

SHIKEN

A Journal of Language Testing and Evaluation in Japan

Volume 26 • Number 2 • December 2022

<https://doi.org/10.37546/JALTSIG.TEVAL26.2>

Contents

1. A content analysis of prefectural senior high school entrance exams: What aspects of English do they assess and how?

Peter O'Keefe and David Allen

<https://doi.org/10.37546/JALTSIG.TEVAL26.2-1>

19. Test-taker self-assessment accuracy using the TOEIC speaking and writing can-do statements

Jonathan Harrison and Ruth Vanbaelen

<https://doi.org/10.37546/JALTSIG.TEVAL26.2-2>



Testing and Evaluation SIG

ISSN 1881-5537

Shiken: A Journal of Language Testing and Evaluation in Japan

Volume 26 No. 2
December 2022

<https://doi.org/10.37546/JALTSIG.TEVAL26.2>

Editor

David Allen
Ochanomizu University

Reviewers

(see editorial board, plus additional reviewers)

Website Editor

Peter O' Keefe
Fujikawa Board of Education

Editorial Board

David Allen
Ochanomizu University

Nat Carney
Kobe College

Trevor Holster
Fukuoka University

Jeff Hubbell
Hosei University

J. W. Lake
Fukuoka Jo Gakuin University

Edward Schaefer
Ochanomizu University

James Sick
Temple University, Japan Campus

A content analysis of prefectural senior high school entrance exams: What aspects of English do they assess and how?

Peter O’Keefe¹ and David Allen²

okeefepeter03@gmail.com

1. Fujikawa Board of Education

2. Ochanomizu University

<https://doi.org/10.37546/JALTSIG.TEVAL26.2-1>

Abstract

Research into high-stakes English tests in Japan has typically focused on entrance examinations to university while few studies have investigated those for high school. In this study, a content analysis was conducted of 24 prefectural English entrance examinations for public high schools administered in 2022. We focused on the language skills and subskills targeted in the exams by referring to the socio-cognitive framework (Weir, 2005), and examined the item types and response formats, and readability of the reading passages. Key findings revealed that in terms of item distribution, the exams focus primarily on reading (62.38%), followed by listening (29.16%), and finally writing (8.46%). In terms of subskills, the exams tended to target comprehension of main ideas at the text level rather than comprehension of specific details or discrete linguistic knowledge at the sentence level. Overall, the findings highlight the convergence between the exam content and the goals of the MEXT course of study. However, the dominant focus on receptive skills remains, which presents a potential barrier to achieving positive washback from these high-stakes exams.

Keywords: Content analysis, English entrance exams, Japanese senior high schools, prefectural entrance exams

Foreign languages form a core component of mainstream education in Japan, and English, which is the most commonly taught foreign language, typically features on entrance exams to senior high schools. These exams are high stakes because entrance to a prestigious high school greatly increases the chances of entry to a prestigious university, which in turn increases the chances of gaining employment and climbing the socio-economic ladder (Rohlen, 1983). Consequently, the number of students attending cram schools (*juku*) for academic subjects reaches its peak in the third year of junior high school as students prepare for the exams (Benness, 2017).

The senior high school entrance exams are intended to assess prior learning achievement and thus are expected to reflect the content of the course of study provided by the Ministry of Education, Culture, Sports, Science, and Technology (MEXT). The revised course of study, which was designed with reference to the Common European Framework of Reference for Languages (CEFR) and implemented in junior high schools in 2021, aims for learners to achieve up to an A2 level of proficiency in English by the end of junior high school (MEXT, 2016)¹. The stated core goals at this level are “(1) To acquire the skills to understand foreign language sounds, vocabulary, expressions, grammar, and linguistic functions, and to utilize this knowledge in actual communication through listening, reading, speaking, and writing; (2) to understand and express simple information and ideas in a foreign language about everyday topics and social topics according to the purpose, scene, and situation of communication. To cultivate the ability to communicate; and (3) to deepen understanding of the culture behind foreign languages, and to cultivate an attitude of proactively trying to communicate using the foreign language while paying attention to listeners, readers, speakers, and writers” (MEXT, 2017, pp. 12-14).

Although several empirical studies have focused on the content of university entrance exams and their (mis-)alignment with the course of study (e.g., Brown & Yamashita, 1995; Fukazawa, 2021; Kikuchi, 2006; Law, 2004; Watanabe, 1997), very few have considered the content of high school entrance exams. This lack of research is concerning because these exams are high stakes and thus likely impact language teaching and learning. The present study, therefore, sought to provide an overview of the content of senior high school entrance exams and to highlight the ways in which they overlap or deviate from the aims of the course of study.

Literature review

As far as we are aware, only a handful of studies have focused on the content of senior high school entrance exams. For instance, Akiyama (2003) discussed the prefectural high school exam in Tokyo, showing that it targets the skills of reading (56%), listening (20%), writing (12%) and indirect speaking (12%). Akiyama’s focus was the assessment of speaking and his survey of junior high school teachers revealed that approximately 80% of them believed a direct speaking component would have a positive impact on their teaching (p. 134). They believed that the entrance exam should adequately reflect the

time and effort that is spent on *all* of the four skills, as stipulated in the course of study, in particularly speaking, which is the only skill that was not assessed directly.

In a more recent study, Minato (2020) analyzed three private senior high school entrance exams in Hokkaido and evaluated to what extent they assess communicative competence. His findings revealed considerable variation in the exams. For example, only one included a listening test while another used solely selective responses which highlights the disparity and lack of standardization in these tests. Most importantly, in terms of the assessment of communicative competence, Minato found the tests to be deficient because many items focused on linguistic knowledge (grammar, phonology, lexis) that often did not require an understanding of the text to answer them correctly, that is, they assess linguistic knowledge discretely and independently of context. Such items included items targeting knowledge of word order, grammatical knowledge, and lexical stress placement. The relative lack of items that assess reading and listening comprehension more broadly is criticized, as is the lack of writing and speaking assessment.

University entrance exams in Japan have also been criticized for assessing certain aspects of knowledge and particular skills at the expense of others. For instance, Brown and Yamashita (1995), Fukazawa (2021), Kikuchi (2006) and Watanabe (1997) all show how reading is the primary focus of the exams, translation ability is commonly assessed, listening is not commonly assessed, and speaking is almost never tested directly. Consequently, senior high school teachers often describe conflict between their desire to teach communicatively in accordance with the course of study and the perceived demands of entrance exams (e.g., Bailey, 2018; Green, 2014; Sakui, 2004; Takagi, 2009; Underwood, 2012). Moreover, learners also comment on the importance of university exams and how their English study is influenced by the content of these exams (e.g., Allen & Nagatomo, 2019; Green, 2014; Takagi, 2009).

In contrast to the lack of academic research into high school exams, there is an abundance of test preparation materials and services available from cram schools and educational publishers. For instance, *Saikō suijun mondaishū: Kōkō nyūshi: Eigo* [Top-level question collection: High school entrance exams: English] published by Buneidō in 2021 aims to prepare learners for high-level entrance exams at both public and private senior high schools. The textbook contains 19 sections: 10 cover grammatical knowledge, two cover lexical and phonological knowledge, three cover reading skills, and three cover composition, listening and indirect speaking skills (i.e., conversational phrases and dialogue reading). Each section includes practice items, many of which are drawn from previously administered high school exams. By analyzing the textbook content, readers may thus conclude that grammatical knowledge is the primary focus of assessment, and that reading is the most important skill.

Without an analysis of the actual tests, however, any conclusions on exam content would be premature. Moreover, while cram schools and educational organizations regularly publish their own test analyses, which may indeed be of use for research, these analyses often fall short in a number of areas. Most notably, because they target the mass market (i.e., test takers, parents, and teachers), the tests are not analyzed from the perspective of contemporary research. In particular, it is not clear to what extent the exams assess the different component subskills specified in modern theoretical models of language use and assessment. To address this issue, we refer to the socio-cognitive model of test validity in our analysis of exam content.

A subskills approach

The socio-cognitive framework developed by Cyril Weir and colleagues (Chalhoub-Deville & O'Sullivan, 2020; O'Sullivan & Weir, 2011; Weir, 2005) was developed to provide researchers with a means to design and evaluate language tests. In its initial iteration, the framework included five aspects of assessment validity: theory-based validity (later referred to as cognitive validity), context validity, scoring validity, consequential validity, and criterion-related validity.

The first of these components, cognitive validity, concerns the psycholinguistic processes that underlie language use in each of the four skills and at different levels of language proficiency. Researchers have further developed theoretical models that explain the underlying processes involved in reading (Khalifa & Weir, 2009), writing (Shaw & Weir, 2007), listening (Geranpayeh & Taylor, 2013) and speaking assessments (Taylor, 2011). Each of these models outlines the relevant subskills and considers the tasks that can elicit them. By referring to these studies, we analyze senior high school exams in terms of the subskills that are assessed.

We must note at the outset, however, that determining which subskills are actually being employed by test-takers when they complete test tasks is a challenge fraught with difficulty (Khalifa & Weir, 2009, p. 40). Empirical investigations are necessary to shed light on the actual processes involved and anything less will inevitably include educated guesswork.

Nevertheless, predicting language learners' behavior when they face test tasks is the first step towards understanding that behavior and towards validating a language assessment. With this in mind, the definitions of the subskills for reading, listening and writing are presented below. Speaking subskills are not listed because the exams do not directly assess speaking ability.

Reading

The skill of reading is divided into two subskills of careful reading and expeditious reading, each of which can be conducted at the local (word, sentence) level or global (paragraph, text) level.

Careful Reading: Local. This involves understanding the propositional meaning at clause and sentence level and is typically performed in a linear manner following the writer's intended path. The reader is thus concerned with the detail at the level of clause and sentence. It is used in all basic reading tasks, such as those that require the test taker to identify key factual information from a sentence or very short text. It is also the primary way in which people begin to learn to read and is typically employed when learning sentence structure.

Careful Reading: Global. This involves reading the text typically in a linear manner, from beginning to end. The concern is with the macro-propositional meaning, or rather, main idea(s) of the text. That is, "The reader accepts the writer's organization of the text and the reader attempts to build up a macro structure on the basis of the majority of information received" (Khalifa & Weir, 2009, p. 60). The aim is to gain a more-or-less complete understanding of the text, which may then be followed by activities that assess and/or build upon that understanding.

Expeditious Reading: Local. This requires the reader to quickly locate specific information on a predetermined topic. It involves the subskills of scanning and searching for information. While scanning is defined very narrowly as searching for a specific word or phrase (i.e., an exact match), search reading is more broadly defined as searching for words and phrases that are semantically related to the target, including synonyms and related words. In language assessment tasks, test takers are often required to quickly search for words and phrases in the text that relate to key words in the question stems and/or response choices, thereby locating the relevant part of the text. Careful reading would typically be employed thereafter to determine whether a response choice is correct or not.

Expeditious Reading: Global. This subskill requires the ability to process large amounts of information quickly and extract key information. It is typically referred to as skim-reading or skimming, and involves reading more quickly and selectively than careful global reading. In educational contexts, this may involve skimming through a long text to gain a general understanding (i.e., the gist) of it. It may also be used to identify which sections of the text are important for a particular purpose (e.g., for citing in an essay or for answering a comprehension question). As this subskill of reading is usually associated with higher level abilities, it is not expected in the high school exams. In fact, Khalifa and Weir (2009) concluded that skim reading was not assessed in any of the Cambridge ESOL Main Suite reading papers (p. 61), suggesting that it is rarely tested even at higher levels of language proficiency.

Listening

There are four main subskills for listening in the socio-cognitive framework: listening for gist, listening for detailed/specific information, listening for main idea/important information/key message, and listening to infer opinion.

Listening for gist. This subskill of listening is defined by Elliot and Wilson (2013) as "the ability to recognize the main point of a text despite limited knowledge of syntax and vocabulary" (p. 176). Thus, listening for gist may be used at lower levels of proficiency. However, it may also be used at higher levels to test for the overall comprehension of the main ideas of a discourse as opposed to the detail. This subskill is not expected in the high school exam because it typically requires processing a large amount of text.

Listening for detailed/specific information. This subskill requires test takers to locate specific, concrete information (Elliot & Wilson, 2013, p. 179). Examples of such information may include how old a person is or where an object is. The focus is on small details usually confined within one sentence. It is typically associated with comprehension of part of an utterance, rather than the entire utterance. A common format for this type of listening is the short response gap-fill item where test takers must listen carefully to identify a word or phrase that precedes or follows a key word printed on the question sheet.

Listening for main idea / important information / key message. This subskill requires identifying the main topic of a story or conversation. This will usually become apparent through a sequence of related words that appear throughout a text. By

recognizing this important information, it is possible to understand the main idea or key message intended. This subskill typically requires comprehension of most of, or the whole, utterance. It is likely to be assessed regularly in the listening sections of high school exams, which often feature short dialogues, for which test takers are required to listen carefully to the entire passage in order to select the correct response and eliminate distractor choices.

Listening to infer opinion. This subskill requires listeners to listen for multiple clues, such as prosodic and phraseological clues, and make use of prior experiential knowledge rather than direct statements to gain an understanding of the intended meaning of a conversation/story. As the ability to perform this subskill develops at higher levels of proficiency, it is unlikely to be assessed in high school exams.

Writing

Shaw and Weir (2007) identify six cognitive processes involved in writing: macroplanning, organization, microplanning, translation, monitoring, and revising (pp. 38-39). It is expected that because most of the writing tasks encountered in the entrance tests for public high school are simple and limited in length, these tasks will not involve all six of these cognitive processes. In particular, the higher-level processes of macroplanning and organization are unlikely to be involved in most tasks. On the other hand, it is expected that all the remaining processes (microplanning, translation, monitoring and revising) are invoked to some extent. That is, while writing tasks for the high school exams may only involve writing a single sentence, or even part of a sentence, these basic processes are essentially required. Of course, the only way to determine whether they are or not invoked in actual learner behavior is to conduct empirical investigation into learner processes (e.g., using eye-tracking and/or stimulated-recall tasks), which is beyond the scope of this study. Consequently, we only consider writing subskill components to the extent that the above predictions are evaluated, that is, macroplanning and organization subskills are rarely employed, while the remaining subskills are always employed.

Test and item characteristics

The cognitive processes invoked during language tasks will typically be determined by the characteristics and demands of the task. For instance, if an item only involves re-ordering elements within a sentence, it will only require careful reading at the local level. In Weir's (2005) socio-cognitive framework, task demands are considered under the heading of context validity, and the interdependence between cognitive and context validity is emphasized. Context validity covers the conditions of test administration (e.g., security conditions), the task setting (e.g., response format, weighting, and time constraints), and task demands which include linguistic demands (e.g., discourse mode, text length, and textual characteristics) and interlocutor demands (e.g., speech rate and accent). Given our aims, we document those features that are most commonly discussed, which allows for comparison with other content analyses. More specifically, we consider item response format (e.g., selective/constructive responses) and textual characteristics (e.g., readability) as part of our content analysis.

The present study

The aim of the present study is to investigate the content, including the subskills that are assumed to be assessed, and format of the English section of high school entrance exams in Japan. In 2020, there were 4884 senior high schools, of which 3564 were public/national and 1320 were private (MEXT, 2020). Here, we focus on the exams that are created by and administered in each of the 47 prefectures for selection purposes at public schools in those prefectures. Although prefectural exam boards are more strictly bound to MEXT guidelines than are private high schools, there is still some degree of autonomy in the design and administration of exams. Therefore, the analysis will reveal general trends and any prefectural idiosyncrasies. In addition to skills and subskills, we analyze the item types and response formats, and the complexity of written texts, which allows us to compare the exam content with that previously reported for other English exams in Japan. Following the analyses, we discuss the extent to which the exams are in line with the content of the course of study.

The following research questions were thus framed:

1. What English skills and subskills are being assessed in the exams?
2. How do the exams vary in terms of test and item characteristics?
3. How complex are the reading passages of the exams?

Method

Materials

Twenty-four exams for entrance into Japanese public high schools administered in 2022 were selected for the analysis. They were sourced from Toshin.com (a cram school website where an incomplete but quite large sample of the exams from around the nation is freely available in PDF format) and *Zenkoku kōkō nyushi* (2022) which provides all of the prefectural tests, answers and scoring information, as well as listening scripts/broadcasts. The sample aimed to provide a broad geographical representation of Japan's 47 prefectures, from Hokkaido to Okinawa. A greater number of prefectural exams was chosen from the Kanto area to reflect the dense population in the area surrounding Tokyo.

Table 1

Prefectural English exams included in the analysis

Hokkaido	Ibaraki	Yamanashi	Osaka**
Aomori	Gunma	Nagano	Hiroshima
Miyagi	Saitama	Gifu	Ehime
Akita	Chiba	Aichi*	Nagasaki
Yamagata	Tokyo	Mie	Kagoshima
Fukushima	Kanagawa	Kyoto	Okinawa

*Note that Aichi prefecture has two tests, A and B, of identical format administered on different days. The test known as Aichi-A was chosen for inclusion in this analysis. ** Osaka has three tests of different ability level A, B and C. Test B was chosen because it represents the regular level observed across all prefectures, whereas Test A was low and Test C was high (i.e., for students who had experienced an extended period of life abroad).

Procedure

Items were first coded for skill (reading, listening, writing) and then by subskill (listed previously), according to the type of cognitive processing expected for successful completion of the item. In addition to subskills, we looked at item response format, which was coded as selective or constructive, then more specifically. Selective items were coded as multiple choice (MC), MC visual (when options were pictures or other visual information such as graphs), MC gap fill or gapped summary, matching, re-ordering a sentence, and re-ordering a paragraph/sequence of events. Constructive items were coded as gap fill, gapped summary, transformation, Japanese response, short response, medium response, extended response, translation from Japanese to English, guided writing, and paragraph writing. In addition, we recorded the number of words required for constructive responses in English. In cases where there are five words or fewer (short response), often the words are taken/copied directly from the text. If the response requires 10 words or fewer (medium response), test takers' ability to produce complete sentences is typically being assessed. Responses longer than 10 words (extended response), typically target the ability to produce short, coherent texts. Finally, for the English responses we additionally noted those that required a personal response, that is, one that required test takers to express their own opinions or views on a given subject.

General information was recorded including the total time provided and the number of items in each section. All passages were coded for discourse type. Listening passages were coded as dialogues or monologues; reading texts were coded as dialogues, email/letter correspondences, informative or narrative texts. The term informative was chosen to represent expository texts that provide information but also contain elements of narrative.

Both researchers worked together on 10 tests to calibrate both the content for analysis and our method of recording information. Each researcher analyzed a test and then the results were compared and discussed until agreement was reached. In this way, the categories for coding were refined and differences in perceptions regarding how to categorize items were ironed out. Following this, the lead researcher coded the remaining 14 tests and the second researcher checked the coding. Agreement reached at least 88% in all categories. All discrepancies were discussed until agreement was reached.

To determine the complexity of the reading passages in the exams, we analyzed each test using Coh-Metrix, a free web-based text-analysis tool. We selected two readability formulae, Flesch Reading Ease Readability and Flesch Kincaid Grade Level, which have been used extensively in similar studies and which are widely used by textbook publishers and examination boards (Khalifa & Weir, 2009). The Flesch Reading Ease Readability considers the average number of words per sentence and the average number of syllables per word. Taken together these fundamental textual properties provide an indication of the lexical and structural complexity of a text. The Flesch Kincaid Grade Level is based on the same textual

properties but assigns a grade level equivalent to formal English medium education (e.g., 4.5 indicates the text is appropriate for the 4th Graders of 9-10 years of age). For descriptive purposes we also report basic textual properties, such as text length and number of sentences per text.

Given that text analysis tools work better with longer texts, only passages with around 200 words or more were analyzed.² Listening passages were excluded not only because the comprehension processes involved in listening and reading are markedly different but also because most fell below this word count. Each text was copied to a word processor document and the embedded items and extraneous textual features were removed (e.g., asterisks, Japanese kana options, and boxes). If a sentence or word had been omitted for a gap-fill task, the appropriate sentence or word was added to the new document to ensure cohesion in the analysis.

Analysis

The *item distribution proportion* was calculated by counting all the items coded for a specific skill in a test and dividing this by the total number of items in a test. For example, if there were 15 items categorized as targeting the skill of reading in a test of 30 items then listening accounted for 50.00%. In the same way, the proportion of subskills was calculated by counting all the items coded as a specific subskill and then dividing this by the total of all items for that main skill found in the test. *Scoring distribution/weighting* was calculated similarly by summing the points allocated for each item of a given skill and dividing this by the total points of the test then converting this to a percentage. Score distribution was calculated only for each of the main skills identified in the tests. *Frequencies* refer to the actual number of items for a skill or subskill in a test, or the actual number of points available for a given item. As not all tests were out of 100, some differences can be seen in Table 2 between the frequencies and proportions (%) of score distribution.

In this study, items such as word reordering and transformation were coded as reading rather than indirect writing. Other researchers may classify such items as indirect-writing items that target linguistic knowledge and the lower order process of micro-planning (Moore, 2015). Other items such as conversation gap fill accompanied with manga type pictures, may be classified by some researchers as indirectly targeting speaking. However, we consider this kind of task to focus on writing because it bears little or no resemblance to what people do when engaging in the act of speaking in their daily lives; conversely, it does to a certain degree resemble what people do when they write. Finally, many of the writing items were found to target two skills. These were reading into writing, and listening into writing tasks. These were classified ultimately as writing as they were considered to be predominantly targeting this skill.

Results

What English skills and subskills are being assessed in the exams?

The skills of reading and listening were found in all prefectural tests analyzed, while writing was assessed directly in all but two of the tests (Ibaraki and Kyoto). Speaking was not found to be tested directly in any of the tests. The average distribution of items for these skills across all tests was 62.38% for reading, 29.16% for listening, and 8.46% for writing items (Table 2). The score weightings for each skill were 61.43% reading, 26.13% for listening, and 12.44% for writing. In terms of both item and score distribution then, reading was found to be the most assessed skill, while writing was the least assessed.

Table 2*Frequency and percentage data for skills, scoring distribution, and subskills*

Item distribution	Frequencies (N)				Proportions (%)			
	M	SD	MIN	MAX	M	SD	MIN	MAX
Listening	8.79	2.38	4.00	14.00	29.16	5.16	20.00	40.74
Reading	18.63	4.07	10.00	27.00	62.38	8.02	37.04	73.91
Writing	2.46	1.44	0.00	6.00	8.46	5.43	0.00	22.22
Total	29.88				100%			
Score distribution (weighting)	M	SD	MIN	MAX	M	SD	MIN	MAX
Listening	22.17	7.84	5.00	35.00	26.13	5.06	15.00	36.00
Reading	52.57	17.57	15.00	75.00	61.43	8.32	40.00	74.00
Writing	10.13	5.36	0.00	23.00	12.44	6.42	0.00	26.00
Total	84.87				100%			
Listening	M	SD	MIN	MAX	M	SD	MIN	MAX
Main Ideas	6.12	1.91	3.00	11.00	70.46	13.76	50.00	100.00
Details	2.67	1.31	0.00	5.00	29.54	13.76	0.00	50.00
Total	8.79				100%			
Reading								
CR: L	6.79	3.59	0.00	17.00	34.98	13.66	0.00	62.96
CR: G	9.29	3.20	4.00	15.00	51.05	17.85	25.00	100.00
ER: L + CR: L	1.67	1.79	0.00	6.00	8.94	10.06	0.00	35.29
ER: L + CR: G	0.88	1.60	0.00	7.00	5.03	9.51	0.00	41.18
Total	18.63				100%			
Writing								
Four lower-order processes *	2.13	1.57	0.00	6.00	79.55	32.40	0.00	100.00
All six processes **	0.33	0.48	0.00	1.00	20.45	32.40	0.00	100.00
Total	2.46				100%			

Note: CR: L = careful reading, local; CR: G = careful reading, global; ER: L = expeditious reading, local. * The four lower-order processes are: microplanning, translation, monitoring, and revising. ** Includes macroplanning and organization in addition to the four lower-order processes.

Listening

Two of the four subskills for listening were found to be assessed in the prefectural tests. These were *listening for main idea/important information/key message* (main ideas), and *listening for detailed/specific information* (details). The subskills of *listening for gist* and *listening to infer opinion* were not observed in any test. This was probably because these two subskills generally require longer passages and a higher level of ability than that expected of junior high school students.

Items targeting comprehension of main ideas were the most common and comprised an average 70.46% of listening items across all tests. Some prefectures tested only for this listening subskill (Aichi and Hiroshima) while the least it was assessed was found to be 50.00% (Yamagata and Nagasaki). These items often required complete understanding of a very short text. This typically involved listening to a short dialogue and selecting the appropriate picture or completing a short monologue or dialogue with an appropriate response. Such items were common in the first two sections of the listening test and targeted comprehension of the whole utterance. Items targeting main ideas were also found regularly in the latter two sections of the listening sections, where test takers are required to listen to a longer monologue or dialogue (typically involving an ALT or visitor to Japan) and comprehend main ideas that are somewhat distributed throughout the text. That is, the correct response can only be confirmed, and distractors correctly eliminated, by comprehending multiple parts of the text.

Items that targeted comprehension of details made up the remaining 29.54% of items in the listening section as there were only two subskills found in the tests. These items were found almost exclusively in the latter sections of the test, where test takers listened to a longer monologue or dialogue. What made these items different from those targeting comprehension of main ideas was that they required test takers to identify specific details stated by the speakers; for example, when or where something happened, or what someone did. These items could typically be answered correctly by understanding only part of an utterance, rather than understanding multiple parts of the text.

Reading

The subskills for reading found in the prefectural tests comprised four patterns: *careful reading: local*; *careful reading: global*; *expeditious reading: local + careful reading: local*; and *expeditious reading: local + careful reading: global*. Another subskill in the socio cognitive framework, *expeditious reading: global*, was not found in any of the tests, probably because it is associated with higher levels of ability and in fact, as was previously mentioned, it is not found to be tested very frequently even in tests for higher levels of ability.

Items found to assess *careful reading: local* comprised an average of 34.98% of items in reading sections, although there was considerable variation in this across tests with one prefecture not testing it at all (Mie) and another testing it frequently at 62.96% of reading item distribution (Okinawa). These items focused mainly on grammatical or lexical awareness. Common item types included re-ordering a jumbled sentence (grammar), single word gap fill in short passages (vocabulary or grammar), and transformation tasks where test takers are required to alter the form of a word to fit a sentence (grammar). As these types of items are the staple of traditional grammar tests, it was encouraging to see that other than Okinawa, only three other prefectures (Akita, Kanagawa, and Miyagi) included more than 50% of reading items that assessed this subskill.

Items that assessed *careful reading: global* were more common and comprised an average 51.05% of reading items across the tests. The minimum for this subskill was 25.00% (Yamagata) and the maximum was 100% (Mie), highlighting significant variation across prefectures. These items focused on meaning rather than form and typically required test takers to form a macro-level (i.e., paragraph, whole text) representation in order to respond correctly. Items in this category usually included gapped-summary tasks which focused on main ideas, multiple-choice questions, matching a given sentence into one of multiple possible locations in a text, and in one instance (Yamanashi), of matching headings from a list of options to the appropriate paragraphs in a text.

Items that assessed *expeditious reading: local* followed by *careful reading: local* comprised an average of 8.93% of all reading items. The maximum for this was 35.29% (Hokkaido) while eight tests did not assess this subskill at all. These items typically involved comprehension questions requiring test takers to locate an area in the text and find the relevant detail within a sentence required to answer the question. Usually there were no more than two of these items within a given test. Finally, items that appeared to involve *expeditious reading: local* followed by *careful reading: global* comprised the smallest proportion of reading items at 5.02% across all tests. Higher proportion outliers included Tokyo (41.18%), Kyoto (20.00%) and Aichi (14.28%). In general, however, this subskill was identified in fewer than 10% of reading items with 15 prefectures not assessing it at all. These items involved more challenging MC tasks or gapped-summary tasks that required searching for a relevant location in the text and then reading carefully through a paragraph to determine the correct response.

Writing

We divided writing into two categories: those that engage the lower four cognitive processes for this skill (*microplanning, translation, monitoring, and revising*) and those that engage all six cognitive processes (*macroplanning, organization, microplanning, translation, monitoring, and revising*). Most writing items engaged only the lower four cognitive processes. These averaged 79.55% of all writing tasks across the tests. Typically, they required sentence-level organization from two or three sentences where the format of what is to be written is provided, or where only one phrase or sentence is required usually in the form of a gap-fill response. Translation from Japanese into English was included as a writing task as it involves the same four lower order cognitive processes as these other tasks (Moore, 2015, p. 248).

Tasks which could be said to incorporate all six cognitive processes for writing were found in only eight of the tests we examined (Gunma, Kagoshima, Nagano, Nagasaki, Okinawa, Tokyo, Yamagata, and Yamanashi) and comprised an average of only 20.45% of all writing items across the tests. These items involved organizing an entire paragraph, that is, test takers were expected to develop a response across a paragraph, exhibiting coherence and cohesion in their writing. Overall, it could be said that the skill of writing was the area of most variance in these tests with two prefectures not testing for it at all and many others only very minimally.

2. How do the exams vary in terms of test and item characteristics?

Overall, the balance of selective and constructive responses varied considerably across prefectures (Table 3), with the maximum for selective responses reaching 96.30% in Kanagawa (followed by 91.30% in Tokyo and 80.95% in Kyoto) while the minimum was recorded at 45.00% (Hiroshima). Selective-response items averaged 64.41% across all prefectures and were thus far more prevalent in general. On the other hand, the maximum for constructive responses was found in Hiroshima (55.00%) followed by Akita (52.78%) and Hokkaido (51.52%). Aichi, Yamagata, and Yamanashi also had many constructive-response items at approximately 48% of their respective totals.

Notwithstanding certain differences in the prefectural exams, a common pattern to the selective-response and constructive-response items is apparent. Selective response items commonly included multiple choice, gap fill, gapped summary, ordering, and matching items. Ordering tasks usually consisted of rearranging jumbled words into a syntactically correct sentence; less often, they consisted of rearranging sentences coherently within a paragraph or events of a story into a chronological order. Matching tasks involved matching graphs, tables, or pictures with text. These matching items assess the ability to apply, relate, and transfer information between sources. That is, they require test takers to read or listen and then integrate information with that in charts. Overall, a great deal of variety of selective-response items were found and the process of neat classification somewhat diminishes this fact.

Constructive responses also exhibited considerable variety. Short responses of five words or less would typically be required for gap-fill, reordering, or comprehension items. Writing tasks, such as translation from Japanese to English and guided writing, usually required 10 words or fewer (i.e., a one-sentence response). No translation tasks from English to Japanese were observed, although eight tests were found to have items that required a Japanese response. These targeted reading comprehension and did not require translation. They were thus classified as Japanese constructive responses to reading items. Two common features among writing tasks were that they often required a personal response³ and typically included an input text that provided background information. These are perhaps thus best described as reading-into-writing tasks. Only one test (Hiroshima) featured a listening-into-writing task, which required test takers to listen to a dialogue and then explain their opinion on the ideas discussed therein.⁴

Table 3*Frequencies and percentages of item response types*

Items	M	SD	MIN	MAX	M	SD	MIN	MAX
Selective	19.21	5.10	9.00	32.00	64.41	13.84	45.00	96.30
Constructive	10.67	4.41	1.00	19.00	35.59	13.84	3.70	55.00
Total	29.88				100%			
Constructive response types								
Short response. – 5 w. or less	6.96	3.83	0.00	14.00	61.96	23.73	0.00	100.00
Medium response. – 10 w. or less	1.63	1.41	0.00	5.00	18.44	22.96	0.00	100.00
Extended response. – 10 + w.	1.17	0.92	0.00	3.00	12.07	12.09	0.00	50.00
Japanese response	0.91	1.56	0.00	5.00	7.53	13.17	0.00	40.00
Total	10.67				100%			
Personal response*	1.25	1.03	0.00	4.00	12.58	12.91	0.00	50.00
Listening								
MC visual	2.17	1.63	0.00	6.00	23.36	16.28	0.00	50.00
MC	4.92	2.30	0.00	11.00	58.13	26.66	0.00	100.00
MC gap fill	0.25	0.68	0.00	2.00	2.93	8.11	0.00	28.57
Gap fill	1.12	1.54	0.00	5.00	11.22	14.18	0.00	41.67
Japanese response	0.12	0.61	0.00	3.00	1.56	7.65	0.00	37.50
Short response	0.21	0.51	0.00	2.00	2.80	6.86	0.00	25.00
Total	8.79				100%			
Reading								
MC Visual	0.25	0.53	0.00	2.00	1.30	2.83	0.00	11.11
MC	4.83	3.05	2.00	14.00	26.21	14.65	10.00	70.60
Matching	1.13	0.85	0.00	3.00	6.24	5.07	0.00	16.67
Re-order sentence	1.33	1.37	0.00	4.00	6.89	6.39	0.00	21.05
Re-order paragraph/events	0.38	0.58	0.00	2.00	2.01	3.21	0.00	11.76
MC gap fill	3.21	1.96	0.00	8.00	17.70	11.46	0.00	44.44
MC gapped summary	0.75	1.26	0.00	4.00	3.71	6.25	0.00	20.00
Gap fill	2.42	2.41	0.00	8.00	12.88	12.78	0.00	41.18
Gapped summary	1.00	1.53	0.00	7.00	6.51	8.42	0.00	33.33
Re-order sentence	0.08	0.28	0.00	1.00	0.35	1.70	0.00	8.33
Transformation	1.04	1.43	0.00	4.00	5.22	6.91	0.00	20.00
Japanese response	0.79	1.32	0.00	4.00	3.28	6.13	0.00	22.22
Short response	1.42	1.28	0.00	4.00	7.70	7.31	0.00	23.54
Total	18.63				100%			
Writing								
Translation (Japanese to English)	0.42	0.83	0.00	3.00	12.88	22.38	0.00	66.67
Guided writing	1.71	1.27	0.00	4.00	66.67	35.17	0.00	100.00
Paragraph writing	0.33	0.48	0.00	1.00	20.45	32.40	0.00	100.00
Total	2.46				100%			

Note. * Personal response is a secondary feature of short/medium/extended constructive responses which were double coded for this attribute and the data presented independently here.

Finally, the length of tests in terms of time allowed, item numbers and total score possible also varied considerably (Table 4). Most tests allowed 50 minutes, while others ranged between 40 and 60 minutes. Interestingly, the time designated for Gunma's test varied from 45 to 60 minutes depending on the school. Within these time frames, however, the length of tests differed considerably in terms of item count. Ibaraki topped the list in terms of items per minute with test takers expected to complete 40 items in 50 minutes, an allowance of 1.25 minutes per item. Perhaps this reflects the fact that there were no writing tasks included. In contrast, Hiroshima, which had four guided writing tasks, had a rate of 2.5 minutes per item, while Hokkaido, which also had four guided writing tasks, had a rate of only 1.52 minutes per item. Naturally, these figures are limited in that they do not reflect item difficulty; however, they further highlight the variation across tests that exists in terms of test and item characteristics.

Table 4

Prefectural test time allowed, total item count and total score available

Prefecture	Time	Items	Score**	Prefecture	Time	Items	Score**
Hokkaido	50	33	100	Yamanashi	45	33	100
Aomori	50	35	100	Nagano	50	35	100
Miyagi	50	30	100	Gifu	50	30	100
Akita	60	36	100	Aichi	50	21	22
Yamagata	50	29	100	Mie	45	27	50
Fukushima	50	33	50	Kyoto	40	19	40
Ibaraki	50	40	100	Osaka	55	29	90
Gunma	45-60*	28	100	Hiroshima	50	20	50
Saitama	50	32	100	Ehime	60	30	50
Chiba	60	31	100	Nagasaki	50	28	100
Tokyo	50	23	100	Kagoshima	50	28	90
Kanagawa	50	27	100	Okinawa	50	38	60

Note. * The length of the test in Gunma is at the discretion of the school where it is taken; ** Score represents maximum number of marks possible for each prefectural test.

3. How complex are the reading passages of the exams?

The statistics for the analysis of reading texts can be seen in Table 5. The average text length for all texts was around 450.61 words, with the longest text encountered being 759.00 words (Nagasaki). Three other prefectures had texts that exceeded 700 words (Kanagawa, Kyoto, and Saitama). The reading passages ranged diversely in difficulty as expressed in Flesch Reading Ease (FRE) from 70.08 to 96.60 ($M = 82.29$, $SD = 5.97$), and Flesch-Kincaid Grade Level (FKGL) from 1.44 to 7.43 ($M = 4.38$, $SD = 1.17$).

To put these textual complexity characteristics into context, we compared them with the figures provided in Hamada (2015) for the EIKEN tests, specifically at Grades Pre-2 and 3. EIKEN Pre-2 measures between A1 and B2 levels and the pass level is the lower-A2 level⁵, while the EIKEN Grade 3 test measures between the A1 and B1 levels and the pass level is at the upper-A1 level (Eiken Foundation of Japan, n.d.). The ability of students entering high school is expected to range primarily between A1 and A2 levels, which makes Grades Pre-2 and 3 most suitable for comparison.

Table 5 shows that EIKEN Grade 3 has an average FRE of 75.15 and FKGL of almost 6, which suggests that they are more complex than those of the high school exams. However, the high school exam passages at the higher end of the ranges for both FRE and FKGL surpass the means for these indices of the EIKEN Grade 3 exam and approach those of the EIKEN Pre-2 exam. In sum, while the mean text complexity scores for the high school exams suggest the reading texts are simpler than those in both EIKEN Grade 3 and Grade Pre-2, the range of complexity in the high school exams is considerable and puts some of the texts in a similar region to the means of the higher-level Grade Pre-2 exam. However, in terms of text length, the average text length of passages in high school exams was 450.61 words, which is much higher than that in EIKEN Grade 3 (255.30 words) and Grade Pre-2 (290.15 words). In fact, it is not until EIKEN Grade 1 that mean passage

length exceeds 450 words (Hamada, 2015). Consequently, regarding text length, the high school exams are considerably more challenging than those of the EIKEN exams.

Considering the university entrance exams reported by Kikuchi (2006), specifically the average figures for public and private university exams and the Center Test in 2004, the mean FRE and FKGL (64.40 and 9.79, respectively) are both considerably higher than those of the high school exams (82.29 and 4.38, respectively). The mean FRE and FKGL in the 2016 university exams reported in Fukazawa (2021) are also considerably higher (at 52.43 and 11.21, respectively). The university exams also exhibit considerably more complex sentences, with sentence length being almost double that of high school exams (e.g., 19.63 words and 10.46 words, respectively, in 2004). In contrast, the mean text length in 2004 was 466.95 words, which is very close to that of the high school exams. In sum, it appears that while text length may be similar, sentence complexity is far higher and readability much lower in the university exams compared to that of the high school exams.

It has been suggested⁶ that the passages in the high school tests are relatively long because, unlike standardized tests such as EIKEN, a wide variety of items and tasks are used within a single passage. These include such things as gap fill, translation, transformation, reordering, find the referent and comprehension tasks, most of which are embedded into the actual passages thus necessitating more length to allow sufficient space between each task. This practice has been observed in some of the prefectural tests to varying degrees in different prefectures. Consequently, although the actual texts may be less complex, students must cope with extraneous features in the text which may interfere with the coherency of the passage and thus pose an additional cognitive challenge (e.g., Negishi, 2013). However, not all prefectural exams included such embedded items yet still had relatively long passages. This can be seen in the Tokyo exam which contains a passage of 675.00 words (FRE - 80.55 and FKG - 4.23) and yet contains no embedded tasks. The Chiba exam, too, has a relatively long passage of 527.00 words (FRE - 73.04 and FKG - 6.19) which contains only one embedded task.

In terms of comparative complexity then, while EIKEN Grade 3 appears more complex using traditional readability indices, the passage length and use of embedded items in some of the prefectural tests may make them more demanding. Further empirical research may shed light on this issue.

Table 5*Prefectural high school entrance exams text readability statistics overview*

					EIKEN 3*	EIKEN Pre-2*	Uni. Exams 2004**	Uni. Exams 2016***
All texts								
	M	SD	MIN	MAX	M	M	M	M
Text length (words)	450.61	161.29	193.00	759.00	255.30	290.15	466.95	NA
Sentence length (words)	10.46	2.07	5.26	15.88	NA	NA	19.63	NA
Word length (syllables)	1.35	0.06	1.22	1.45	NA	NA	1.45	NA
Flesch reading ease	82.29	5.97	70.08	96.61	75.15	67.25	64.40	52.43
Flesch-Kincaid grade	4.38	1.17	1.44	7.43	5.99	7.94	9.79	11.21
Dialogue texts								
	M	SD	MIN	MAX				
Text length (words)	395.67	163.12	193.00	755.00				
Sentence length (words)	8.73	1.66	5.26	12.12				
Word length (syllables)	1.34	0.06	1.22	1.42				
Flesch reading ease	84.80	5.92	76.21	96.61				
Flesch-Kincaid grade	3.59	1.12	1.44	5.30				
Narrative, informative, and letter texts								
	M	SD	MIN	MAX				
Text length (words)	482.52	153.89	211.00	759.00				
Sentence length (words)	11.46	1.58	8.41	15.88				
Word length (syllables)	1.35	0.06	1.23	1.45				
Flesch reading ease	80.83	5.59	70.08	91.59				
Flesch-Kincaid grade	4.83	0.95	3.22	7.43				

Note. * Data taken from Hamada (2015). ** Data from Kikuchi (2006) for all exams (i.e., public, private and the Center Test). *** Data taken from Fukuzawa (2021) for all exams (i.e., public, private and the Center Test).

Discussion

In terms of skill distribution and weighting, the prefectural high school entrance exams were found to be testing skills in order of reading, listening and writing skills, while speaking was not assessed directly at all. The weightings do not align well with the goals of the course of study, particularly in regard to the focus on reading assessment over productive skills. While there are undoubtedly practical reasons for focusing on assessment of receptive skills (i.e., they can be assessed quickly and reliably using selective-response items), these practical concerns do not outweigh the importance of assessing productive skills, especially when the course of study stipulates the development of spoken and written communicative ability as goals of junior high school English education (See O'Sullivan et al., 2022 for a similar argument). Moreover, given the role of entrance exams as key short-term goals for language learners and teachers, these entrance exams are unlikely to generate much positive washback on the teaching and learning of productive skills in schools. As Akiyama reported back in 2003, a common complaint of junior high school teachers was that the exams should mirror the effort that is spent on teaching conversation in class. The implementation of the *English Speaking Achievement Test for Junior High School Students* (ESAT-J) in Tokyo is therefore likely to be welcomed by many junior high school teachers, at least in terms of the principle it stands for, that is, the importance of learning, teaching and assessing speaking ability. However, whether such speaking tests become a standard feature of high school exams across the country remains to be seen.

As regards subskills, the present study found that the exams tested a reasonable range of them, which should support the curriculum goal of communicative language learning. The majority of reading items required careful reading at the global level, while a minority (approximately one in three) targeted careful reading at the local level. Given the close association of local careful reading with assessment of grammatical and lexical knowledge, it could be argued that candidates' basic linguistic knowledge is directly targeted in only a minority of reading items. Similarly, in listening tasks, main ideas and listening for overall comprehension was targeted in the majority, whereas listening for details was targeted again in approximately one in every three items. These trends for reading and listening items appear to reflect the primary goal of English education in Japan, that is, the development of communicative competence. In other words, rather than focusing on the development of knowledge *about* language assessed through sentence-level grammar/vocabulary items, these exams appear to primarily target the ability to *use* such linguistic knowledge to comprehend the main ideas of texts and conversations. Amid the often-heard criticisms of English education in Japan being doggedly traditional, these exams in fact reveal the evolution from a structural orientation toward a contemporary communicative one. However, we must also note that our findings for public school exams run contrary to those of Minato's (2020) for private school exams in Hokkaido. His study reported a major focus on the assessment of linguistic knowledge through careful reading at the local level, while we found such items to be in the minority. This divergence of findings appears to be suggestive of a difference between the public and private education sectors in terms of their commitment to the national course of study; however, a broader study of private school exams throughout Japan would be needed to confirm this.

Not all subskills described in the socio-cognitive framework were observed in the exams or were featured only rarely. In contrast to careful reading, expeditious reading subskills were assessed comparatively rarely. Just under one in 10 reading items required test takers to scan for a key word and then read carefully around it, indicating a combination of expeditious and careful reading at the local level. Likewise, approximately one in 20 items assessed test takers' ability to locate key words then read the sections in which they were identified carefully to comprehend main ideas in the text. While these patterns were observed less often than other subskills, this is not necessarily detrimental to evaluation of the test design. In fact, it is encouraging that expeditious reading is involved at this level of proficiency, given that it requires making fast eye-movements across long stretches of text. Making students aware that there are different reading subskills, which should be employed according to the purpose of reading, is undoubtedly beneficial for their ongoing development as language learners. Moreover, it will prepare them for developing the ability to read quickly at the global level, that is, skimming texts for general comprehension. In fact, skim reading (i.e., global expeditious reading) was not observed at all in the high school exams, most likely because it is a cognitively demanding subskill that can only be utilized once learners have reached a higher level of proficiency.

The subskills of listening for gist and listening to infer attitude were also not observed at all in the exams. This is again most likely due to the target proficiency level of the exams. Listening for gist items typically requires a relatively long text that contains multiple clues to its general topic or purpose, while listening to infer opinion items typically require lengthy texts or multiple turns in a dialogue, in which an opinion can be deduced from multiple clues, particularly prosodic and phraseological ones. Given the relatively high linguistic demands of these subskills, it is unsurprising that they are not observed at this level. Writing subskills were also typically restricted to lower-order processes, while the higher-order ones of macro-planning and organization were absent from most tests. This was clearly due to the short length of the required written response, which made the use of the higher-order subskills redundant. Again, the target proficiency level may explain the absence of extended written responses.

A number of additional features stand out as reflective of the goals of the course of study. Firstly, many of the tests included personal responses which made up approximately an eighth of constructive responses. These allowed students the opportunity to express their opinions, desires, or memories, highlighting the importance of the communication of ideas and providing motivation for students to develop their self-expression through the use of a foreign language. Secondly, a recurring feature in the exams was the necessity for test takers to transfer information between texts and visual information, which reflects the way much information is presented in the digital age and hence the way most students are likely to encounter English in their daily lives. Thirdly, the inclusion of integrated skill tasks, such as reading-into-writing and listening-into-writing, is likely to promote the use of classroom activities that move from input to output. In these ways, there seems to be alignment with the MEXT goals, which focus on the communication of ideas and developing the desire to proactively communicate in a foreign language.

Finally, a comparison of the high school exams with the 2004 public university entrance exams examined in Kikuchi (2006) is revealing. Firstly, considering the items in the university exams by skill, reading was by far the most assessed, while only a handful of public exams assessed listening, and writing appeared to be assessed rather minimally. Thus, although the main focus on reading and relatively minimal focus on writing appear convergent, the fact that all high school exams assess

listening ability reveals a notable step forward towards more balanced assessment of skills. Perhaps more striking is the breakdown of items in the university entrance exams into those that assessed receptive, productive, and translation abilities. According to Kikuchi's (2006) data, public university exams assessed translation (46.68%) more than receptive (31.01%) or productive (22.31%) abilities. Moreover, translation was generally in the direction of English to Japanese, thus requiring test takers to write responses in Japanese, which were then assessed by raters. In the writing items of the high school exams, in contrast, there were no English to Japanese translation items at all, and less than 8% of all constructive responses (i.e., including those for reading and listening) were written in Japanese. These findings reveal a marked difference between the high school and university entrance exams and no doubt reflect the goals of the course of study for foreign languages, which aims to develop communicative competence in English, not to fine-tune writing competence in Japanese. Consequently, because Japanese responses are not being evaluated in high school exams, there is little reason to focus on English to Japanese translation in junior high school English classrooms. On the other hand, if Japanese university entrance exams continue to include such items, as the recent replication study of Fukazawa (2021) suggests they do, there is likely to remain a tendency for learners and teachers to practice the skill of English-to-Japanese written translation in English classes, especially in the third year of high school when the entrance exams gain prominence.

Limitations and future directions

This study analyzed the content of prefectural exams for public senior high schools from approximately half of the 47 prefectures in Japan. As the majority of Japanese teenagers will attend such schools, this analysis provides a broadly representative overview of the exam content that most test-takers will encounter. Nevertheless, we did not investigate the exams for highly-selective national schools or the multitude of private schools in Japan. Research by Minato (2020) suggests that the exams of private schools may differ considerably from those analyzed here and thus future research in this area is necessary. Another limitation is the simple treatment of readability, and the lack of consideration for the many contextual parameters that can influence performance in EFL listening tests, such as familiarity with accents, the number of times allowed for listening, speech rates, and syntactic complexity (Yanagawa, 2013). Future research could explore such characteristics of the audio texts, and further examine the dimensions of the written texts, such as features related to cohesion, coherence, and grammatical and lexical range. Finally, other aspects of validity including the reliability of scoring, the relationship between exam scores with those from other exams or future academic performance, and impact of the exams in society are all important areas that are yet to be investigated.

Considering scoring validity, around a third of responses were constructive, which perhaps reflects the course of study goal of producing language. At the same time, this high proportion of constructive responses necessitates investigating the reliability of scoring. We cannot investigate this issue because we have no access to score data or grading criteria. However, if the raters are well trained and use validated rubrics for assessing written responses, the inclusion of constructive responses does not preclude reliability and fairness. Because most constructive responses are short, it is very likely that grammatical and lexical accuracy, spelling, and mechanics are all critical assessment criteria. Nevertheless, it would be beneficial if basic information about the grading criteria were to be made public so that learners and teachers are not left in the dark as to how the written responses will be assessed (Kowata, 2015).

Conclusions

It is still commonly heard in the comments of students and teachers, that English teaching in Japan is heavily skewed towards building linguistic knowledge, particularly grammatical knowledge, without sufficiently focusing on application of this knowledge in the four skills. This view is also somewhat supported by high school entrance exam preparation materials, which devote much space to grammatical constructions. However, the prefectural high school entrance exams analyzed here do not support this view. The tests appear to reflect a view that language exists in context, and communicative language use requires learners to process text beyond the sentence level, and to integrate information across sentences. Overall, the exams appear to reflect the MEXT goals of achieving actual communication of ideas and developing the desire to proactively communicate in a foreign language. While the imbalance in productive and receptive skills deserves ongoing attention in future reforms, the exams in their current state at least demonstrate a commitment to the assessment of the aims of the course of study.

Acknowledgements

We are grateful to the two expert readers who commented upon earlier drafts of this article. Any errors or omissions remain our own.

Notes

¹ However, the goal of A2 level should be considered an ideal and is unlikely to be achieved by many students: According to a survey by MEXT (2018) of third-year junior high school students' achievement, fewer than half reached the CEFR A1 level; also, fewer than a fifth of third-year senior high school students reached the A2 level.

² In the case of Nagano, there were not two individual passages that neared this limit, so it was decided to combine three short passages from one of the sections for the analysis. Miyagi had a similar section to Nagano that used three short passages. As these seemed intuitively more difficult than either of the two longer passages in the test, they were also combined to include in this analysis.

³ In this study, a personal response is a dual characteristic of a short/medium/extended response. This additional information has been presented separately to establish how much these tests encourage students to express themselves in a foreign language while maintaining table integrity for the precise length of each written response.

⁴ One other test (Akita) had a writing question in the listening section; however, this was an orally delivered writing prompt that simply required test takers to answer in two sentences what they did during their winter vacation.

⁵ It was pointed out by an anonymous reviewer that there is a difference between the *content* and *score* target range of each EIKEN level. For example, while the EIKEN Grade Pre-2 score reflects a level of between A1-B2, its content, however, is targeted at the A2 level.

⁶ We are grateful to an anonymous reviewer who pointed out the issue of the comprehensive problem (*sōgō mondai*).

References

- Akiyama, T. (2003). Assessing speaking: Issues in school-based assessment and the introduction of speaking tests into the Japanese senior high school entrance examination. *JALT Journal*, 25(2), 117-141. <https://doi.org/10.37546/JALTJJ25.2-1>
- Allen, D. (2020). Proposing change in university entrance examinations: A tale of two metaphors. *Shiken*, 24(2), 23-38. <https://doi.org/10.37546/JALTSIG.TEVAL24.2-2>
- Allen, D., & Nagatomo, D. H. (2019). Investigating the consequential validity of TEAP: Washback to high school learners of English. *Eiken Research Report*. Eiken Foundation of Japan. https://www.eiken.or.jp/center_for_research/pdf/bulletin/vol99/vol_99_21.pdf
- Bailey, J. L. (2018). A study of the washback effects of university entrance examinations on teaching pedagogy and student learning behaviour in Japanese high schools. *British Journal of Education*, 6(6), 50-72. <http://www.eajournals.org/wp-content/uploads/A-Study-of-the-Washback-Effects-of-University-Entrance-Examinations-on-Teaching-Pedagogy-and-Student-Learning-Behaviour-in-Japanese-High-Schools.pdf>
- Bennesse. (2017). *Gakkō-gai kyōiku katsudō ni kansuru chōsa 2017* [Survey on Extracurricular Activities 2017]. https://berd.benesse.jp/up_images/research/2017_Gakko_gai_tyosa_web.pdf
- Brown, J. D., & Yamashita, S. O. (1995a). English language entrance examinations at Japanese universities: What do we know about them? *JALT Journal*, 17(1), 7-30. <https://jalt-publications.org/sites/default/files/pdf-article/jj-17.1-art1.pdf>
- Chalhoub-Deville, M., & O'Sullivan, B. (2020). *Validity: Theoretical developments and integrated arguments*. British Council Monographs on Modern Language Testing (3). Equinox.
- Eiken Foundation of Japan. (n.d.). *The EIKEN CSE Score*. <https://www.eiken.or.jp/eiken/en/eiken-tests/overview/cse/>
- Elliot, M. & Wilson, J. (2013). Context validity. In A. Geranpayeh and L. Taylor (Eds), *Examining listening: Research and practice in assessing second language listening*. (pp. 152-241). UCLES/Cambridge University Press.
- Fukazawa, M. (2021). Have English entrance examinations at Japanese universities changed over the last two decades? *KASELE (Kyushu Academic Society of English Language Education) Bulletin*, 49, 51-60.
- Geranpayeh, A., & Taylor, L. (Eds). (2013). *Examining listening: Research and practice in assessing second language listening*. Studies in Language Testing (35). UCLES/Cambridge University Press.

- Green, A. (2014). *The Test of English for Academic Purposes (TEAP) Impact Study: Report 1 - Preliminary Questionnaires to Japanese High School Students and Teachers*. Eiken Foundation of Japan. https://www.eiken.or.jp/teap/group/pdf/teap_washback_study.pdf
- Hamada, A. (2015). Linguistic variables determining the difficulty of Eiken reading passages. *JLTA Journal*, 16, 57-77. https://doi.org/10.20622/jltajournal.18.0_57
- Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge University Press.
- Kikuchi, K. (2006). Revisiting English entrance examinations at Japanese universities after a decade. *JALT Journal*, 27(1), 77-96. <https://doi.org/10.37546/JALTJJ28.1-5>
- Kowata, T. (2015). *Washback effects of university entrance examination writing tasks on learning and teaching*. [Unpublished doctoral dissertation]. Tokyo University of Foreign Studies. <http://hdl.handle.net/10108/80582>
- Law, G. (2004). College entrance exams and team teaching in high school classrooms. In M. Wada & T. Cominos (Eds.), *Studies in team teaching* (pp. 90-102). Kenkyusha.
- McNamara, D. S., Graesser, A. C., & Louwse, M. M. (in press). Sources of text difficulty: Across the ages and genres. In J. P. Sabatini & E. Albro (Eds.), *Assessing reading in the 21st century: Aligning and applying advances in the reading and measurement sciences*. R&L Education.
- MEXT. (2016). *Eigo no 4 ginō ni kansuru genjō kadai kongo no hōkō-sei* [Current status, issues, and future direction of the four skills of English]. https://www.mext.go.jp/content/20191224-mxt_daigakuc02-000003411_6.pdf
- MEXT. (2017). *Chūgakkō gakushū shidō yōryō (Heisei 29-nen kokujū) kaisetsu: Gaikoku-go-hen*. [Junior High School Curriculum Guidelines (Announced 2017) Commentary: Foreign Language Edition.] https://www.mext.go.jp/component/a_menu/education/micro_detail/_icsFiles/afieldfile/2019/03/18/1387018_010.pdf
- MEXT. (2018). *Heisei 29-nendo eigodjikara chōsa kekka (chūgaku 3-nensei kōkō 3-nensei) no gaiyō* [Overview of 2017 English proficiency survey results (third year junior high school and third year high school)]. https://www.mext.go.jp/a_menu/kokusai/gaikokugo/_icsFiles/afieldfile/2018/04/06/1403470_01_1.pdf
- MEXT. (2022). *Shiritsu gakkō gakkō hōjin kiso dēta* [Basic data for private schools and school corporations]. https://www.mext.go.jp/a_menu/koutou/shinkou/main5_a3_00003.htm#topic1
- Minato, S. (2020). Shiritsu kōkō nyūshimondai eigo-ka ni miru shidō yōryō: Shiritsu kōkō wa chūgakkō shidō yōryō (eigo-ka) o rikai shite iru ka [Private high school entrance exam questions: Do private high schools understand the junior high school curriculum guidelines (English)?]. *Hokusei Review, the School of Humanities*, 57(2), 27-36. https://hokusei.repo.nii.ac.jp/?action=repository_uri&item_id=2406&file_id=45&file_no=1
- Moore, Y. (2015). An evaluation of English writing assessment in Japanese university entrance examinations. *Writing and Pedagogy*, 7(2-3), 233-260. <https://doi.org/10.1558/wap.v7i2-3.26227>
- Negishi, M. (2013). Sayonara, sōgō mondai [Goodbye, comprehensive problem]. *Teaching English Now*, 24, 14-15. https://tb.sanseido-publ.co.jp/english/newcrown/pdf/ten024/TEN_vol24_04.pdf
- O' Sullivan, B., Motteram, J., Skipsey, R., & Dunlea, J. (2022). *The importance of the four skills in the Japanese context*. British Council Perspectives on English Language Policy and Education. <https://www.britishcouncil.org/exam/aptis/research/publications/english-language-policy-and-education>
- O' Sullivan, B., & Weir, C. J. (2011). Test development and validation. In O' Sullivan, B. (Ed.), *Language testing: Theories and practices*. Palgrave Macmillan.
- Rohlen, T. P. (1983). *Japan's high schools*. University of California Press.
- Sakui, K. (2004). Wearing two pairs of shoes: Language teaching in Japan. *ELT Journal*, 58(2), 155-163. <https://doi.org/10.1093/elt/58.2.155>
- Sato, K. (2002). Practical understandings of communicative language teaching and teacher development. In S. J. Savignon (Ed.), *Interpreting communicative language teaching: Contexts and concerns in teacher education* (pp. 41-81). Yale University Press.
- Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge University Press.

- Takagi, A. (2010). *A critical analysis of English language entrance examinations at Japanese universities* [Unpublished doctoral dissertation]. University of Exeter. <http://hdl.handle.net/10036/117893>
- Taylor, L. (2011). *Examining speaking: Research and practice in assessing second language speaking*. Cambridge University Press.
- Toshin.com. (2022). *Zenkoku kōritsu kōkō nyūshi kaitō*. [Nationwide national high school entrance exam answers]. https://www.toshin.com/koukou_nyushi/
- Underwood, P. G. (2012). Teacher beliefs and intentions regarding the instruction of English grammar under national curriculum reforms: A theory of planned behavior perspective. *Teaching and Teacher Education*, 28, 911-925. <https://doi:10.1016/j.tate.2012.04.004>
- Watanabe, Y. (1997). *The washback effects of the Japanese university entrance examinations of English: Classroom-based research* [Unpublished doctoral dissertation]. Lancaster University.
- Weir, C. J. (2005). *Language test validation: An evidence-based approach*. Palgrave Macmillan.
- Yanagawa, K. (2013). Factors which affect listening comprehension test performance: A comprehensive framework. *Japanese Journal for Research on Testing*, 9, 107-127. https://doi.org/10.24690/jart.9.1_107
- Zenkoku kōkō nyūshi mondai seikai (eigo): 2023 juken-yō* [Nationwide high school entrance tests correct answers (English): For 2023 use]. (2022). Obunsha.

Test-taker self-assessment accuracy using the TOEIC speaking and writing can-do statements

Jonathan Harrison¹ and Ruth Vanbaelen²

jon.harrison@nihon-u.ac.jp

1. *Nihon University, College of Science and Technology*

2. *University of Tsukuba*

<https://doi.org/10.37546/JALTSIG.TEVAL26.2-2>

Abstract

Can-do statements have been integrated for decades into research and marketing strategies related to standardized tests of English for non-native language learners. For instance, while ETS indicated that TOEIC can-do statements are specifically for “decision making in organizations,” the Council of Europe created user-friendly statements that all stakeholders, including test takers, can use and understand. This study discusses a partial replication of a large-scale validity study with the purpose of developing a scale which can be used by examinees in combination with the TOEIC SW can-do statements for self-assessment. As part of an extracurricular program, 23 graduate-level EAL university students took the TOEIC SW and speaking and writing can-do self-assessment surveys. Results indicated that both self-assessment scales used in this study had moderate to strong correlations with TOEIC SW scores ($r = 0.62-0.73$) and TOEIC proficiency levels ($r = 0.47-0.66$). Self-assessment results using a Level Mean Average (LMA) scale correlated significantly with both speaking and writing test results and produced slightly higher percentages of test takers self-assessing within the true score zones of 1 SEM. For the most part, self-assessment predictive accuracy using these scales and true score zones was 52–57% for speaking and 43% for writing. Other findings were that EAL learners tended to underestimate their scores, and they more accurately self-assessed speaking than writing. If LMA scale accuracy at a particular level can be verified in future studies, claims could be made that TOEIC SW can-do statements have value for examinees.

Keywords: Assessment, can-do statements, English language learning, self-assessment, speaking, TOEIC, predictive accuracy, writing

Student self-assessment is a common educational practice, and can-do statements, which can be a form of self-assessment, were integrated into research and marketing related to international standardized tests of English language skills for English as a second language (ESL) and English as a Foreign Language (EFL), hereafter referred to as English as an Additional Language (EAL). Can-do statements were adopted by Educational Testing Service (ETS) and the Council of Europe (CoE) around 2000. However, ETS states that the can-do statements are for “decision-makers in organizations” (The Chauncey Group International, 2000), and CoE (2001) states that the can-do statements are designed to be user-friendly and educational to test takers as well as being useful for other stakeholders, e.g., potential employers. The TOEIC Speaking and Writing (SW) literature does not directly state that can-do statements can be used for self-assessment.

Self-assessment and can-do statements

There is a plethora of studies on self-assessment, and for those interested in summative literature Brantmeier, Vanderplank, and Strube (2012) and Andrade (2019) are recommended. Student self-assessment is a common educational practice that has been defined as a “descriptive and evaluative act carried out by the student concerning his or her own work and academic abilities” (Brown & Harris, 2013, p. 368) and as a “wide variety of mechanisms and techniques through which students describe (i.e., assess) and possibly assign merit or worth to (i.e., evaluate) the qualities of their own learning processes and products” (Panadero, Alonso-Tapia, & Huertas, 2012, p. 804). Andrade (2010, 2018, 2019) writes that self-assessment is feedback which should be used to inform adjustments to processes and products that deepen learning and enhance performance. Therefore, self-assessment should be formative, meaning for ongoing learning instead of being summative at the end of a project, unit, or course, because if it is not, then for the most part it lacks value.

When self-assessment is used in the classroom, researchers say that can-do statements should describe concrete tasks (Blanche & Merino, 1989) that are familiar to the target group, and self-assessment of language learner achievement needs to be designed according to specific curricular content to achieve “clear potential for predictive accuracy of criterion skills based on self-assessment measures” (Ross, 1998, p. 17). Therefore, self-assessment using can-do statements based on specific curricular content that is familiar to the learners may have better predictive accuracy and may be a formative part of the learning process.

Self-assessment in EAL

The self-assessment literature on EAL learners indicates that speaking is more accurately assessed than writing (Rasch, 1979). Also, there is research that suggests EAL learner proficiency level has an effect on the accuracy of self-assessment, and that advanced learners know which skills they are better and worse at doing (Brantmeier et al., 2012; Heilenman, 1990; Oskarsson, 1984). In regard to correlations between test scores and self-assessments, the literature shows that correlations tend to be moderate or moderately strong (Brantmeier et al., 2012; Edele, Seuring, Kristen, & Stanat, 2015). Finally, research suggests EAL learners tend to underestimate their self-assessments, especially more proficient students underestimate their abilities (Heilenman, 1990; Oskarsson, 1984). Other research suggests self-assessed language skills are systematically biased in certain groups (Edele et al., 2015); specifically, Japanese EAL learners may self-assess more harshly than might be expected because of cultural issues with modesty (Matsuno, 2009). International comparison data support the finding that Japanese students report low self-efficacy (OECD, 2019).

TOEIC can-do statements and proficiency levels

The TOEIC Can-Do Guide (2000), published by The Chauncey Group International in coordination with ETS, states that (underscore added for emphasis):

This *TOEIC Can-Do Guide* allows users of the TOEIC test to link TOEIC scores to the activities that examinees may or may not be able to do in English. The tables in the *Guide* provide examples of the activities that examinees are likely to be able to perform in English given certain Reading Comprehension scores and Listening Comprehension scores (p. 1).

In this statement, a distinction is made between users and examinees. Later, these roles are more concretely defined.

The tables on the following pages allow users of the TOEIC test to determine which English-language activities a test taker can be expected to perform given a certain TOEIC score. This is generally how decision-makers in organizations use TOEIC test scores: Test scores help them make decisions about such things as employee selection, placement, and training (p. 5). Note that “examinees” are referred to as “test takers” and “users” are more thoroughly defined as “decision-makers in organizations” who use the test scores to make human resources decisions.

The Chauncey Group International (2000) offers three caveats regarding “using the score interpretation information”, which are summarized here: Examinees self-rated their abilities to perform tasks, and “it is still unclear whether or not test takers can actually perform the English-language activities which they reported being able to do” (p. 1). The tables are based on average ratings of examinee groups. Therefore, “proficiency scales are useful for describing the abilities of a group of examinees, but can only describe what a given individual is likely to do, not necessarily what that person can actually do” (p. 3, underlined emphasis by The Chauncey Group International). This is soon followed by a similar remark: “There will be people in a given score range who may be able to do more, or fewer, tasks than are indicated. These tables are intended to serve as guidelines only and will not necessarily apply equally to every individual.” The sample is completely Japanese, yet it “may not be representative of the population of test-takers in Japan.” This is followed with “The information presented here may not apply to test takers from other countries, educational backgrounds, or industries. Decision-makers should be cautious when applying these score interpretations to other groups of employees” (p. 1).

Given these statements, (1) it is unclear if examinees can actually perform the activities, (2) these are guidelines and individual task performance will vary, and (3) the usefulness of can-do survey responses for validity purposes may be in question as the population the research was based on may not have been representative. To use an analogy, if a company employee was tasked with buying a fleet of vehicles for company use, and the salesman told him that the cars under consideration may or may not be able to be used for certain tasks, individual car performance would generally be fine but could vary individually, and the research on the cars may not represent these specific cars, it is doubtful that the employee would purchase those vehicles.

To summarize, the above statements from The Chauncey Group International focus on two main issues: the users of the can-do statements and the disclaimers regarding the research. From the above, it can be clearly stated that can-do statements are not for the examinees (test takers) but for decision-makers in organizations that use TOEIC results, and that these decision-makers “should be cautious” as examinees may or may not be able to actually perform the activities, as research methods studied groups and did not account for differences in individual performance, and as the research may or may not be representative of people inside or outside of Japan.

In 2008, a large-scale validity study was undertaken in coordination with ETS in Korea and Japan (ETS, 2018, 2019, 2021; Powers, Kim, Yu, Weng, & VanWinkle, 2009, 2010). The study used self-assessment questionnaires, a speaking test, and a writing test, and the results reportedly showed “TOEIC scores related relatively strongly to test-taker self-reports for both speaking and writing tasks” (Powers et al., 2009, p. i). This was later reworded as “moderately strong relationships of TOEIC scores with test takers’ self-reports... for both the speaking and writing language domains” (ETS, 2019, p. 20). TOEIC speaking scores and self-assessments of the ability to perform speaking tasks are correlated at ($r = 0.54$), and writing scores and the ability to perform everyday writing tasks correlate at ($r = 0.52$) and noted as correlating “reasonably strongly” (ETS, 2019, p. 20). To affirm these correlations, the ETS User Guide cites Cohen (1988) to support the belief that correlations of 0.5 can be seen as large, and then it cites Kuncel and Hezlett’s (2007) use of a 0.45 correlation between tests and first-year GPA.

One point of confusion with the original studies (Powers et al., 2009, 2010) is that both end with the following sentences. It is stated that “In the present study, we are not able to evaluate the soundness of test-taker self-reports as a validity criterion. However, in comparable studies that we have conducted recently in similar contexts, can-do self-reports have exhibited several characteristics that suggest reasonably trustworthy validity criteria” (2009, p. 14). However, the summation in the next paragraph seems to contradict what was previously stated with “In conclusion, the current study provides evidence of the validity of the TOEIC Speaking and Writing tests’ scores by linking them to test takers’ assessments of their ability to perform a variety of everyday (often job-related) English-language activities” (2009, p. 15).

CoE can-do statements and the common reference levels

The CoE (2001), which created the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR), clearly states numerous times that the can-do items are for all stakeholders, including the test takers, and they were purposely designed to be easily understandable by the test takers. They write that “the categorisation and description needs to be theoretically grounded... the description must also remain user-friendly – accessible to practitioners. It should encourage them to think further about what competence means in their context” (p. 21). The document also states that the can-do descriptors are provided for the three metacategories of reception, interaction and production (p. 25). Readers should refer to the Common Reference Levels self-assessment grid (p. 26) for examples. In Appendix D, the CoE states again that the can-do items are “user-oriented scales” that “assist communication between stakeholders in the testing process” and that they are a tool which can be used as a checklist of what language users can do and thus define the stage they are at” (p. 244).

Non-ETS related TOEIC can-do research

ETS has many publications referring to can-do statements and their tests. However, non-ETS related publications on the use of TOEIC can-do statements predictive accuracy of TOEIC scores or other can-do statements and their predictive accuracy on other standardized tests are somewhat limited.

In regard to reading and listening, Yoshizawa (2009) redesigned TOEIC can-do statements to fit the curriculum and researched test score correlations for first-year EAL university students ($n = 151$), which were 0.39 and 0.31, respectively. These were lower than the 0.46 and 0.53 in Powers, Kim, and Weng (2008), but this could be due to a neutral response being added to the TOEIC can-do survey tool. Runnels (2016) noted that some universities have shifted away from TOEIC use, for reasons including that interpretability of its scores is problematic and limited in its usefulness for test takers (p. 2), toward the CEFR “to create a synergy between the three areas of learning, teaching and assessment, resulting in a positive learning and assessment culture” (p. 3). Results from this correlation study on university students’ ($n = 57$) TOEIC scores and CEFR can-do self-assessment scores indicated that a moderate correlation was found for listening, but not for reading.

A few studies have called into question the validity of the TOEIC SW. Irwin and Nagy (2011) question the validity of the TOEIC SW based on can-do questionnaires saying, “Relying on can-do self-evaluations as a measure of validity is untenable. Future TOEIC SW Test research should explore the extent to which this test actually assesses the skills and constructs that ETS claims it measures” (p. 14). Uenishi et al. (2017) question the validity of the tests due to the fact that they are taken using a computer, which removes factors of human interaction (e.g., communication strategies and negotiation of meaning).

The present study

To sum up the introduction, the TOEIC SW literature does not directly state that can-do statements can be used by examinees for self-assessment. Therefore, because no studies have sought to replicate the Powers et al. (2009) validation study, the authors decided to partially replicate it on a small scale using the original survey questions for speaking and writing and newly collected test score results. The following research questions were thus formulated:

1. When replicated, what will the correlations be between test-taker TOEIC SW results and scaled self-assessment can-do statement survey results?
2. Using scaled self-assessment can-do surveys, to what degree can test takers accurately self-assess their speaking and writing skills to predict their TOEIC SW scores?
3. Will any other patterns emerge from the data?

Methods

Participants

The participants in this study took part in a voluntary extra-curricular graduate school English program (Harrison et al., 2022) in which the TOEIC Speaking and Writing test was offered. All 32 test takers were contacted by email and asked to take the skills self-assessment surveys (described below) via Google Forms. In total, 23 participants (22 male, 1 female) who were in their first or second year of a master's degree program in Japan completed the surveys and then the test.

The self-assessment survey tool, response analysis and response scaling

The Powers et al. (2009) survey had test takers of the TOEIC Speaking and Writing tests perform a bilingual self-evaluation, of “their ability to perform a variety of everyday (often job-related) English-language activities” (Powers et al., 2009, p. 15). As the original survey could not be procured from ETS, the self-assessment tool in Appendices A and B was recreated from ETS (English) and IIBC (Japanese) publications (Powers et al., 2009; IIBC, 2018). The tool was updated due to issues with jargon and clarity as well as discrepancies such as missing items, altered items, and translation discrepancies in items and choices. These are given in Appendix C. The tool had two parts: a 39-item speaking skills self-assessment survey and a 29-item writing skills self-assessment survey.

The authors followed the design of the original study and participants responded on a 0–5 scale:

- 0 = 答えられない (私に関係ない・判断できない) cannot answer: does not apply to me/unable to judge
- 1 = 全くできない not at all
- 2 = 非常に難しい with great difficulty
- 3 = ある程度難しい with some difficulty
- 4 =それほど難しくなく with little difficulty
- 5 = 簡単 easily

The original study matched activities that test takers in certain score proficiency levels claimed to be “able to perform, are likely to be able to perform with difficulty, and are unlikely to be able to perform” (p. 12). This was done by coding (0–5) the responses of participants with certain test scores into three categories: PND (Probably cannot do, responses 0–2), PCDwdiff (Probably can do with difficulty, responses 3–4), and PCD (Probably can do, response 5). The original study tabulated responses of test takers in each proficiency level resulting in a total percent of test takers who probably can do, probably can do with difficulty, and probably cannot do particular activities.

The results of the original research and subsequent publications (Powers et al., 2009; IIBC, 2018) placed certain skills within certain proficiency levels. Table 1 indicates the score ranges for each proficiency level. Using those results, the authors reverse engineered scales which could potentially be used to predict scores based on self-analyses using the can-do statements. The self-assessment categories were scaled to find the method which correlated best with the test-taker TOEIC SW scores and TOEIC proficiency level results. The reason that correlations were analyzed for both scores and proficiency levels was to see if differences existed. ETS reports both as results to test takers. Therefore, both seem to have value in the eyes of ETS, possibly test takers, and other stakeholders.

Self-assessment survey scaling methods

The original study (Powers et al., 2009, pp. 7-13) indicated which can-do statements were associated with each proficiency level as well as the rationale and verification process for the placements. Using those can-do item placements and the proficiency levels given by ETS and IIBC, the following scales were reverse engineered. The participant self-analysis results and the test results were used to verify which scale is potentially more accurate. Descriptions of the two scaling methods are given below.

Level midpoint average scale

The first of the two scales in this study is the Level Midpoint Average (LMA) scale. The rationale behind this scale is that the values of the midpoints of the given ETS/IIBC proficiency levels for each self-assessment item can be averaged to indicate a value that is potentially near the actual test result (score or proficiency level).

To calculate the LMA scaled scores, the midpoints of each level, Levels 1–8 for speaking and Levels 1–9 for writing, on the 200-point scale were used as a two-tier scale with responses for “Probably can do” and “Probably can do with difficulty” being awarded points, as shown in Table 1. The total points for all levels were then averaged. For example, if survey takers marked that they “Probably can do” a speaking item under Level 4 (scaled score 80–100), then they received 90 points for that item as that is the midpoint of that level. The same item is listed under “Probably can do with difficulty” in Level 2. Therefore, survey takers who marked that choice would receive 45 points for that survey item as that is the midpoint score of Level 2 (scaled score 40–50). Points for all items were then averaged to calculate the LMA self-assessment scores.

The same method was used for the scaling of proficiency levels. The actual proficiency level numbers were used for PCD and PCDwdiff calculations. These were then averaged to get the LMA self-assessment levels. For example, if survey takers marked that they “Probably can do” a speaking item under Level 4, then they received 4 points for that item. The same item is listed under “Probably can do with difficulty” in Level 2. Therefore, that choice would receive 2 points. Points for all items were then averaged to calculate the LMA self-assessment proficiency levels.

Table 1

Level midpoint average scale (number of survey items at that level)

Speaking (39 items)			Writing (29 items)		
Prof. level (Range)	PCD	PCDwdiff	Prof. level (Range)	PCD	PCDwdiff
8 (190–200)	195 (5)		9 (200)	200 (10)	
7 (160–189)	175 (15)		8 (170–199)	185 (10)	
6 (130–159)	145 (5)	145 (1)	7 (140–169)	155 (9)	155 (6)
5 (110–129)	120 (12)	120 (10)	6 (110–139)		125 (11)
4 (80–109)	95 (2)	95 (8)	5 (90–109)		100 (12)
3 (60–79)		70 (15)	4 (70–89)		
2 (40–59)		50 (5)	3 (50–69)		
1 (0–39)			2 (40–49)		
			1 (0–39)		

Note. See IIBC (2018) and Powers et al. (2009) for which items are situated at each level.

Equal weight sum scale

The second scale used in this research assigned equal weight to each item and then took the sum; hence it was termed the Equal Weight Sum (EWS) scale. The rationale behind the EWS scale is that, given a number of skills and a group of participants, a bell curve will form in regard to those who probably can do and probably cannot perform each skill. If the self-assessment items are equally weighted and then summed, the scale could potentially predict the number of skills that respondents indicated they can possibly perform. Those with fewer skills would have lower scores and those with more skills would have higher scores. A sum of the indicated skills could produce a value that is expected to correlate with actual test results (score or proficiency level).

There were 39 items on the 200-point speaking test and 29 items on the 200-point writing test, so to calculate the EWS scale each PCD item on the speaking test was 5.128 points and each of the 29 writing items was 6.896 points. PCDwdiff

scores were assigned half weight, 2.564 points and 3.448 points, respectively. Survey takers who marked that they probably could do items or probably could do items with difficulty were given their respective points, and then the scores were summed. For example, if survey takers marked that they probably could do 15 speaking items and probably could do 10 items with difficulty, then their EWS speaking score would be 103 ($[15 \times 5.128] + [10 \times 2.564]$).

For proficiency level analysis, the value of PCD for each of the 39 speaking items was 0.231 points, and for each of the 29 writing items it was 0.31 points. PCDwdiff level scores were assigned half weight, 0.116 points and 0.155 points, respectively. Survey response totals were summed as above.

Methods of obtaining TOEIC SW results

ETS and their affiliates score the speaking and writing tests separately. ETS provides test takers with both a speaking score and a writing score, and they also provide test takers with both a speaking proficiency level and a writing proficiency level. In this way, test takers can use their scores to understand where they stand within a proficiency level. Test scores and proficiency levels were used to scale the results as well as for correlations and other analyses.

Results

Statistical descriptions and distributions

Test and self-assessment results for both scores and proficiency levels for this research sample ($n = 23$) are shown in Table 2. For speaking scores and proficiency levels, the EWS scale results were more consistent with the test scores in regard to ranges and means. The LMA scale produced lower maximums and lower means, but the SDs and SEMs were more consistent with test scores. For writing scores and proficiency levels, the LMA scale produced ranges that were more consistent with the test ranges. The mean score of the EWS scale was slightly closer to the mean of the test, but for proficiency levels the LMA scale mean was closer to the test results. The SDs and SEMs with these scales were similar.

Table 2

Descriptive statistics for TOEIC SW results and both self-assessment scales

	Scores			Proficiency levels		
	TOEIC	LMA	EWS	TOEIC	LMA	EWS
	Speaking					
Range	30–150	2–128	3–156	1–6	0–5	0–7
Mean	69	49	64	3.17	2.09	2.91
SD	28	33	42	1.17	1.35	1.88
SEM	6	7	9	0.24	0.28	0.39
	Writing					
Range	30–140	3–163	3–172	3–7	0–7	0–8
Mean	88	72	68	4.35	3.38	3.07
SD	33	52	52	0.87	2.37	2.36
SEM	7	11	11	0.18	0.49	0.49

Normal distributions have 68% of the scores within 1 SD, 95% within 2 SDs, and 99% within 3 SDs of the mean. The test speaking score distribution was near-normal with 61% of scores within 1 SD, but with writing scores 57% were within 1 SD, and 100% were within 2 SDs. For speaking self-assessments, both self-assessment scales revealed score results with near-normal distributions having 74% of scores within 1 SD. For writing scores, the LMA scale results had the same distribution as the test. The EWS scale writing score results were closer to a normal distribution with 61% within 1 SD, 96% within 2 SDs and 100% within 3 SDs.

Regarding the proficiency level distributions for speaking, the test results formed a near-normal distribution, but the data were closer to the mean with 83% within 1 SD. Both self-assessment proficiency level results were distributed the same with near-normal distributions with 74% within 1 SD. For the writing test results, the proficiency level distributions within 1, 2 and 3 SDs were 30, 70, and 96%, respectively, which means the data points were further from the mean than in the case of a normal distribution. Both of the proficiency level self-assessment result distributions showed 74% within 1 SD

and 100% within 2 SDs, which means overall the data were closer to the mean.

Correlations

The correlation (r) between TOEIC speaking and writing scale scores was 0.66 (ETS, 2021, p. 22), and the correlation of the participants' ($n = 23$) speaking and writing scores in this report was 0.78. This was perhaps due to the limited sample size and a narrower range of test-taker proficiencies.

ETS reported correlations between test results and self-assessments as 0.54 for speaking and 0.52 for writing (ETS, 2019). To answer RQ1 (When replicated, what will the correlations be between test-taker TOEIC SW results and scaled self-assessment can-do statement survey results?), correlations between LMA and EWS self-assessments and TOEIC SW scores and the proficiency levels for the current sample are shown in Table 3. Both the LMA and EWS scaled scores had higher correlations with the test scores than the self-assessment proficiency levels had with the test proficiency levels. The EWS scale scores and proficiency levels correlated slightly higher with test results than those of the LMA scale. The LMA speaking, LMA writing, and EWS writing self-assessments correlated significantly ($p < .05$), indicated with an asterisk in Table 3. Only the EWS speaking score and proficiency level self-assessments did not correlate significantly with the test results. Therefore, both LMA and EWS scales correlated highly with test results, and the LMA scale seems to be the better of the two scales as it correlated significantly with both speaking and writing scores and proficiency levels.

Table 3

Correlations of test and self-assessment results

	Scores				Proficiency levels			
	r	t	df	p	r	t	df	p
LMA speaking	0.73	4.18	22	0.0003*	0.65	4.78	22	0.00009*
EWS speaking	0.73	0.77	22	0.45	0.66	0.87	22	0.39
LMA writing	0.62	2.74	22	0.006*	0.47	2.16	22	0.02*
EWS writing	0.63	2.26	22	0.017*	0.54	2.97	22	0.004*

Note. Significant at $p < .05$

“True score” zone, proficiency level, and self-assessment zone results

ETS reports TOEIC SW score results on scales of 0–200 and by proficiency levels, and they also provide information on the standard error of measurement (SEM) in the *Score User Guide TOEIC Speaking & Writing Tests*. For the TOEIC SW, the SEM for speaking is 13 scale points, and for writing it is 17 points (ETS, 2021, p. 21). According to the guide, this “can be used to estimate the expected variation of each test taker’s observed scores around their ‘true score.’ ... the true score can be thought of as the average score a test taker would earn on a very large number of forms... and across a number of testing dates.” The guide further explains that scores “would be expected” to be within the SEM 68% of the time (p. 22).

To answer RQ2 (Using scaled self-assessment can-do surveys, to what degree can test takers accurately self-assess their speaking and writing skills to predict their TOEIC SW scores?), the percentage of survey takers who self-assessed within the true score zone (1 SEM) of their test scores was 26–39%, and 9–35% self-assessed within the proficiency levels reached through the tests, shown in Table 4.

ETS used “true score” zones of 1 SEM. If the SEMs of the self-assessment scales, given in Table 2, are considered as “true score” zones, then 52–57% of test takers accurately self-assessed speaking and 43% accurately self-assessed writing.

Table 4

Percentage of participants within the true score zone or proficiency level

	Speaking			Writing		
	True zone (SEM 13)	Proficiency level	True + SA zone	True zone (SEM 17)	Proficiency level	True + SA zone
LMA	39	35	57	30	9	43
EWS	35	35	52	26	17	43

Note. Self-assessment is abbreviated as SA.

Although not shown in the table, if self-assessments for both speaking and writing are considered together, both the LMA and EWS score results indicated that 17% of survey takers self-assessed both tests accurately, and those that assessed accurately were of high, middle, and low proficiencies. Regarding proficiency level assessment with the LMA scale, no one self-assessed accurately for both speaking and writing. With the EWS scale, 9% of self-assessments were accurate for both speaking and writing, and they were in the middle- and high-proficiency ranges.

When TOEIC true score zones and speaking and writing self-assessment scale true score zones of 1 SEM were considered together, 26% of participants on the LMA scale and 22% of participants on the EWS scale self-assessed both speaking and writing accurately. These participants were in the top and bottom, not middle, proficiency groupings. This suggests that in total 22–26% of participants self-assessed accurately on both speaking and writing, and an additional 17–31% of participants assessed either speaking or writing accurately.

Speaking result visualizations

Answers to RQ3 (Will any other patterns emerge from the data?), are discussed forthwith. Figure 1 shows test-taker TOEIC scores, the “true score” zones, and the LMA and EWS self-assessment scores, and Figure 2 shows the proficiency levels. Participant order is from high to low based on speaking score test results. As can be seen, self-assessment scores with both methods fluctuated above and below the true score zone.

Figure 1

Speaking test scores and true score zone with LMA and EWS score self-assessments

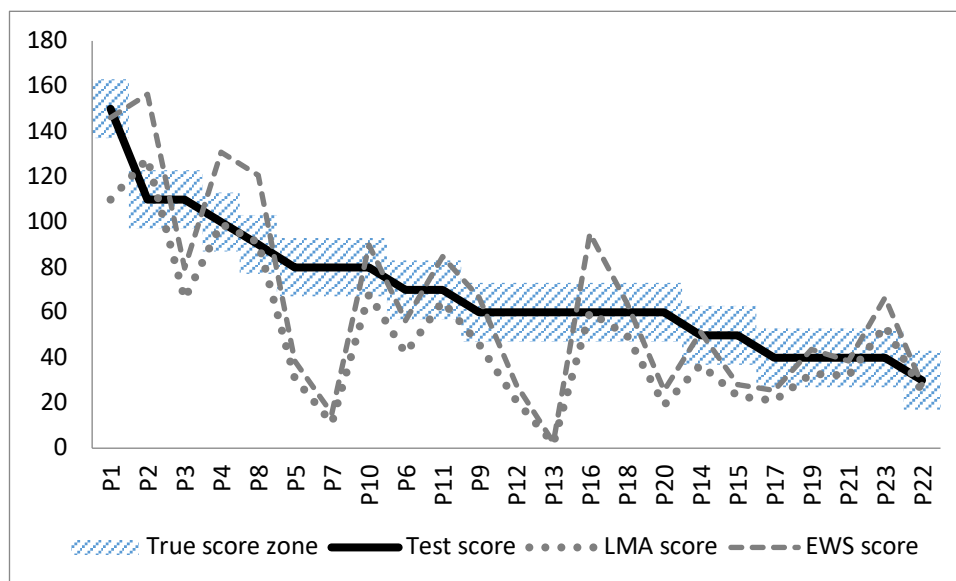
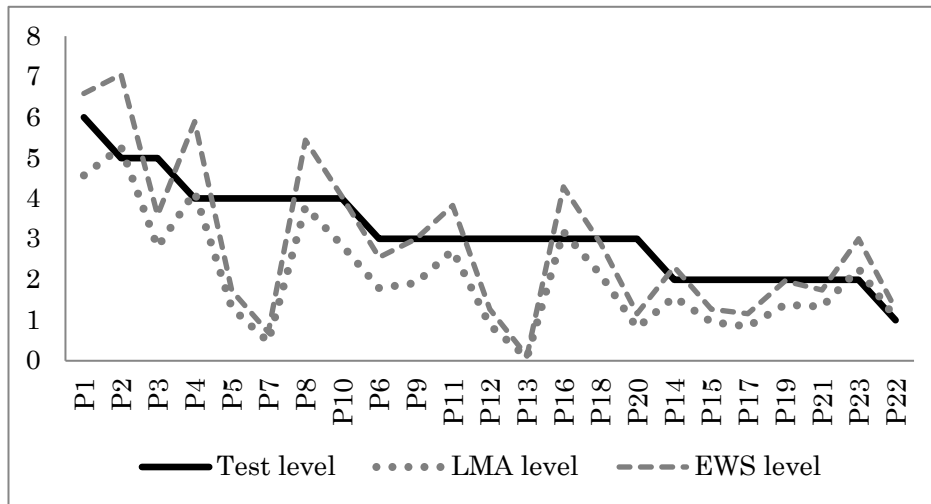


Figure 2

Speaking test, LMA, and EWS proficiency levels



Writing result visualizations

Results for writing are shown in Figures 3 and 4. As the speaking and writing tests were separate, the order of participants is from high to low based on writing test results, but participant numbers (e.g., P1, P2, are in order by writing score). Writing self-assessment scores with both methods fluctuated above and below the true score zone.

Figure 3

Writing test scores and true score zone with LMA and EWS score self-assessments

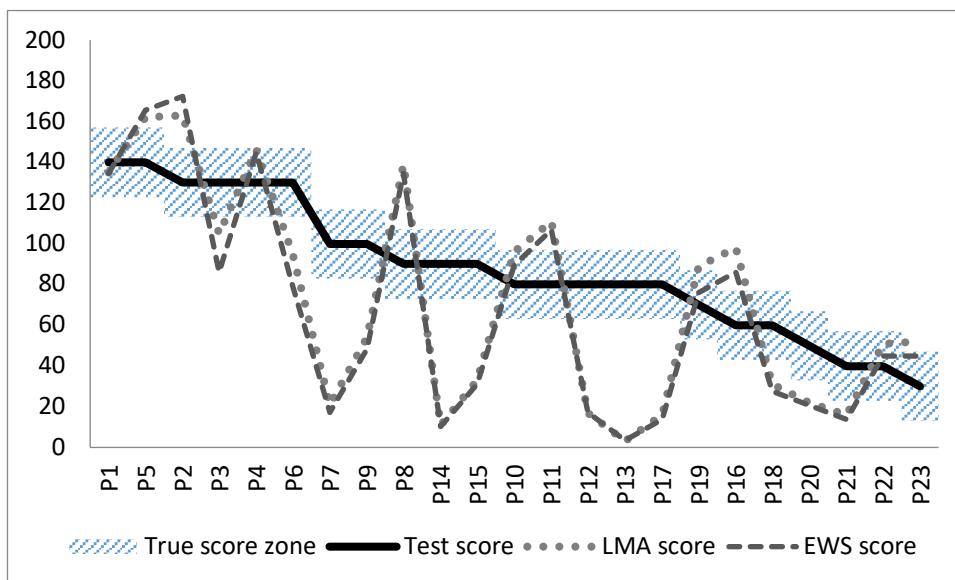


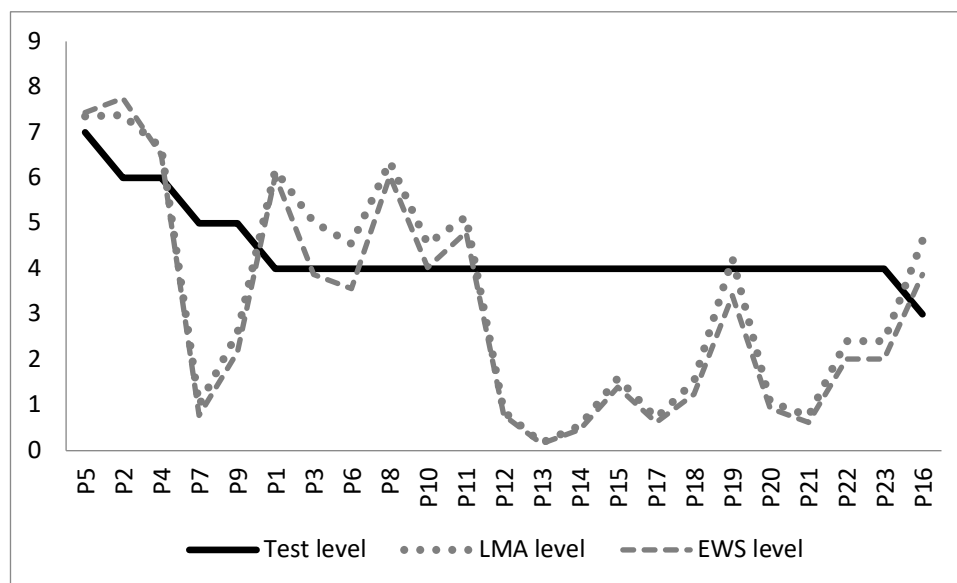
Figure 4*Writing test, LMA, and EWS proficiency levels*

Figure 4 shows that the majority of participants in the middle and bottom proficiency groupings self-assessed very low, but the reasons were unclear. The majority of items for writing are set at the higher proficiencies (Levels 7–9), and it could be that survey takers became tired with the length of the survey or discouraged with the difficulty of the items and filled in answers leading to scores of zero under “Probably cannot do.”

Differences between self-assessment and test results

The differences between self-assessments and test results per participant were calculated to understand if participants were underestimating or overestimating their results. The results in Table 5 indicated that overall, on both scales participants underestimated their results. The minimums of the ranges, which were negative, were further from zero than the maximum values, and the mean values for scores and proficiency levels were all negative.

Table 5*Differences between self-assessments and test results*

	Scores		Proficiency levels	
	LMA	EWS	LMA	EWS
Speaking				
Range	-69–18	-65–46	-3.54–0.26	-3.31–2.05
Mean	-20	-5	-1.08	-0.26
SD	22	29	1.06	1.42
Writing				
Range	-81–38	-83–44	-4.00–2.36	-4.23–2.05
Mean	-22	-20	-0.97	-1.28
SD	38	41	2.1	2.02

The results in Table 5 indicated that the self-assessment results generated with these scales tended to be underestimations, and for both scales raw data tallies indicated more students underestimated in all cases. With the LMA scale, 87% ($n = 20$) for speaking and 57% ($n = 13$) for writing self-assessed lower than their scores. With the EWS scale, respectively, 52% ($n = 12$) and 57% ($n = 13$) self-assessed low. Regarding proficiency level assessment for speaking with the LMA scale 65% ($n = 15$) underestimated, and 35% ($n = 8$) were perfect. For speaking levels with the EWS scale, 35% ($n = 8$) were below, 35% ($n = 8$) were in alignment with the true score zone, and 30% ($n = 7$) were high. For proficiency levels of writing with the LMA scale, 52% ($n = 12$) were below, 9% ($n = 2$) were perfect, and 39% ($n = 9$) were high. For writing levels with the

EWS scale, 57% ($n = 13$) were below, 17% ($n = 4$) were perfect, and 26% ($n = 6$) were high.

To analyze the data by proficiency groupings, the participants were divided into roughly equal groupings based on their test results. These groupings of top, middle, and bottom are relative to each other in this analysis and not to proficiency levels in general or as defined by any organization. Table 6 shows results with groupings of 7, 9, and 7 participants; however, groupings of 8, 7, and 8 participants yielded similar results. Results in Table 6 indicated that when analyzed by proficiency groupings, all proficiency groups underestimated their test results for both scores and proficiency levels. Regarding scores, for both speaking and writing the middle-proficiency groups underestimated their scores the most, followed by the top and then the bottom groups. Regarding proficiency levels, the EWS speaking scale followed this trend, but the LMA speaking scale showed the amount of underestimation in descending order from top to bottom groups. For writing proficiency levels, the bottom and middle group participants underestimated more than the top proficiency groups, and this was believed to be because the majority of these participants tested into the same proficiency level (Level 4).

Table 6

Differences between self-assessment and test results by proficiency grouping

	Speaking				Writing			
	LMA scores	EWS scores	LMA proficiency level	EWS proficiency level	LMA scores	EWS scores	LMA proficiency level	EWS proficiency level
Top ($n=7$)	-22	-3	-1.38	-0.14	-20	-14	-0.12	-0.35
Middle ($n=9$)	-23	-8	-1.29	-0.54	-37	-35	-1.28*	-1.57*
Bottom ($n=7$)	-9	-2	-0.51	-0.04	-4	-5	-1.43*	-1.84*

Note. Proficiency level is abbreviated as *prof. lvl.* *Test results place the majority of these participants at proficiency level 4.

Results for absolute values of the differences between self-assessments and tests

The absolute values of the differences between self-assessments and test results were calculated to measure the distance by which participants were either underestimating or overestimating. Regarding scores, for speaking, self-assessments were on average 23 points off and for writing 36–38 points off the mark. Regarding proficiency levels, for speaking they were on average 1.11–1.16 levels off and for writing 2.02 levels different from the test results. This demonstrated that participants more accurately self-assessed their speaking than their writing in regard to both scores and proficiency levels.

Table 7

Differences between self-assessment and test result absolute values

	Scores		Speaking	Proficiency levels	
	LMA	EWS		LMA	EWS
Range	0–69	1–65		0.05–3.54	0.01–3.31
Mean	23	23		1.16	1.11
SD	19	18		0.98	0.92
			Writing		
Range	2–81	5–83		0.22–4.00	0.03–4.23
Mean	36	38		2.02	2.02
SD	25	24		1.14	1.29

When results were calculated by proficiency levels with nearly equal numbers of participants in each group, the results, in Table 8, indicated that for speaking scores and proficiency levels, the bottom group self-assessed most accurately, then the middle group and then the top group. Regarding self-assessment scores for writing, the results showed that the bottom group self-assessed most accurately, then the top group, and then the middle group. Yet, for proficiency levels the top group was the most accurate, followed by the bottom group, and then the middle group. However, the bottom and the middle group were mostly the same level according to the test results. For this reason and due to the limited sample size, these results were not considered conclusive for writing. Groupings of 8, 7, and 8 participants yielded similar results.

Table 8

Differences between self-assessment and test result absolute values by proficiency grouping

	Speaking				Writing			
	LMA scores	EWS scores	LMA proficiency level	EWS proficiency level	LMA scores	EWS scores	LMA proficiency level	EWS proficiency level
Top (n=7)	29	32	1.50	1.85	33	40	1.71	1.70
Middle (n=9)	22	24	1.33	1.02	46	51	2.30*	2.21*
Bottom (n=7)	17	13	0.60	0.49	27	26	1.95*	2.09*

Note. *Test results placed the majority of these participants at proficiency level 4.

Discussion

Based on the results presented in this study, strengths and weaknesses of the self-assessments can be discussed. First, to review the answer to RQ1, self-assessments using both the LMA and the EWS scales correlated highly with TOEIC SW scores and proficiency levels, and the correlations were significant for both speaking and writing using the LMA scale. Correlations between test scores and self-assessments in this study were moderate to strong, which is in line with previous studies.

Second, as a review of the answer to RQ2, we will focus on the LMA scale because it outperformed the EWS scale. The percentage of survey takers who self-assessed within the true score zone (1 SEM) of their test scores was 30-39% and within their proficiency level it was 9-35%. When “true score” zones of 1 SEM were considered for both the tests and the self-assessments, then 57% of test takers accurately self-assessed speaking and 43% accurately self-assessed writing. These results are consistent with the literature (Rasch, 1979), as EAL learners more accurately assessed speaking than writing.

Finally, regarding RQ3, there were two other patterns that emerged from the data. The first was that with writing, the majority of participants in the middle and bottom proficiency groupings self-assessed very low, but the reasons were unclear. The self-assessment results generated with both scales tended to be underestimations. With the LMA scale, 87% ($n = 20$) for speaking and 57% ($n = 13$) for writing self-assessed lower than their scores. This underestimation by EAL learners is consistent with other studies (Edele et al., 2015; Heilenman, 1990; Matsuno, 2009; OECD, 2019; Oskarsson, 1984).

The second was a pattern of irregularity regarding participant results when grouped by proficiency (top, middle, and bottom). For writing proficiency levels in this study, the top group self-assessed most accurately, which follows other studies (Brantmeier et al., 2012; Heilenman, 1990; Oskarsson, 1984) that suggest that EAL learner proficiency level has an effect on accuracy of self-assessment and that advanced learners know which skills they are better and worse at doing. However, contradicting previous studies, writing score results and speaking score results in this study demonstrate that the bottom group self-assessed most accurately. This could be due to the small sample size and that in this study proficiency groupings were used and levels of proficiency were not defined by any outside criteria.

Limitations

This study has multiple limitations. First, this is only a partial replication study, as the original study was used to set the can-do items within the proficiency levels to indicate to stakeholders what skills test takers may or may not actually possess at those levels. This study used the can-do items and participant test results to reverse engineer scales which could potentially be used by learners and educators in combination with the can-do items to predict test results.

Second, results may differ because the original survey in Powers et al. (2009) had to be recreated and differences in wording and translations between the original English and IIBC publications in Japanese had to be reconciled. Items also had to be adjusted due to changes in society and technology. This could have affected the results.

Third, in this study the sample size was limited, and the testing situation was extracurricular. The self-assessment accuracy would probably improve with a larger sample size of learners who are familiar with the concrete tasks listed in the can-do items, via curricular content and experience. If the self-assessment could be designed to be a formative part of the learning process, it would benefit the learners more.

Conclusion

In summary, the background literature indicated that TOEIC can-do statements are for decision-makers in organizations and not specifically for test takers. This study indicated that TOEIC SW can-do statement survey responses scaled using the LMA scale could be used for predictive purposes because they correlated strongly with TOEIC SW scores and moderately with proficiency levels. These correlations were significant at $p < .05$. Regarding predictive accuracy of self-assessments using the LMA scale and true score zones of 1 SEM, participants accurately predicted their TOEIC SW scores 57% of the time for speaking and 43% for writing. Further research is needed with a larger sample of participants to confirm the accuracy and reliability of the LMA scale developed in this study. In the future, if such research supports the use of the LMA scale, then claims could be made by educators and by ETS that for test takers the can-do statements can be useful at a particular degree of predictive accuracy.

References

- Andrade, H. (2010). Students as the definitive source of formative assessment: academic self-assessment and the self-regulation of learning. In H. Andrade and G. Cizek (Eds.) *Handbook of Formative Assessment* (pp. 90-105). Routledge.
- Andrade, H. L. (2018). Feedback in the context of self-assessment. In A. Lipnevich and J. Smith (Eds.) *Cambridge Handbook of Instructional Feedback* (pp. 376-408). Cambridge University Press.
- Andrade, H. L. (2019). A critical review of research on student self-assessment. *Frontiers in Education*, 4, 87, 1-13. <https://doi.org/10.3389/educ.2019.00087>
- Blanche, P., & Merino, B. (1989). Self-assessment of foreign-language skills: Implications for teachers and researchers. *Language Learning*, 39 (3), 313-338. <https://doi.org/10.1111/j.1467-1770.1989.tb00595.x>
- Brantmeier, C., Vanderplank, R., & Strube, M. (2012). What about me?: Individual self-assessment by skill and level of language instruction. *System*, 40 (1), 144-160.
- Brown, G. T., & Harris, L. R. (2013). Student self-assessment. In J. H. McMillan (Ed.) *Sage Handbook of Research on Classroom Assessment* (pp. 367-393). Sage. <https://dx.doi.org/10.4135/9781452218649.n21>
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Erlbaum.
- Council of Europe. (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.
- Edele, A., Seuring, J., Kristen, C. & Stanat, P. (2015). Why bother with testing? The validity of immigrants' self-assessed language proficiency. *Social Science Research*, 52, 99-123.
- ETS. (2018). *TOEIC User Guide Speaking & Writing Tests*. Educational Testing Services.
- ETS. (2019). *Score User Guide TOEIC Speaking & Writing Tests*. Educational Testing Services.
- ETS. (2021). *Score User Guide TOEIC Speaking & Writing Tests*. Educational Testing Services.
- Harrison, J., Suzuki, T., & Falout, J. (2022). 2018-2019 Graduate School English Speaking and Writing Proficiency Program. *日本大学理工学部一般教育教室彙報 (Nihondaigaku Rikougakubu Ippan Kyouiku Ihou, Bulletin of the Department of General Education of the College of Science and Technology at Nihon University)*, 112, 1-8.
- Heilenman, L. K. (1990). Self-assessment of second language ability: the role of response effects. *Language Testing*, 7(2), 174-201. <https://doi.org/10.1177/026553229000700204>
- IIBC. (2018). *TOEIC Speaking & Writing Tests のご案内* (No goannai, A guide to). The Institute for International Business Communication.
- Irwin, B., & Nagy, P. (2011). The 2010 revision of the TOEIC speaking test. *SHIKEN: JALT Testing & Evaluation SIG Newsletter*, 15(2), 10-19.
- Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science*, 315, 1080-1081.
- Matsuno, S. (2009). Self-, peer- and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing*, 26 (1), 75-100.
- OECD. (2019). *PISA 2018 Results (Volume III): What School Life Means for Students' Lives*. PISA, OECD Publishing. <https://doi.org/10.1787/acd78851-en>
- Oskarsson, M. (1984). *Self-Assessment of Foreign Language Skills: A Survey of Research and Development Work*. Council for Cultural Cooperation.

- Panadero, E., Alonso-Tapia, J., & Huertas, J. A. (2012). Rubrics and self-assessment scripts effects on self-regulation, learning and self-efficacy in secondary education. *Learning and Individual Differences*, 22, 806–813. <https://doi.org/10.1016/j.lindif.2012.04.007>
- Powers, D. E., Kim, H.-J. & Weng, V. Z., (2008). *TOEIC Compendium Study: The Redesigned TOEIC (Listening and Reading) Test: Relations to Test-Taker Perceptions of Proficiency in English*. (ETS Research Rep. RR-08-56). Educational Testing Services. <https://www.ets.org/Media/Research/pdf/TC-10-06.pdf>
- Powers, D. E., Kim, H.-J., Yu, F., Weng, V. Z., & VanWinkle, W. (2009). *The TOEIC Speaking and Writing Tests: Relations to Test-Taker Perceptions of Proficiency in English*. (ETS Research Rep. RR-09-18). Educational Testing Services. <https://doi.org/10.1002/j.2333-8504.2009.tb02175.x>
- Powers, D. E., Kim, H.-J., Yu, F., Weng, V. Z., & VanWinkle, W.H. (2010). *The TOEIC Speaking and Writing Tests: Relations to Test-Taker Perceptions of Proficiency in English*. Compendium Study. Educational Testing Services.
- Rasch, G. (1979). *Probabilistic Models for Some Intelligence and Achievement Tests*. Danish Institute for Educational Research (Expanded edition, 1980. University of Chicago Press, Chicago).
- Runnels, J. (2016). Self-assessment accuracy: Correlations between Japanese English learners' self-assessment on the CEFR-Japan's can-do statements and scores on the TOEIC. *Taiwan Journal of TESOL*, 13(1), 105-137.
- Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experimental factors. *Language Testing*, 15, 1-20.
- The Chauncey Group International. (2000). *The TOEIC Can-Do Guide: Linking TOEIC Scores to Activities Performed Using English*. https://www.ets.org/Media/Research/pdf/TOEIC_CAN_DO.pdf
- Uenishi, K., Sakaue, T., Lauer, J., Davies, W., Fraiser, S., Howell, P., Selwood, J., Song, K., Morita, M., & Kida, S. (2017). Integrating the TOEIC speaking test with small group classes. *広島外国語教育研究 (Hiroshima Gaikokugo Kyouiku Kenkyu, Hiroshima Studies in Language and Language Education)*, 20, 1-15. <https://doi.org/10.15027/42610>
- Yoshizawa, K. (2009). To what extent can self-assessment of language skills predict language proficiency of EFL learners in school context in Japan? *関西大学外国語教育研究 (Kansai Daigaku Gaikokugo Kyouiku Kenkyu, Kansai University Research in Foreign Language Education)*, 17, 65-82.

Appendices

Appendix A: Self-Report on English Language Speaking Skills for Business

ビジネス場面における英語のスピーキングスキルに関する自己報告

Self-Report on English Language **Speaking** Skills for Business

- 0 = 答えられない (私に関係ない・判断できない) cannot answer: does not apply to me/unable to judge
 1 = 全くできない not at all
 2 = 非常に難しい with great difficulty
 3 = ある程度難しい with some difficulty
 4 =それほど難しくなく with little difficulty
 5 = 簡単 easily

英語のスピーキングなら、私は: Regarding my **English speaking**, I can:

1. メニューを見ながら、喫茶店やレストランで料理を注文することができる
use a menu, order food at a café or restaurant
2. 人と会う約束を取る・変更する・キャンセルすることができる
make/change/cancel an appointment to see a person
3. 非公式 (インフォーマル) の場面で、自分の趣味や故郷、興味のあることについて話すことができる
discuss my hobby, hometown, or interests when introducing myself in informal situations
4. トイレの場所の説明等、来訪者の案内をすることができる
show a guest around (e.g., where the restroom is)
5. 留守番電話に、折り返し電話をもらえるよう伝言を残すことができる
leave a message on an answering machine/voicemail to ask a person to call me back
6. 電話で自分の名前・用件を告げ、担当者への取り次ぎを依頼することができる
give my name and the reason I called and ask to connect me to a person in charge
7. レストランの営業時間を問い合わせたり、夕食の予約をしたりすることができる
request information about business hours of a restaurant and make dinner reservations
8. 最寄り駅から、特定の目的地 (オフィス等) までの道順を説明することができる
give directions on how to get to some destinations (e.g. the office) from the nearest station
9. 航空会社に電話して、飛行機の予約を変更することができる
telephone the airline to change my flight plans
10. 電話で、商品の注文や注文状況の問い合わせをすることができる
telephone a company to place (or follow-up) an order for an item
11. 社交の場面で、適切な挨拶・自己紹介をすることができる
use appropriate greeting expressions and introduce myself in social situations
12. 来訪者と商談前に、天気等の一般的な話題について、世間話をするすることができる
have “small talk” with a guest about topics of general interest (e.g. the weather) before discussing business
13. 海外からの来訪者に、ちょっとした観光案内をすることができる
give a brief sightseeing tour to a guest from abroad
14. 外国人の同僚や新人に、日常業務の手順を口頭で教えることができる
tell a foreign colleague or newly employed person how to perform a routine task
15. 自分の会社 (または大学) について (例: 主要事業は何か、どのような履修コースがあるか)、資料を見ながら手短かに説明することができる
briefly explain about my company (or university) while looking at some data (e.g. what the main business is or what courses it offers)
16. 将来のキャリア上の目標について話すことができる (例: 来年自分が何をしようとしているか話すことができる)
talk about my future professional goals (e.g., what I plan to be doing next year)
17. 今起きているトラブル (例: フライト・宿に関するトラブル) を係員に説明し、問題を解決するよう要求

- することができる
- explain ongoing troubles (e.g., about flight or hotel accommodations) and make a request to settle a problem
18. 自分の学歴や現在の職務を、フォーマルな場面で説明することができる
describe my academic training or my present job responsibilities in formal settings
19. 自分の意見を、異なる意見を持つ相手に説明することができる
express my opinion to others who may have a different view
20. 同僚や仲間に、使用方法を知っている機械や機器（例：コピー機、PC、オーディオプレーヤー）の操作の仕方を説明することができる
explain (to a co-worker or colleague) how to operate a machine or device (e.g., photocopier, PC, audio player) that I am familiar with
21. 非公式（インフォーマル）の場面で、会話等の通訳をすることができる
translate (e.g., conversations) in an informal setting
22. 外国人の従業員に、会社の歴史や規則について説明することができる
explain company history and rules (in English) to foreign employees
23. 自分がよく知っている仕事関連のプロセスについて、説明することができる
describe a familiar work-related process
24. その場に適した暗記したフレーズを使って、質問したり会話を行ったりすることができる
ask a question and talk by using memorized phrases and expressions correctly in appropriate situations
25. 上司や同僚と、顧客サービスや製品の質の向上について話し合うことができる
discuss the improvement of customer service/product quality with my boss/co-workers
26. 事業計画や方針について、新しい仲間や同僚に説明することができる
explain a business plan or policy to a new colleague or co-worker
27. 議論や会議の場で、自分の意見を述べたり力説したりすることができる
state and emphasize my opinion during a discussion or meeting
28. 議論の場で、相手の意見に対して自分の意見を述べたり反論したりすることができる comment on or react to someone's opinion during a discussion
29. 事前に準備をした上で、新規プロジェクトや新製品に関する 20～30 分のプレゼンテーションや、公式（フォーマルな）スピーチを行うことができる
give a prepared 20-30 minute presentation or formal talk (e.g., about a new project or product)
30. プレゼンテーションやスピーチで議論されている内容について、質問をしたり質問に答えたりすることができる
ask or answer questions about an issue being discussed in a presentation or speech
31. 講義やプレゼンテーションの際、不明瞭な問題点についてその場で質問することができる
ask questions to understand unclear or problematic points made in a lecture or presentation
32. 教師や同僚、友人、仲間等、聞き手の経歴や英語力に応じて話し方を変えることができる
adjust my speaking for different learners (e.g., instructors, co-workers, friends, colleagues, according to their backgrounds and English proficiency)
33. 事前に用意した資料を用いて、他社の製品と比較しながら自社の製品やサービスの品質、効率性、価格などを説明することができる
using prepared materials, discuss my company's products/services, comparing them with other companies' products in terms of quality, efficiency, and prices
34. 来訪者と世界の出来事について話すことができる
discuss (in English) world events with a guest
35. 少人数の歓送迎会で、事前準備なしに 2～3 分程度の短い挨拶やスピーチをすることができる
without any preparation, give a short (a few minutes) greeting or speech at a small welcome/farewell party
36. 講演やプレゼンテーション等、フォーマルな場面で通訳することができる
translate (e.g., lectures or presentations) in a formal setting
37. 自分が行った研究や調査に関して、結果を報告することができる
report on the outcomes of my research or investigation
38. 自分の専門分野に関係する話題について、分かりやすく伝えることができる communicate in an understandable way about topics related to my field or specialty
39. 不良品や不十分なサービスに対する苦情を処理することができる
deal with complaints about defective product or inadequate service

Appendix B: Self-Report on English Language Writing Skills for Business

ビジネス場面における英語のライティングスキルに関する自己報告

Self-Report on English Language **Writing** Skills for Business

- 0 = 答えられない (私に関係ない・判断できない) cannot answer: does not apply to me/unable to judge
 1 = 全くできない not at all
 2 = 非常に難しい with great difficulty
 3 = ある程度難しい with some difficulty
 4 =それほど難しくない with little difficulty
 5 = 簡単 easily

英語のライティングなら、私は: Regarding my English writing, I can:

1. ホテルの宿泊設備に関する情報を求める E メールを書くことができる
write an email requesting information about hotel accommodations
2. 仕事上の知り合い宛てに、プレゼントや食事に対する短い礼状を書くことができる
write a short thank-you note to a business acquaintance for a gift or dinner
3. 同僚または教師宛てに、会議または授業に出席できなかった理由を説明した簡潔なメモを書くことができる
write a brief note to a co-worker (or instructor) explaining why I was not able to attend a meeting (or class)
4. 歓送迎会のご案内の E メールを書くことができる
convey information in an email about a welcome or farewell party
5. 業務上または授業に関する確認のためのメモを書いたり、E メールを送ったりすることができる
write a memorandum or send an email confirming some information related to my job or class
6. 上司または教師宛てに、この一週間で達成したことを説明する E メールを書くことができる
write an email to my supervisor (or instructor) explaining what I have accomplished during the past week
7. 自分のオフィスへの道順を分かりやすく書くことができる
write clear directions on how to get to my office
8. 商品やサービスに対する、苦情の E メール・文書を企業に対して書くことができる
write a letter or send an email to a company to complain about a product or service
9. ある特定の商品やサービスに関する情報を得るために、企業に対し E メール・文書で問い合わせることができる
write a letter or send an email to a company to request information about a certain product or service
10. 公的機関に対し、必要な情報を得るために E メール・文書で問い合わせることができる send an email or write a letter to a public organization requesting necessary information
11. ビジネスレターの返事を書くことができる
write a reply to a business letter
12. 求人に応募するにあたり、自己紹介し、その仕事に自分が適任であることを説明する手紙を書くことができる
write a letter introducing myself and describing my qualifications for a job
13. 自分が作成した報告書内の図や表で示された情報について、説明文を書くことができる
describe (in writing) information presented in a table/chart of a report that I have written
14. 自分の専門分野に関連したグループ会議で話し合われた要点を、簡単なレポートやメモにまとめることができる
write a brief report (or memo) summarizing the main points discussed in a group meeting related to my field or specialty
15. ビジネスレターやマニュアル等の書類を、英語に翻訳することができる
translate documents (e.g., business letters, manuals) into English
16. 同僚に、使用方法を知っているコピー機や PC 等、オフィス機器の操作の仕方を説明する短いメモを書くことができる
write a short note (to a co-worker) describing how to operate an office machine or device (e.g., photocopier, PC) that I

am familiar with

17. 会議または授業中に議論された内容をメモし、要点をまとめることができる
write discussion notes during a meeting or class and summarize them
18. 上司または教師宛てに、プロジェクトまたは課題の進行状況を説明するメモを書くことができる
write a memorandum to my supervisor or instructor describing progress on a current project or task
19. 潜在顧客宛てに、自社の製品・サービスを説明する手紙を書くことができる
write a letter to a potential client describing the services or products of my company
20. 自分がよく知っている専門書に関して、1~2ページの要約を書くことができる
write a 1-2 page summary of a technical book that I am familiar with
21. 顧客に対する、フォーマルな礼状を書くことができる
write a formal letter of thanks to a client
22. 自分の専門分野内のよく知っている話題に関して、専門的な報告書を書くことができる
write a technical report on a familiar topic within my area of expertise
23. 自分の仕事に関連する、計画書や提案書を書くことができる
write a plan or proposal related to my work
24. 専門家による会議（または学会で）のプレゼンテーションで使用する文章やスライドを作成することができる
prepare text and slides (in English) for a presentation at a professional conference
25. 上司または教師から割り当てられた、調査報告書を書くことができる
write a research report assigned by a supervisor or instructor
26. 商品やサービスのセールスマニュアルを作成することができる
write a manual (for sales purposes) for a product or service
27. 新しい業務プロセスや新商品・新サービスを提案する企画書を書くことができる
write a proposal to suggest a new business process, a new product, or a new service
28. プロジェクトの進行状況を説明する、数ページにわたる正式な報告書を書くことができる
write a brief, several-page (formal) report explaining the progress being made on a current project
29. 新商品や新サービスを紹介するプレスリリースを書くことができる
create a press release to introduce a new product or service

Appendix C: Can-Do Item Edits in Self-Reports due to Variations in Sources

Regarding recreation of the survey, the Japanese Can-Do List (IIBC) was aligned with the English Can-Do List (Powers et al., 2009). The Can-Do Writing Lists by Powers and IIBC (2018) do contain the same 29 items, but the English and Japanese Can-Do Lists vary in the following ways (A-E):

- A. As the authors did not have the original format of the surveys, the answer choices were taken from Powers et al. (2010, p. 11.4) and translated by the authors, as shown below. The “0” choice was created to account for respondents being allowed to “omit a task statement if they thought it did not apply to them or they were unable to make a judgment” (Powers et al., 2010, p. 11.4). Item numbers represent the final randomized version order, and Japanese precedes English in these bilingual surveys as it is the first language of the majority of survey takers.

0=答えられない（私に関係な・判断できない） cannot answer: does not apply to me/unable to judge

1=全くできない not at all

2=非常に難しい with great difficulty

3=ある程度難しい with some difficulty

4=それほど難しくなく with little difficulty

5=簡単 easily

- B. Item 31 was in IIBC (2018), not in Powers et al. (2009), but something similar in Powers et al. (2010, Appendix B). 講義やプレゼンテーションの際、不明瞭な問題点についてその場で説明することができる

- C. Item 37 was in IIBC (2018), not in Powers et al. (2009), but in Powers et al. (2010, Appendix B). 自分が行った研究や調査に関し、結果を報告することができる report on the outcomes of my research or investigation

- D. Item 38 was in IIBC (2018), not in Powers et al. (2009), but in Powers et al. (2010) Appendix B. 自分の専門分野に関係する話題について、分かりやすく伝えることができる communicate in an understandable way about topics related to my field or specialty

- E. Item 39 was in IIBC (2018), not in Powers et al. (2009), but in Powers et al. (2010) Appendix B. 不良品や不十分なサービスに対する苦情を処理することができる deal with complaints about defective product or inadequate service

- F. Item 40 was in Powers et al. (2009, 2010), but not in IIBC (2018). The English was translated into Japanese, but then not included in the survey as not “everyday” item. ビジネス交渉や表敬訪問など、さまざまな場面で経営陣の通訳を務めることができる serve as an interpreter for top management on various occasions such as business negotiations and courtesy calls

A-F are the differences between the lists. However, upon inspection of the ETS and IIBC lists, the authors made the following adjustments to what they believe is more accurate translation, easier to understand, or for continuity of style. The translation changes were verified as “correct” by 2 other native Japanese professionals who said these changes “should be made.”

On the ビジネス場面における英語のスピーキングスキルに関する自己報告 / Self-Report on English Language Speaking Skills for Business, the following changes were made.

- A. In Item 1 the English was changed from “using” to “use.”
メニューを見ながら、喫茶店やレストランで料理を注文することができる using a menu, order food at a café or restaurant
- B. In Item 3 the Japanese was changed to “非公式（インフォーマル）...” as the IIBC (2018) translation uses フォー

マル instead of 公式, see Item 18.

非公式の場面で、自分の趣味や故郷、興味のあることについて話すことができる discuss my hobby, hometown, or interests when introducing myself in informal situations

- C. In Item 21 the Japanese was changed to “非公式（インフォーマル）...”, for the same reason as B above.
非公式の場面で、会話等の通訳をすることができる translate (e.g., conversations) in an informal setting
- D. In Item 23 the Japanese was changed to “自分が...”.
自分やよく知っている仕事関連のプロセスについて、説明することができる describe a familiar work-related process
- E. In Item 27 the Japanese was changed to “述べたり力説したりする”.
議論や会議の場で、自分の意見を述べたり力説することができる state and emphasize my opinion during a discussion or meeting
- F. In Item 29, the Japanese was changed to “公式（フォーマルな）”.
事前に準備をした上で、新規プロジェクトや新製品に関する 20～30 分のプレゼンテーションや、公式スピーチを行うことができる give a prepared 20-30 minute presentation or formal talk (e.g., about a new project or product)
- G. Item 30 was very similar to Item 31, but different. For our study, we used the Japanese from Item 31 and made a proper English translation.
プレゼンテーションやスピーチで議論されている内容について、質問をしたり質問に答えたりすることができる ask or answer questions about an issue being discussed in a presentation or speech
- H. In Item 31 the Japanese was changed to “質問”.
講義やプレゼンテーションの際、不明瞭な問題点についてその場で説明することができる (IIBC, 2018) ask questions to understand unclear or problematic points made in a lecture or presentation (Powers et al., 2010)
- I. In Item 37, the Japanese was changed to “関して”.
自分が行った研究や調査に関し、結果を報告することができる report on the outcomes of my research or investigation

On the “ビジネス場面における英語のライティングスキルに関する自己報告/ Self-Report on English Language Writing Skills for Business, the following changes were made.

- J. In Item 15, the English was English changed to “translate”.
ビジネスレターやマニュアル等の書類を、英語に翻訳することができる transcribe documents (e.g., business letters, manuals) into English
- K. In Item 21, the Japanese was changed to “フォーマル”.
顧客に対する、正式な礼状を書くことができる write a formal letter of thanks to a client
- L. In Item 24, “（または学会で）” was added to the Japanese both for accuracy of translation and as it suited the group of engineers being surveyed.
専門家による会議のプレゼンテーションで使用する文章やスライドを作成することができる prepare text and slides (in English) for a presentation at a professional conference