data like this and not otherwise?*) in purpose, and further stimulate scholarly debate, to hopefully improve existing applied linguistic scholarship around the world. Areas of interest include but are not limited to the following:

- Bilingualism and multilingualism
- Classroom-based language education
- Cognitive linguistics
- Contrastive linguistics
- Conversation/discourse/critical discourse analysis
- Critical language pedagogy
- Curriculum design and teaching methods
- Intercultural communicative competence
- Language acquisition/learning
- Language policy and planning
- Language testing/evaluation
- Phonetics
- Pragmatics
- Psycholinguistics
- Semantics
- Sociolinguistics
- Syntax
- Teacher training
- Translation and interpretation
- Vocabulary

Guidelines

Style

Authors are encouraged to submit manuscripts in five categories: (1) full-length articles, (2) short research reports or papers addressing specific theoretical and/or methodological issues in applied linguistics research (*Research Forums*), (3) theory-grounded essays which may include analysis of primary or secondary data (*Perspectives*), (4) comments on previously published *JALT Journal* articles (*Point-to-Point*), and (5) book/media reviews (*Reviews*) either requested by the Book Reviews editor or suggested by the author. Articles should be written for a general audience of language educators; therefore, statistical techniques and specialized terminologies must be clearly explained and their use clearly justified.

Authors are responsible for obtaining permissions for any copyrighted material included in their manuscript submission. When submitting a manuscript based on a thesis or dissertation published in an institutional repository, authors are responsible for checking with the copyright holder for permission to publish portions of the original text.

JALT Journal follows the *Publication Manual of the American Psychological Association, 7th edition* (available from <https://apastyle.apa.org/products/publication-manual-7th-edition>). A downloadable copy of the *JALT Journal* style sheet is also available on our website at <https://jalt-publications.org/jj/>. Instructions for online submissions can also be found at this website.

Format

Full-length articles must not be more than 8,000 words, including references, notes, tables, and figures. *Research Forum* submissions should not be more than 4,000 words. *Perspectives* submissions should not be more than 5,000 words. *Point-to-Point* comments on previously published articles should not be more than 1,000 words in length, and *Reviews* should generally be around 2,000 words. All submissions must be word processed in A4 or 8.5" x 11" format with line spacing set at 1.5 lines. **For refereed submissions, names and identifying references should appear only on the cover sheet**. Authors are responsible for the accuracy of references and reference citations and for obtaining any permissions for copyrighted material contained in the manuscript.

Submission Procedure

Please submit the following two documents in MS Word format to the appropriate editor indicated below:

1. Cover sheet with the title and author name(s), affiliation(s), and contact information of the corresponding author.
2. A blinded manuscript, including title, abstract, and keywords, with no citations referring to the author. Do not use running heads and do not include your biographical information. Follow the *JALT Journal* style sheet. Blinded manuscripts should not include meta-data.

If the manuscript is accepted for publication, a Japanese translation of the abstract will be required. Authors will also be asked to provide biographical information. Insert all tables and figures in the manuscript. Do not send them as separate files.

Submissions will be acknowledged within 1 month of their receipt. All manuscripts are first reviewed by the Editor to ensure they comply with *JALT Journal* Guidelines. Those considered for publication are subject to blind review by at least two readers, with special attention given to (1) compliance with *JALT Journal* Editorial Policy, (2) the significance and originality of the submission, and (3) the use of appropriate research design and methodology. The first round of review is usually completed within 3 months. Each contributing author of published articles and *Book Reviews* will receive one complimentary copy of the *Journal* and a PDF of the article (*Book Reviews* are compiled together as one PDF). *JALT Journal* does not provide off-prints. Contributing authors have the option of ordering further copies of *JALT Journal* (contact JALT Central Office for price details).

Restrictions

Papers submitted to *JALT Journal* must not have been previously published, nor should they be under consideration for publication elsewhere. *JALT Journal* has First World Publication Rights, as defined by International Copyright Conventions, for all manuscripts published. If accepted, the editors reserve the right to edit all copy for length, style, and clarity without prior notification to authors. Plagiarism, including self-plagiarism, will result in articles not being published or being retracted and may also result in the author(s) being banned from submitting to any JALT publication. For further information, see the JALT Publications Statement of Ethics and Malpractice at: https://jalt-publications.org/jalt-publications-statement-ethics-and-malpractice

Full-Length Articles, Research Forum, Perspectives, and Point-to-Point Submissions

Please upload submissions in these categories to the following website:

https://jalt-publications.org/content/index.php/jj

Manuscripts should follow the **American Psychological Association (APA) 7th Edition style**. Please indicate if your submission is a 1) Full-length article; 2) *Perspectives* article; 3) *Research Forum* article or 4) *Point-to-Point* submission.

For any general inquiries about English-language submissions, please contact:

Dennis Koyama, *JALT Journal* Editor
jaltpubs.jj.ed@jalt.org

Japanese-Language Manuscripts

JALT Journal welcomes Japanese-language manuscripts on second/foreign language teaching and learning as well as Japanese-language reviews of publications. Submissions must conform to the Editorial Policy and Guidelines given above. Authors must provide a detailed abstract in English, 500 to 750 words in length, for full-length manuscripts and a 100-word abstract for reviews. Refer to the Japanese-Language Guidelines (following page) for details. Please send Japanese-language manuscripts to:

Masayuki Kudo, *JALT Journal* Japanese-Language Editor
jaltpubs.jj.ed.j@jalt.org

Reviews

The editors invite reviews of books and other relevant publications in the field of language education. A list of publications that have been sent to JALT for review is published bimonthly in *The Language Teacher* and can be found online in each issue at <https://jalt-publications.org/tlt/>. Review authors receive one copy of the *Journal*. *JALT Journal*'s latest *Reviews* guidelines can be found in Melodie Cook's *Expositions* piece published in the November 2023 issue. Please send submissions, queries, or requests for books, materials, and review guidelines to:

Melodie Cook, *JALT Journal* Reviews Editor
jaltpubs.jj.reviews@jalt.org

Special Issues

Special issues often make an important contribution to the development of academic discourse in a specific field, because they allow researchers and practitioners to (a) identify an issue or topic of particular relevance to the context in which the journal is read, (b) summarize the key concepts and debates shaping that issue, (c) bring further sophistication to existing academic discourse and identify new research possibilities, and (d) identify key readings for the journal readership. Special issues can also attract new authors and readers to an academic journal, and can be an effective means of finding new editors for that journal.

We strongly encourage *JALT Journal* readers to submit special-issues proposals. When submitting such proposals, please make sure that they adhere to the aims and scope of *JALT Journal*. Proposals should include: (1) a title which clearly captures the special issue topic, (2) a brief description of the special issue, (3) an account of the motivation behind the special issue and its importance to the field at large, (4) a list of no more than three guest editors with short biographical information, including editorial work experience, and (5) a list of article contributors, with a short description of each article contribution. Special-Issues editors are also responsible for reviewing and editing contributions. However, the *JALT Journal* editorial team reserves the right to accept or reject individual contributions.

Inquiries about Subscriptions, Ordering *JALT Journal*, or Advertising

JALT Central Office
Marunouchi Trust Tower Main Building 20F
1-8-3 Marunouchi, Chiyoda-ku, Tokyo 100-0005 JAPAN
Tel.: (+81) 3-5288-5443
Email: jco@jalt.org URL: https://jalt.org

日本語論文投稿要領

*JALT Journal*では日本語で執筆された（a）論文、（b）研究報告、（c）展望論文、（d）*JALT Journal*に掲載された著作物へのコメント・考察、（e）書評を募集しています。（a）論文と（b）研究報告の違いは、以下の通り字数制限による違いです。（c）展望論文は、言語教育研究に関する課題に焦点をあてた短い論文で、先行研究の検証、理論や1次2次データに基づく議論などを含むものです。文体:一般的な学術論文のスタイルを用い、章立ての仕方や参考文献のデータの書き方などは、*Publication Manual of the American Psychological Association*（7th edition）の定める方式に合わせて下さい。*JALT Journal*書式シート（日本語原稿用）を以下からダウンロードできます<https://jalt-publications.org/jj/>。なお、*JALT Journal*の読者は現場の教師が主なので、特殊な専門用語や統計的手法は、わかりやすく定義するか説明を加えるなどして下さい。原稿:長さは、参考文献リストも含め、（a）論文は25,000字、（b）研究報告は13,000字、（c）望論文は16,000字、（d）*JALT Journal*に掲載された著作物へのコメント・考察は2,000字、（e）書評は1,500~3,000字以内です。A4の用紙に横書きで、1行40字、1ページ30行で印刷して下さい。手書きの原稿は受け付けません。

提出するもの:
*JALT Journal*書式シート（日本語原稿用）を参考に作成の上、電子メールの添付書類でお送りください。　なお、上記（a）論文~（e）書評のどのカテゴリーへの投稿かを明記ください。審査を経て掲載の認められた草稿は、図表などを全て写植版にしたものにして提出願います。

査読: 編集委員会で投稿要領に合っているかどうかを確認したあと、少なくとも二人の査読者が査読を行います。査読者には執筆者の名前は知らされません。査読の過程では特に、原稿が*JALT Journal*の目的に合っているか、言語教育にとって意味があるか、独創性はあるか、研究計画や方法論は適切か等が判定されます。査読は通常二か月以内に終了しますが、特に投稿の多い場合などは審査にそれ以上の時間がかかることがあります。

注意: *JALT Journal*に投稿する原稿は、すでに出版されているものや他の学術雑誌に投稿中のものは避けて下さい。*JALT Journal*は、そこに掲載されるすべての論文に関して国際著作権協定による世界初出版権を持ちます。なお、お送りいただいた原稿は返却しませんので、控を保存して下さい。

投稿原稿送り先またはお問い合わせ:

〒001-0016 北海道札幌市北区北16条西2丁目 藤女子大学
JALT Journal 日本語編集者　工藤 雅之
電話: 011-736-5368
jaltpubs.jj.ed.j@jalt.org

JALT Journal 第 46 巻　第 1号

2024年 4月20日　印刷
2024年 5月1日　発行
編集人　小山デニス
発行人　クレア・カーネーコー
発行所　全国語学教育学会事務局
〒100-0005 東京都千代田区丸の内1-8-3 丸の内トラストタワー本館20階
TEL（03) 5288 5443
印刷所　コーシンシャ株式会社
〒530-0043　大阪市北区天満1-18-4天満ファーストビル301 TEL（06）6351-8795

JALT2024

全国語学教育学会
第50回年次国際大会教材展示会

2024年11月15日～2024年11月18日

静岡県・静岡コンベンションアーツセンター
グランシップ

50th Annual International Conference on
Language Teaching and Learning & Educational
Materials Exhibition

November 15 – 18, 2024

Shizuoka Granship, Shizuoka, Japan