

# The Case for Rubric-Scored Writing Items on Japanese University Entrance Examinations

Kenneth BROWN

Department of English Language Education and Research

(Received September 30, 2024 / Accepted December 23, 2024)

## Abstract

Given the extremely high-stakes nature of Japanese university entrance examinations (UEEs) on test takers, a correspondingly high level of reliability and validity is essential. The washback effect that these tests have also needs to be addressed. Unfortunately, numerous studies have pointed to problems in all three of these issues when it comes to the English sections of UEEs. In this paper, the case is made for the more widespread adoption of rubric-scored writing questions on UEEs as one way to address these issues. First of all, including the productive and communicative skill of writing would increase validity. Also, the use of scoring rubrics would increase reliability. Finally, including writing would have the potential to increase positive washback by encouraging a more communicative approach to English instruction, as hoped for by MEXT, as well as other stakeholders. Barriers to this change, such as time constraints and institutional resistance, are acknowledged to exist, but it is posited that these barriers are not insurmountable.

*Key words:* entrance examinations, scoring rubrics, validity, reliability, washback

In Japan, no tests have a bigger impact on the lives of more people than the ones that high school graduates take to enter university. Considering the very high-stakes nature of these tests, a corresponding very high level of reliability accompanied by abundant evidence of validity would be called for (Bachman & Palmer, 1996). The issue of positive washback is also a major concern because these tests are *superordinate tests* that, due to their associated prestige and their huge effect on stakeholders, dominate language instruction (Brown, 1999). Unfortunately, concerns have been raised over all three issues. One way to address these issues would be to include productive writing items on university entrance examinations (UEEs) and to score those items using scoring rubrics.

## Background Information

Generally speaking, two major tests are the steppingstones to becoming a university student in Japan. The first of these is the Common Examination for University Admission, or simply the Common Test (Allen, 2020), which is administered by the National Center for University Entrance Examinations under the auspices of the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) (Dokuritsu Gyōsei Hōjin Daigaku Nyushi Center [National Center for University Entrance Examinations], n.d.). This test replaced the Center Test in 2021 (“Unified Univ. Entrance Exams Begin Across Japan as No. of Test-Takers Dips Below 500,000”, 2024). The test consists of the six subject areas of Japanese language, geography and history, civics, mathematics, science, and, pertinent here, foreign languages, that

is, English. Most national and public universities, as well as many private universities, require applicants to take this test.

The second of these are the UEEs, which are designed and administered by the universities themselves. Most national and public universities require that applicants take these exams after they have taken the Common Test. Private universities will often do the same or require these tests in lieu of the Common Test (Allen, 2020). English is commonly included in these tests (Guest, 2008).

### Issues with Reliability and Validity

The UEEs have long been criticized for questionable *reliability*, i.e., “the extent to which the results can be considered consistent or stable” (Brown, 2005, p. 175) and *validity*, i.e., “the degree to which a test measures what it claims, or purports, to be measuring” (Brown, 2005, p. 295). Nearly 30 years ago, Brown and Yamashita (1995a; 1995b) examined the English sections of UEEs from both public and private universities as well as the Center Test, predecessor of the Common Test, and identified serious shortcomings in terms of both reliability and validity due to the following issues: (1) Reading passages are typically at an inappropriately high level, (2) item types within the tests are too varied, (3) productive items are severely lacking, (4) many items test the overly specialized skill of translation, (5) many of the exams are too short, and (6) statistical analyses of either reliability or validity on the part of the test developers are severely lacking. Ten years later, Kikuchi (2006) replicated Brown and Yamashita’s (1995b) study and, while noting the addition of summaries of reading and/or listening passages on some of the tests as well as the addition of listening components to the Center Test, found no truly significant changes.

As for writing items, both Moore (2015) and Kowata (2016) found that only around 20% of UEEs included writing compositions of a productive nature, with Moore also reporting that 33% of them did not include any writing tasks at all. Watanabe (2016) reports that of the writing items that are included on UEEs, 43% of them are expositions and 41% of them are personal reflections. Sequential explanations and descriptive texts, on the other hand, were, respectively speaking, practically and completely nonexistent. Watanabe also points out that personal reflections would be analogous to what Rothery (1994, as cited in Watanabe, 2016) refers to as personal responses and describes as being the least beneficial toward the development of the sort of more sophisticated writing that demands explaining and/or interpreting. This obviously calls into question these tests’ effectiveness in validly assessing writing skills.

### Issues with Washback

In addition to serving the purpose of assessment, high-stakes language tests, such as entrance examinations, have been valued as a way to positively affect the way in which foreign language is taught and learned in the classroom (Shohamy, 1992). This is referred to as *measurement-driven instruction* or, more commonly in the field of second language acquisition, positive (as opposed to negative) *washback*.

MEXT has made its hope for positive washback that is supportive of the national course of study quite clear with its proposed implementation of four-skills testing (Paxton et al., 2022). More specifically, MEXT has stressed the need for UEEs to make changes (Kowata, 2016) so that they assess the ability to communicate in all four skills (Saito, 2019). It seems MEXT is not alone in hoping that a redesign of UEEs will lead to this sort of positive washback. High school students and teachers also seem to believe it (Green, 2014), as do business leaders (Keizai Doyukai [Japan Association of Corporate Executives], 2013).

MEXT took concrete action in this direction when, in 2016, it announced its intention to encourage universities to use one of a number of approved privately produced four-skills tests (e.g., EIKEN, GTEC, IELTS, TEAP, TOEFL) instead of the English section of the Common Test (MEXT, 2016, as cited in Allen, 2020). This plan, which was well-intentioned but ill-conceived, was shelved in 2019 (Allen, 2020) after pushback over concerns regarding the difficulty in comparing the scores of tests with so many

different purposes, proficiency levels, and formats, as well as the lower degree of access to the tests that rural residents or those of lower socio-economic means would have (Allen, 2020; Butler et al., 2022).

As to the degree to which UEEs induce washback in Japan, Paxton et al.'s (2022) systematic review ascertained that evidence of their causing washback on learners' behavior was tentative at best. They also report that UEEs that include items testing communicative ability do not guarantee positive washback on teachers' classroom instruction. They further detail how numerous studies have identified clear evidence of UEEs causing negative washback on what teachers do in the classroom.

The unpredictability of washback compels Andrews (2004) to caution against "the dangers of an oversimplistic, naive reliance on high-stakes tests as a primary change strategy" (p. 48). However, to say that positive washback is not automatic is not to say that it is not worth pursuing as one tool among many in the toolbox. As Popham (1987) says of measurement-driven instruction, i.e., washback, "Of course, any effective tool can always be misused. A scalpel that can save lives when used by a skilled neurosurgeon can become a murder weapon in the wrong hands. That possibility, however, should not incline us to outlaw scalpels" (p. 681). It seems worthwhile, then, to pursue the use of improved UEEs to engender positive washback even in light of, or even because of, the spotty record of UEEs ability to do this.

### **Rubric-Scored Writing Items as a Partial Solution**

Clearly, for UEEs to give a well-rounded and, therefore, valid assessment of the ability to communicate in English, they would need to assess not only the receptive skills of reading and listening but also the productive skills of writing and speaking. However, from a practical standpoint, the inclusion of a speaking component is particularly daunting for the reason that oral exams often require more time and administrators and/or electronic devices to implement (Ockey, 2017). Additionally, Saito (2019) has observed that the major privately produced four-skills tests only rarely include more communicative speaking skills such as turn-taking or initiating and ending conversations. This raises the specters of both negative washback (Saito, 2019) and a lack of validity. On the other hand, including writing items on UEEs is feasible, albeit not without its challenges.

From a validity standpoint, writing items that are productive and communicative should be included on the UEEs. Watanabe (2016) further suggests that, ideally, different UEEs would feature different writing genres to more effectively engender positive washback that leads to MEXT's stated goal of students being able to write English texts that are appropriate to various contexts. It is difficult to imagine the sort of coordination that would be necessary for all UEEs to provide a full and balanced variety of writing genres among themselves. However, there is no doubt that the situation would be greatly improved if the percentage of UEEs featuring productive writing increased from the roughly 20% that it is now to at least a majority and, further, if that productive writing eschewed the less demanding mini-genre of personal reflections.

Introducing productive writing items to UEEs is not enough, however. To address the issues of reliability and validity, it is highly advisable to employ a scoring guide—most commonly referred to as a *rubric*—when evaluating test takers' responses on the writing items. The use of rubrics can generally be expected to lead to higher reliability, even if this may not always be the case, as well as higher validity in the sense of increased positive washback (Jonsson & Svingby, 2007).

These rubrics would need to include the criteria to be used to evaluate the response, descriptions of the differences between said criteria, and which of the two types of approaches, holistic or analytic, is to be used (Popham, 1999). The difference between the two types is that *holistic rubrics* give a single score for each level of ability, with each level describing any number of criteria, whereas *analytic rubrics* differentiate not only between levels but also between various criteria that are to be evaluated.

Both holistic and analytic rubrics have their advantages and disadvantages. The biggest advantage of holistic rubrics is their perceived usability, particularly in the context of large-scale tests

(Jonsson & Svingby, 2007). Holistic scoring is generally much less time consuming because raters using holistic rubrics tend to read a response only once, as opposed to raters using analytic rubrics who tend to read the response several times so as to assess each criterion separately (Weigle, 2002). Analytic rubrics, on the other hand, usually have increased reliability due to the greater number of categories (Ghalib & Al-Hattami, 2015; Ohta et al., 2018; Weigle, 2002). Test administrators will naturally consider these tradeoffs when selecting which type of rubric to use.

Such a rubric could be developed from scratch, or a preexisting rubric could be adopted or adapted. One of the most enduring analytic rubrics is the one designed by Jacobs et al. (1981). This rubric, with its five weighted criteria of content, language use, organization, vocabulary, and mechanics, is often used in writing programs at the university level (Weigle, 2002). Closer to home, Kinshi et al. (2011) enhanced both reliability and validity of a rubric they had designed earlier (Kuru et al., 2009, cited in Kinshi et al., 2011). This rubric is intended to be used at the tertiary level in Japan and consists of the five criteria of content and idea development, organization, grammar, vocabulary, and mechanics.

Actual use of whichever rubric is decided upon would ideally be preceded by rating training for the purpose of improving reliability via higher *interrater agreement*, i.e., the degree to which raters agree with each other and *intrarater agreement*, i.e., the degree to which raters are consistent with themselves (Bachman & Palmer, 1996; Brown, 2005; Jonsson & Svingby, 2007; Weigle, 2002). Weigle (2002) describes a training process wherein the team of raters read through a set of representative writing responses that are illustrative of the points on the rubric. Once raters have familiarized themselves with the rubric in this way, they can move on to trying their hand at scoring responses at different, random levels. During this process, raters can work as a group to bring their level of agreement to an acceptable level.

The inclusion of productive writing items on the UEEs that are rubric scored might also increase the likelihood of positive washback on the educational process. As a case in point, Kowata (2016) found that 129 Japanese high school teachers self-reported that they rated students' writing based on the factors they assumed were typically included in writing rubrics used in UEEs; this despite the fact that the universities do not actually release these rubrics. An increase in writing items on UEEs accompanied by transparency regarding the scoring rubrics might logically lead to similar but more pronounced results.

All of this is not to say that there would not be barriers to making the aforementioned changes. The scoring of writing is clearly more time consuming than the scoring of, for example, a multiple-choice exam. This issue is exacerbated by the limited amount of time in which to administer and score the exams as well as the large number of test takers, although the latter is less of an issue in the case of national universities where applicants are fewer due to having been screened based on their performance on the Common Test (Moore, 2015). There can also be internal resistance to making such changes. Murphey (2004), for example, describes his ultimately fruitless attempts to convince his university's administration to apply data-driven item analyses to the improvement of their entrance examinations. Admittedly, these issues are formidable, but they cannot be said to be insurmountable.

## Conclusion

This paper has argued for a concerted effort to include productive writing tasks that are scored using rubrics on UEEs. The productive and communicative nature of these writing tasks would increase validity, and the use of scoring rubrics would increase reliability. Furthermore, the inclusion of these writing tasks should at least increase the odds of positive washback taking place. While implementing these changes are undeniably not without challenges, taking actions toward this goal would be well worth the effort.

## References

- Allen, D. (2020). Proposing change in university entrance examinations: A tale of two metaphors. *Shiken*, 23–38. <https://doi.org/10.37546/JALTSIG.TEVAL24.2-2>

- Andrews, S. (2004). Washback and curriculum innovation. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing* (pp. 59–72). Routledge. <https://doi.org/10.4324/9781410609731-11>
- Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.
- Brown, J. D. (1999). The roles and responsibilities of assessment in foreign language education. *JLTA Journal*, 2, 1–21. [https://doi.org/10.20622/jlta.2.0\\_1](https://doi.org/10.20622/jlta.2.0_1)
- Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. McGraw-Hill College.
- Brown, J. D., & Yamashita, S. O. (1995a). English language entrance examinations at Japanese universities: What do we know about them? *JALT Journal*, 17 (1), 7–30. <http://jalt-publications.org/files/pdf-article/jj-17.1-art1.pdf>
- Brown, J. D., & Yamashita, S. O. (1995b). English language entrance examinations at Japanese universities: 1993 and 1994. In J. D. Brown, & S. O. Yamashita (Eds.), *Language testing in Japan*, 86–100.
- Butler, Y. G., Lee, J., & Peng, X. (2022). Failed policy attempts for measuring English speaking abilities in college entrance exams: Cases from China, Japan, and South Korea. *English Today*, 38(4), 271–277. <https://doi.org/10.1017/s0266078420000346>
- Dokuritsu Gyōsei Hōjin Daigaku Nyushi Center [National Center for University Entrance Examinations]. (n.d.). *Kyōtsū tesuto no yakuwari [Role of the Common Test]*. [https://www.dnc.ac.jp/kyotsu/shiken\\_gaiyou/yakuwari.html](https://www.dnc.ac.jp/kyotsu/shiken_gaiyou/yakuwari.html)
- Ghalib, T. K., & Al-Hattami, A. A. (2015). Holistic versus analytic evaluation of EFL writing: A case study. *English Language Teaching*, 8(7), 225–236. <https://doi.org/10.5539/elt.v8n7p225>
- Green, A. (2014). *The Test of English for Academic Purposes (TEAP) impact study: Report 1—Preliminary questionnaires to Japanese high school students and teachers*. Eiken Foundation of Japan. [https://www.eiken.or.jp/teap/group/pdf/teap\\_washback\\_study.pdf](https://www.eiken.or.jp/teap/group/pdf/teap_washback_study.pdf)
- Guest, M. (2008). Japanese university entrance examinations: What teachers should know. *The Language Teacher*, 32(2), 15–19.
- Jacobs, H., Zingraf, A., Warmuth, D., Hartfiel, V. & Hughey, J. (1981). *Testing ESL composition: A practical approach*. Newbury House.
- Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: reliability, validity and educational consequences. *Educational Research Review*, 2, 130–144. <https://doi.org/10.1016/j.edurev.2007.05.002>
- Keizai Doyukai [Japan Association of Corporate Executives]. (2013). *Achieving university entrance exams that test practical English abilities*. <https://www.doyukai.or.jp/en/policyproposals/2013/pdf/130422a.pdf>
- Kikuchi, K. (2006). Revisiting English entrance examinations at Japanese universities after a decade. *JALT Journal*, 28(1), 77–96. <https://doi.org/10.37546/jaltjj28.1-5>
- Kinshi, K., Kuru, Y., Masaki, M., Yamanishi, H., & Otoshi, J. (2011). Revising a writing rubric for its improved use in the classroom. *LET Kansai Chapter Collected Papers*, 13, 113–124.
- Kowata, T. (2016). Do imagined writing rubrics used in university entrance examinations affect scoring in classroom? *ARCLE Review*, 10, 29–38.
- Moore, Y. (2015). An evaluation of English writing assessment in Japanese university entrance examinations. *Writing and Pedagogy*, 7(2–3), 233–260. <https://doi.org/10.1558/wap.v7i2-3.26227>
- Murphey, T. (2004). Participation, (dis-)identification, and Japanese university entrance exams. *TESOL Quarterly*, 38(4), 700–710. <https://doi.org/10.2307/3588286>
- Ockey, G. J. (2017). Approaches and challenges to assessing oral communication on Japanese entrance exams. *JLTA Journal*, 20, 3–14. [https://doi.org/10.20622/jltajournal.20.0\\_3](https://doi.org/10.20622/jltajournal.20.0_3)
- Ohta, R., Plakans, L. M., & Gebril, A. (2018). Integrated writing scores based on holistic and multi-trait scales: A generalizability analysis. *Assessing Writing*, 38, 21–36. <https://doi.org/10.1016/j.asw.2018.08.001>
- Paxton, S., Yamazaki, T., & Kunert, H. (2022). Japanese university English language entrance exams and



- the washback effect: A systematic review of the research. *Journal of Pan-Pacific Association of Applied Linguistics*, 26(2), 1–20. <https://doi.org/10.25256/PAAL.26.2.1>
- Popham, W. J. (1987) The merits of measurement-driven instruction. *Phi Delta Kappa*, 68, 679–82.
- Popham, W. J. (1999) *Classroom assessment: What teachers need to know*. Allyn and Bacon.
- Saito, Y. (2019). Impacts of introducing four-skill English tests into university entrance examinations. *The Language Teacher*, 43(2), 9–14. <https://doi.org/10.37546/jalttl43.2-2>
- Shohamy, E. (1992). Beyond proficiency testing: A diagnostic feedback testing model for assessing foreign language learning. *The Modern Language Journal*, 76(4), 513–521. <https://doi.org/10.2307/330053>
- Unified univ. entrance exams begin across Japan as no. of test-takers dips below 500,000. (2024, January 13). *Mainichi Japan*. <https://mainichi.jp/english/articles/20240113/p2a/00m/0na/007000c>
- Watanabe, H. (2016). Genre analysis of writing tasks in Japanese university entrance examinations. *Language Testing in Asia*, 6(1), 1–14. <https://doi.org/10.1186/s40468-016-0026-8>
- Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press. <https://doi.org/10.1017/cbo9780511732997>