Koizumi *et al. Language Testing in Asia* (2022) 12:53
https://doi.org/10.1186/s40468-022-00203-3

Language Testing in Asia

RESEARCH Open Access

Skill profiles of Japanese English learners and reasons for uneven patterns

Rie Koizumi[1*], Toshie Agawa[1], Keiko Asano[2] and Yo In'nami[3]

*Correspondence:
rkoizumi@seisen-u.ac.jp

[1] Research Institute for Language Education, Seisen University, 3-16-21 Higashi Gotanda, Shinagawa-ku, Tokyo 141-8642, Japan
[2] Juntendo University, 1-1 Hirakagakuendai, Inzai, Chiba 270-1695, Japan
[3] Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan

Abstract

Background: Variations in the skill profiles of learners have become an important research area in recent years. However, there is a lack of empirical research on this topic in Japan. We conducted three studies to address this gap.

Methods: Study 1 investigated the characteristics of the flat and uneven skill profiles of Japanese learners of English using 10 datasets from five standardized four-skill second-language English proficiency tests. Studies 2 and 3 examined the reasons behind learners having these uneven profiles using a convergent mixed-methods approach (Creswell & Plano Clark, Designing and conducting mixed methods research, 2018) that consisted of a questionnaire and an interview, respectively.

Results: The results of Study 1 suggested that a flat profile is uncommon, and that various types of uneven profiles exist across datasets. The most frequently observed uneven profiles were as follows: (a) listening, speaking, and writing are lower than reading (LSW<R); (b) speaking and writing are lower than listening and reading (SW<LR); and (c) speaking is lower than the other three skills (S<LRW). The results of Studies 2 and 3 suggested three key reasons, namely, (a) insufficient practice, (b) particular subskills or processes required to accomplish test tasks, and (c) unfamiliarity with test formats and test-taking environments.

Conclusions: The results demonstrated the importance of considering uneven skill profiles in L2 research and practice. These results have implications for test development, admission, placement, and pedagogy. For example, skill profile information can benefit admission and placement officers when they make a decision and can aid teachers and administrators in planning remedial instructions.

Keywords: Japanese learners of English, Second-language English proficiency, Listening, Reading, Score report, Skill imbalance, Skill profiles, Speaking, Writing

Introduction

Second-language (L2) learners have flat or uneven profiles in terms of their listening, reading, speaking, and writing skills. The Council of Europe (2001, 2009, 2020) acknowledged different types of uneven skill profiles. Learners with a flat (i.e., even) profile have acquired the four skills to an equal level, whereas those with uneven (i.e., jagged or nonflat) profiles have one or more skills that are substantially higher or lower than the others. For example, a learner may have very good reading skills, average listening and writing skills, and poor speaking skills. The use of proficiency levels, such as those in the Common European Framework of Reference for Languages (CEFR), enables researchers and practitioners to compare profiles across tests using the same standard; however, test users must be cognizant that each test differs in terms of test constructs and purposes, and that the test results may not be strictly comparable (Deygers et al., 2018; Harsch, 2018).

According to North (2021), uneven profiles have long been noted and are more commonly observed than flat profiles, even among advanced language users. In the introduction to his workshop that focused on uneven profiles, North (2021) stated that communicative tests are designed to measure skills separately because testing institutions have recognized profile variability. Moreover, a variety of CEFR scales are provided "to facilitate creating a differentiated needs profile——at an appropriate degree of detail" (North, 2021, 0:13:10). Additionally, he observed that little research has been conducted on the percentages of L2 learners with flat and uneven profiles. Uneven profiles can be divided into various types based on their skill levels, such as those with high skills in reading and writing and low skills in listening and speaking, and vice versa. These differences in skill levels may be the product of a variety of reasons, and the exploration of such reasons could provide insights into the factors affecting learners' varied profiles, such as educational backgrounds, and affective and cognitive processes. The current study focuses on the breakdown of the flat and uneven skill profiles of Japanese learners studying English as a foreign language (EFL) and the reasons for the uneven profiles (see Ma & Winke, 2022, and Pang & Skehan, 2021, for subskill profiles within each skill).

Literature review

Skill profiles can be regarded as one of the important research areas (Harsch, 2014; Hulstijn, 2015; Hulstijn et al., 2012). Skill profiles, or concise descriptions of each learner's L2 skills, can provide valuable information on learners' skill balance and imbalance, as well as learning, teaching, program evaluation, and policy making (Choi, 2017). For example, feedback based on skill profiles can inform learners of their strengths and weaknesses and prompt them to modify their learning routines. Skill profiles can be found in the score report of standardized four-skill tests, such as the Test of English as a Foreign Language® (TOEFL) Internet-based test (iBT). Score reports typically include total scores and skill scores in skill-based sections, the latter of which can be translated into skill profiles.

Since the differences in skill levels are expected, the four skills are usually reported separately and are typically only moderately related to each other (Liao et al., 2010; Sawaki & Sinharay, 2018). Such a moderate degree of relationship among skills mirrors the varied and dynamic nature of how they operate together when processing and completing tasks and how they help researchers define L2 proficiency. For example, Powers (2013) reported that listening self-assessment scores were best predicted using the Test of English for International Communication® (TOEIC) four-skill scores. The prediction rate decreased when speaking, writing, or both were removed. Similar findings were obtained for the prediction of the self-assessments of reading, speaking, and writing. These results suggest that the differences across skill scores allow researchers to more comprehensively assess learners' ability and predict the overall L2 proficiency (Powers, 2013; see also Li & Zhang, 2021).

The recognition of the importance of L2 learners' variability in the four-skill profiles has led to increased academic interest into this topic. For example, Ginther and Yan (2018) reported based on their cluster analysis that Chinese international undergraduates (2011 and 2012 cohorts) at an American university showed three-skill profiles from the TOEFL iBT, namely, (a) speaking (S) lower than the other three skills (listening [L], reading [R], and writing [W]), expressed as S<LRW; (b) SW<LR (the difference between SW and LR was 21 to 23 points on average[1]); and (c) all four skills being relatively low equally (balanced low). The university's admission policy until 2011 was to accept international students with a minimum TOEFL iBT total score of 80. However, the university noticed a problem with the (b) group, which had a lower grade point average (GPA) in their first year than the group that had the other skill profiles. Consequently, an additional admission requirement of obtaining minimum speaking and writing scores of 18 each was implemented in 2013. Thus, applicants with the (b) SW<LR profile were no longer admitted. The results from the 2013 cohort showed three profiles: (a) S<LRW, (c) balanced low, and (d) balanced high, where the four skills were relatively and equally high. The skill profiles were related to first-year academic performance as the (b) SW<LR group had lower GPA means with larger variations than the other groups. The authors speculated that these uneven profiles may have been caused by an intense TOEFL preparation for the reading and listening sections or by participants engaging in cheating.

Other studies also reported some groups of learners having uneven skill profiles (e.g., Bridgeman et al., 2016; Harsch et al., 2017; Vahed, 2021). Bridgeman et al.'s (2016) study reported an extreme TOEFL iBT score difference (about 16 points) between productive and receptive skills (SW<RL) among some Chinese international undergraduates with business majors at an American university. The TOEFL scores were found to better predict students' GPA when the scores of such learners were not included in the analysis. The authors also inferred that intensive test preparation could yield unbalanced skill profiles.

Harsch et al. (2017) reported that Chinese and Indian groups, who were undergraduate and graduate students in the UK, had slightly higher scores for reading and listening than writing and speaking (14 points at most and an approximate average of 2 points) in the TOEFL iBT. These differences were not observed in the German group.

The review of the abovementioned studies can be summarized as follows. First, it revealed uneven skill profiles with a divergence between productive and receptive skills (SW<LR). Second, such an uneven profile could be a problem because adequate levels of productive abilities are required for success in academic settings (Ginther & Yan, 2018). It could also pose a challenge in predicting learners' academic performances (Bridgeman et al., 2016). Third, uneven skill profiles were observed among particular groups of learners (Harsch et al., 2017). Fourth, the admission policy of minimum skill score requirements (e.g., of at least 18 each in the TOEFL iBT speaking and writing sections) can affect students' post-entry academic performance (Ginther & Yan, 2018).

---

[1] To facilitate understanding, = and the spaces before and after < and = were omitted.

While previous studies have provided insights into the flat and uneven profiles of the four skills among L2 learners of English, and the potential negative effects of uneven profiles on academic performance, its effects among Japanese EFL learners of English warrant additional examination. Japan's Ministry of Education, Culture, Sports, Science and Technology (MEXT, 2018) observed that at the group level, some Japanese learners had an uneven profile of SW<LR; however, few empirical studies have extensively examined this topic at the individual level. In the MEXT's study, L2 four-skill English tests were conducted in 2017 among lower and upper secondary school students to examine how the four English skills can be fostered in a balanced way. MEXT administered listening, reading, and writing tests to approximately 60,000 junior and 60,000 senior high school students and speaking tests to 20,000 junior and 10,000 senior high school students (from 600 national and 300 public schools that were randomly selected from the population). Although the results did not focus on skill profiles, they suggested that some students had an uneven profile. For example, approximately 30% of senior high school students obtained a CEFR level of A2 in listening (30.2%) and reading (29.4%), whereas less than 20% obtained the same level in speaking (11.7%) and writing (19.3%). The small percentage of students with B1 or more in all the skills (i.e., 0.4% to 4.1%) suggests that some students had an uneven profile of SW<LR, which may lead to future problems with academic performance (Ginther & Yan, 2018). However, a more detailed examination was needed as the MEXT (2018) analyzed the skill profiles at the group level, and only individual-level analyses could reveal a detailed overall trend of the learners' profiles. Furthermore, previous studies (e.g., Ginther & Yan, 2018) including by MEXT did not examine the reason behind the uneven profiles among learners with uneven profiles. Therefore, studies examining the underlying factors behind skill profiles are needed.

Research questions and overall study design

We examined Japanese learners' skill profiles of all possible skill combinations and learners' self-perceived reasons for the uneven profiles to investigate the general trend of the varied four-skill profiles of Japanese EFL learners of English at the individual level and the underlying reasons for uneven profiles. Since the breakdown of skill profiles would differ across tests and contexts, we systematically examined all possible combinations of such breakdowns using multiple datasets (from five tests and nine different groups of participants) by posing the following two research questions (RQs).

- RQ1: What are the characteristics of the four-skill profiles of Japanese EFL learners that are frequently observed across datasets?
- RQ2: What are the reasons perceived by learners for their uneven profiles?

We conducted three studies, as visually presented in Fig. 1. RQ1 was examined in Study 1 (quantitatively) by comparing skill profiles using 10 datasets from five standardized four-skill English proficiency test scores. The datasets were examined to identify the overall patterns in the flat and uneven skill profiles as obtaining common profiles across datasets and tests could indicate the typical characteristics of the target learners. We made our utmost efforts into obtaining as many datasets as possible by contacting

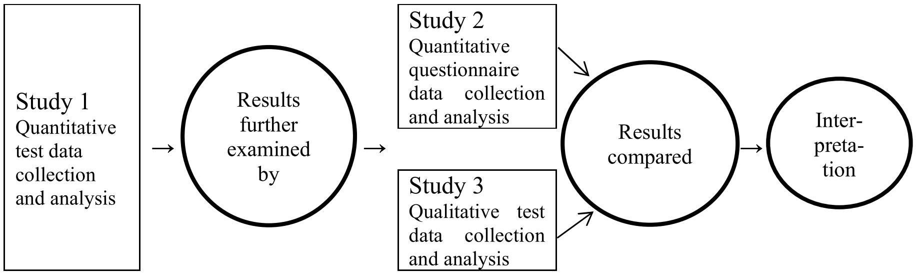
Study 1
Quantitative test data collection and analysis
Results further examined by
Study 2
Quantitative questionnaire data collection and analysis
Study 3
Qualitative test data collection and analysis
Results compared
Inter-preta-tion

Fig. 1 Overall study design based on Creswell and Plano Clark (2018)

Table 1 Ten datasets in Study 1

Dataset	Test used	Participants	Reference
1	GTEC CBT	Senior high school students	Koizumi et al. (2019)
2	TOEFL Junior Comprehensive	Senior high school students	Koizumi & In'nami (2017)
3		Undergraduate students	
4	TEAP	Undergraduate students	In'nami et al. (2016)
5	TOEFL iBT		
6	TOEIC	Undergraduate and graduate students	Koizumi (2015a, 2015b, reanalyzed)
7	TOEFL iBT	Undergraduate students in 2014	Koizumi et al. (2018)
8		in 2015	
9		in 2016	
10		in 2017	

agencies and collecting data ourselves, and we analyzed all the datasets in which we could obtain permission for use.

RQ2 was investigated in Studies 2 (quantitatively) and 3 (qualitatively) using a questionnaire and an interview. The combined use of quantitative and qualitative methods followed a convergent mixed-methods design (Creswell & Plano Clark, 2018; see Riazi, 2017 for a summary of the different terminology). The convergent mixed-methods design allowed us to utilize the strengths of the quantitative and qualitative approaches. The quantitative-oriented questionnaire in Study 2 enabled us to identify the overall reasons for uneven skill profiles, which were further examined in the qualitative interview in Study 3. Collectively, the insights from the current article would help researchers clarify the importance of skill profiles in L2 research and practice.

Study 1: investigation into skill profiles

Method for Study 1

Study 1 used 10 datasets to identify whether any patterns could be observed in skill profiles across datasets (RQ1). While they were from different tests and participants, they were all test-score data from learners studying English in Japan that were provided by testing institutions. As shown in Table 1, the 10 datasets contained skill scores from five tests: (a) Global Test of English Communication (GTEC) Computer-Based Testing (CBT; https://www.benesse.co.jp/gtec/en/); (b) TOEFL Junior® Comprehensive (TOEFL Junior, hereafter; https://www.ets.org/toefl_junior/content/); (c) the Test of English for Academic Purposes (TEAP; https://www.eiken.or.jp/teap/);

Table 2 Five tests used in Study 1

	GTEC CBT	TOEFL Junior	TEAP	TOEFL iBT	TOEIC
Listening & reading	Computer adaptive, multiple choice	Computer based, multiple choice	Paper based, multiple choice	Computer based, multiple choice	Paper based, multiple choice
Speaking	Computer based	Computer based	Face-to-face interview	Computer based	Computer based
Writing	Computer based	Computer based	Paper based	Computer based	Computer based
Tested CEFR levels	A1 to C1	A1 to B2	A1 to C1	A2 to C2	A1 to C1
Purpose	General and academic	General and academic	Readiness for academic work	Academic	Business and general
Conversion table	Center for Entrance Examination Standardization (2017)	Tannenbaum and Baron (2015)	Internal score comparison table	Papageorgiou et al. (2015)	Educational Testing Service (2019)

TOEFL Junior, which was used in the current study, was a test of four skills but now includes listening, reading, and speaking while excluding writing

(d) TOEFL iBT (https://www.ets.org/toefl/test-takers/ibt/about); and (e) TOEIC (https://www.ets.org/toeic). All tests consisted of the four-skill sections but differed in delivery modes, range of proficiency levels, and purposes (see Table 2). They had a conversion table that related each test skill score to CEFR.

Datasets

Ten datasets of four-skill scores were analyzed separately. Participants in each dataset were a convenience sample of those who took any one or two of those tests and agreed to participate in the data collection. A majority of the participants (95% or more) were Japanese English learners who studied English primarily in Japan. We used all the data in each dataset, and the number of participants in each dataset varied as a result.

Dataset 1 The dataset contained scores from GTEC CBT from senior high school (public and private) students in Japan. A total of 1805 students took the test in March 2018.

Datasets 2 and 3 The datasets contained scores from TOEFL Junior from 2799 Japanese senior high school (public and private) students and 234 Japanese university (public and private) undergraduates. The data were collected between September 2015 and January 2016.

Datasets 4 and 5 The datasets contained scores from TEAP and TOEFL iBT from 100 Japanese private university undergraduates, who completed both tests in December 2013. The datasets differ from the others in that the same group of students took both tests. This allowed us to rigorously examine the skill profiles of the same participants across the tests.

Dataset 6 The dataset contained scores from TOEIC. In total, 106 Japanese undergraduates and graduates from national universities completed the TOEIC L&R and TOEIC S&W between July and May 2009.

Datasets 7 to 10 The datasets contained scores from TOEFL iBT. The data were collected from 2014 to 2017, between August and November ($N = 521$). Each year, 126 to 138 undergraduate students at a Japanese private medical university participated in this study. The datasets were unique in that they were all derived from students at one university, which enabled us to examine skill profiles from datasets collected at the same school across 4 years.

Analysis

We identified the CEFR levels for each skill to construct learners' skill profiles and computed their percentage in two steps. First, four-skill scores were converted into the corresponding CEFR levels using a conversion table provided by each test developer. For example, to convert the TOEFL iBT test scores, we used Table 3 that was adapted from Papageorgiou et al.'s (2015) comparison table of TOEFL iBT and CEFR (see Table 1 of their paper); when a learner received 16 points for the listening section in the TOEFL iBT, the corresponding level in the CEFR level was judged to be at the B1 level for listening. This conversion process was repeated for each learner, skill, and dataset.

Second, the CEFR levels for each skill were summarized manually. For example, if all four skills belonged to the same CEFR level, the skill profile was considered flat or LRSW. All other profiles were considered to be uneven. If R was one level higher than the other skills, the profile was labeled as LSW<R. If R was two levels higher than the other skills, the profile was labeled as LSW<<R (e.g., R is at the B2 level; the other skills are at the A2 level). After all the profiles were summarized into one table, they were classified to examine whether any patterns could be observed among them.

The CEFR levels for each skill, rather than the test scores, were used to compare the analyses across datasets, using the percentages of skill profiles in each dataset due to the unsuitability of the other two other alternatives. The first alternative method is to use percentile scores or standard scores (e.g., *z*-scores) for each skill. However, this was difficult to apply here because not all tests reported information on percentile scores, and the target test-takers were not the same across the tests. Therefore, standard scores were not comparable across tests.

Table 3 Comparison table between TOEFL iBT scores and CEFR levels in Study 1

	L	**R**	**S**	**W**	**Total**
C1 or higher	22–30	24–30	25–30	24–30	95–120
B2	17–21	18–23	20–24	17–23	72–94
B1	9–16	4–17	16–19	13–16	42–71
A2 or lower[a]	0–8	0–3	0–15	0–12	0–41

[a] The current study did not distinguish between the "A1 or lower" level and the A2 level in S and W, unlike Papageorgiou et al. (2015), in order to compare the four skills on an equal footing

We also did not use cluster or latent profile analyses or finite mixture modeling (Choi, 2017; Dunn & Iwaniec, 2021; Ginther & Yan, 2018; In'nami & Koizumi, 2021; Roohr et al., 2022). These methods take a person-centered approach and classify learners according to the similarity in their score patterns. These methods contrast with factor analysis, which takes an item-centered approach and classifies items that receive similar responses (see Flaherty & Kiff, 2012; Oberski, 2016). Cluster and latent profile analyses differ in that the former aims to identify the observed groups, whereas the latter aims to identify the latent groups. Finite mixture modeling includes cluster and latent profile analyses and others (Masyn, 2013). Cluster and latent profile analyses were not suitable for this study for two reasons. First, when the standard scores for each skill in a test were analyzed, the skill profiles were likely to be flat at different proficiency levels. For instance, Sawaki and Sinharay (2013) examined learners' skill profiles for TOEFL iBT data using cluster analysis (e.g., 14,495 test-takers from the April 2007 administration) and found flat profiles at different proficiency levels. Learners who did well on one section were likely to do equally well on another. Furthermore, cluster or latent profile analyses show an overall pattern of the group, with each profile group including various sub-profiles. Since the current study aimed to examine a general pattern of skill profiles of all possible skill combinations, cluster or latent profile analyses were not used. Second, when the difference between skill standard scores was computed, six scores were derived (L-R, L-S, L-W, R-S, R-W, and S-W scores). These scores were interdependent and likely to result in multicollinearity. Therefore, they were not suitable for cluster or latent profile analyses or finite mixture modeling.

In sum, percentile or standard scores (e.g., *z*-scores) for each skill were not applied in the current study due to the lack of publicly available information on percentile scores or the different groups of examinees taking tests (making the score comparison difficult). Cluster and latent profile analyses or finite mixture modeling was also not applied as flat profiles were more likely to be observed with standard scores. Furthermore, different scores between skills were expected to be highly correlated with each other, a situation that is not suitable for the use of cluster and latent profile analyses or finite mixture modeling. Hence, the use of CEFR conversion tables and percentages was judged to be the most appropriate for this study. Additionally, since our study was descriptive, we did not use statistical significance testing.

Results and discussion for Study 1

Table 4 shows the results of the CEFR levels in each dataset. A wide range of English proficiency of learners was observed, with the highest percentage at the A2 or B1 level for each skill. There were four exceptions to this: In Dataset 2, most learners had an A1 or lower level in speaking (50.88%); in Datasets 6, 8, and 9, most learners had a B2 level in writing (40.57%, 38.89%, and 33.08%, respectively). This indicates variations in the level of mastery across the skills and datasets.

We found 75 skill profiles across 10 datasets (see Table 13 in Appendix). From the 75 profiles, 15 skill profiles, with 5% or more of the participants having at least one dataset, were selected and are shown in Table 5 in the order of GTEC CBT. For example, learners with higher reading skills than listening, speaking, and writing skills (LSW<R) were found to be 24.42% and 12.65% of all learners in Datasets 1 and 2, respectively.

Table 4 Percentages of participants' CEFR levels for each dataset in Study 1

Dataset		A1 (or lower)	A2	B1	B2	C1 or higher
1: GTEC CBT Senior	L	4.99	44.93	42.16	6.15	1.77
	R	0.61	18.17	62.60	16.18	2.44
	S	4.04	58.12	34.79	2.71	0.34
	W	2.27	56.95	38.73	2.05	0
2: TOEFL Junior Senior	L	26.01	42.23	26.69	5.07	--
	R	15.97	53.34	22.65	8.04	--
	S	50.88	40.16	7.22	1.75	--
	W	34.66	50.59	11.4	3.36	--
3: TOEFL Junior Univ.	L	6.84	28.21	49.15	15.81	--
	R	0.43	28.21	41.45	29.91	--
	S	30.34	49.15	16.67	3.85	--
	W	13.25	48.72	29.06	8.97	--
4: TEAP Univ.	L	1.00	21.00	63.00	15.00	--
	R	1.00	11.00	64.00	24.00	--
	S	1.00	33.00	50.00	16.00	--
	W	1.00	21.00	64.00	14.00	--
5: TOEFL iBT Univ.	L	--	23.00	50.00	16.00	11.00
	R	--	3.00	66.00	29.00	2.00
	S	--	34.00	32.00	31.00	3.00
	W	--	55.00	24.00	18.00	3.00
6: TOEIC Univ.	L	0	25.47	56.60	16.04	1.89
	R	0	33.02	56.60	8.49	1.89
	S	16.98[a]	41.51	38.68	1.89	0.94
	W	1.89	19.81	34.91	40.57	2.83
7: TOEFL iBT 2014 Univ.	L	--	26.77	50.39	12.6	10.24
	R	--	1.57	55.12	35.43	7.87
	S	--	75.59	18.90	4.72	0.79
	W	--	26.77	43.31	27.56	2.36
8: TOEFL iBT 2015 Univ.	L	--	27.78	38.89	21.43	11.90
	R	--	0	53.17	38.89	7.94
	S	--	72.22	20.63	4.76	2.38
	W	--	23.02	35.71	38.89	2.38
9: TOEFL iBT 2016 Univ.	L	--	23.85	43.85	17.69	14.62
	R	--	3.08	56.92	26.15	13.85
	S	--	65.38	21.54	7.69	5.38
	W	--	27.69	33.08	33.08	6.15
10: TOEFL iBT 2017 Univ.	L	--	33.33	39.13	14.49	13.04
	R	--	2.17	59.42	28.26	10.14
	S	--	75.36	14.49	6.52	3.62
	W	--	33.33	36.23	26.09	4.35

[a] Including below A1, 0.94%

Those with the same CEFR levels across skills were interpreted as having a flat profile (LRSW). Such learners were observed at small percentages, ranging from 1.59 (Dataset 8) to 29.00% (Dataset 4). This indicates that learners with a flat profile are in the minority, corroborating North's (2021) argument that uneven profiles are more common. Furthermore, compared with other datasets, the percentages were relatively higher for the data with GTEC CBT taken by senior high school students (Dataset 1, 18.94%), with

Table 5 Skill profile types with 5% or more of participants (%) in Study 1

Dataset	1	2	3	4	5	6	7	8	9	10		
Pattern	**GTEC CBT: Senior[a]**	**TOEFL Junior: Senior[b]**	**TOEFL Junior: Univ.[c]**	**TEAP:Univ.[d]**	**TOEFL iBT: Univ.[d]**	**TOEIC: Univ.[e]**	**TOEFL iBT: Univ. 2014[f]**	**TOEFL iBT: Univ. 2015[g]**	**TOEFL iBT: Univ. 2016[h]**	**TOEFL iBT: Univ. 2017[i]**	**Lowest skill**	**Highest skill**
LSW<R	*24.42*	*12.65*	*11.54*	*12.00*	*17.00*	1.89	7.87	*15.87*	*13.85*	*19.57*	LSW	R
Flat	*18.94*	*17.90*	9.40	*29.00*	9.00	9.43	3.15	1.59	*10.00*	6.52	LRSW	LRSW
SW<LR	*11.79*	*14.29*	*13.25*	8.00	*14.00*	6.60	*11.81*	7.94	*13.85*	*12.32*	SW	LR
S<LRW	6.58	*12.40*	*15.81*	9.00	2.00	20.75	*20.47*	*16.67*	*13.58*	*10.87*	S	LRW
L<RSW	5.53	1.89	0.43	5.00	1.00	0	0	0	0	0	L	RSW
W<LRS	5.53	3.75	2.99	4.00	7.00	0	0	0	0	0	W	LRS
LS<RW	5.42	5.22	3.42	6.00	1.00	4.72	0	0	0	0	LS	RW
LW<RS	3.16	1.43	0	1.00	7.00	0	0	0	0	0	LW	RS
RSW<L	2.55	7.68	2.56	5.00	2.00	3.77	0	0	0	0	RSW	L
S<LW<R	2.16	2.64	7.69	2.00	0	0	8.66	5.56	3.08	7.97	S	R
LRW<S	1.22	0.75	0	7.00	7.00	0	0	0	0	0	LRW	S
LRS<W	1.16	2.00	0	4.00	1.00	*15.09*	3.15	2.38	4.62	0.72	LRS	W
RS<LW	0.22	2.54	0	2.00	1.00	8.49	0	0	0	0	RS	LW
R<LSW	0.06	1.04	0.43	0	5.00	1.89	0	0	0	0	R	LSW
S<LR<W	0.06	0.39	1.28	1.00	0	*10.38*	0	0	0	0	S	W

Out of the 75 skill profiles in Table 13 in Appendix, 15 typical profiles, of which 5% or more of all the participants had in at least one dataset, were selected. Arranged in the order of GTEC CBT. Senior, senior high school student data. Univ., university student data. [a]$n = 1805$. [b]$n = 2799$. [c]$n = 234$. [d]$n = 100$. [e]$n = 106$. [f]$n = 127$. [g]$n = 126$. [h]$n = 130$. [i]$n = 138$. ([f] to [i] total $N = 521$). 10% or more were *italicized*. This also applies to Tables 6, 7 and 8

TOEFL Junior taken by senior high school students (Dataset 2, 17.90%), and with TEAP taken by first-year university students (Dataset 4, 29.00%). Since senior high school or first-year university students are the main target group of such tests, their difficulty level is likely to be adequate for most test-takers. If this is the case, the results might indicate that if learners take a test that matches their proficiency level, they are more likely to obtain a flat skill profile than when they take a test that is too challenging or easy for them. This can be tested in future studies.

Among the various uneven profiles, there were four that 10% or more of learners had across datasets: LSW<R (in eight datasets out of 10), SW<LR (in seven datasets), S<LRW (in seven datasets), and flat (in four datasets). The highest percentages of each skill profile were less than 30% (24.42%, 29.00%, 14.29%, and 20.75%, respectively). A limited number of skill profiles (four) across datasets with a small percentage suggests that skill profiles vary across learners and datasets, with no dominant, single profile of skills explaining performance across learners and datasets.

Further analysis across datasets highlighted a lack of consistency in the skill profiles. For example, the data from two different tests from the same participants (Datasets 4 and 5) produced different results. The data from the same tests with different participants also produced different results (within Datasets 2 and 3 and between Datasets 5 and 7 to 10). These results suggest that the skill profiles vary across learners and tests. The results may be explained by each test having different constructs and purposes (Deygers et al., 2018; Harsch, 2018).

Table 6 shows the results according to the lowest skills. For instance, the lowest skill group of LSW includes LSW<R and LSW<<R (not shown in Table 6; see Table 13 in Appendix). The results show that the lowest skill groups with 10% or more of learners across datasets were LSW in nine datasets (e.g., 25.20% in Dataset 1), S in eight datasets, SW in seven datasets, and LRSW (flat) in three datasets. The groups included S in the four-skill profiles and W in the three-skill profiles. The results of having lower productive skills than receptive skills were consistent with MEXT's 2018 study. The reasons for the lower levels of S and W were explored in Studies 2 and 3.

Table 7 displays the results of the groups with the highest skills. For example, the highest-skill group of R included LSW<R, LS<W<R, and L<S<W<R (the latter two not shown in Table 7; see Table 13 in Appendix). The results indicate that the groups with 10% or more across datasets were R in nine datasets (e.g., 33.42% in Dataset 1), LR in eight datasets, LRW in seven datasets, and LRSW (flat) in three datasets. The groups included R in the four-skill profiles and L in the three-skill profiles. The reasons for the higher R and L are explored below.

Some unique features were found for TOEIC in Dataset 6, where the first to third highest-level skills were W (33.96% in Table 7), LRW (20.75%), and LW (11.32%). This was reflected in a higher percentage of LRS as the lowest skill in TOEIC than in the other datasets (16.98% in Table 6). W had the highest scores among the skills, probably because the writing tasks in TOEIC were easier than those in other tests. The TOEIC W section has three task formats, namely, describing a picture (in one sentence, five items, in eight min), writing a reply to an email (length not specified, two items, in 20 min), and writing an argumentative essay (300 words, one item, in

Table 6 Skill profile types classified according to the *lowest* skills (%) in Study 1

Dataset		1	2	3	4	5	6	7	8	9	10
Lowest skill	**Profile type example**	**GTEC CBT: senior**[a]	**TOEFL Junior: Senior**[b]	**TOEFL Junior: Univ.**[c]	**TEAP: Univ.**[d]	**TOEFL iBT: Univ.**[d]	**TOEIC: Univ.**[e]	**TOEFL iBT: Univ. 2014**[f]	**TOEFL iBT: Univ. 2015**[g]	**TOEFL iBT: Univ. 2016**[h]	**TOEFL iBT: Univ. 2017**[i]
LSW	LSW<R	*25.20*	*13.11*	*13.25*	*12.00*	*18.00*	1.90	*10.23*	*16.67*	*15.34*	*21.03*
LRSW	Flat	*18.94*	*17.90*	9.40	*29.00*	9.00	9.43	3.15	1.59	9.95	6.52
SW	SW<LR	*14.51*	*17.15*	*18.38*	8.00	*20.00*	6.60	*18.12*	9.53	*16.88*	*15.94*
S	S<LW<R	9.65	*20.40*	*40.60*	*12.00*	3.00	*39.62*	*44.07*	*50.80*	*32.78*	*36.23*
W	W<S<L<R	8.97	5.79	7.69	6.00	*17.00*	0	0.79	0.79	0.74	0
L	L<W<RS	7.04	2.14	0.85	6.00	2.00	0	2.37	2.38	0.74	0
LS	LS<RW	6.03	5.72	3.85	6.00	1.00	6.60	14.96	11.9	10.74	13.05
LW	LW<S<<R	3.44	1.50	0	1.00	7.00	0	0	0	0	0
RSW	RSW<L	2.61	7.82	2.56	5.00	2.00	3.77	1.58	1.59	3.82	2.17
LRW	LRW<S	1.22	0.75	0	7.00	7.00	0	0	0.79	0	0
LRS	LRS<W	1.16	2.00	0	4.00	1.00	*16.98*	3.15	2.38	4.59	0.72
LR	LR<SW	0.50	0.51	0	1.00	3.00	0.95	0	0	0	0
RW	RW<LS	0.39	0.75	2.14	0	4.00	0.95	0	0.79	0	0
RS	RS<LW	0.28	2.82	0.85	2.00	1.00	9.43	1.58	0.79	2.28	1.45
R	R<LSW	0.06	1.64	0.43	1.00	5.00	3.77	0	0	2.14	2.89

Table 7 Skill profile types classified according to the *highest* skills (%) in Study 1

Dataset		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Highest skill	**Profile type example**	**GTEC CBT: senior**[a]	**TOEFL Junior: senior**[b]	**TOEFL Junior: Univ.**[c]	**TEAP: Univ.**[d]	**TOEFL iBT: Univ.**[d]	**TOEIC: Univ.**[e]	**TOEFL iBT: Univ. 2014**[f]	**TOEFL iBT: Univ. 2015**[g]	**TOEFL iBT: Univ. 2016**[h]	**TOEFL iBT: Univ. 2017**[i]
R	LSW<R	*33.42*	*19.29*	*32.05*	*15.00*	*23.00*	1.89	*28.34*	*28.58*	*20.73*	*34.79*
LRSW	Flat	*18.94*	*17.90*	9.40	*29.00*	9.00	9.43	3.15	1.59	9.95	6.52
LR	SW<LR	*13.40*	*16.97*	*18.38*	8.00	*16.00*	8.49	*16.54*	*13.50*	*17.65*	*16.67*
LRW	S<LRW	6.64	*12.75*	*16.67*	9.00	2.00	*20.75*	*21.26*	*19.05*	*15.07*	*11.60*
LRS	W<LRS	5.59	3.79	2.99	4.00	7.00	0	0.79	0.79	0.74	0
RSW	L<RSW	5.59	1.89	0.43	5.00	1.00	0	0	1.59	0	0
RW	LS<RW	5.53	5.57	4.70	6.00	1.00	4.72	*13.39*	*12.69*	*14.59*	*11.61*
L	RSW<L	3.61	*11.79*	9.83	5.00	9.00	5.66	6.29	7.94	6.89	7.96
RS	LW<RS	3.33	1.46	0	1.00	7.00	0	0	0	0	0
W	L<RS<W	1.50	2.61	2.14	6.00	1.00	*33.96*	8.66	8.73	7.65	6.51
S	LRW<S	1.28	0.86	0.43	8.00	*12.00*	0	0	0.79	0	0
SW	LR<SW	0.50	0.43	0	1.00	3.00	0.94	0	0	0	0
LS	RW<LS	0.39	0.75	2.14	1.00	2.00	0.94	0	0.79	0	0
LW	RS<LW	0.22	2.86	0.43	2.00	1.00	*11.32*	1.58	3.96	4.59	2.17
LSW	R<LSW	0.06	1.07	0.43	0	5.00	1.89	0	0	2.14	2.17

30 min). While the argumentative-essay task could be of appropriate difficulty for university students who are the target audience of the TOEIC test, the former two tasks (picture description and email response) may be relatively easy for them (see https://www.ets.org/toeic/test-takers/speaking-writing/about/content-format/). Such a difference in task difficulty may not be an issue, since the target language used in the domain of TOEIC includes global workplaces, a feature that differentiates TOEIC from other tests. Still, it would be useful to consider this difference when skill profiles are compared across TOEIC and other tests.

We also focused on how many levels of differences commonly existed between skills (e.g., two-level disparities) by classifying all the skill profiles according to the number of level differences. A one-level difference was found for SW<LR and LSW<R. A four-level difference was found in S<<W<L<R. Table 8 indicates that a one-level difference was most frequently observed, ranging from 55.11 (for Dataset 7) to 70.00% (for Dataset 5). This was followed by a two-level difference (7.00% for Dataset 4, to 40.16% for Dataset 7) and a zero-level difference (i.e., a flat profile; 1.60% for Dataset 8, to 29.00% for Dataset 4). A three-level difference (0.00% for Datasets 4 and 5, to 5.99% for Dataset 3) and a four-level difference (0.00% for Datasets 2 to 10, to 0.06% for Dataset 1) were rarely observed.

One reason for level differences may be the measurement error of tests, as it is caused by differences in test content, suboptimal test-taking environments, learners' personal issues, and others (e.g., Green, 2020). If the standard error of measurement (SEM) of a test or section is available, it is possible to calculate a range in which learners' true scores are included at the 95% probability when they repeatedly take the same test. The range can be computed using the score ± (1.96 × SEM). For example, assume that a learner received 16 points in the TOEFL iBT listening section (equivalent to the B1 level in the CEFR, following Table 3), the SEM of the listening section was reported to be 2.38 (Educational Testing Service, 2018, p. 7), and the learner's true score was estimated to fall between 11 and 21 by 95% chance ($16 \pm (1.96 \times 2.38) = 16 \pm 4.66 = 11.34, 20.66$; rounding them to the nearest integer, we get 11 and 21). According to Table 3, the learner's true CEFR level could fall on either the B1 or B2 level. These values suggest that four-skill and CEFR levels may fluctuate between plus or minus one level due to measurement error, which could lead to level differences in skill profiles. Therefore, it could be argued that a one-level difference is not very worrisome because it is within the margin of measurement error. In contrast, two- or three-level differences may indicate skill imbalances and/or other issues related to cognitive and/or affective factors (Goh & Vandergrift, 2021) and may need to be considered carefully.

Study 2: questionnaire study of learners with uneven skill profiles

Study 1 showed that uneven profiles were common, and that they varied in the mastery level of each skill. To examine the reasons for having uneven profiles (RQ2), Study 2 examined the test-takers' perceptions of profiles via a questionnaire, focusing on a group with clear uneven profiles in TOEFL iBT.

Table 8 Skill-level differences and percentages of participants in each category (%) in Study 1

Dataset		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Level difference	**Profile type example**	**GTEC CBT: senior**[a]	**TOEFL Junior: senior**[b]	**TOEFL Junior: Univ.**[c]	**TEAP: Univ.**[d]	**TOEFL iBT: Univ.**[d]	**TOEIC: Univ.**[e]	**TOEFL iBT: Univ. 2014**[f]	**TOEFL iBT: Univ. 2015**[g]	**TOEFL iBT: Univ. 2016**[g]	**TOEFL iBT: Univ. 2017**[h]
0	Flat	*18.94*	*17.90*	9.40	*29.00*	9.00	9.43	3.15	1.60	9.89	6.53
1	SW<LR; LSW<R	*68.53*	*66.74*	*52.56*	*64.00*	*70.00*	*65.09*	*55.11*	*56.34*	*64.89*	*57.98*
2	S<LW<R; S<W<LR	*12.01*	*14.93*	*32.05*	7.00	*21.00*	*23.58*	*40.16*	*38.89*	*22.24*	*34.03*
3	S<W<L<R; W<S<L<R	0.46	0.43	5.99	0	0	1.90	1.58	3.17	2.98	1.46
4	S<<W<L<R	0.06	0	0	0	0	0	0	0	0	0
Total		100	100	100	100	100	100	100	100	100	100

Table 9 The gap group's perceptions toward their skill scores in Study 2

	n	Q1 (Any sections in which you received lower scores?)		Q2 (In what section did you receive lower scores?)			
		Yes (*n*)	**No**	**L**	**R**	**S**	**W**
Overall							
2014	53	67.92 (36)	18.87	50.00[a]	5.56	55.56	19.46
2015	53	83.02 (44)	13.21	38.64	6.82	70.45	2.27
2016	33	84.85 (29)	12.12	44.83	3.45	55.18	6.90
2017	49	73.47 (38)	16.33	44.74	10.52	68.42	7.89
High prof.							
2014	14	50.00 (7)	14.29	28.57	28.57	57.14	28.57
2015	16	93.75 (15)	6.25	0	0	100.00	6.67
2016	8	100.00 (8)	0	12.50	0	87.50	12.50
2017	14	85.71 (12)	14.29	0	25.00	83.33	25.00
Low prof.							
2014	39	74.36 (29)	20.51	55.17	0	55.17	17.24
2015	37	78.38 (29)	16.22	58.62	10.34	55.17	0
2016	25	84.00 (21)	16.00	57.14	4.76	42.86	4.76
2017	32	74.29 (26)	17.14	65.38	3.85	61.54	0

Percentages of participants in the gap group (i.e., those with two- or more-level differences in mastery of skills). High prof. and low prof., higher- and lower-proficiency gap groups. This also applies to Table 10. [a] 18/36 (the number of those who selected L divided by the number of those who answered "yes" in question 1). Yes and no responses in Q1 do not add up to 100% due to missing responses. In Q2, learners were able to select more than one skill

Method for Study 2

The participant groups were the same as those in Datasets 7 to 10 from Study 1. They were the 2014 to 2017 cohorts at a medical university. Study 2 used the data of the skill scores used in Study 1 as well as the questionnaire collected for Study 2.

We first selected students with uneven skill profiles, that is, those with two- or more-level differences across skills ($n = 33$ to 53; see Table 9), which were based on their skill profiles derived from their TOEFL iBT scores. Hereafter, these students are referred to as the gap group. Students with a one-level difference in skill profiles were not selected because a one-level difference could be due to measurement errors (see Study 1).

The gap group was further divided into a higher-proficiency gap group (achieving the B2 or higher level judged from their total scores) and a lower-proficiency gap group (achieving the B1 or lower level judged from their total scores) in order to investigate the possible effects of L2 proficiency on perceived reasons for uneven skill profiles. By comparing gap groups of different levels of proficiency, we intended to examine why some learners performed better in one skill than in others.

To examine the reasons behind the uneven profiles (RQ2), we developed a questionnaire (see Additional file 1) which was conducted for 4 years after the participants took the TOEFL iBT and received their score reports each year. The responses were analyzed separately for each year. As our study was descriptive, we did not use statistical significance testing.

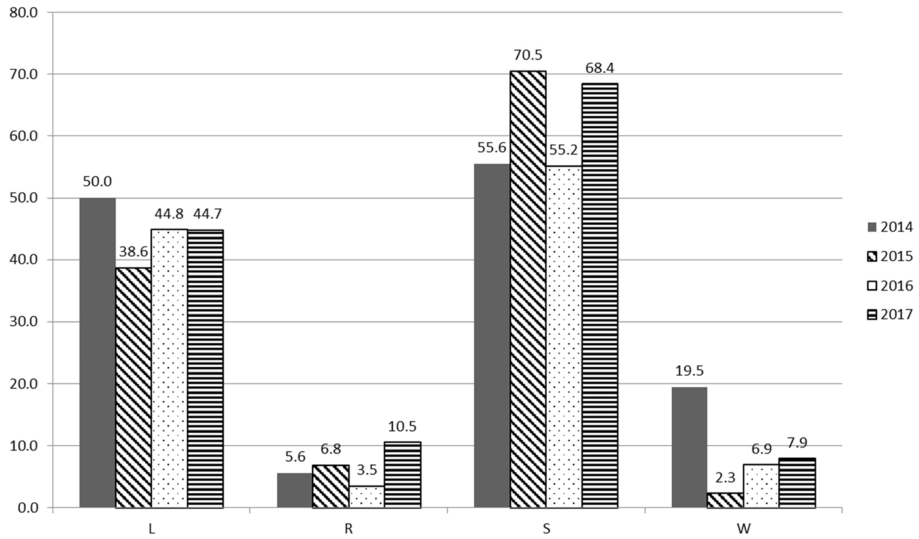

Fig. 2 Skills that those in the gap group thought received lower scores in TOEFL iBT

Results and discussion for Study 2

Table 9 shows whether learners in the gap group—a selected group of learners with two- or three-level differences in the mastery of skills—thought that there were sections with lower skill scores than other sections (question 1). Overall, more than two-thirds of the learners reported that their scores were uneven across various skills (e.g., 67.92% in 2014 to 84.85% in 2016). As seen in Table 9 and Fig. 2, among such learners, the largest number of learners reported receiving lower scores for S (e.g., 55.56% in 2014). The second-largest number reported receiving lower scores for L (e.g., 50.00% in 2014). When the gap group was subdivided into the higher- and lower-proficiency groups, different patterns were observed. The high-proficiency gap group reported receiving lower scores in S the most and in L, R, and/or W the second most (L, R, and W in 2014; W in 2015; L and W in 2016; and R and W in 2017). The lower-proficiency gap group reported receiving lower scores in L the most (e.g., 57.14% in 2016) and in S the second most (e.g., 42.86% in 2016) or L and S at the same percentage (e.g., 55.17% in 2014). These results suggest that, according to learners' perceptions, S is their weak skill regardless of the level of their proficiency; L was frequently the weakest skill among lower-proficiency learners relative to reading and writing skills.

Table 10 and Fig. 3 outline the percentages of learners in the gap group who thought they had sections with lower scores (i.e., responded "yes" to question 1) and who answered question 3 (why do you think the scores in these sections were lower? Select all the reasons that apply to your case). The results show that most learners selected options 6 (I was poor at this skill; up to 68.18% in 2015) and 7 (I have not studied the skill much; up to 52.78% in 2014). They indicate that learners with wide skill gaps felt that they received lower scores in certain skills due to their low skills and/or lack of a sufficient amount of time to study. Options 1 (I did not know the iBT test format well; up to 24.32% in 2017) and 2 (I was tense or confused; up to 36.36% in 2015) were also selected consistently.

Table 10 Test-takers' views toward reasons of having uneven profiles in Study 2

Option	1	2	3	4	5	6	7	8	9
	I didn't know formats	**I was tense**	**I was tired**	**I was sick**	**I was hungry**	**I was poor at the skill**	**I have not studied the skill much**	**I don't know why**	**Others**
Overall									
2014	16.67[a]	22.22	16.67	5.56	0	*63.89*	52.78	8.33	5.56
2015	11.36	36.36	4.55	9.09	0	*68.18*	50.00	0.00	4.55
2016	24.14	31.03	24.14	10.34	10.34	*62.07*	24.14	3.45	13.79
2017	24.32	29.73	35.14	5.41	2.70	*62.16*	43.24	2.70	5.41
High prof.									
2014	42.86	*71.43*	14.29	0	0	28.57	57.14	0	14.29
2015	13.33	53.33	6.67	6.67	0	*66.67*	33.33	0	0
2016	25.00	25.00	25.00	0	25.00	*75.00*	25.00	0	0
2017	16.67	41.67	16.67	8.33	8.33	*50.00*	41.67	8.33	0
Low prof.									
2014	10.34	10.34	17.24	6.90	0	*72.41*	51.72	10.34	3.45
2015	10.34	27.59	3.45	10.34	0	*68.97*	58.62	0	0
2016	23.81	33.33	23.81	14.29	4.76	*57.14*	23.81	4.76	0
2017	26.92	23.08	42.31	3.85	0	*69.23*	42.31	0	0

Percentages of those who selected one option in Q3. (Why do you think the scores in these sections were lower? Select all the reasons.) [a]6/36 (the number of those who selected this option divided by the number of those who answered "yes" to question 1). Examples of option 9 (other reasons): because the listening section was long ($n = 1$ in 2014, lower-proficiency group), because I was 1-h late for the test ($n = 1$ in 2014, higher-proficiency group), due to the lack of study ($n = 1$ in 2015, lower-proficiency group), because of the lack of effort ($n = 1$ in 2016, lower-proficiency group), because I was poor at typing ($n = 1$ in 2016, higher-proficiency group), because I was sleepy ($n = 1$ in 2016, lower-proficiency group), because I fell in sleep during the test ($n = 1$ in 2017, lower-proficiency group), because I had not been exposed to English for a while because of the summer vacation ($n = 1$ in 2016, higher-proficiency group), and because the listening and speaking sections were much more difficult than other sections ($n = 1$ in 2017, lower-proficiency group)

As with the overall group, similar response patterns were observed for the lower-proficiency gap group. For example, most learners selected options 6 (I was poor at this skill; up to 72.41% in 2014) and 7 (I have not studied the skill much; up to 51.72% in 2014). Options 1 (I did not know the iBT test format well; up to 26.92% in 2017) and 2 (I was tense or confused; up to 33.33% in 2016) were also selected consistently.

The response patterns were slightly different among the higher-proficiency gap group. Options 1 (I didn't know the iBT test format well; up to 42.86% in 2014) and 2 (I was tense or confused; up to 71.43% in 2014) were selected by a large percentage of the learners in this group along with options 6 (I was poor at this skill; up to 75.00% in 2016) and 7 (I had not studied the skill much; up to 57.14% in 2014).

These results indicate that learners with uneven skill profiles attribute their lower scores to their low skills and/or lack of a sufficient amount of time to study. This is likely observed in overall-, higher-, and lower-proficiency learners. In other words, most learners with skill gaps see skill scores as a reflection of their ability and/or effort. High-proficiency learners also see them as a reflection of their knowledge about the test format and their own emotional conditions.

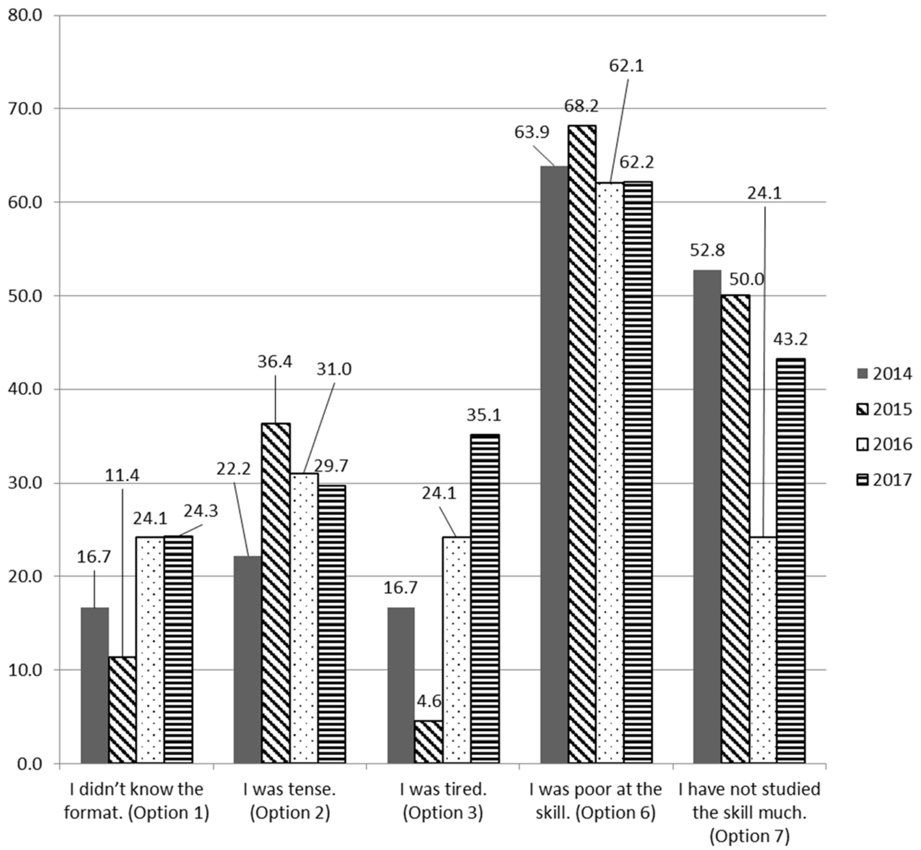

Fig. 3 Major reasons why those in the gap group thought received lower scores in TOEFL iBT

Study 3: interview with learners exhibiting uneven skill profiles

Based on quantitative responses to a questionnaire conducted on learners with uneven skill profiles, Study 2 has reported on learners' perceptions of the reasons for having such profiles. Study 3 further investigated RQ2 by qualitatively analyzing the interview data from learners exhibiting uneven skill profiles.

Method for Study 3

Participants

We selected six participants from the gap group in Study 2 for Dataset 7 (TOEFL iBT in 2014). They exhibited two- or three-level differences in the mastery of skills in the CEFR levels, with wide gaps across skills. Table 11 summarizes the information of the six students. The first four achieved the B1 level in the CEFR as judged by their actual (not self-reported) total scores in the TOEFL iBT. The other two achieved the B2 level. All took the TOEFL iBT between August and November 2014, when they were first-year university students. They were interviewed between July and August 2015, when they were second-year university students, and the gratuity was paid after the interviews.

Table 11 Six participants' characteristics in Study 3

Pseudonym	Gender	CEFR levels	iBT skill profile	Q2: Lower skills	Q3: Reasons
Subaru	Male	L = A2, R = B2, S = A2, W = B1, total = B1	LS<W<R (2-level difference)	L	I don't know why
Takeru	Male	L = C1, R = B1, S = A2, W = A2, total = B1	SW<L<<R (3-level difference)	L, S, W	I was tense I was tired
Gaku	Male	L = B1, R = A2, S = A2, W = B2, total = B1	LS<R<W (2-level difference)	S	I don't know why
Arata	Male	L = B1, R = B2, S = A2, W = B2, total = B1	S<L<RW (2-level difference)	S	I didn't know formats
Yamato	Male	L = C1, R = C1, S = B1, W = B2, total = B2	S<W<LR (2-level difference)	S, W	I was tense
Hina	Female	L = C1, R = B2, S = A2, W = B2, total = B2	S<<RW<L (3-level difference)	S	I didn't know formats I was tense I was tired

Names were anonymized. All responded "Yes, there were sections in which you received lower scores than other sections." (Question 1 in the questionnaire). Q2, in what section did you receive lower scores? Q3, why do you think the scores in these sections were lower?

Interview procedures

The interview was conducted in a semi-structured, face-to-face format in which an interviewer (one of the authors) and an interviewee (i.e., learner) talked in a quiet room. The interviewer obtained informed consent from the learner and recorded their responses using a voice recorder, along with a video when necessary.

The interview consisted of four phases (see Fig. 4). First, learners received a copy of the questionnaire administered in Study 2, with their responses filled in. They were asked to double-check the responses and revise them as needed. Second, they were asked the following questions: (a) How have you studied English from the time you started until your first year at university? (b) Why do you think that you have lower skills, and how have you studied them? (c) What particular aspects of the skills do you think are the most difficult to study? The interviewer asked questions to clarify what the learners meant in their responses in order to elicit more details.

Third, learners took a TOEFL iBT practice test (Educational Testing Service, 2012). Those who had lower scores in the L section in the practice test also took a

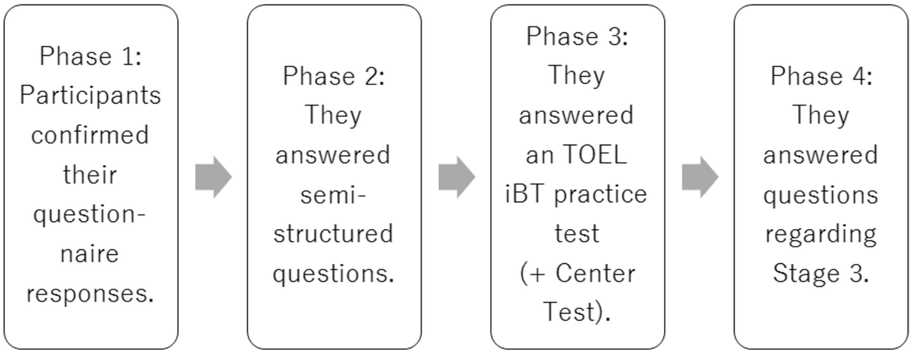
Phase 1: Participants confirmed their questionnaire responses.
Phase 2: They answered semi-structured questions.
Phase 3: They answered an TOEL iBT practice test (+ Center Test).
Phase 4: They answered questions regarding Stage 3.

Fig. 4 Interview procedures in Study 3

listening section in the National Center Test for University Admissions (Center Test, hereafter; see National Center for University Entrance Examinations, 2017; Watanabe, 2013) to further highlight their weak areas. The Center Test is a high-stakes university entrance examination used by almost all Japanese universities, including the participants. The Center Test administered in January 2015 was used in the current study as this version was administered nationwide after the current participants entered university. No participants were expected to have taken it prior to data collection. As learners took the tests, they verbally explained their thought processes. The interviewer recorded their behaviors and responses using video and voice recorders. She also observed and took notes to describe how the learners responded to these tests.

Fourth, learners were asked about distinctive test-taking behaviors noticed by the interviewer and were asked to explain why they behaved in such a particular manner. Additionally, when necessary, they were asked to explain how they arrived at the answers to questions that they had answered incorrectly in the test.

Analysis

The interview responses were transcribed verbatim. One of the authors extracted points related to skills and skill profiles from the transcript, summarized them into a table, and described the summary in bulleted sentences. This was repeated while examining all of the data, based on a thematic analysis method (Nowell et al., 2017). Two other authors double-checked the summary to examine if all relevant points were included while excluding irrelevant ones. Two of the authors then described them in a passage.

Results and discussion for Study 3

By analyzing the results of the interviews, we have identified three main factors that explained the learners' uneven skill profiles with two- or three-level CEFR-level differences across skills: insufficient practice, particular subskills or processes required to accomplish test tasks, and unfamiliarity with test formats and test-taking environments.

Insufficient practice

Speaking (S) All the interviewees indicated that they spent less time on learning skills on which they scored lower than on the ones on which they scored higher. This tendency was particularly notable for S, which all six participants perceived as (one of) the English skills that yielded exceptionally poor results. Subaru, Takeru, Gaku, and Hina reported that they had insufficient S training in English. Subaru, Yamato, and Hina pointed out that proficiency in S is not usually required to enter a university in Japan. This could have caused students and teachers at senior high schools and cram schools to allot little time to improving their speaking skills. Indeed, Hina said that the EFL classes at her senior high school did not cover S. Arata also indicated that he had few opportunities to discuss daily issues or express opinions in English in EFL classes at his senior high school.

Unlike Hina and Arata, Yamato went to a comprehensive (or combined) junior and senior high school that emphasized improving all four skills related to students' English ability. Therefore, he had many opportunities to improve his S and other skills. However, as the university entrance examination season neared, English classes focused more on preparing students using past examination papers. As a result, toward the end of his senior high school year, Yamato felt that the opportunities to practice S were severely limited.

The participants' responses illustrate that the learners and their schools did not see much need or allocate much time to practicing S in English due to the lack of an S section in university entrance examinations. This finding demonstrates that university entrance examinations affect which English skill(s) senior high school students and teachers focus on. In other words, an insufficient amount of time to practice in speaking English was potentially caused by the negative impacts of entrance examinations on teaching and learning (see Green, 2020; Tsushima, 2011). This seemed to be a factor that led to uneven skill profiles.

Listening (L) Subaru and Gaku perceived their L skills to be the lowest among the four skills, whereas Yamato and Hina perceived them to be the highest. The interview results suggest that this individual difference may have been caused by whether learners spent a long time improving their L skills to pass an advanced level of an English proficiency test, which was more challenging than the Center Test. It should be noted that, except for the Center Test, it is not common for Japanese universities to require applicants to demonstrate their listening skills via tests or certificates.

Both Subaru and Gaku explicitly reported that they practiced L to perform just sufficiently to achieve the level required by the Center Test. As they did not see the need to improve their L beyond that level, they did not allocate the time to do so. Thus, similar to the case of S, Subaru's and Gaku's responses indicate that university entrance examinations influence senior high school students' decisions on which English skill(s) to focus on.

On the other hand, Yamato, Arata, and Hina worked hard on their L to reach a higher level than that required by the Center Test. It seems that they had different learning goals than Subaru and Gaku. For example, Yamato studied to pass the EIKEN Grade Pre-1 (equivalent to the B2 level in the CEFR; https://www.eiken.or.jp/eiken/en/grades/), which is aimed primarily at Japanese university students. Yamato said that preparing for the test enhanced his L. Hina also took the EIKEN test when she was in senior high school. She took grade 2 (equivalent to the B1 level), which is aimed at Japanese senior high school graduates. Finally, Arata's senior high school required students to take the GTEC for STUDENTS (a paper-based version of GTEC CBT). Although L was not covered in regular EFL classes at his senior high school, he spent time improving his L to pass the test, along with studying R and W.

Reading (R) Subaru, Takeru, Yamato, and Arata received higher scores in R (relative to other skills) and perceived their R to be the strongest among the four skills. All six participants said that they worked hard to improve their R skills as they were measured during university entrance examinations. Subaru articulated the importance of R to get into a university as follows: "For university entrance examinations, we can confidently say you need the R skill only. Other than that, you sometimes need the W skill."

The analysis of interview results repeatedly showed that the contents of university entrance examinations affected students' and schools' choices regarding which English skill(s) to focus on. This may have led to an imbalance in the development of the four skills. Although the participants took English courses in their first year at the university, they mentioned them little. They intensively studied English for university entrance examinations to attend a medical school, and that experience may have had a stronger impact on them than their first-year courses at the university.

Particular subskills or processes required to accomplish test tasks

To further investigate learners' imbalanced English skills, they were asked about the skill(s) that were two or three levels lower (S, L, and/or W) than their highest-ranked skill(s). Specifically, they were asked to explain why and in what ways they had difficulty answering questions correctly in sections that required them to use their lower-ranked skill(s) of the TOEFL iBT (see Xi & Norris, 2021, for the test constructs).

Speaking (S) All six participants had lower S scores than the highest-ranked skill(s) by at least two levels. Takeru and Gaku said that they were not used to speaking extensively in English. This skill is required in the TOEFL iBT S section, wherein 45 s is allotted to independent tasks and 60 s to integrated tasks. Gaku said that he was not used to producing a long monologue in English. Similarly, Takeru said that although he was capable of sustaining a dialogue by speaking one or two sentences during his turn, he could not make a series of sentences and speak for a longer time. Takeru further explained that he was not able to produce an impromptu monologue because he did not know how to organize his logic to construct a speech. Such difficulties may have stemmed from the EFL education he received at secondary and perhaps tertiary levels. According to a survey conducted among junior and senior high school teachers and principals by the Benesse Educational Research and Development Institute (2016), secondary school EFL classrooms in Japan often or sometimes implement pair or group conversations in which students talk in English (79.9% and 46.4% at junior and senior high schools, respectively). Impromptu speeches, in which students talk about themselves or express their opinions, are not as common (42.7% and 29.4%, respectively). At the university where all six participants attended, some instructors had given students opportunities to practice constructing and producing impromptu speeches. However, more could be done to encourage teachers to shift their attention to oral activities.

Gaku, Arata, and Hina mentioned translating Japanese (their first language) into L2 English when speaking English. Hina told the interviewer that she would structure her speech in Japanese first and then translate it into English during situations that provided her with preparation time and required her to speak extensively (i.e., iBT speech or classroom presentation; she would not use a translation method in L, R, or W). She would use this S-translation strategy because she had not studied S enough. Gaku reported that he could not even transform his thoughts easily into Japanese before trying to speak in English. He seemed to be aware that directly translating Japanese into English would not create sensible outcomes. He seemed to think that he should reorganize his thoughts in plainer Japanese to facilitate the translation process and produce a more accurate English outcome. Arata also mentioned the translation, saying that even if he could speak about the topic in Japanese, it was difficult to translate his ideas into English.

One possible reason why some learners used Japanese when developing English speech is the way that English is taught in secondary schools in Japan. According to the 2016 Benesse Educational Research and Development Institute survey, 68.3 to 68.8% of Japanese secondary school teachers of English either often or sometimes use translation (English to Japanese) exercises in classes (pp. 4–5). Most of them (89.4 to 96.1%) often or sometimes provide grammar explanations. As learners have become familiar with the translation method but have had little experience planning and delivering impromptu speeches, it may be natural for them to use Japanese in the process of producing English sentences. Arata also stated that he was not good at speaking even in Japanese in that he could not speak simply and tended to structure his thoughts in a complex manner with much time spent when possible. This suggests that lower S scores can be attributed not only to L2 skills but also to L1 (i.e., Japanese) proficiency.

Listening (L) Subaru, Takeru, and Gaku stated that when they did not understand part of what was said, they tried to derive the missing information and overall meaning of the speech from the partial information they were able to capture and their background knowledge. However, such an approach was not always successful for long passages or conversations, such as in the TOEFL iBT L section. Subaru said that he could catch parts of words, but he could not always put them together to understand the whole text. Takeru said that sometimes he could understand a part of a speech and would try to guess the whole story based on what he could. However, as he revealed, such an L strategy often resulted in an overly imagined story. He also mentioned that his concentration usually lasted briefly, as he would start to think of something else when he lost interest in what he was listening to.

Another common point among the three learners was that as senior high school students, they practiced L to pass the Center Test but stopped practicing after they reached the passing level. As speeches used for the Center Test are generally shorter than those used in TOEFL iBT, it is not surprising that the learners found the latter's listening section challenging.

Writing (W) Takeru and Yamato seemed to have their own explanation for the difficulties presented by W in English. Yamato said that he attempted to write perfectly from the beginning, by writing a complete introduction, and then completing the body and the conclusion. It is not difficult to imagine that if he flexibly started from the body or stopped writing the introduction, and skipped to a different section when he could not come up with "perfect" sentences for the introduction, he could have completed his essays more strategically. He mentioned that similar inflexibility could be observed in his daily life, and that his friends tended to say that he lacked flexibility. Based on his responses, the reason why Yamato struggles with W may be connected to his approach when he writes in English, as well as when he performs tasks both related and unrelated to language.

Takeru could not write a summary of the passage well. He revealed that because of his poor listening skills, he rarely used a listening text, primarily summarizing a reading text in an R-L-and-then-W task. Furthermore, he stated that he could not perform the same R-L-and-then-W task in Japanese. Therefore, Takeru's lower W score may be related not only to his English skills but also to his W and L skills in Japanese.

Unfamiliarity with test formats and test-taking environments

Some participants were not familiar with the formats for the S, L, and/or W sections, which made it difficult for them to correctly respond to questions.

Speaking (S) All participants except for Subaru found that the integrated tasks were difficult, partly because they were not familiar with the format. They seemed to be aware of such a format, but they did not practice enough to be comfortable with it. Takeru said he practiced the L-and-then-S type of integrated tasks but did not practice the R-L-and-then-S task before he took the actual TOEFL iBT. Hina said that she took the test without fully understanding the format of the S section. Regarding the integrated tasks, three students (Takeru, Gaku, and Arata) said that they had trouble speaking because they did not understand the L and/or R inputs that they were required to orally summarize and/or give their opinions.

Yamato and Hina revealed that they were not used to accomplishing tasks, such as preparing a structured speech in a short time (i.e., 15 or 30 s). They made similar comments when explaining the difficulties they experienced while answering questions in the S section.

The physical test-taking environments in the test room and technical issues may have also affected the participants' S performance and scores. Arata said he felt nervous with other examinees sitting near him, close enough for them to hear him. Yamato also hesitated to articulate his speeches because doing so could disturb the examinees sitting close to him. Gaku could not concentrate well because a person nearby was not wearing a headset and spoke loudly. In terms of technical issues, Takeru's microphone could not pick up his voice during the microphone check. It turned out that his voice was too soft for the microphone to detect; however, the "failure" of the microphone check made him nervous during the S section.

Listening (L) Subaru and Gaku pointed out that the formats of the TOEFL iBT L section were dissimilar from those of the Center Test, which made it difficult to score higher in the former. Both Subaru and Gaku said that listening was much faster and longer on the iBT than on the Center Test. Another challenging factor was that, in the iBT, answer choices were not shown until the end of the speech. In the Center Test and the TOEFL Institutional Testing Program (ITP), both of which the participants had taken, answer choices were printed on a test book, so they could read them before the speech started to get an idea of what it would be about. Gaku said that, during the iBT, speeches are read only once, whereas during the Center Test, they are read twice. He explained that he would often translate the English text into Japanese when listening to (and even reading) English texts; thus, he could not catch up with the fast, long speeches while translating in the iBT L section. In contrast, he was able to catch up with L texts and managed to answer questions in the Center Test due to the repetition of the speech (see He & Jiang, 2020; Pusey, 2020, for previous studies).

The order of the skill sections may also have slightly affected the performance. Takeru mentioned that in an R-and-then-L integrated task, he was not able to concentrate well on L and used too many cognitive resources to complete the R section.

Writing (W) Only one point was raised regarding the test formats used in the W section. Takeru could not type fast, which slowed down his answering process in the W section. Although Japanese students are familiar with texting on smartphones using an on-screen keyboard, they are typically less familiar with conventional physical keyboards. At the university, they wrote essays using a laptop. However, some learners like Takeru may have needed more typing practice. Additionally, for TOEFL iBT, a standard English-language (QWERTY) computer keyboard was used (https://www.etsglobal.org/hu/en/test-type-family/toefl-ibt-test), which has a slightly different layout from that of a Japanese keyboard. Thus, typing speed and keyboard layout could have affected the learners' W performance (see Ling, 2017, for previous studies).

Conclusion

The current study examined characteristics of the four-skill profiles that are frequently observed across datasets (RQ1 in Study 1 [via a quantitative approach]) and the learners' perceived reasons for having uneven profiles (RQ2 in Studies 2 and 3 [via a quantitative and qualitative approach]) among Japanese English learners. The findings are summarized in Table 12. Study 1 used 10 datasets from five standardized four-skill tests to search for general patterns in profiles. We found 75 various types of skill profiles, and learners with uneven profiles were more common (over 30%) than those with a flat profile. Skill profiles frequently observed across databases included LSW<R (in eight datasets), SW<LR (in seven datasets), S<LRW (in seven datasets), and flat (in four datasets). A one-level difference across skills accounted

Table 12 Joint display of the current study's results

Study 1 Quantitative (RQ1)	Study 2 Quantitative (RQ2)	Study 3 Qualitative (RQ2)	Mixed-methods meta-inference
Flat profile (1.59 to 29.00%)	--	--	--
Uneven profiles (71.00 to 98.41%) Frequently observed uneven profiles: LSW<R, SW<LR, S<LRW	--	--	--
1-level difference (55.11 to 70.00%)	--	--	--
2-level difference (7.00 to 40.16%) 3- or more-level difference (0.00 to 5.99%)	Typical reasons for having uneven profiles (1) I was poor at this skill (62.07 to 63.89%) (2) I had not studied the skill much (24.14 to 52.78%) (3) I didn't know formats (11.36 to 24.32%) (4) I was tense (22.22 to 36.36%) or tired (4.55 to 35.14%)	Emergent themes from the interview (a) Insufficient practice (b) Particular subskills or processes required to accomplish test tasks (c) Unfamiliarity with test formats and test-taking environments A more detailed environments and contexts were noted	Convergent Quantitative and qualitative data matched well (1) and (2) = (a) and (b) (3) = (c)

for the highest percentage (up to 70.00%), followed by a two-level difference (up to 40.16%) and a flat profile (up to 29.00%).

Based on a quantitative approach using a questionnaire, Study 2 examined why some learners had uneven profiles (RQ2). Those with substantially uneven skill profiles reported that they were poor at the skill (e.g., up to 68.18%); they had not studied the skill sufficiently (up to 52.78%), and they did not know the test formats (up to 24.32%) or were tense (up to 36.36%).

Study 3 further examined RQ2 by qualitatively analyzing responses from interviews with six learners exhibiting uneven skill profiles. Their responses revealed that their skill imbalances can be explained by three main factors: (a) insufficient practice, (b) particular subskills or processes required to accomplish test tasks, and (c) unfamiliarity with test formats and test-taking environments. There were also underlying factors behind these factors, such as the impact of entrance examinations on teaching and learning, and the difficulty in understanding a long listening text played only once. Study 3's qualitative findings supported Study 2's quantitative findings, because (a), (b), and (c) as found in Study 3 were three of the four key reasons reported in Study 2.

The current study reinforced the importance of considering uneven profiles of L2 proficiency, as suggested by previous studies (e.g., North, 2021; Hulstijn, 2015). However, we found far more uneven profiles (i.e., 75 profiles in Study 1) than those reported in previous studies (e.g., Ginther & Yan, 2018, showing three profiles) as we computed the percentage of each skill profile to examine detailed skill profiles. We also reported more detailed explanations of learners' self-perceived reasons for having uneven profiles (in Studies 2 and 3, e.g., insufficient practice) than previous studies that suggested that extensive test preparation is only one of the factors that produce them (e.g., Ginther & Yan, 2018).

In future studies, in addition to expanding learner groups and tests to examine the generalizability of the current findings, we suggest examining how skill profiles affect

L2 performance in general, academic, and other specific contexts. In particular, learners with uneven profiles of SW<LR and SW<<LR should be examined along with their L2 test or real-life performance. These two-skill profiles were identified frequently in the current study and were considered problematic by Ginther and Yan (2018), and Bridgeman et al. (2016) as learners with these profiles performed more poorly than expected in academic contexts. It remains to be examined whether and how those learners could have compensated for their weak skills by effectively using their strong skills (Harsch, 2014; Hulstijn, 2015). There might be minimally required degrees of skill or an optimal balance of four skills for learners to perform effectively. These insights would help explore how students with extremely uneven profiles can be taught and how this information can be integrated into score reports.

As for the implications of our study, we argue that effective use of skill profiles in score reports would help test developers, admission and placement officers, learners, teachers, and other test users understand learners' strengths and weaknesses. Test developers can utilize these results to consider how to present skill profiles and convey the interpretations in score reports and supplementary materials provided to test-takers and users to enhance the usefulness of their tests. For example, they can flag two- or more-level differences between skills on the score reports, so that test users can consider them when making decisions. Those who are involved in admission and placement can also incorporate the use of both total and skill scores, especially scores of important skills (Ginther & Yan, 2018), or pass on the skill profile information and recommendations to teachers and administrators who may plan remedial instructions.

When test-takers are shocked to see their uneven profiles in the score reports, we suggest that a skill imbalance, specifically a skill difference of one CEFR level, is not surprising. For those who had two- or more-level differences, we suggest that they fully utilize the score report to understand their strengths and weaknesses, and that they review possible factors affecting the uneven profiles, such as insufficient practice, particular subskills or processes required to accomplish test tasks, and unfamiliarity with test formats and test-taking environments. To fully demonstrate L2 skills and avoid obtaining uneven skill profiles due to the lack of test format knowledge, it may be necessary to do prior familiarization practice with the test format, for example, by watching videos (e.g., https://www.youtube.com/user/TOEFLtv). Furthermore, when the uneven profiles come from true skill imbalance, learners and teachers may need to decide whether to focus on weaker or stronger skills for further study based on their contexts and needs. While one direction is to improve a weaker skill, a stronger skill could be further improved to compensate for the weaker one. Teachers can also make informed decisions on how to provide students with useful feedback, remedial instructions for lower skills, and training for strengthening higher skills, according to the learner's needs. These activities are in line with the global trends in learning-oriented assessment (Gebril, 2021) by relating assessment results with instruction and learning.

Appendix

Table 13

Table 13 All skill profile types that appeared in Study 1 (%)

Dataset	1	2	3	4	5	6	7	8	9	10		
Pattern	**GTEC CBT: Senior**[a]	**TOEFL Junior: Senior**[b]	**TOEFL Junior: Univ.**[c]	**TEAP:Univ.**[d]	**TOEFL iBT: Unit.**[d]	**TOEIC: Univ.**[e]	**TOEFL iBT: Univ. 2014**[f]	**TOEFL iBT: Univ. 2015**[g]	**TOEFL iBT: Univ. 2016**[h]	**TOEFL iBT: Univ. 2017**[i]	**Lowest skill**	**Highest skill**
Flat	*18.94*	*17.90*	9.40	*29.00*	9.00	9.43	3.15	1.59	*10.00*	6.52	LRSW	LRSW
L<RS<W	0.28	0.04	0	0	0	0	0.79	0	0.77	0	L	W
L<<RSW	0.06	0	0	0	0	0	0	0	0	0	L	RSW
L<RSW	5.53	1.89	0.43	5.00	1.00	0	0	1.59	0	0	L	RSW
L<RW<S	0.06	0	0	1.00	1.00	0	0	0	0	0	L	S
L<S<RW	0	0	0	0	0	0	0.79	0	0	0	L	RW
L<SW<R	1.05	0.21	0.43	0	0	0	0.79	0.79	0	0	L	R
LR<S<W	0	0.04	0	0	0	0	0	0	0	0	LR	W
LR<SW	0.50	0.43	0	1.00	3.00	0.94	0	0	0	0	LR	SW
LR<W<S	0	0.04	0	0	0	0	0	0	0	0	LR	S
LRS<<W	0	0	0	0	0	1.89	0	0	0	0	LRS	W
LRS<W	1.16	2.00	0	4.00	1.00	*15.09*	3.15	2.38	4.62	0.72	LRS	W
LRW<S	1.22	0.75	0	7.00	7.00	0	0	0.79	0	0	LRW	S
LS<<RW	0	0	0	0	0	0	0	0.79	0.79	0	LS	RW
LS<R<W	0	0.04	0	0	0	1.89	1.57	1.59	0	0.72	LS	W
LS<RW	5.42	5.22	3.42	6.00	1.00	4.72	9.45	7.14	*10.00*	8.70	LS	RW
LS<W<R	0.61	0.46	0.43	0	0	0	3.94	2.38	0	3.62	LS	R
LSW<<R	0.78	0.46	1.71	0	1.00	0	2.36	0.79	1.54	1.45	LSW	R
LSW<R	*24.42*	*12.65*	*11.54*	*12.00*	*17.00*	1.89	7.87	*15.87*	*13.85*	*19.57*	LSW	R
LW<R<S	0	0.04	0	0	0	0	0	0	0	0	LW	S
LW<RS	3.16	1.43	0	1.00	7.00	0	0	0	0	0	LW	RS
L<W<RS	0.06	0	0	0	0	0	0	0	0	0	L	RS
LW<S<<R	0.06	0	0	0	0	0	0	0	0	0	LW	R
LW<S<R	0.22	0.04	0	0	0	0	0	0	0	0	LW	R
R<<LSW	0	0.04	0	0	0	0	0	0	0	0	R	LSW
R<<S<LW	0	0.04	0	0	0	0	0	0	0	0	R	LW
R<LS<W	0	0.07	0	1.00	0	1.89	0	0	0	0	R	W

Table 13 (continued)

Dataset Pattern	1 GTEC CBT: Senior[a]	2 TOEFL Junior: Senior[b]	3 TOEFL Junior: Univ.[c]	4 TEAP:Univ.[d]	5 TOEFL iBT: Unit.[d]	6 TOEIC: Univ.[e]	7 TOEFL iBT: Univ. 2014[f]	8 TOEFL iBT: Univ. 2015[g]	9 TOEFL iBT: Univ. 2016[h]	10 TOEFL iBT: Univ. 2017[i]	Lowest skill	Highest skill
R<LSW	0.06	1.04	0.43	0	5.00	1.89	0	0	2.17	2.17	R	LSW
R<S<LW	0	0.11	0	0	0	0	0	0	0	0	R	LW
R<SW<L	0	0.36	0	0	0	0	0	0	0	0.72	R	L
RS<<LW	0	0.04	0	0	0	0	0	0	0	0	RS	LW
RS<L<W	0	0.04	0.85	0	0	0.94	0	0	0	0	RS	W
RS<LW	0.22	2.54	0	2.00	1.00	8.49	0.79	0.79	2.31	1.45	RS	LW
RS<W<L	0.06	0.21	0	0	0	0	0.79	0	0	0	RS	L
RSW<<L	0.06	0.14	0	0	0	0	0.79	0	0	0	RSW	L
RSW<L	2.55	7.68	2.56	5.00	2.00	3.77	0.79	1.59	3.85	2.17	RSW	L
RW<LS	0.39	0.68	2.14	0	2.00	0.94	0	0.79	0	0	RW	LS
RW<S<L	0	0.07	0	0	2.00	0	0	0	0	0	RW	L
S<<L<RW	0	0	0	0	0	0	0	0.79	0	0	S	W
S<<LR<W	0	0	0	0	0	0.94	0	0	0	0	S	W
S<<LRW	0	0.36	0.85	0	0	0	0.79	2.38	1.54	0.72	S	LRW
S<<LW<R	0	0.04	0.43	0	0	0	0	1.59	0	0	S	R
S<<RW<L	0	0	0	0	0	0	0.79	0	0	1.45	S	L
S<<W<LR	0	0.04	0	0	0	0	0	0	1.54	0	S	LR
S<L<RW	0.11	0.36	1.28	0	0	0	3.15	4.76	3.85	2.90	S	RW
S<L<W<R	0	0	0.43	0	0	0	0	0	0	0	S	R
S<LR<W	0.06	0.39	1.28	1.00	0	10.38	3.15	3.97	2.31	5.07	S	W
S<<LRW	0.06	0	0	0	0	0	0	0	0	0	S	LRW
S<LRW	6.58	*12.40*	*15.81*	9.00	2.00	*20.75*	*20.47*	*16.67*	*13.58*	*10.87*	S	LRW
S<LW<<R	0.06	0.04	0.85	0	0	0	0	0	0.77	0	S	R
S<LW<R	2.16	2.64	7.69	2.00	0	0	8.66	5.56	3.08	7.97	S	R
S<R<L<W	0	0	0	0	0	0.94	0	0	0	0	S	W
S<R<LW	0	0.14	0.43	0	0	2.83	0.79	3.17	2.31	0.72	S	LW

Table 13 (continued)

Dataset	1	2	3	4	5	6	7	8	9	10		
Pattern	**GTEC CBT: Senior**^{a}	**TOEFL Junior: Senior**^{b}	**TOEFL Junior: Univ.**^{c}	**TEAP:Univ.**^{d}	**TOEFL iBT: Unit.**^{d}	**TOEIC: Univ.**^{e}	**TOEFL iBT: Univ. 2014**^{f}	**TOEFL iBT: Univ. 2015**^{g}	**TOEFL iBT: Univ. 2016**^{h}	**TOEFL iBT: Univ. 2017**^{i}	**Lowest skill**	**Highest skill**
S<RW<L	0.06	1.89	4.27	0	0	1.89	2.34	5.56	2.31	2.90	S	L
S<<W<L<R	0.06	0	0	0	0	0	0	0	0	0	S	R
S<W<L<R	0.22	0.14	2.99	0	0	0	0	0	0	0	S	R
S<W<<LR	0	0	0	0	0	0	0	0	0.77	0	S	LR
S<W<LR	0.28	1.96	3.85	0	0	1.89	3.94	5.56	0.77	3.62	S	LR
S<W<R<L	0	0	0.43	0	0	0	0	0.79	0	0	S	L
SW<<L<R	0	0.04	0	0	0	0	0	0	0	0	SW	R
SW<<LR	0.61	0.25	0.85	0	0	0	0.79	0	0.77	0.72	SW	LR
SW<LR	*11.79*	*14.29*	*13.25*	8.00	*14.00*	6.60	*11.81*	7.94	*13.85*	*12.32*	SW	LR
SW<L<<R	0.06	0	0	0	0	0	0.79	0	0	0	SW	R
SW<L<R	1.72	1.71	2.99	0	2.00	0	3.94	1.59	1.54	2.17	SW	R
SW<R<L	0.33	0.86	1.28	0	4.00	0	0.79	0	0.77	0.72	SW	L
W<<LRS	0.06	0.04	0	0	0	0	0	0	0	0	W	LRS
W<L<RS	0.11	0.04	0	0	0	0	0	0	0	0	W	RS
W<LR<S	0	0.04	0.43	0	4.00	0	0	0	0	0	W	S
W<LRS	5.53	3.75	2.99	4.00	7.00	0	0.79	0.79	0.77	0	W	LRS
W<LS<<R	0.06	0	0	0	0	0	0	0	0	0	W	R
W<LS<R	1.94	0.75	1.71	1.00	3.00	0	0	0	0	0	W	R
W<R<LS	0	0.07	0	1.00	0	0	0	0	0	0	W	LS
W<RS<L	0.55	0.57	1.28	0	1.00	0	0	0	0	0	W	L
W<S<L<R	0	0.11	0.85	0	0	0	0	0	0	0	W	R
W<S<LR	0.72	0.43	0.43	0	2.00	0	0	0	0	0	W	LR
Total	100	100	100	100	100	100	100	100	100	100	100	100

Senior, senior high school student data. Univ., university student data. $^a n = 1805$. $^b n = 2799$. $^c n = 234$. $^d n = 100$. $^e n = 106$. $^f n = 127$. $^g n = 126$. $^h n = 130$. $^i n = 138$. (^{f} to ^{i} total $N = 521$). 10% or more were *italicized*

The percentages do not always add up to 100.00% due to rounding

Abbreviations

CBT	Computer-based testing
CEFR	Common European Framework of Reference for Languages
Center Test	National Center Test for University Admissions
EFL	English as a foreign language
GPA	Grade point average
GRE	Graduate record examination
GTEC	Global Test of English Communication
iBT	Internet-based test
L	Listening
MEXT	Japan's Ministry of Education, Culture, Sports, Science and Technology
R	Reading
RQ	Research question
S	Speaking
SEM	Standard error of measurement
TEAP	Test of English for Academic Purposes
TOEFL	Test of English as a Foreign Language®
TOEFL Junior	TOEFL Junior® Comprehensive
TOEIC	Test of English for International Communication®
W	Writing

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40468-022-00203-3.

Additional file 1. Questionnaire regarding the TOEFL iBT (originally in Japanese; translated into English for presentation purposes).

Acknowledgements
We would like to thank the people and institutions that supported us in Study 1: Yuko Kashimada, Takuya Akimoto, and Yumiko Tashiro for Dataset 1; the Global Communication & Testing for Datasets 2 and 3; Keita Nakamura and Eiken Foundation of Japan for Datasets 4 and 5; Shin'ichiro Ishikawa, Mineo Mitsuhashi, Institute for International Business Communication, and Yuka Ishikawa for Dataset 7; and Ryoko Fujita and Yumi Koyamada for Dataset 8.

Authors' contributions

RK and YI planned Study 1 and collected the datasets. The first three authors planned Studies 2 and 3 and collected and transcribed the data. All the authors analyzed the data. RK, TA, and YI wrote the initial draft, which was extensively revised, and the final manuscript was approved by all the authors.

Funding
This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant-in-Aid for Scientific Research (C), Grant Numbers 17K02983 and 26370737, and the Juntendo University General Education Joint Research Grants 2015 and 2016.

Availability of data and materials

The datasets for the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 7 March 2022 Accepted: 1 November 2022

References

Benesse Educational Research and Development Institute. (2016). *Chuko no eigo shidou ni kansuru jittai chosa 2015.* [Field survey on English education at junior and senior high schools, 2015]. https://berd.benesse.jp/up_images/research/Eigo_Shido_all.pdf.

Bridgeman, B., Cho, Y., & DiPietro, S. (2016). Predicting grades from an English language assessment: The importance of peeling the onion. *Language Testing*, *33*(3), 307–318. https://doi.org/10.1177/0265532215583066.

Center for Entrance Examination Standardization. (2017). *GTEC sukoa to CEFR reberu kanrenduke chosa hokoku* [Standard setting study of relating GTEC scores with CEFR levels]. https://www.benesse.co.jp/gtec/schoolofficials/research/.

Choi, I. (2017). Empirical profiles of academic oral English proficiency from an international teaching assistant screening test. *Language Testing*, *34*(1), 49–82. https://doi.org/10.1177/0265532215601881.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.

Council of Europe (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR): A manual*. Language Policy Division. https://www.coe.int/en/web/common-european-framework-reference-languages/relating-examinations-to-the-cefr.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment: Companion volume*. https://www.coe.int/en/web/common-european-framework-reference-languages/home

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed., International student ed.). Sage.

Deygers, B., Van Gorp, K., & Demeester, T. (2018). The B2 level and the dream of a common standard. *Language Assessment Quarterly*, *15*(1), 44–58. https://doi.org/10.1080/15434303.2017.1421955.

Dunn, K., & Iwaniec, J. (2021). Exploring the relationship between second language learning motivation and proficiency: A latent profiling approach. *Studies in Second Language Acquisition*. Advance online publication. https://doi.org/10.1017/S0272263121000759.

Educational Testing Service (2012). *Official guide to the TOEFL test with CD-ROM* (4th ed.). McGraw-Hill.

Educational Testing Service (2018). *Reliability and comparability of TOEFL iBT® scores*, *TOEFL® research insight series* (vol. 3) https://www.ets.org/toefl/research/insight-series/.

Educational Testing Service. (2019). *Mapping the TOEIC® tests on the CEFR*. https://www.ets.org/s/toeic/pdf/toeic-cefr-flyer.pdf.

Flaherty, B. P., & Kiff, C. J. (2012). Latent class and latent profile models. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology: Vol. 3: Data analysis and research publication*, (pp. 391–404). American Psychological Association.

Gebril, A. (Ed.) (2021). *Learning-oriented language assessment: Putting theory into practice*. Routledge.

Ginther, A., & Yan, X. (2018). Interpreting the relationships between TOEFL iBT scores and GPA: Language proficiency, policy, and profiles. *Language Testing*, *35*(2), 271–295. https://doi.org/10.1177/0265532217704010.

Goh, C. C. M., & Vandergrift, L. (2021). *Teaching and learning second language listening: Metacognition in action* (2nd ed.). Routledge.

Green, A. (2020). *Exploring language assessment and testing: Language in action* (2nd ed.). Routledge.

Harsch, C. (2014). General language proficiency revisited: Current and future issues. *Language Assessment Quarterly*, *11*(2), 152–169. https://doi.org/10.1080/15434303.2014.902059.

Harsch, C. (2018). How suitable is the CEFR for setting university entrance standards? *Language Assessment Quarterly*, *15*(1), 102–108. https://doi.org/10.1080/15434303.2017.1420793.

Harsch, C., Ushioda, E., & Ladroue, C. (2017). Investigating the predictive validity of TOEFL iBT® test scores and their use in informing policy in a United Kingdom university setting. *ETS Research Report Series*, *1*, 1–80. https://doi.org/10.1002/ets2.12167.

He, L., & Jiang, Z. (2020). Assessing second language listening over the past twenty years: A review within the socio-cognitive framework. *Frontiers in Psychology: Language Sciences*, *11*(Article 2123), 1–15. https://doi.org/10.3389/fpsyg.2020.02123.

Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research*. John Benjamins.

Hulstijn, J. H., Schoonen, R., de Jong, N. H., Steinel, M. P., & Florijn, A. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR). *Language Testing*, *29*(2), 203–221. https://doi.org/10.1177/0265532211419826.

In'nami, Y., & Koizumi, R. (2021). *Factor structure and four-skill profiles of the Aptis test* (ARAGs research reports, AR-G/2021/2). British Council. https://www.britishcouncil.org/factor-structure-and-four-skill-profiles-aptis.

In'nami, Y., Koizumi, R., & Nakamura, K. (2016, August 20–21). *Nihonjin eigo gakushusha no yongino reberu no zure no tokucho* [Characteristics of four-skill imbalances of Japanese learners of English: Cases of TEAP and TOEFL iBT; Paper presentation]. 42nd JASELE Annual Conference, Dokkyo University, Saitama, Japan.

Koizumi, R. (2015a). Factor structure and four-skill profiles of the TOEIC® tests among Japanese university learners of English. *ARELE, 26*, 109–124. https://doi.org/10.20581/arele.26.0_109.

Koizumi, R. (2015b, March 21–24). *Gaps among four skills measured by the TOEIC and TOEIC SW Tests: Investigating learner groups and predictor variables* [Paper presentation] Joint American Association for Applied Linguistics and Association Canadienne de Linguistique Appliquée 2015 Conference, Fairmont Royal York, Toronto, Ontario, Canada.

Koizumi, R., Agawa, T., & Asano, K. (2018, March 24–27). *Deriving useful information on skill imbalance from TOEF iBT® scores* [Paper presentation] American Association for Applied Linguistics 2018 Conference, Sheraton Chicago Hotel and Towers, Illinois, U.S.A.

Koizumi, R., & In'nami, Y. (2017, August 19–20). *Nihonjin eigo gakushusha no yongino reberu no zure no tokuchou* [Characteristics of four-skill imbalances of Japanese learners of English: Case of TOEFL Junior® Comprehensive; Paper presentation]. 43rd JASELE Annual Conference, Shimane University, Japan.

Koizumi, R., Kashimada, Y., & Akimoto, T. (2019, August 17–18). *Nihonjin eigo gakushusha no yongino reberu no zure no tokucho: GTEC CBT taipu no baai* [Characteristics of four-skill profiles of Japanese learners of English: Case of GTEC CBT; Paper presentation] 45th Japan Society of English Language Education (JASELE) Conference, Hirosaki University, Aomori, Japan.

Li, M., & Zhang, X. (2021). A meta-analysis of self-assessment and language performance in language testing and assessment. *Language Testing*, *38*(2), 189–218. https://doi.org/10.1177/0265532220932481.

Liao, C.-W., Qu, Y., & Morgan, R. (2010). *The relationships of test scores measured by the TOEIC® Listening and Reading test and TOEIC® Speaking and Writing tests (Compendium study)*. Educational Testing Service. https://www.ets.org/Media/Research/pdf/TC-10-13.pdf.

Ling, G. (2017). Are TOEFL iBT® writing test scores related to keyboard type? A survey of keyboard-related practices at testing centers. *Assessing Writing*, *31*, 1–12. https://doi.org/10.1016/j.asw.2016.04.001.

Ma, W., & Winke, P. (2022). An investigation of the impact of jagged profile on L2 speaking test ratings: Evidence from rating and eye-tracking data. *Language Assessment Quarterly*, *19*(4), 394–421. https://doi.org/10.1080/15434303.2022.2078720.

Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in psychology (Vol. 2: Statistical analysis)* (pp. 551–611). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199934898.013.0025.

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2018). *Heisei 29 nendo eigo kyoiku kaizen notameno eigoryoku chosa jigyo hokoku* [2017 survey on English ability to improve English education]. https://www.mext.go.jp/a_menu/kokusai/gaikokugo/1403470.htm.

National Center for University Entrance Examinations (2017). *National Center for University Entrance Examinations*. https://www.dnc.ac.jp/albums/abm00033004.pdf.

North, B. (2021). *An introduction to the theme 'Uneven profiles'* [Language Assessment for Migrants' Integration (LAMI) Workshop]. https://www.youtube.com/watch?v=oh9oryJv4fk&t=1s.

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, *16*(1), 1–13. https://doi.org/10.1177/1609406917733847.

Oberski, D. (2016). Mixture models: Latent profile and latent class analysis. In J. Robertson, & M. Kaptein (Eds.), *Modern statistical methods for HCI* (pp. 275–287). Springer. https://doi.org/10.1007/978-3-319-26633-6_12.

Pang, F., & Skehan, P. (2021). Performance profiles on second language speaking tasks. *The Modern Language Journal*, *105*(1), 371–390. https://doi.org/10.1111/modl.12699.

Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). The association between TOEFL iBT test scores and the Common European Framework of Reference (CEFR) levels. *Research Memorandum*, *ETS RM-15-06*. https://www.ets.org/Media/Research/pdf/RM-15-06.pdf.

Powers, D. E. (2013). Assessing English-language proficiency in all four language domains: Is it really necessary? In D. E. Powers (Ed.), *The research foundation for the TOEIC tests: A compendium of studies*, (vol. II, pp. 1.1–1.7). Educational Testing Service. https://www.ets.org/research/policy_research_reports/publications/chapter/2013/jroa.

Pusey, K. (2020). Assessing L2 listening at a Japanese university: Effects of input type and response format. *Language Education & Assessment*, *3*(1), 13–35. https://doi.org/10.29140/lea.v3n1.193.

Riazi, A. M. (2017). *Mixed methods research in language teaching and learning*. Equinox.

Roohr, K., Olivera-Aguilar, M., Bochenek, J., & Belur, V. (2022). *Exploring GRE® and TOEFL® score profiles of international students intending to pursue a graduate degree in the United States*, *ETS research report series*. Advance online publication. https://doi.org/10.1002/ets2.12343.

Sawaki, Y., & Sinharay, S. (2013). Investigating the value of section scores for the TOEFL iBT® test. *ETS Research Report Series*, *2*, i–113. https://doi.org/10.1002/j.2333-8504.2013.tb02342.x.

Sawaki, Y., & Sinharay, S. (2018). Do the TOEFL iBT® section scores provide value-added information to stakeholders? *Language Testing*, *35*(4), 529–556. https://doi.org/10.1177/0265532217716731.

Tannenbaum, R. J., & Baron, P. A. (2015). Mapping scores from the TOEFL Junior® Comprehensive Test onto the Common European Framework of Reference (CEFR). *Research Memorandum*, *ETS RM-15-13*. https://www.ets.org/Media/Research/pdf/RM-15-13.pdf.

Tsushima, R. (2011). *The mismatch between educational policy and classroom practice: EFL teachers' perspective on washback in Japan*. Master's thesis, McGill University https://escholarship.mcgill.ca/concern/theses/44558j191.

Vahed, S. T. (2021). *The use of language proficiency test scores in graduate admissions.* Ph.D. dissertation, Purdue University. https://hammer.purdue.edu/articles/thesis/THE_USE_OF_LANGUAGE_PROFICIENCY_TEST_SCORES_IN_GRADUATE_ADMISSIONS/15054471.

Watanabe, Y. (2013). The National Center Test for University Admissions. *Language Testing*, *30*(4), 565–573. https://doi.org/10.1177/0265532213483095.

Xi, X., & Norris, J. (Eds.) (2021). *Assessing academic English for higher education admissions*. Routledge.

Publisher's Note