THE JOURNAL OF ASIA TEFL
Vol. 18, No. 1, Spring 2021, 208-222
http://dx.doi.org/10.18823/asiatefl.2021.18.1.12.208

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511

An analysis of Macau's Joint Admission Examination–English

Ally Oi Kuan Ho
University of Macau, Macau, China

Don Yao
University of Macau, Macau, China

Antony John Kunnan
University of Macau, Macau, China

This study examined the evidence for the validity of the content of the Joint Admission Examination-English (JAE-E) developed for university admissions in Macau. The areas of investigation were the linguistic characteristics of the test, the alignment between the test and CEFR levels, and the comparison between JAE-E and other admission examinations. Five experienced English teachers from Macau were recruited to provide expert judgments. In addition to the judgments, two instruments used were Coh-Metrix and CEFR scales with descriptors. Results showed that generally the JAE-E was valid in terms of its content. The Coh-Metrix results suggested that the tests were suitable in terms of narrativity, syntactic simplicity, word concreteness, and deep cohesion (although not in terms of referential cohesion). In terms of CEFR alignment, the test was at B1 level for reading and C1 level for writing. Finally, suggestions for improvements were put forward in terms of integrated skills, scenario-based assessment, and technology.

Keywords: Joint Admission Examination-English (JAE-E), content validity, expert judgment, Coh-Metrix, CEFR levels

Introduction

There was no centralized university admission examination system in Macau prior to the implementation of the Joint Admission Examination (JAE) in 2017. Admission to local universities was based on school-level curriculum and examinations which prepared students for different university entrance examinations. Students passing the individual secondary school assessments (equivalent to Grade 12) could choose to take the university entrance examination of the university they wanted to enter. According to Kin (2002), some of the schools set their goals towards the university entrance examinations in mainland China, the Hong Kong Certificate of Education Examination, or the Hong Kong Advanced Level Examination; others aimed for the U.S. SAT, the SAT Subject Test, or the Test of English as a Foreign Language (TOEFL).

Due to the universal dissatisfaction of this system of admission into the four Macau universities (i.e., University of Macau, Macau Institute for Tourism Studies, Macau Polytechnic Institute, and Macau University of Science and Technology), a new admission examination scheme entitled the *Joint Admission Examination for Macau's Four Higher Education Institutions* (JAE) was implemented for four subjects: Chinese, Portuguese, English, and Mathematics in 2017. This test has been jointly designed and

administered by a committee of about ten academic representatives from the four institutions. This study focused on an examination of the linguistic characteristics of the *Joint Admission Examination – English* (JAE-E), the alignment between JAE-E and the Common European Framework of Reference (CEFR) levels; and the comparison between JAE-E and other admission examinations.

Review of Literature

Constructs

Assessing language use (grammar and vocabulary)

The traditional discrete-point approach to grammar assessment - the most conventional way of assessing grammar - uses selected-response (SR) tasks to assess manipulations of discrete units of grammar. Specifically, SR tasks of grammatical knowledge measuring recognition or recall (i.e., receptive knowledge) usually involve one area of grammatical knowledge at a time but fail to capture dynamic and complex understandings of resources needed for communication. This is the type of task that is favored in the JAE-E.

Purpura (2014) stated that such conceptualizations of language knowledge were generally defined by second language (L2) educators "as a set of structural rules, patterns, norms, or conventions that govern the construction of well-formed and meaningful utterances with respect to specific language use contexts" (p. 2). Purpura also argued that language knowledge is the "interaction between grammatical knowledge and pragmatic knowledge" (p. 6). Grammar knowledge, he argued, is defined as "a range of linguistic forms" (e.g., *-s*, affix, word order, phonological, lexical, and morphosyntactic knowledge) and semantic meanings (such as the logical representation of literal and intended meanings conveyed by individual forms) associated with these forms, either individually (e.g., plurality with a noun) or collectively (e.g., the overall literal meaning of the utterance)…these forms and meanings occur at sub-sentential, sentential, and suprasentential or discourse levels" (p. 6). These are the targets of assessment in the JAE-E.

However, many L2 assessment experts believe "performance tasks" are the best for the assessment of grammatical ability. Purpura (2014) stated "where examinees are presented with input in the form of a prompt and are required to produce extended amount of spoken or written data, of which the quality and quantity can vary considerably among test takers" (p. 18).

In terms of vocabulary assessment, Read and Chapelle (2001) presented eight lexical measures with different inferences, uses and impacts. These include the Vocabulary Levels Test, the Lexical Frequency Profile, English as a Second Language (ESL) Composition Profile, a multiple-choice cloze test for proficiency testing, the C-test, the Vocabulary Knowledge Scale and the Lexical Density Index. They argue "a vocabulary test should require learners to perform tasks under contextual constraints that are relevant to the inferences to be made about their lexical ability" (p. 1).

Assessing reading

Grabe and Jiang (2014) stated "the construct of reading comprehension abilities, as well as the relevant component subskills, knowledge bases, and cognitive processes, had not been well thought out and convincingly described in assessment contexts until the 1990s" (p. 1). Reading assessment options typically used were often limited to multiple choice, true or false, matching, and brief open-ended response. The JAE-E had these task types.

From 1990 onward, research on the roles of various component subskills on reading performance, reading for different purposes (e.g., reading to learn, reading for general comprehension, and expeditious reading) have been categorized. With numerous empirical research studies on reading comprehension abilities, the consensus on reading assessment has emerged regarding componential skills, knowledge sources, and general cognitive abilities. Grabe and Jiang (2014) argued that key factors with strong

impact on reading abilities in both L1 and L2 contexts were "efficient word recognition processes", "a large recognition vocabulary", "the abilities to formulate the main ideas of a text", and "the ability to recognize discourse structuring and genre patterns" (p. 4).

Assessing writing

The assessment of writing for the purpose of university entrance examination has changed over the last three decades. The field which began with one-line prompts such as "Discuss whether the pen is mightier than the sword" or "Discuss the effects of air pollution in your city" has moved to slightly less cryptic "a short paragraph" type prompt (as in the JAE-E) and made clear the purpose and audience of the writing. In the last decade, writing experts have promoted the "two points of view" and "facts and argument" type prompts and source-based reading and writing tasks that use authentic reading with tables, charts followed by writing. The main argument that has been put forth is that source-based writing best reflects actual university academic tasks (Cumming, 2014) and therefore it ought to be the way writing is assessed. The JAE-E tasks are typically the one-line or short prompt types.

Content Analysis

Content analysis has become an especially useful tool in L2 assessment investigations, for carrying out the investigations into "construct and content validation of tests and assessment scales (e.g., Bachman et al., 1995). Content analysis is also practical and feasible for the comparability studies of the target language use (TLU) domains, of items, tasks, and assessment scales, especially for specific purposes contexts. Graesser et al. (2011) claimed that "an automatic analysis is unquestionably more reliable and objective than approaches that involve humans annotating and rating texts by hand" (p. 223).

The Cambridge-TOEFL comparability study (Bachman et al., 1995) included content analysis focusing on items, tasks, and prompts and a quantitative analysis of the test performance on the Cambridge English: First (FCE) and the TOEFL paper-and-pencil version. The research steps were two-fold. First, a qualitative content analysis of two tests for items, tasks, and prompts was conducted. The authors stated that "the intention of the content analysis is to derive numerical values, based on expert judgment, for all relevant facets for each test and item" (p. 20). Therefore, experts used the Communicative Language Ability (CLA) instrument developed for the study. Second, a quantitative investigation of patterns of relationships between test content and test performance. Bachman et al. (1995) concluded that there were more similarities than differences between the two tests in general.

The Coh-Metrix

The Coh-Metrix software (2011) is among the broadest and most sophisticated automated textual assessment tools available. McNamara et al. (2014) argued that it is a computational tool providing a wide range of language and discourse measures. It serves as a linguistics workbench for researchers, teachers, and students of many different disciplines. Crossley et al. (2007) further asserted that it measures cohesion as well as text difficulty at different levels of languages.

Kunnan and Carr (2017) conducted a content analysis study of two English language proficiency tests, i.e., the General English Proficiency Test-Advanced (GEPT-A, developed by Language Training and Testing Center [LTTC], Taipei) and the TOEFL iBT®. The study focused on two specific analyses. First, a content analysis of the reading passages in the GEPT-A form and on the sample TOEFL iBT® reading passages was conducted. The analysis was conducted with Coh-Metrix to examine the cohesion, syntax, and vocabulary of the reading texts. The text analysis indicated that the reading texts of the two tests were comparable but differed in several key regards. The task analysis revealed that the construct coverage, item scope and task formats of the two tests were patently distinct. GEPT-A included only one careful reading main idea items, extensive coverage of skimming and scanning, to read for specific details, and to

extract the necessary information from a large portion of the test. On the other hand, the TOEFL iBT® focused on targeting vocabulary knowledge or the ability to infer the meaning of unfamiliar vocabulary from context, to read for major points and specific details. As for response formats, the reading section of TOEFL iBT® entirely relied on the selected response items, primarily multiple choice. In the GEPT-A, in addition to multiple-choice questions, short-answer and matching questions were used extensively.

In a recent example, Latifi and Gierl (2020) reported on the use of Coh-Metrix to evaluate language features of Coh-Metrix for a novel AES program that would score junior and senior high school students' essays from their large-scale assessments.

The CEFR Analysis

The CEFR, although rooted in Europe, has become a worldwide benchmark. It has provided a common basis for the elaboration of language syllabi, curriculum guidelines, examinations, and textbooks across Europe (CCC, 2001, p. 2). The main aims of CEFR are to achieve a greater unity from the Council of Europe members and overcome the barriers to communications in the field of modern languages arising from various educational systems in Europe. CEFR is also used for language learning, teaching and assessment in a more effective way, either for a language learner or a profession for language teaching and assessment. The CEFR describes foreign language proficiency at six levels: A1 & A2 (Basic User), B1 & B2 (Independent User), and C1 & C2 (Proficient User). Most English medium universities around the world prefer to recruit students at the C1 level, although B2 level students would have sufficient level of English to start university study. Students below the B2 level would require hundreds of hours of English language instruction to benefit from courses conducted in English.

Ho (2018) analyzed five test papers of previous university entrance examinations used in Macau using Coh-Metrix and CEFR linking analyses as the main methods. It was found similarities and differences among the previous examinations. Some of the JAE-E parts of Ho's study are reported and discussed in this article.

The JAE-E Link to Macau's BAA

An examination of how the JAE-E linked to Macau government's Basic Academic Attainments (BAA) for senior secondary English-L2 (Education and Youth Affairs Bureau [DSEJ], Macau SAR, 2017) was conducted. The BAA document, similar to the Curriculum Standards, includes the basis for the curriculum, textbooks and teaching in secondary schools in Macau for the ESL section. The requirements of basic academic attainments were listed for different domains in a kind of taxonomy of "subskills", i.e., Listening (a total of 18 "subskills"), Speaking (18), Reading (20), and Writing (12). These "sub-skills" are listed for each domain with no governing goals and principles at the overall or domain levels. Here is a sample of the "subskills" listed for each of the domains. Only the first five are included for each domain.

For listening, the list included: A-1: Understand more complex descriptions of things, people, and places; A-2: Understand basic expressions of certainty and obligation in spoken discourse; A-3: Understand a range of questions from the teacher; A-4: Take down the main points and some details or examples of a simple spoken text on a familiar topic if given a framework; A-5: Identify the purpose of a spoken text on a familiar topic.

For speaking, the list included: B-1: Pronounce a monosyllabic or multisyllabic new word from its spelling with reasonable accuracy; B-2: Use a wider range of vocabulary and language patterns in speaking; B-3: Use a level of pronunciation that can be understood easily; B-4: Use intonation to support more effective communication; B-5: Employ more complex sentence structures, especially in prepared presentations. As is obvious these lists do not properly account for how language learning is accomplished in these domains as they are discrete and do not have a progressive sequence either. *But despite the BAA for these domains, there are no listening or speaking items or tasks in the JAE-E.*

For reading, the list included: C-1: Accumulate new vocabulary from reading and organize them with proper use of vocabulary building skills; C-2: Infer the meaning of new words by making use of the co-

text and context, background knowledge, and common word affixes; C-3: Set a purpose for reading and meet it by selecting appropriate texts; C-4: Understand how punctuation organizes meaning in a written text and explain how the punctuation affects the reader's interpretation of it; C-5: Read a text at an appropriate pace with accuracy and fluency to achieve the purpose of reading.

For writing, the list included: D-1: Use a range of vocabulary, sentence patterns and verb forms with consistency and accuracy; D-2: Use a range of resources, including paragraphs, connecting phrases, and word order, to organize text; D-3: Write texts with structures and language features appropriate to their purpose; D-4: Write with a clear understanding of the purpose, context, and audience of a range of texts; D-5: Establish cohesion through lexical and grammatical choices. Only the first 5 out of the 12 are included here. But the list does not account for how writing ability progresses.

Overall, the discrete, sequential, and mostly ambiguous nature of these lists did not seem to serve as a useful way of linking with the JAE-E constructs and test items. They do not allow for clear identification of which of the BAA domain "subskills" were being targeted in the JAE-E test and items. Thus, the decision to link the JAE-E with the BAA's attainments for reading and writing did not work. Alderson and Lukmani (1989) argued that such attempts to link curricula to tests would not work.

Research Questions

The research questions were:

(1) What is the content of the JAE-E in terms of construct coverage and task formats?
(2) What are the linguistic characteristics of the JAE-E reading texts as examined by Coh-Metrix?
(3) How does the JAE-E align with the CEFR for languages?
(4) How does the design of the JAE-E compare with other contemporary university admission examinations?

Methodology

Data Collection

The analysis from one JAE-E test paper reported here was based on the downloaded version from the Higher Education (see GAES websites).[1] The JAE-E structure is composed of three sections: language use (40%), reading (30%) and writing (30%).

The language use section has three parts including vocabulary and grammar with 15 isolated sentences or language expressions with multiple-choice response format; a cloze passage of 200 words with ten rational deletions in multiple-choice response format; and joining sentences with five two-sentence items in an open-ended response format.

The reading section has three parts containing everyday English with five items in multiple-choice response format; a short text of about 300 words with five items in multiple-choice response format; and a long text of about 550 words with ten items partly in multiple choice and partly in open-ended response format.

The writing section has three prompt choices from which test takers are required to choose one to write a short essay of 200 words. Table 1 presents the overall structure of the test.

[1] Analyses of tests designed and administered by the other four institutions are not reported here (see Ho [2018] for analyses).

TABLE 1
Overall Structure of the JAE-E

Sections	Task type	# of Items	Maximum marks
Language use	A. Vocabulary and grammar	15	15
	B. Cloze passage	10	15
	C. Joining sentences	5	10
Reading	A. Everyday English	5	5
	B. Short passage	5	7.5
	C. Long passage	10	17.5
Writing	A. Short essay	1	30
Total		51	100

Linguistic and Task Analyses

A content analysis of the JAE-E was conducted to examine various aspects of the test through automated linguistic analysis of texts using the Coh-Metrix web tool. Further, task analysis of the constructs, task types, response formats, and match with CEFR levels were conducted by five qualified Macau English teachers invited to analyze the test. A modified CEFR can-do descriptors were provided to the teachers so that they could assign an appropriate level for the JAE-E tasks. The levels assigned by each teacher for the tasks were assigned a score of 1 to 6 (i.e., 1 for the A1 level to 6 for the C2 level). These scores were then averaged, and an overall level was assigned for each section of the test.

Results

Construct and Response Format

Language use

This section assessed test takers' knowledge of grammar and vocabulary. Table 2 shows the 30 items in the 15 specific categories. Table 3 presents the tasks and task formats; the first two task types are in the multiple-choice format and the third task is the open response format.

TABLE 2
Focus Areas

	Focus areas	# of items	% of items
1.	Article (+ Possessive Case)	2	6.7
2.	Auxiliary Verb/Modal verb	2	6.7
3.	Comparative/Superlative adjective	1	3.3
4.	Conjunction/Connecting word	5	16.7
5.	Preposition	1	3.3
6.	Quantifier	1	3.3
7.	Transition word	1	3.3
8.	Verb tense	5	16.7
9.	Verb voice	2	6.7
10.	Vocabulary (Adjective)	2	6.7
11.	Vocabulary (Compound Noun/Noun)	1	3.3
12.	Vocabulary (Phrasal verb)	2	6.7
13.	Vocabulary (Verb)	2	6.7
14.	Word form (Adjective)	1	3.3
15.	Word form (Noun)	2	6.7
Total		30	100

TABLE 3
Tasks and Task Formats

Tasks and task formats	# of Items	% of Items
1. Sentence completion – m-c	15	50.0
2. Cloze passage – m-c	10	33.3
3. Joining sentences – open-ended	5	16.7
Total	30	100

Note. m-c = multiple-choice

Reading

In this section, test takers were required to read texts and then answer questions. The sub-skills and the number of items for each are listed in Table 4. In terms of task format, most of the tasks in the reading section were in the multiple-choice response format. And a few were in the open response format as shown in Table 5.

TABLE 4
Sub-skills Assessed

Sub-skills	# of items	% of items
1. Search processes	5	25
2. Vocabulary knowledge	2	10
3. Inferences about text information	5	25
4. Strategic processing abilities	2	10
5. Summarization abilities (paraphrasing)	2	10
6. Synthesis skills	4	20
Total	20	100

TABLE 5
Task Formats

Task format	# of items	% of items
Multiple-choice	15	75
Open-ended	5	25
Total	20	100

Writing

In this section, test takers were expected to write an essay based on one of three prompts. Test takers were expected to organize ideas and logically develop an argument. A summary of tasks and requirements is depicted in Table 6. Overall, the writing task in the test was the most difficult given its genre and the number of words. Particularly, the three prompts provided were related to the international trends such as the latest technology: effect of self-driving cars or taxis on society; a prompt related to Macau society such as the offer to tourists of a greater variety of activities; or an agreement or disagreement on using the Internet in classrooms as a learning resource. Additionally, test takers needed to be familiar with various types of genres.

TABLE 6
Tasks

Type of essay	# of prompts provided	Words required minimum
Cause and effect Persuasive Argumentative	3	200 words

Coh-Metrix Analysis

Reading texts from the JAE-E were analyzed through Coh-Metrix Text Easability Assessor in terms of narrativity, syntactic simplicity, word concreteness, referential cohesion, deep cohesion, and Felsch Kincaid grade level. The three passages had a total of 1,089 words with an average of 363 words.

Narrativity

A narrative text tells a story with characters, events, places, and things that are familiar to the readers. Narrativity is closely affiliated with everyday oral conversations. This robust component is highly affiliated with word familiarity, world knowledge, and oral language. Non-narrative texts on less familiar topics would lie at the opposite end of the continuum (McNamara et al., 2014, p. 85). The three texts in JAE-E were at 88%, 30% and 36% for an average of 51%, indicating the difficulty level varied. The first text was quite easy to be understood but the other two were not story-like and contained many unfamiliar words.

Syntactic simplicity

This component reflects the degree to which the sentences in the text contain fewer words and use simpler, familiar syntactic structures that are less challenging to process. At the opposite end of the continuum are texts that contain sentences with more words and that use complex, unfamiliar syntactic structures (McNamara et al., 2014, p. 85). The three texts in the JAE-E were 99%, 40% and 46% for an average of 62%, implying that except for the first text which was extremely simple, the others were of average syntactic simplicity.

Word concreteness

Texts containing content words that are concrete and meaningful and evoking mental images are easier to process and understand. In contrast, texts containing more abstract words are more challenging to understand as abstract words are difficult to represent visually. (McNamara et al., 2014, p. 85). The three texts were at 35%, 83%, and 57% for an average of 58%, showing that except for the second text which was easy, the others had an average level of word concreteness.

Referential cohesion

A text with high referential cohesion contains words and ideas that overlap across sentences and the entire text, forming explicit threads that connect the text for readers. A low cohesion text is typically more difficult to process because there are fewer connections that tie the ideas together for the readers (McNamara et al., 2014, p. 85). The three texts in the JAE-E were at 50%, 24%, and 7% for an average of 27%, showing that the texts had less overlap in words and ideas between sentences. Such cohesion gaps would require the test taker to make inferences and could be challenging or even unsuccessful without sufficient prior knowledge.

Deep cohesion

This dimension reflects the degree to which the text contains causal and intentional connectives when there are causal and logical relationships within the text. These connectives help readers form a deeper and more coherent understanding of the causal events, processes, and actions in the text. When a text contains many relationships but does not contain those connectives, readers must infer the relationships

between the ideas in the text. If the text is high in deep cohesion, then those relationships and global cohesion are more explicit (McNamara et al., 2014, p. 85). The three texts in the JAE-E were at 18%, 50%, and 75% for an average of 48%, which shows that except for the third text was relatively easy, the others had fewer connecting words to help clarify the relationship between events, ideas and information.

Flesch Kincaid grade level

Graesser et al. (2011) claimed that the metrics of Flesch Kincaid (FK) were based on the length of words and length of sentences. Word length was measured as the mean number of syllables per word while sentence length was the mean number of words per sentence. The FK-Level was computed as "11.8 *Syllables + .39 * Words − 15.59", which "is a robust predictor of the amount of time it takes to read a passage, a result that offers an impressive validation of the metric" (Graesser et al., 2011, p. 224). Moreover, "infrequently used words in a language tend to be longer" and "long sentences tend to place more demands on working memory and to have more complex syntax" (ibid). The three texts in the JAE-E had a wide range of levels: 2.0, 7.9 and 10.9 for an average of 6.9. This indicated that the texts were relatively easy to read as the average level was at Grade 7 level, but the third text was quite difficult with the level of Grade 11 (see Table 7).

TABLE 7
Coh-Metrix Indices

Categories	%			Average
	Text 1	Text 2	Text 3	
Narrativity	88	30	36	51
Syntactic Simplicity	99	40	46	62
Word Concreteness	35	83	57	58
Referential Cohesion	50	24	7	27
Deep Cohesion	18	50	75	48
Flesch Kincaid Grade Level	2.0	7.9	10.9	6.9

In summary, the results of the Coh-Metrix analysis suggested that overall indices of the five variables of the reading section, except for referential cohesion, were about average, meaning the *texts were neither too difficult nor too easy to* test takers**.** The Flesch Kincaid Grade Level indicated the texts were quite low in relation to the level university students are expected to be at.

CEFR of Languages Analysis

The mean scores from teachers' analyses were averaged for the JAE-E sections. Based on these scores, CEFR levels were assigned to the test sections. Table 8 displays the scores and assigned levels.

TABLE 8
CEFR Levels Assigned to JAE-E Sections

Test section	Scores	CEFR Level
Language use	2.39	A2
Reading	2.67	A2
Writing	5.00	C1
Overall	3.35	B1

The Journal of Asia TEFL
Vol. 18, No. 1, Spring 2021, 208-222

Discussion

RQ1: What is the content of the JAE-E in terms of construct coverage and task formats?

The results showed that the language use section reflected previous conceptualizations of grammar assessment, where "a set of structural rules, patterns, norms, or conventions" (Purpura, 2014, p. 2); "interaction between the grammatical knowledge (including a range of linguistic forms and semantic meaning) and pragmatic knowledge" (p. 6) were not incorporated. In addition, according to Purpura (2014), selected response tasks of grammar knowledge measure recognition or recall (i.e., receptive knowledge) and fail to capture dynamic and complex understandings of resources needed for communication, but they are useful in the situations aiming at observing test takers' receptive knowledge of isolated language features.

Moreover, as Purpura (2014) also argued that "performance tasks" are the best way for assessing the grammatical ability where test takers are required to respond to a prompt with their input of the forms and produce an extended amount of spoken or written data. The language use section in the JAE-E focused instead on discrete point grammatical points in the multiple-choice response format for all tasks including the cloze passage task. Total dependence on this response format would clearly result in test taker scores reflecting knowledge of grammar and vocabulary with little transfer to the actual use of the language. In addition, the joining sentences task required manipulation of grammatical features in the open response format. However, this task is both inauthentic and irrelevant to students' work in the university context. Finally, the scoring of these items was on a credit-no (0-1 credit) basis with no provision for partial credit, a much better indicator of test takers as it provides gradation of test takers with no grammar knowledge from some grammar knowledge to perfect grammar knowledge.

As for Reading, Grabe and Jiang (2014) identified 14 sub-skills. This range of sub-skills is reflected in many international tests. For example, the International English Language Testing System (IELTS) includes reading for specific information and main ideas, to evaluate or identify a topic or a theme. The TOEFL iBT® includes basic comprehension, inferencing, and reading to learn. Cambridge English to Speakers of Other Language (ESOL)'s suite of exams (FCE now called B2 First, CAE now called C1 Advanced and CPE now called C2 Proficiency) require greater recognition of the discourse structure of texts, main ideas, careful readings abilities, facility in reading multiple text genres, and a large amount of reading itself. When the JAE-E reading texts and tasks are compared to these sub-skills, it is clear the JAE-E does not have the breadth and depth it should have for a university entrance examination.

In terms of the response format, the test used mainly multiple choice. According to Kintsch and Yarbrough (1982), open-ended questions could best measure test takers' comprehension of the main ideas of the texts. For the task formats of Cambridge FCE, CAE, and CPE, "complex matching tasks of various types, multiple choice items, short response items and summary writing" (Grabe et al., 2014, p. 5) were involved. Also, in the academic version of IELTS, "short response items of multiple types, matching of various types, several complex readings with diagrams and figures and innovative fill-in tasks" were used (Grabe et al., 2014, p. 5). In the TOEFL iBT®, two new tasks, prose summary and schematic table were included where the longer, more complex texts were used in comparison with ones in the traditional TOEFL. In comparison, the JAE-E had an extremely limited variety which was mainly multiple choice.

Regarding writing, the test had a choice of three prompts from different writing genres (i.e., cause and effect, persuasive, and argumentative essay). These tasks reflect the tasks that students would need to carry out on a regular basis in a university. In terms of essay length, the number of words required was 200 words minimum. This requirement is lower than the B2 First (280-380 words for two pieces of writings). To sum up, the writing task was almost at the appropriate level, as the task was at the C1 level as assigned by the teachers.

RQ2: What are the linguistic characteristics of the JAE-E reading texts?

The results of Coh-Metrix analysis of the reading texts indicated that the distribution of the five variables varied in the texts: 2.0, 7.9 and 10.9 for an average of 6.9 which is about Grade 7 school level in the U.S. This would map to about A2 or B1 level on the CEFR. If test takers reached this level, it was likely that they would need additional English language instructions for many semesters before reaching the required B2 or higher level for studying at an English-medium university.

RQ3: How does the JAE-E align with the CEFR for Languages?

The scores from the Macau English teachers placed the JAE-E at 3.35, which is just above B1 level. The CEFR posits that successful test takers at the B1/B2 levels will have language skills to live and work independently in an English-speaking country or study on courses taught in English. At the C1 level, successful test takers will have language skills that employers and universities are looking for, such as following an academic course at the university level, communicating effectively at a managerial and professional level, participating with confidence in workplace meetings or academic tutorials and seminars, or expressing him/herself with a high level of fluency. Thus, the overall level of JAE-E is low except for the writing task which is at the C1 level.

RQ4: How does the test design of the JAE-E compare with other contemporary university admission examinations?

In terms of language use, the JAE-E reflected an outdated understanding of the assessment of grammar and vocabulary. None of the well-known contemporary tests (e.g., IELTS and TOEFL) for the same purpose as the JAE-E have the assessment of grammatical features in the multiple-choice response format. In addition, the cloze passage with rational deletion is clearly an indirect measure of reading, although it may provide a fair measure of test takers reading proficiency. Further, the joining sentences task is both inauthentic and irrelevant in relation to what students would be doing in university classes.

In the reading section, the JAE-E assessed a restricted range of sub-skills mainly in the multiple-choice response format. None of the task types that are popular with contemporary tests such as readings with diagrams and figures, fill-in tasks, and creating tables are included.

In the writing section, the JAE-E provided a choice of three writing prompts on topics of general interest. These were too traditional in approach; contemporary approaches, to list a few, include source material with multiple texts (with opposing views), and the use of diagrams, charts, and tables.

Suggestions for Improvements

Based on the findings from the content analysis, the following suggestions are made for improvement of the test from use of integrated abilities, use of scenario-based assessments, use of technology, and use of clearer BAA domains and "subskills".

Use of integrated skills

Over the last two decades, the most enduring idea in language assessment is the integration of skills such as combining listening and speaking, reading and writing, or listening, reading, and writing. Such tasks will require test takers, according to Cumming (2014), "to integrate source reading or listening material into their writing or speaking performance in ways that stimulate the cognitive, communication, and literacy demands of real-life academic or vocational tasks" (p. 1). The rationale for these initiatives is that abilities to write or speak coherently for relevant ideas, to handle source documents appropriately,

and to demonstrate the knowledge in relevant ways are primary abilities required for the successful performance in universities, colleges, high schools and in many workplaces.

Examples of integrated writing tasks are in the TOEFL iBT® and the Canadian Academic English Language Assessment that require test takers to read a lengthy text on a specific topic, hear segments of a lecture or other discussion about it, and then write and/or speak about the information from the source materials. The Diagnostic English Language Needs Assessment from New Zealand asks doctoral students to read several short excerpts from a variety of sources about a topic, and then to write an essay in answer to a specific question about the topic by referring to the statements in the readings. In short, these tasks reflect the types of work students would be expected to carry out to perform competently in their academic courses. Similar integration of skills can also be attempted with listening and speaking or listening, reading, and writing. Therefore, the JAE-E would benefit from such an approach; it would help university admissions and placement officers with more accurate decision makings.

Use of scenario-based assessments

Language tests developed for any skill area or a combination of skills could be housed in a scenario that offers (a) individual goal-oriented with purposeful multi-stage activities, (b) motivating and engaging tasks not focused on language, and (c) integration and synthesis of language skills. This approach is favored over the traditional approach of activities based on random sets of listening, speaking, reading, and writing activities. As Sabatani et al. (2011) stated these "task sets are composed of a series of related tasks that unfold within an appropriate social context. The goals include to communicate how the tasks fit into a larger social activity system; to set standards for performance; to give test takers a clearer idea of how to allocate attention and give focus to their deliberations; to provide opportunities to apply strategic processing and problem solving; and to have learners evaluate and integrate multiple sources of information in a meaningful, purpose-driven context" (p. 10). Shore et al. (2017) delineated two types of scenario-based assessments in reading: "the Global, Integrated Scenario-based Assessment (GISA) and the English Learner Formative Assessment (ELFA)" (p. 235). Their illustrations showed the clear applications that are possible through this approach.

Recent illustrations included: Carroll et al. (2015) demonstrated the feasibility of a scenario-based language assessment design intended for placement decisions. Jang et al. (2016) developed a set of multi-staged, multimodal tasks to elicit learners' L2 proficiency and cognitive processing through simulated academic learning scenarios. Banerjee (2019) demonstrated the effectiveness of a scenario-based approach in an investigation of topical knowledge.

Therefore, a scenario-based assessment provides an opportunity to create authentic-like situations that require the use of appropriate language abilities in realistic contexts that involve interlocutor discourse and/or multi-media texts. The JAE-E would certainly benefit from this approach.

Use of technology

It is obvious now that there is a widespread use of technology (e.g., computers, tablets, and smart phones) among students, teachers, and the school community in Macau. Almost college-bound students can use smart phones, tablets, and computers. Besides, it is time to embrace the concept of digital literacy and find ways of using it in the teaching, learning and assessment activities.

However, the JAE-E has stayed away from technology completely. There is no listening or speaking section in the test and the other sections are also paper-and-pencil based. The usual argument that practicality issues do not make it conducive for listening and speaking tasks is not insurmountable if there is motivation to incorporate technology into the test. A test that is computer-delivered or available on-line would bring in listening and speaking activities and reading and writing can have authentic sources and realistic activities. For example, task types that involve test takers in click and drag operations for reordering sentences in a paragraph or matching information in columns are can benefit from the use of

technology. Scoring responses through automated scoring of short answers, if not essay length responses, would be another benefit.

Brunfaut et al. (2018) argued that the effect of mode of delivery on performances and perceptions on a writing test does not affect writing scores at B2/C1 level and has a small effect on integrated skills tasks. Further, automated item generation and automated scoring of writing and speaking are innovative features that are possible with the use of technology. In previous research on automated scoring of writing, it has been shown that there is relative strong correlation (about 0.70 and above) between automated scoring and human scoring of essays (see Hoang and Kunnan, 2018, and Liu and Kunnan, 2018, for examples). Many tests including the Pearson Academic and General English Tests and the Duolingo English Test use innovative testing using technology.

Clearer BAA domains and "subskills"

A taxonomy that is a clearer set of "subskills" for the four domains in the Macau BAA would help in designing, developing, and scoring the JAE-E performances and in placement decisions based on section or total scores. This in turn can help link the JAE-E to the BAA and provide a clear diagnosis of the strengths and weaknesses of test takers. Kunnan and Yao (2020) evaluated the alignment between JAE-E items and BAA domains. A total of 28 M.A. students majoring in second language acquisition (SLA) were invited to map the JAE-E items with the BAA domains and their subskills. The instruments adopted were BAA for senior secondary English (2017) and JAE-E (2017/2018). And data collected from students were analyzed by SPSS 24.0. Results showed both JAE-E items and BAA domains need further improvement. For JAE-E, it is to be hoped that listening and speaking skills should be included to complete and authenticate the test. Additionally, it is recommended to cover more subskills from the BAA to fit the requirements prescribed by DSEJ. In terms of BAA domains, abilities such as grammatical accuracy are not specified in the reading domain. And it is better to distinguish whether domain subskills should be achieved in a test or during the process of learning English.

Conclusion

The content analysis of the JAE-E showed that the constructs were representative of traditional testing in language use, reading, and writing. A more contemporary design that includes all skill areas with some skill integration would provide accurate scores based on which error-free placement decisions can be made. A few suggestions that can improve the JAE-E are: First, newer conceptualizations of these constructs of interest as discussed in the review of literature need to be made for the JAE-E.

Second, except for writing, most of the test items use the multiple-choice response format. This is a serious flaw in the design and needs to be changed to include a variety of response formats. If the test design includes tasks that are in the direction of performance tasks, there will be more response formats naturally.

Third, the Coh-Metrix analysis and the CEFR analysis placed the language use and reading test sections much below the university entrance level; the writing section was at the C1 level, the appropriate level. Thus, the level of the test needs to be raised in language use and reading so that the test can provide useful information for admission and placement purposes.

Fourth, a more meaningful JAE-E needs to be in place for direct diagnosis and feedback for test takers. See Kunnan and Jang (2009) of how such efforts can be offered within a standardized language assessment. Providing generic feedback at the CEFR levels for the different skills could be an improvement if not individual feedback.

Fifth, the use of technology cannot be avoided by educational authorities in Macau. A research and operational agenda need to be in place soon so that Macau's JAE-E can make use of some of the advantages of technology listed earlier.

One of the limitations of this study is the lack of test taker performance data analysis for a more comprehensive picture of the JAE-E, as argued in many publications (Kunnan, 1992; Kunnan and Grabowski, 2013; Kunnan, 2018). Given the JAE-E design, test performance analysis will likely reveal other deficiencies such as poor construct definitions and operationalizations into items and tasks, response format bias, lack of internal consistency and scoring consistency and errors in placement decisions.

Acknowledgments

We thank the teachers in Macau schools and students at the University of Macau who took part in the study.

The Authors

At the time of the research, Ally Oi Kuan Ho was an M.A. student, Don Yao was beginning his PhD studies and Antony John Kunnan was a professor at the University of Macau. In 2020, Ho completed her MA studies, Don became a Ph.D. candidate and Antony joined Duolingo, Inc., as a Principal Assessment Scientist.

Corresponding author:
Don Yao
E-mail address: yb87710@connect.um.edu.mo

References

Alderson, J. C., & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language, 5*, 253-270.

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I-C. (1995). *An investigation of the comparability of the two tests of English as a foreign language: The Cambridge-TOEFL comparability study.* Cambridge University Press.

Banerjee, H. (2019) Investigating the construct of topical knowledge in second language assessment: A scenario-based assessment approach. *Language Assessment Quarterly, 16,* 133-160,

Brunfaut, T., Harding, L., & Batty, A. (2018). Going online: The effect of mode of delivery on performances and perceptions on an English L2 writing test suite. *Assessing Writing, 36*, 3-18.

Carroll, B., Liu, H., & Oh, S. (2015). *Design considerations in the assessment of L2 integrated skills through scenarios*. The first TC/ETS forum on teaching, learning, and assessment of English language learners. New York.

Council for Cultural Co-operation, Education Committee, Modern Languages Division, & Strasbourg (2001, 2009 printing). *Common European framework of reference for languages: Learning, teaching, assessment.* Press Syndicate of the University of Cambridge.

Crossley, S. A., Dufty, D. F., McCarthy, P. M., & McNamara, D. S. (2007). Toward a new readability: A mixed model approach. In *Proceedings of the 29th annual conference of the Cognitive Science Society*, 197-202.

Cumming, A. (2014). Assessing integrated skills. In A. J. Kunnan (Ed., Vol. 1), *The companion to language assessment* (pp. 1-13). Wiley-Blackwell.

Grabe, W., & Jiang, X. (2014). Assessing reading. In A. J. Kunnan (Ed., Vol. 1), *The companion to language assessment* (pp. 1-16). Wiley-Blackwell.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*, 223-234.

Ho, O. K., (2018). *A comparability study between the Joint Admission Examination and the earlier admission examination papers in Macau: A content analysis* [Unpublished M.A. thesis]. Department of English, University of Macau.

Hoang, G., & Kunnan, A. J. (2016). Automated writing instructional tool for English language learners: A case study of *MY Access*. *Language Assessment Quarterly, 13*, 359-376.

Jang, E. E., Wagner, M., & Dunlop, M. (2016, June). *Construct validation of multimodal scenario-based language assessment (SBLA) tasks for diagnostic placement purposes*. Plenary presented at the 38th LTRC Conference, Palermo, Italy.

Liu. S., & Kunnan, A. J. (2016). Investigating the application of automated writing evaluation to Chinese undergraduate English majors: A case study of *WriteToLearn. CALICO Journal, 33,* 71-91.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge University Press.

Kin, T. K. (2002). Examination policies and practices: A comparative study in (Mainland) China, Taiwan, Hong Kong, and Macau. [*Durham theses*]. Durham University.

Kintsch, W., & Yarbrough, J. C. (1982). Role of rhetorical structure in text comprehension. *Journal of Educational Psychology, 74*, 828-834.

Kunnan, A. J. (1992). An investigation of a criterion-referenced test using G-theory, factor and cluster analyses. *Language Testing, 9*, 30-49.

Kunnan, A. J. (2018). *Evaluating language assessments*. Routledge.

Kunnan, A. J., & Yao, D. (2020). *Teacher ratings of JAE-E test items.* Paper circulated at the LASeR group, Department of English, University of Macau.

Kunnan, A. J., & Carr, N. (2017). A comparability study between the general English proficiency test: Advanced and the internet-based test of English as a foreign language. *Language Testing in Asia, 7,* 1-16.

Kunnan, A. J., & Grabowski, K. (2013). Large scale second language assessment. In M. Celce-Murcia et al. (Eds.), *Teaching English as a second or foreign language* (4th ed., pp. 304-319). Heinle Cengage.

Kunnan, A. J., & Jang, E. (2009). Diagnostic feedback in language assessment. In M. Long & C. Doughty (Eds.), *The handbook of language teaching* (pp. 610-627). Wiley.

Latifi, S., & Gierl, M. (2020). Automated scoring of junior and senior high essays using Co-Metrix features: Implications for large-scale language testing. *Language Testing, 38,* 62-85.

Purpura, J. E. (2014). Assessing grammar. In A. J. Kunnan (Ed. Vol. 1), *The companion to language assessment* (pp. 1-25). Wiley-Blackwell.

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18,* 1-32.

Rosson, M. B., & Carroll, J. M. (2009). Scenario-based design. In A. Sears & J. A. Jacko (Eds.), *Human computer interaction development process* (pp. 1032-1050*)*. CRC Press.

Shore, J., Wolf, M., O'Reilly, T., & Sabatini, J. (2017). Measuring 21st century reading comprehension through scenario-based assessments. In M. Wolf & Y. Butler (Eds.), *English language proficiency assessments for young learners* (pp. 234-252). Routledge.

Websites

Coh-Metrix. (2011). http://cohmetrix.com/.

DSEJ. (2017). The Requirements of Basic Academic Attainments for Senior Secondary English. https://portal.dsej.gov.mo/webdsejspace/internet/Inter_main_page.jsp?id=21349.

GAES. (2017). General Information and Rules for JAE (Subjects: Languages and Mathematics). http://www.gaes.gov.mo/unification/regulations_e.pdf.

(Received January 20, 2021; Revised February 25, 2021; Accepted March 20, 2021)