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

Peter O'Keefe[1] and David Allen[2]
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 (Bennesse, 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 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

	Frequencies (N)				Proportions (%)			
Item distribution	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 response[3] 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 level[5], 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 suggested[6] 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., & Louwerse, 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 kokuji) 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.