Chapter 5
Fairness in College Entrance Exams in Japan and the Planned Use of External Tests in English

Yuko Goto Butler and Masakazu Iino

Abstract The Japanese government recently decided to replace the English section of the nationwide college entrance exam with external proficiency tests. This policy was motivated by the desire to improve the speaking proficiency of students by directly assessing it in college entrance examinations. However, in Japan, an English-as-a-foreign-language context, students' English proficiency, and speaking ability in particular, is greatly influenced by socioeconomic status (SES) because students need to seek greater opportunities to develop English-speaking skills. The accessibility and affordability of taking external tests are also influenced by students' SES. Issues regarding the fairness of this policy need to be carefully examined. In this paper, we consider a series of potential rebuttals that would weaken the fairness of assessment in the validity arguments regarding the use of external tests in this policy. We also identify fairness issues that are critical for major stakeholders in this reform. And finally, we raise questions concerning the basic premises underlying this policy, including arguments for a positive washback effect caused by the speaking tests on primary and secondary school English education and the importance of English-speaking abilities for a globalizing world.

5.1 Introduction

Beginning in 2020, the Japanese government will test high school students' English-speaking skills as part of the nationwide college entrance exam (referred to as the *Common Test*) using external standardized proficiency tests. To satisfy this requirement, students can choose from among eight external assessments: TOEFL iBT, TOEIC, IELTS, Cambridge English tests, Eiken tests, GTEC, TEAP-PBT, and TEAP-CBT (Ministry of Education, Culture, Sports, Science and technology: MEXT

Y. G. Butler (✉)
University of Pennsylvania, Philadelphia, PA, USA
e-mail: ybutler@upenn.edu

M. Iino
Waseda University, Tokyo, Japan
e-mail: iino@waseda.jp

B. Lanteigne et al. (eds.), *Challenges in Language Testing Around the World*,
https://doi.org/10.1007/978-981-33-4232-3_5

2018).[1] The first four tests are international proficiency tests, and the rest are domestic exams. This new policy was motivated by the desire to improve students' English-speaking skills by directly assessing these skills in college entrance examinations. Relying on external tests was a solution to the logistical challenges of measuring the speaking performance of a large number of students in a single day as part of the Common Test (MEXT 2017). Japanese national universities make admission decisions based on a two-step selection procedure that involves the results of both the nationwide college entrance exam (administered once a year) and in-house exams that are developed by individual universities. As part of the new policy, the Japanese Ministry of Education, Culture, Sports, Science, and Technology (MEXT) asked national universities to accept any of these designated external tests either as a stand-alone screening for taking the universities' in-house tests or as part of a combined score with the Common Test, or both. A growing number of private universities are following this policy as well. All institutions following this policy need to decide which external tests to accept, how to determine the cut-off score for each designated test, and how to use the test results. Without reliable information about interpreting and using the scores of these external tests, however, universities are experiencing difficulty in making these decisions. Even more troubling, many universities appear to have made their decisions almost arbitrarily, without relying on any justifiable criteria (Kim and Mizuto 2019).

This policy is a great example of what McNamara and Roever (2006) described as "the manipulation of test consequences in the service of political goals, such as accountability or systematic reform, and the unintended fallout from the test" (p. 203). To illustrate their point, McNamara and Roever described Akiyama's 2004 study in which Japanese high school teachers resisted a proposal to introduce external English-speaking proficiency tests as part of high-stakes high school entrance exams. According to Akiyama, at the time the Japanese believed strongly in meritocracy and egalitarianism, and they expected tests to function purely on merit and apply equally to everybody. Moreover, it was believed that test scores should reflect one's diligence and effort, which are valued highly in Japanese society as characteristics possessed by everybody regardless of innate talent and background. High school teachers' resistance to the policy proposal in Akiyama's study stemmed, in part, from their perception that speaking performance is not a sign of diligence and effort because, for example, students can acquire English-speaking skills without making much effort if they have a chance to live in an English-speaking country. Given that English was treated as an academic subject rather than a practical subject in Japanese schools, testing students' English-speaking skills did not meet teachers' expectations for what should be tested by an entrance exam.

While there is still some expectation that entrance exam scores should reflect Japanese students' diligence and effort, in the 15 years since Akiyama's study there

[1] After we finished writing this chapter, the Japanese government announced on November 1, 2019 that they postponed the implementation of this policy (MEXT 2019). In the meantime, universities are still allowed to use these tests at their discretion. MEXT indicated that they will make a final decision on implementing this policy in a year.

have been substantial changes in educational practices and social perceptions. In schools, English teachers have gradually spent more time on "practical English" instead of focusing solely on grammar and reading instruction (Miyamoto 2018). In the broader society, there is a growing recognition that the widely held belief that Japan is an egalitarian and homogenous society is a myth. In fact, Japan has a high relative poverty rate (the 10th highest in 2015) (OECD 2015), and socioeconomic disparities are frequent topics of public discourse (Moriguchi 2017). Honda (2005) identified a new type of meritocracy—*hyper-meritocracy*—in Japanese postmodern society. In hyper-meritocracy, abilities that are viewed as highly desirable are unfortunately deeply rooted in one's upbringing. Finally, Kariya (2008) empirically showed that students' effort-making is not independent of their socioeconomic status (SES) in Japan.

It is with this background in mind that we argue that Japan's new policy of testing English-speaking skills as part of college entrance exams further imperils *fairness* in Japan's rapidly changing society. The new policy will most likely produce unintended fallouts because the external tests reflect students' SES more than their diligence and effort; it will also likely contribute to widening social disparities. Moreover, the types of speaking abilities that are measured in the external tests are mostly irrelevant to the actual needs of the majority of Japanese students. In this chapter, we examine issues that potentially threaten the fairness and validity of the external assessments being used as part of this new policy. We also raise questions about a basic premise underlying this policy: that there is a universal, measurable (via a single test) oral communicative ability in our globalizing world.

5.2 Testing Problem Encountered

This new policy was implemented as part of a larger reform of college entrance examinations to make them more problem-solving oriented. For English, however, the main goal was to shift to measuring all four skill domains, based on the assumption that incorporating speaking into the exam will lead to greater emphasis on oral communicative skills in English education. There was also strong pressure from the business and political communities to take radical action to improve citizens' English-speaking skills, which are viewed by many as necessary for the nation to be competitive in the global economy (Abe 2017). MEXT plans to completely replace the English portion of the Common Test (which currently mostly assesses receptive skills) with external tests starting in 2024.

The eight proficiency tests were chosen in 2018, but how they were selected is unclear. In 2017, a series of criteria for selecting external tests were released by the government, but the chosen external tests do not meet many of their criteria. For example, one of the critical requirements is that a test should be aligned with Japan's national high school curriculum, but none of the international tests meet this requirement. Even for the four tests developed in Japan, the degree of alignment

with the national curriculum is largely unclear because the test developers have not provided sufficient validity information.

Under this new policy, students can take any test(s) of their choice, but only up to two scores from the same test obtained during the 12th grade can be used for admission purposes. One could argue that having multiple opportunities to take tests will create less anxiety than having only one chance (which is the policy under the current format). But in reality, students already feel pressured to start preparing for tests early because they can practice taking tests an unlimited number of times before Grade 12 (Miyamoto 2018). Since universities can use any of these eight tests, students have to be strategic about which tests to prepare for in order to maximize their chance of being accepted by universities of their choice.

One of the biggest challenges for test users is to identify how to compare the results of the multiple tests, which vary substantially in terms of the test formats and goals as well as the targeted domains, abilities, and proficiency levels. MEXT made a conversion table based on the Common European Framework of Reference (CEFR) for test users (MEXT 2018). Critically, the table was not based on MEXT's own validation efforts; instead, MEXT simply put together information reported by the test developers, but the credibility of some of that information (i.e., validity evidence) is questionable. Curiously, MEXT modified the table a couple of times without clearly explaining the changes. For example, TOEIC has a listening and reading test (TOEIC L&R, 990 points in total) and a speaking and writing test (TOEIC S&W, 400 points in total), and the sum of the scores of these two tests (1390 points) was used in the table released by MEXT in July 2017. In the version released in March 2018, however, the TOEIC speaking and writing score was multiplied by 2.5 (1000 points) and added to the TOEIC L&R score, resulting in a total of 1990. Moreover, MEXT simply replaced the old numbers with the new aggregated scores without verifying their compatibility with CEFR (Hato 2018). Unexplained changes were made in all four domestic tests as well.

The problems discussed above are firmly rooted in the fairness of this new policy. Fairness, or the absence of bias, is a complicated notion, yielding multiple interpretations and definitions. Traditionally, fairness in Japan has often been discussed in the collectivist cultural framework, in which fairness means ensuring equal treatment of *all* members of society. This approach to fairness is often contrasted with Western-oriented conceptions of fairness, which frequently focus on equal treatment of the *individual*. However, empirical investigations do not necessarily support such dichotomous conceptualizations of fairness. For example, Kobayashi and Viswat (2007) compared Japanese and American students' perceptions of fairness in educational settings and reported "diverse viewpoints" (p. 1) in the respective groups. In any event, under either a collective or an individual view of fairness this new policy can be considered "unfair"—both because it does not ensure equal access to test takers (due to regional and socioeconomic differences) and because it does not ensure an evidence-based comparison of the scores of different tests.

In language assessment, test fairness is often discussed in relation to validity, but the relationship between fairness and validity can be conceptualized differently depending on how one defines fairness and validity (Kane 2010). For example, for

Kunnan (2004), validity is part of the quality of fairness, whereas Xi (2010) discusses fairness issues within an argument-based validity framework. Kane (2010) takes the position that fairness and validity are closely related (they essentially concern the same question) and that either one is part of the other. In this chapter, we subscribe to Kane's position because his broad conceptualization of fairness appears to fit the current complex policy context where multiple tests are involved.

According to Kane (2010), fairness can be conceived of as a combination of *procedural fairness* and *substantive fairness*. Procedural fairness stands on a core notion of fairness—"everybody should be treated in the same way"—and concerns "a lack of bias for or against any individual or group" in testing (p. 178). Substantive fairness demands that "score interpretation and any test-based decision rule be reasonable and appropriate" and, most importantly, that "they be equally appropriate for all test takers" (pp. 178–179). Procedural fairness is a necessary condition for fair and valid assessment but does not sufficiently ensure it. For test developers, procedural fairness is largely controllable, but substantive fairness is not entirely controllable.

Many of the problems with the new college entrance exam policy can be organized according to the procedural and substantive fairness frameworks. With respect to procedural fairness, first of all, basic validity and fairness information—including the results of differential item functioning (DIF, a statistical analysis detecting unexpected behaviors for certain subgroups at the item level)—is not fully available for all eight tests. For some tests, insufficient validation/fairness analyses have been carried out or reported. Second, there are a number of concerns related to test accessibility and administration. Some tests have a small number of test locations, which tend to be concentrated in large cities. This means that the accessibility of testing locations differs according to students' place of residence. Moreover, it is not uncommon for some domestic tests to have school-wide administration (students in a given school take the test together at their school). But such administration is a potential threat to fairness/validity of the tests if they are used for high-stakes admission purposes. Therefore, the test agencies need to secure sufficient locations and proctors outside high schools. Another threat to procedural fairness concerns test examination fees, which students are responsible for paying and which can vary substantially, ranging from 5800 to 26000 yen (approximately from US$52 to US$235). Such fees can be a potential hurdle for lower SES students and may influence which test they take and how many times they take it. In addition to paying test fees, students in rural areas far from testing sites might have to pay transportation and accommodation costs. In response to such concerns, some wealthier local governments are considering covering the examination fees for their residents, but this, in turn, can yield an additional potential bias by region. Finally, there are also fairness/validity concerns with respect to test scoring. For some domestic speaking tests, a large number of high school English teachers and college instructors have served as raters. Such practices are no longer acceptable from a fairness point of view, and the testing agencies must secure well-trained raters in a short period of time, which will likely be a tremendous challenge.

As pointed out by Kane (2010), resolving procedural fairness issues such as those described above is not sufficient for achieving a fair and valid assessment practice. Even if these problems are fixed, serious issues concerning substantive fairness remain. One such issue is differing access to test preparation materials and practices. The targeted test domains and proficiency levels in some of the external tests deviate substantially from national curriculum targets; if students want to perform well on those tests, they will likely need to obtain additional materials and learning opportunities beyond normal practices at school. High schools and families have varying capacities to offer additional support to these students. Even for the tests that are more or less aligned with the national curriculum, access to test preparation materials—which are often published by test agencies as well as other private entities and available for a fee—and opportunities to practice speaking English for the test within and outside of school likely differ by SES. Moreover, the misuse of test results, such as conducting inappropriate score aggregations, using invalidated and inappropriate cut scores for admission decisions, and comparing multiple test scores based on CEFR (also see Green 2018), are all serious issues of substantive fairness.

5.3 Unsolved Problems

Due to a chaotic rollout process, stakeholders have experienced tremendous frustration and confusion even before the policy implementation date. As mentioned, universities are having a difficult time deciding how to use the external test results for their admissions procedures (Kim and Mizuto 2019). Notably, a few top national universities, including the University of Tokyo, announced that they would not make the external test scores obligatory for applicants (Ujioka 2018). In Japan's highly centralized educational system, it is very unusual for schools not to follow MEXT's decisions; the fact that some of the most prestigious institutions are not falling in line indicates their strong opposition to the policy. Social network sites are full of students' remarks expressing their confusion and frustration about conflicting or insufficient information about the policy. A recent survey shows that high percentages of high schools in Japan have already started providing special instruction to help students prepare for the external tests (i.e., 68.6% for Eiken tests, 58.1% for GTEC) (Shibasaki 2018), although the nature of that instruction is not known. Meanwhile, select local boards of education have started offering workshops for English teachers at public high schools to provide the educators with information on the external tests as well as instructional tips for helping their students prepare for the tests. Again, the details of such workshops are unknown, but this could be a sign of a potentially undesirable test-driven washback effect. Finally, as of January 2019, MEXT has not proposed any explicit guidelines for accommodations or special considerations for students with disabilities or special needs; instead, all such considerations are left up to the individual test agencies, whose practices differ substantially.

5.4 Insights Gained

After observing preliminary unintended fallout from the use of external tests in the college entrance exam system in Japan, we can see that a number of the fairness issues addressed above appear to originate in the very assumptions that underlie this policy. Such assumptions include that (1) English-speaking skills, as an important global competence, should be used as a gatekeeper for everybody who seeks higher education; and (2) such skills should be understood and measured uniformly against a global framework or standard such as CEFR. But what particular English-speaking skills does MEXT expect students to develop? Are these particular skills really a global competence that Japanese students need? Should we evaluate Japanese students' English-speaking skills using a global standard and, if so, should CEFR be that standard?

Communicative competence—one's ability to use language appropriately in social situations—was originally proposed by Hymes (1972) and has had a tremendous influence on language teaching and assessment. There are various models for communicative competence, but many models conceptualize it as a composition of some sort of linguistic and social/pragmatic knowledge and the ability to use such knowledge in performance. In assessment theory, communicative competence has largely been conceived of as an individual's capability that can be inferred from his or her independent performance on tasks that are representative of language use in the target domain. In assessment practice, "the ability to use" component in the original Hymes model has not been seriously discussed due to its complexity, which goes beyond linguistic elements (various cognitive, social, and affective elements are also involved) (McNamara 1996). In many standardized proficiency tests, the knowledge components in communicative competence are organized into four skill domains and assessed separately. The "appropriateness" aspect of communicative competence has largely been judged based on the performance of "native speakers." In the context of Japan, the speaking skill domain is often considered the ultimate manifestation of communicative competence (Abe 2017).

In the past decade or two, however, there has been growing interest in socio-interactional approaches to conceptualizing language abilities. In those approaches, language abilities are considered to be embedded in social contexts and constructed in fluid and dynamic interaction. The field of English-as-a-lingua-franca (ELF) challenges the very notion of native-speaker norms and questions the static view of language ability that has been conventionally accepted in the assessment community (e.g., Canagarajah 2009; Harding and McNamara 2017; Jenkins 2006). ELF's emphasis on communicative effectiveness, rather than correctness and appropriateness, highlights the role of "the ability for use" in language abilities, which presumably varies substantially in communication in people's first language as well as their second language. Reflecting such a fluid conceptualization of language abilities in assessment is not easy, especially in standardized tests (Harding and McNamara 2017), but this new conceptualization of language abilities better fits the realistic needs of Japanese students who largely interact in English-as-a-lingua-franca contexts in the globalizing world.

What MEXT promotes and tries to measure through standardized tests, therefore, are the knowledge-based components of communicative competence that are sliced into skills under the old static view of competence based on native norms, even though that is not the kind of ability Japanese students need in a globalizing world. Measuring English-speaking skills and using those measurements as a qualification for higher education is particularly problematic because these are the skills where students' SES and regional backgrounds are most likely manifested, no matter how hard test developers work to control procedural fairness issues. In a society where Japanese is used almost exclusively, students must make a special effort to create opportunities to speak English and get feedback to develop their speaking skills, and those opportunities usually require financial and regional resources. Foreign language learning is a huge and fast-growing business in Japan, with an 867-billion-yen market in 2017 (Yano Economic Research Institute 2018). Parents with higher educational backgrounds and who reside in larger cities invest significantly more in their children's English-speaking practice and do so earlier in their children's lives (Benesse General Research Center for Education 2014).

Meanwhile, the assessment community has yet to develop an unbiased strategy for capturing the kinds of language abilities needed for a globalizing world (the newly conceptualized language abilities). One may even wonder if such abilities are measurable through a standardized test. Similarly, it is not clear if they can even be evaluated and compared against some sort of universal framework (besides the fact that CEFR was not developed for such purposes in the first place). Perhaps the language abilities necessary for a globalizing world are not competencies that can be made uniform or standardized across the globe. Because such language abilities are highly context dependent, fluid, and complex, quantification based on any uniform standards or frameworks is misleading, whether or not it is done through a standardized test. Until the assessment community can come up with a fair and valid remedy, quantified evaluation of such language abilities should not be implemented for high-stakes purposes such as college admission.

5.5 Conclusion: Implications for Test Users

We have discussed Japan's new policy decision to use external English proficiency tests for college admission, and argued that the central problem is one of fairness. Based on Kane's (2010) distinction between procedural fairness and substantive fairness, we examined a series of potential issues that appear to weaken the fairness and validity of these assessments. If MEXT wants to implement this policy, it must address these procedural fairness problems. However, there remain a number of serious substantive fairness issues as well. These substantive fairness issues are difficult to solve, even if test users could gain sufficient assessment literacy, because the premise underlying the MEXT policy not only rests on a misperception of the language abilities needed for Japanese students in a globalizing world but also structurally works against students with lower SES. Without a fair and valid

assessment that captures the language abilities that students really need, making high-stakes college admission decisions based on existing quantification methods is highly misleading and potentially contributes to widening socioeconomic disparities in Japan.

References

Abe, M. (2017). *Shijosaiaku-no eigoseisaku [The worst English education policy in history]*. Tokyo: Hitsuji Shobo.

Benesse General Research Center for Education. (2014). *Eigo-kyoiku dainikai* [English education Part 2]. https://berd.benesse.jp/berd/data/dataclip/clip0014/index2.html. Accessed 18 January 2019.

Canagarajah, S. (2009). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Language Assessment Quarterly, 16,* 229–242.

Green, A. (2018). Linking tests of English for academic purposes to the CEFR: The score user's perspective. *Language Assessment Quarterly, 15*(1), 59–74.

Harding, L., & McNamara, T. (2017). Language assessment. In J. Jenkins, W. Baker, M. Dewey (Eds.), *The Routledge handbook of English as a lingua franca.* https://doi.org/10.4324/9781315717173.ch45

Hato, Y. (2018). Minkanshiken-no nani-ga mondai nanoka [What is the problem with external tests?]. In T. Haebara (Ed.), *Kenho—meiso-suru eigo nyushi [Verification: Troubled college entrance examinations in English]* (pp. 41–68). Tokyo: Iwanami.

Honda, Y. (2005). *Tagenka-suru nouryoku-to nihon shakai [Diversifying competencies and Japanese society]*. Tokyo: NTT Publications.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269–293). Harmondsworth: Penguin Books.

Jenkins, J. (2006). The spread of EIL: A testing time for testers. *ELF Journal, 60*(1), 42–50.

Kane, M. (2010). Validity and fairness. *Language Testing, 27*(2), 177–182.

Kariya, T. (2008). *Gakuryoku-to kaiso [Academic achievement and social class]*. Tokyo: Asahi Shinbun Publications.

Kim, S., & Mizuto, K. (2019, January 7). Eigo minkanshiken katsuyo-ni annun [English external tests in trouble]. *Mainichi Newspaper*, p. 14.

Kobayashi, J., & Viswat, L. (2007). An exploratory study of "fairness" in educational settings: American and Japanese university students. *Journal of Intercultural Communication, 14*. https://www.immi.se/intercultural/nr14/kobayashi.htm. Accessed 27 July 2019.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. J. Weir (Eds.), *European language testing in a global context: Proceedings of the ALTE Barcelona Conference* (pp. 27–48). Cambridge, UK: Cambridge University Press.

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

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension.* Malden, MA: Blackwell.

MEXT. (2017). *Koudai setsuzoku kaikaku-no shinchoku jokyo-ni tsuite* [Progress report on the reform connecting high-school and college education]. http://www.mext.go.jp/b_menu/houdou/29/05/1385793.htm. Accessed 17 January 2019.

MEXT. (2018). *Daigaku nyushi kaikaku* [College entrance examination reform]. http://www.mext.go.jp/a_menu/koutou/koudai/detail/1408564.htm. Accessed 17 January 2019.

MEXT. (2019). Daijin meseji eigo minkan shiken-ni tsuite [A message from the Minister regarding the external English tests]. https://www.mext.go.jp/a_menu/other/1422381.htm. Accessed 20 March 2020.

Miyamoto, H. (2018). Koko-kara-mita eigo nyushi kaikaku-no mondaiten [Problems with the reform of college entrance examinations from a perspective of high schools]. In T. Haebara (Ed.), *Kensho – meiso-suru eigo nyushi [Verification: Troubled college entrance examinations in English]* (pp. 26–40). Tokyo: Iwanami.

Moriguchi, C. (2017). Nihon-wa kakusya syakai-ni nattanoka? [Has Japan had a disparity in wealth?]. *Discussion Paper Series A*, No. 666. http://www.ier.hit-u.ac.jp/Common/publication/DP/DPS-A666.pdf. Accessed 17 January 2019.

OECD. (2015). *Poverty rate.* https://data.oecd.org/inequality/poverty-rate.htm. Accessed 18 January 2019.

Shibasaki, O. (2018, November 6). Daigaku nyugakusya senbatsu kaikaku-ni kansuru anketo cyosa hokokusyo [A survey report on the college entrance exam reform]. http://company.sanpou-s.net/press/pdf/181106.pdf. Accessed 18 January 2019.

Ujioka, M. (2018, September 25). *Tokyodai, eigo minkanshiken-no seiseki teisyutsu hissyutosezu*, shindaigaku nyushi [The University of Tokyo, not requiring English external test scores, the new entrance exam]. Asahi Newspaper. https://www.asahi.com/articles/ASL9T5T8HL9TUTIL031.html. Accessed 20 January 2019.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147–170.

Yano Economic Research Institute. (2018). *Gogaku bijinesu tettei chosa repoto* [Report on extensive research of language business]. https://www.yano.co.jp/press-release/show/press_id/2013. Accessed 18 January 2019.