Item Analysis of the EDC Reading Quizzes

Davey Young and Christopher Nicklin

ABSTRACT

Multiple-choice quizzes are a common way to assess reading comprehension. In Rikkyo University's English Discussion Class, such quizzes are used to encourage students to complete a short homework reading designed to provide students with content to discuss in the lesson. The present study utilizes proportionate stratified random sampling to determine the extent to which these quizzes do or do not meet their intended aim, as well as to identify problematic items for revision. Results show that the reading quizzes generally serve their purpose, but that the lowest proficiency students have more difficulty identifying the correct answer to multiple-choice questions. These results are discussed in detail, as are specific examples of the ten problematic items identified by the analysis. The paper concludes with suggestions for additional principles for writing quiz items, as well as considerations for future reading quiz administration in EDC.

INTRODUCTION

Rikkyo University's English Discussion Class (EDC) is a mandatory EFL course for all first-year students that aims to develop students' speaking fluency and ability to discuss a range of topics in English (Hurling, 2012). The course serves between 4,500 and 4,700 students each year, and does so via a strongly unified syllabus that entails intensive teacher training to ensure standardized lesson delivery and assessment using a suite of level-specific textbooks created specifically for the course. Students are streamed into one of four proficiency levels based on TOEIC tests administered upon commencement to the university. Class sizes range from seven to nine students, with most containing eight. EDC instructors employ a communicative approach that emphasizes high student-to-student interaction and the explicit instruction of functional language used to perform various strategic and discourse competencies. The syllabus is organized around a theme-based CLIL approach, a variety of 'soft' CLIL that utilizes carrier topics to provide the context in which students can acquire and practice target language forms (Brinton & Snow, 2017).

Before class each week, EDC students read a brief article from the textbook about that lesson's topic. These readings' use and presentation of content can be described within Koda and Yamashita's (2019) *reading to learn* framework, which consists of overlapping processes:

> (a) Constructing text meanings based on linguistic information presented in a text (*text-meaning building*), (b) connecting text information to the reader's personal experiences and prior knowledge (*personal-meaning construction*), and (c) reflecting on what the reader has learned from the two proceeding operations (*knowledge refinement*). (p. 3)

Students' completion and comprehension of the text itself constitute the first operation (text-meaning building). As they are presented in the textbook, each reading is bracketed by three 'Before Reading' and three 'After Reading' questions that invite readers to connect their personal experience and knowledge to the reading, a connection made more explicit in a fluency activity that occurs early in the lesson (personal-meaning construction). Content from the homework readings is then recycled in lesson content via discussion prompts and opinion gap tasks that comprise the majority of the lesson (knowledge refinement).

There are two versions of each homework reading, one for the upper two proficiency levels in EDC and a more simplified one for the lower two levels. To ensure a high degree of comprehensibility, these texts are elaborated by repeating key words and using rhetorical signaling

devices, a type of text modification that has been shown to increase comprehensibility among English language learners (Oh, 2001; Yano, Long, & Ross, 1994). Both the higher and lower proficiency level readings are also written and revised by EDC Program Managers (PMs) with level-appropriate readability targets and lexical thresholds in mind. These targets were derived in the 2015 academic year by analyzing TOEIC intake scores and conducting a detailed survey of recent research into the readability of EFL textbooks commonly used in secondary schools in Japan (Young, 2016).

At the beginning of each regular (non-test) EDC lesson, students complete an eight-item multiple-choice (MC) quiz about the homework reading. As EDC is a theme-based speaking class, the aim of both the readings and subsequent quizzes is not to assess reading ability, but rather to prepare students to discuss the topic in accordance with Nation's (2001) principle of meaning-focused output. In theory, the quiz does this by making students accountable for completing the reading before the lesson, as well as by providing positive washback related to topic content by facilitating text-meaning building. As such, each quiz item should be easily answerable by students across all faculties and proficiency levels provided that they read and understood the homework reading. Despite the fact that quiz scores account for 20% of regular lesson grades, or approximately 11.5% of the entire course grade, there has not been a rigorous evaluation of these assessment instruments since the course began in 2010. The present study seeks to fill this gap in the Center for English Discussion Class' ongoing program evaluation and development.

Question stems are crafted in light of the following principles: (a) the format should be sentence completion or direct questions, (b) negative questions should be avoided, (c) the questions should follow the order of the ideas as they appear in the text, and (d) there should be only one idea per question. Correct answers should: (a) be balanced in their placement across questions, (b) be found in the source readings, (c) be clear, and (d) not be guessable based on common knowledge. Distractors should be plausible and not overlap. Although MC questions used to assess reading comprehension typically use distractors that are plausible misinterpretations of information in the text (Alderson, 2000), such distractors in the present context would undermine the wider purpose of the homework reading and subsequent quizzes. The principles outlined above are all followed to further facilitate text-meaning building.

Using Pearson and Johnson's (1978) taxonomy, EDC's MC quiz questions are principally *textually explicit* (the question information and the correct answer appear in the same sentence), with some that are *textually implicit* (the information and key appear across sentences), and none that are *script-based* (text information must be combined with background knowledge to identify the correct answer). Framed within Day and Park's (2006) taxonomy of comprehension types, all of the EDC reading quiz questions assess *literal comprehension*, or a straightforward, surface-level understanding of the text. When writing the quizzes, PMs avoid reorganization, inference, prediction, evaluation, and personal response question types due to increased difficulty, a need for higher-order thinking skills, the need for specific background knowledge, or some combination therein. The narrow scope of predominantly textually explicit, literal comprehension questions is intended to help ensure that students from all proficiency levels and faculties, provided they completed and comprehended the homework reading, are able to readily identify the correct answer to each MC item without relying on higher-order skills or background knowledge.

MC quizzes are used over other formats because they can be quickly and efficiently administered and scored within the first five minutes of the lesson. In regular lessons, each quiz has eight MC questions with one correct answer, also known as the *key*, and three distractors. In the three test lessons (Lessons 5, 9, and 13), this format is simplified. The number of quiz questions is increased to ten but there is only one distractor in addition to the key. Students typically have three to four minutes to complete each quiz and do not have access to the homework reading while

doing so. During this time, instructors monitor to help prevent cheating. Once the quiz has been completed by all students or the time limit has elapsed (whichever happens first), students are instructed to exchange quizzes with a partner and the instructor displays the correct answers on the board. Students then mark their classmates' quizzes and return them to the original quiz taker to view their score before the instructor collects them. For grading purposes, the number of correct answers is converted to a score ranging from zero to four by dividing it in half and rounding up (e.g. seven or eight correct answers both result in the maximum score of four).

As each quiz is taken by 4,500 to 4,700 students each week, there is a very real risk of quiz answers being memorized and passed along by students who take the quiz earlier in the week to students from other classes who will take it later. In an attempt to mitigate this type of cheating, in 2016 it became standard practice within EDC for roughly half of the instructors to administer one version of the quiz, and the other half to administer another. The only substantive difference between the quiz versions is that the sequence of answer options has been rearranged for each question from one version to another. To avoid confusion among instructors, these different versions are color-coded in alternating colors from week to week, resulting in a "Blue & Red" and a "Green & Yellow" version of each quiz. These two quiz versions are further delineated by the letters "B&R" or "G&Y" appearing in one corner of the quiz. This naming convention stems from how EDC's 42 full-time instructors are divided into four shared office spaces, each of which is designated a color for easy reference. Instructors are assigned a quiz version based on their assigned office space, though they are free to use the alternate version if they suspect a student or students of routinely attaining the quiz answers to their usual version in advance.

With the reading quiz purpose and format in mind, the following research questions were posed in advance of data collection to improve this simple but critical component of course assessment.

1. How do students from different proficiency levels compare on quiz performance?
2. How do students from different faculties compare on quiz performance?
3. How do students taking different versions of the quiz compare on quiz performance?
4. Which EDC quiz questions or distractors are problematic for students?

METHOD

In order to answer the research questions, a sample of quiz responses was required for analysis. The sample was obtained through the implementation of proportionate stratified random sampling, which involved dividing the entire first-year student population into subgroups, or strata, and taking a sample from each strata that represented the entire population. This form of sampling requires a sampling fraction of just 10% of the population, yet still allows the use of statistics to make predictions about the entire population (Trochim, Donnelly, & Arora, 2016)

The strata were determined based upon student proficiency level and faculty membership. In total, there were initially 4,539 first-year students enrolled in the course, consisting of 342 (7.5%) Level I students (combined listening and reading TOEIC score of 680 and above), 1,678 (37%) Level II students (TOEIC 480 to 679), 2,154 (47.5%) Level III students (TOEIC 280 to 479), and 365 (8%) Level IV students (TOEIC below 280). These figures included 37 students who filed for administrative leave before the first lesson of the academic year. The population were spread across all 10 Rikkyo faculties, consisting of Arts, Business Administration, Economics, Human Services, Intercultural Communication, Law and Politics, Psychology, Science, Sociology, and Tourism. In order for the sample to be representative of the student population, the sample was chosen based upon the faculty distribution of each level. For each faculty, the population number and percentage of students in each level was calculated. For example there were 56 Level I Arts faculty members, constituting 16% of the Level I population.

Table 1. Original Population Size and Sample Population Sizes for Level I & II Strata

	Level I			Level II		
Faculty	Pop.	%	Sample	Pop.	%	Sample
Arts	56	16.4	7	355	21.2	44
Business Administration	70	20.5	9	158	9.4	20
Economics	21	6.1	3	245	14.6	31
Human Services	7	2.0	1	73	4.4	9
Intercultural Communication	83	24.3	10	53	3.2	7
Law	22	6.4	3	226	13.5	28
Psychology	14	4.1	2	103	6.1	13
Science	7	2.0	1	63	3.8	8
Sociology	32	9.4	4	226	13.5	28
Tourism	30	8.8	4	176	10.5	22
Total	342	100.0	44	1678	100.0	210

Table 2. Original Population Size and Sample Population Sizes for Level III & IV Strata

	Level III			Level IV		
Faculty	Pop.	%	Sample	Pop.	%	Sample
Arts	425	19.7	53	56	15.3	7
Business Admin	121	5.6	15	17	4.7	2
Economics	344	16.0	43	58	15.9	7
Human Services	249	11.6	31	81	22.2	10
Intercultural Communication	17	0.8	2	16	4.4	2
Law	265	12.3	33	57	15.6	7
Psychology	169	7.8	21	24	6.6	3
Science	195	9.1	24	24	6.6	3
Sociology	225	10.4	28	32	8.8	4
Tourism	144	6.7	18	0	0.0	0
Total	2154	100.0	268	365	100.0	45

The population was divided by eight to determine the number of students required to represent that strata in the sample. For example, seven students (56/8) were chosen to represent Level I members of the Arts faculty (see Tables 1 & 2). The population number was divided by eight instead of ten to account for missing data. Even if a large number of the sample dropped out, the sample would still contain a sample fraction larger than 10%. In total, 562 students were required to meet the requirements of the sample.

The resulting number of students required per strata was divided by eight again to determine the number of intact, eight-member classes required to fulfill that strata. For example, seven students were required to fulfill the requirement for the Level IV Law and Politics strata, and so seven students were randomly selected from one intact class. For the Level III Law and Politics strata, 43 students were required, so five intact classes were chosen along with three students randomly selected from a sixth class. Intact classes were chosen where possible to make the retrieval of the quiz answer sheets easier. It was presumed that using intact classes would not affect the sample because each intact class represented a different instructor, and it was unlikely that an instructor would affect student quiz scores. Once this process was complete, the sample students required for each strata were selected at random from a list of the EDC student population.

Once the students for the stratified sample were selected, the data collection began. Due to their modified format, the three quizzes used for test lessons were excluded from the study. The data in the current study therefore comes from the quizzes administered in Lessons 2-4, 6-8, and 10-12 of the spring 2018 semester. The student name, faculty, level, day and period of EDC lesson, instructor name, and gender were recorded on a Google spreadsheet. Each student was also coded with a three-digit number (person) so that the name could be erased and each entry anonymized once the data collection was complete. Each set of quiz answers for each student in the sample was then obtained and the students' answers for each quiz item was recorded, with an *X* marked for any missing data. Since there were two separate sets of answers depending on the version of the quiz administered, each spreadsheet entry was also highlighted by the corresponding color of the quiz version. Once all of the data was recorded, 10% of the data was double checked for errors, revealing only one data entry error (0.22%) from 448 quiz items checked. The data was then adjusted so that all of the recorded answers adhered to one single answer key, as opposed to having two separate sets of data. This resulted in a set of data from which the first three research questions could be answered.

In order to answer research question four, an Excel spreadsheet was constructed for Rasch analysis using Winsteps 4.0.0 (Linacre, 2018). The data from the Google sheet was copied into an Excel spreadsheet because the former cannot currently be processed by Winsteps. Once copied into Excel, the data was edited so that all that remained was a row with a reference code for each quiz item (e.g. 04.2 indicated quiz 4, item 2), a column with the three-digit person numbers, and the responses of each person to each quiz item. This spreadsheet was then imported into Winsteps and a Winsteps control sheet was created for analysis using the dichotomous Rasch model (Rasch, 1960), which was chosen due to the dichotomous nature of the quiz items.

Rasch analysis is a method of analyzing the results of an assessment in order to investigate the performance of the test takers (persons) and their responses to the assessment questions (items). The data is fitted to a statistical model using a mathematical formula that predicts the likelihood of a person's success on an item based on the item's difficulty and the person's ability. For example, a person with high ability will have a higher probability of correctly answering a medium-level difficulty item than a person with low ability. The Rasch model converts the raw scores of the persons and items into *logits*, which makes them comparable. The score in logits is usually spread from around -4 at the bottom of the scale up to 4 at the top, so it is perfectly normal for participants to receive a minus score in logits on an assessment.

A dichotomous Rasch analysis assumes that the construct under investigation is *unidimensional*, which means that it can be measured on a continuum from high to low, where high demonstrates a high level of sophistication on the construct and vice versa. The quiz items analyzed in this study are arguably not measuring a unidimensional construct, since factor, such as memory, receptive vocabulary size, persistence and others, potentially confound the results. However, the ability of the Rasch model to indicate which items were difficult for the sample can be used for a more qualitative follow-up examination of items. Additionally, the distractor frequencies report from Winsteps can be utilized to determine distractors that might be confusing large numbers of students. From this information, problematic quiz items and distractors can be identified and attended to. The information discovered can also be used to assist future quiz item writing.

RESULTS

In order to answer the first two research questions, regarding quiz performance by proficiency level and faculty, boxplots were produced with R (R Core Team, 2018) using the ggplot2 package (Wickham, 2016). Boxplots are an intuitive way of displaying a data set. The median of a group is displayed as a black line inside a box that contains 50% of the data points. The length of the box above the black line is the upper quartile and the length below the box is the lower quartile, and when added together they constitute the inter-quartile range (IQR). The length of the whiskers protruding from the boxes are 1.5*IQR, which in symmetrically distributed data should encompass 75% of the data (Tukey, 1977). Anything outside of 3*IQR (99% of the data) can be considered outliers. If there is a gap between where the bottom of one box ends and the top of another box begins, there is more than likely a significant difference between the two groups (Larson-Hall, 2016).

For research question one, the boxplots revealed that total quiz scores over the course, which was calculated from a maximum score of 72 (9 quizzes with 8 questions on each), decreased with proficiency level (see Figure 1). Although there were no significant differences observed between Levels I through III, there was a significant difference observed between Levels I and IV, $t(48.47) = -2.45$, $p = .02$. The effect size for the comparison, $d = 1.30$, [0.83, 1.77], was large (Plonsky & Oswald, 2014), which indicates that despite the measures in place to make the quiz easier for the Level IV students, they were still performing significantly lower than the Level I students.

For research question two, the boxplots reveled that although Business Administration and Intercultural Communication students' scores were generally higher than the other faculties, there was no significantly different performance between the faculties (see Figure 2).

For research question three, the boxplots (see Figure 3) revealed that there was no significant difference between the two versions of the quiz administered. Therefore, re-sequencing the distractors had no discernable effect on the student's total scores.

For research question four, the distribution of the item and person measures on the Wright map (see Figure 4) suggested that the majority of the items were answered easily by the sample. The left-hand side of the Wright map displays the participants, with each "#" representing five sample members and the M representing the mean score of the test takers, which in this case is 1.40 logits. The right-hand side of the Wright map displays the quiz items, with the most difficult items located towards the top of the scale. Figure 4 shows item 07.3 (week 7, item 3) to be the most challenging for this group, and items 06.1, 02.2, and 02.1 to be the easiest. The mean of the items is always calculated to be 0.00 logits, and is therefore 1.40 logits below the mean of the participants, indicating that the items were generally easy for the sample.

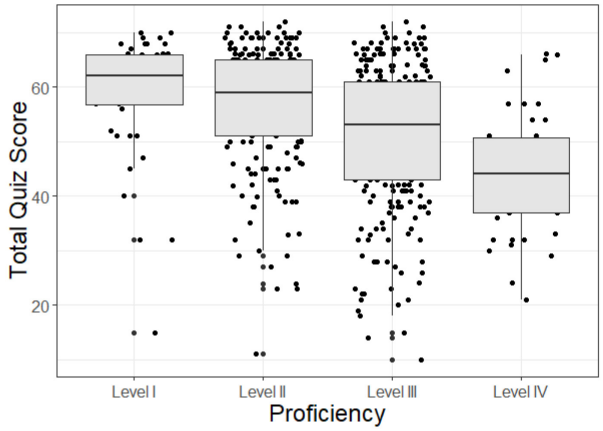

Figure 1. Boxplots representing total quiz score by proficiency level.

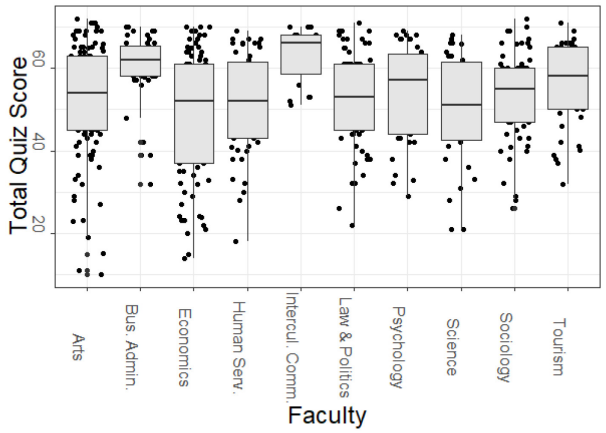

Figure 2. Boxplots representing total quiz score by faculty.

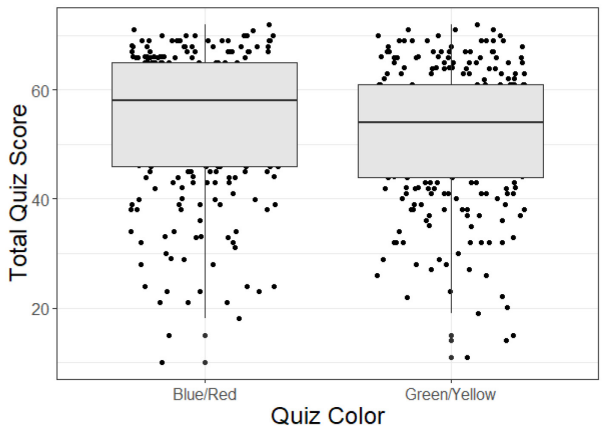

Figure 3. Boxplots representing total quiz score by version of quiz used.

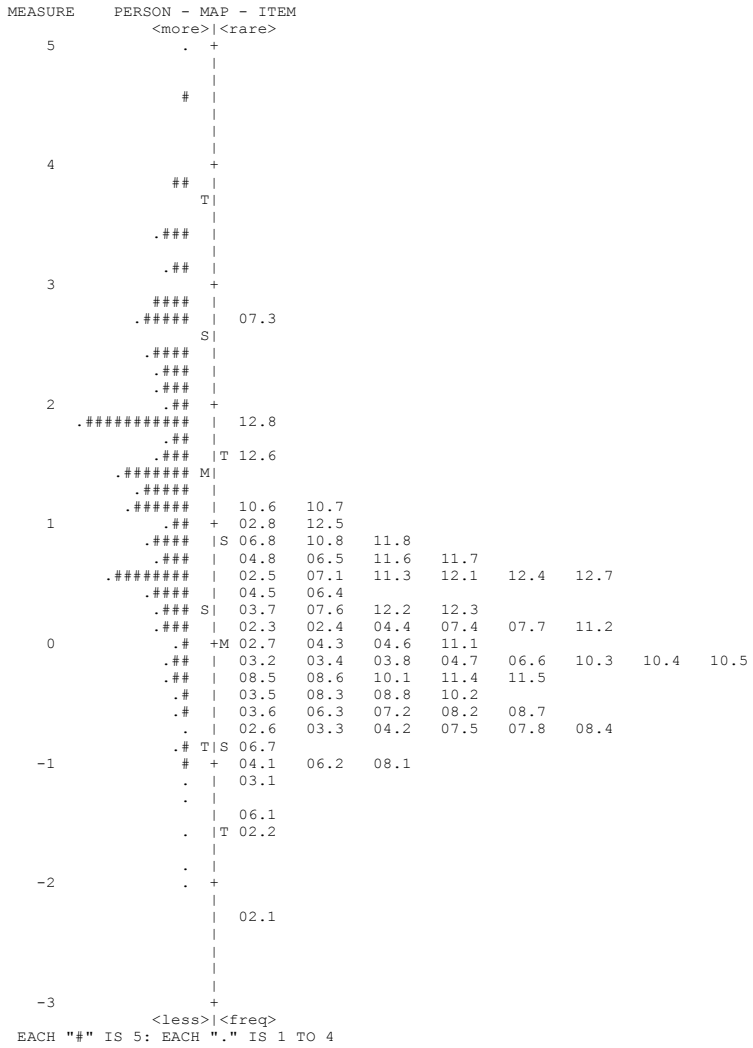

Figure 4. Wright map displaying the distribution of the quiz scores against the distribution of the quiz items.

In order to investigate the quiz answers more deeply, item analysis was implemented on the most challenging items. The easiest items were not considered problematic as no single item was answered by every participant. The easiest item according to the Rasch model, item 02.1, was still answered incorrectly by 14 participants from 543 (2.58%). Therefore, item analysis was only performed on the ten items that were more than one standard deviation above the mean, indicated by an *S* on the Wright map. For each item, the question and answers, and the number, percentage, and mean proficiency in logits of people who selected the answer were recorded for analysis (see Table 3).

Table 3. Distractor Frequency Report for Challenging Items with Mean Logit (Person)

Item	x/o	Questions and Answers	Score	%	*M* Logit
7.3		When did many Japanese people begin to move to the city?	1	0	-0.26
a	x	In the Edo period.	9	2	-0.08
X	x	*Missing data*	59	10	0.22
d	x	During the bubble economy	37	7	1.08
b	x	After the Meiji restoration.	309	55	1.54
c	o	In the 1950s.	147	26	1.83
12.8		What is not always easy to decide?			
X	x	*Missing data*	123	22	0.48
b	x	How values change through our lives.	51	9	0.90
d	x	What values we should teach children.	46	8	1.23
c	x	Why values are different in other countries and cultures.	109	19	1.56
a	o	Which values are most important.	233	41	1.99
12.6		What values can schoolchildren learn from playing and studying together?			
X	x	*Missing data*	121	22	0.47
d	x	Being honest and helpful.	34	6	1.01
b	x	Following rules and working hard.	61	11	1.53
a	x	Being on time and being prepared.	88	16	1.63
c	o	Teamwork and sharing.	258	46	1.82
10.6		A recent survey shows that people who use social media…			
X	x	*Missing data*	77	14	0.28
d	x	are not good at making friends.	14	2	0.82
c	x	spend too much time online.	131	23	1.05
b	x	never meet online friends face-to-face.	30	5	1.25
a	o	have more online friends than real friends.	310	55	1.90
10.7		Professor Yasuda thinks that posting personal information on the internet…			
X	x	*Missing data*	77	14	0.28
b	x	is dangerous.	78	14	0.73
a	x	is not real communication.	66	12	1.12
d	x	is the best way to communicate.	37	7	1.22
c	o	makes people trust each other more.	304	54	1.97
2.8		Many organizations in Japan help *hikkikomori* do activities such as…			
X	x	*Missing data*	24	4	0.00
d	x	going to karaoke.	6	1	0.45
a	x	playing video games.	128	23	0.97
b	x	working.	74	13	1.23
c	o	hiking.	329	59	1.75
12.5		Naoki's parents often ask him to…			
d	x	exercise more.	8	1	0.06
X	x	*Missing data*	122	22	0.46
c	x	take care of his little sister.	48	9	0.51
b	x	do his homework.	65	12	1.47
a	o	wash the dishes.	319	57	1.94

Item	x/o	Questions and Answers	Score	%	*M* Logit
6.8		In order to increase renewable energy, Yokohama is using more…			
X	x	*Missing data*	47	8	0.17
c	x	hydroelectric energy.	45	8	0.75
d	x	geothermal energy.	26	5	1.06
a	x	solar energy.	110	20	1.17
b	o	wind energy.	334	59	1.79
10.8		How many worldwide users does Line now have?			
X	x	*Missing data*	77	14	0.28
c	x	800 million.	69	12	1.04
d	x	40 million.	32	6	1.21
a	x	25 million.	42	7	1.36
b	o	600 million.	342	61	1.77
11.8		Why do many people worry about artificial intelligence (AI)?			
X	x	*Missing data*	80	14	0.32
a	x	AI is expensive to develop.	17	3	0.66
b	x	AI may make people lazy.	41	7	1.06
d	x	AI will reduce human interaction.	87	15	1.23
c	o	AI could become too powerful to control.	337	60	1.80

DISCUSSION

The relatively high number of correct answers across proficiencies and faculties, as well as the low number of problematic questions and distractors, suggest that EDC's reading quizzes are, for the most part, meeting their intended aims within Koda and Yamashita's (2019) reading to learn framework. In other words, a high number of students appear able to correctly answer the majority of the reading quiz questions, which suggests that these students are reading and comprehending the homework reading, and that the literal comprehension MC quiz questions might, in theory, be assisting to some degree with text-meaning building. However, there is a small but not insubstantial number of concerns raised by the results described above.

Firstly and with respect to the first research question, Level IV students appear to be at a distinct disadvantage when identifying the correct answers on the reading quizzes (see Figure 1). Levels I and II receive a longer and more complex version of the homework reading than the one which appears in the textbooks for Levels III and IV. However, the relatively high performance of Level III students on the reading quizzes suggests that Level IV's low performance does not stem from differences in complexity between the higher and lower level readings. Furthermore, Level IV students are no more or less likely to be absent or tardy to lessons than the other levels, a factor that could otherwise shed light on this discrepancy. Level IV's low performance in the present study likely stems from a mismatch between that proficiency level's relatively poor reading proficiency and the readability of the Levels III and IV homework reading, and/or the readability of the quizzes themselves. Additionally, Level IV students might on average have lower motivation to learn English and do well in the course when compared to higher proficiency students. If this is the case, then Level IV students would logically be less inclined to complete and fully comprehend the readings.

As all EDC students should have equal opportunities to succeed in their respective levels, Program Managers should take significant steps to remedy this discrepancy in performance. Such

steps could include the creation and implementation of formalized principles for writing MC items, further text elaboration for the lower level homework readings, and a readability analysis of each reading quiz during their creation and revision.

With respect to the second research question, the results demonstrate that the reading quizzes are no more or less difficult for students from different faculties (see Figure 2). Students from the Business Administration and Intercultural Communication faculties performed noticeably but insignificantly better, a difference that is perhaps explained by these faculties' disproportionately high number of Level I students (see Tables 1 & 2) compared to other faculties. As Level I students have the highest combined TOEIC scores in the course, it is no surprise that they would outperform other levels on the reading quizzes (see Figure 1).

The results shown in Figure 3 alleviate any concern around reordering the sequence of answer options for each MC question between the B&R and G&Y versions of the quiz. As such, the practice of using two versions of the quiz across the course as a means to mitigate cheating should continue.

Regarding the fourth research question, the lack of negative discrimination in the item analysis indicates that all questions are credibly answerable from the information in the source texts. Furthermore, the vast majority of MC items included in the present study do not appear to be problematic for EDC students. However, the set of ten problematic items identified by the Wright map in Figure 4 and subjected to item analysis (see Table 4) provide meaningful data from a materials development standpoint.

Five of the ten problematic items identified were the eighth and last item on their respective quizzes, suggesting that the time limit imposed on the quiz might be rushing students to guess on the final item without fully processing it. These had a relatively high number of missing data points, meaning that on these quizzes a larger than average number of students failed to complete the entire quiz. Looking at each MC item independent of the source information, the most-selected distractor for all five of these questions seems to be a logical best-guess if the information in the paragraph was not read, comprehended, or retained. Therefore, these five final MC items could have been identified as problematic due to overall quiz and/or reading difficulty resulting in a slower total processing time for the eight questions, the number of late arrivals on a given day or days during the respective lesson, or a combination of both. For these five questions, then, MC item or source information revision may only further problematize students' ability to identify the correct answer.

Grouping the list of ten problematic items by quiz suggests that, in some cases, the homework readings' readability may be a more likely source of confusion than an individual MC item's construction. For example, three of the ten problematic items come from Lessons 10 and 12 each, which may indicate that there are broader comprehensibility concerns with these readings. If so, these readings deserve closer scrutiny during textbook revision. However, poor performance on items that appear later in the semester may also be due to student fatigue or an otherwise less rigorous approach to completing the course as time goes by.

The MC question revealed in the item analysis that merits the closest scrutiny is the third question on the Lesson 7 quiz (7.3). This is the only item in which more students (309) selected a particular distractor more often than they selected the correct answer (147). The question reads "When did many Japanese people begin to move to the city?" The favored distractor reads "After the Meiji Restoration," while the key reads "In the 1950s". It should be noted that although only 26% of students correctly identified the key, its 1.83 logits in relation to fewer logits for all of the distractors indicates that those respondents most skilled at answering MC questions on the reading quizzes overall were able to correctly answer this question.

One potential source of confusion for this question is the difference in source information

between the higher and lower level readings. The source information in the higher-level reading reads "In the 1950s, when Japan's economy grew stronger, most young people left their small hometowns to start a new life in the city" (Brereton, Lesley, Schaefer, & Young, 2018a; 2018b). However, in the lower level reading, the source information reads:

> In the 1950s, when Japan's economy grew stronger, young people wanted to live in the city. Cities have good schools, many jobs, and chances to meet new people, so a lot of people left their small hometowns to start a new life there. (Brereton et al., 2018c; 2018d, p. 43)

For Level I and II, therefore, this question was textually explicit (the stem and key appear within the same sentence), while for lower level students, the question was textually implicit (the stem and key are found across sentences). Ironically, the lower level students in the spring 2018 semester suffered an unintended consequence of text simplification and elaboration that is intended to boost comprehension for less proficient students. In essence, the less proficient students had a more difficult time identifying the correct answer to the question independent of their difference in proficiency.

A closer examination of the readings reveals why the distractor, "After the Meiji Restoration" was selected most often. In both the higher and lower level readings, the second paragraph begins with the sentence "Before the Meiji Restoration in 1868, less than 30% of Japan's population lived in cities" (Brereton et al, 2018a; 2018b; 2018c; 2018d). The distractor "After the Meiji Restoration" now appears to be a plausible misinterpretation of information in the text. Therefore, the compounding factors of a textually implicit question for Levels III and IV and a misleading distractor resulted in the presentation of an exceptionally problematic question. There are four aspects of each item that can be revised to improve students' ability to identify the correct answer: (a) the question stem, (b) the set of distractors, (c) the key, and (d) the source information. However, given the different complicating factors in item 7.3, the easiest way to make this question more answerable may be to pose a new MC question altogether, and to take care that it is textually explicit for all levels and contains no misleading distractors.

CONCLUSION

The findings of the current study suggest that the EDC reading quizzes are successful in encouraging students to complete the homework reading, and that the majority of students are able to achieve high scores regardless of faculty or level. There are, however, discrete MC items from the spring semester that item analysis reveal are in need of revision. A similar item analysis will be conducted on the fall 2018 reading quizzes for revision purposes as well. More importantly, steps should be taken to help improve Level IV students' ability and/or motivation to complete the homework reading and comprehend it to a level commensurate with their more proficient peers.

Beyond these findings, it is worth noting that MC questions as a format are problematic. There is an ample body of research demonstrating that readers adjust their reading strategies and comprehension processes to suit the means of comprehension assessment (Rupp, Ferne, & Choi, 2006). In other words, readers will approach and process text differently if they know they will be assessed with an essay question, a cloze quiz, an MC quiz, and so on. Therefore, the narrow EDC multiple-choice question-type of textually explicit, literal comprehension questions may promote only a surface-level understanding of the text, encouraging students to read for details likely to appear in a quiz question and not broader meaning. Additionally, students who have attended test-coaching schools, an exceedingly common occurrence among Japanese university students, will be more test-wise and are likely to have been trained to answer MC questions (Alderson, 2000). Students may therefore be able to answer literal comprehension questions correctly without even

a surface-level understanding of the text. With respect to the role that content plays in EDC, such potential negative washback is hopefully offset by the in-class speaking activities that, in theory, facilitate personal-meaning construction and knowledge refinement in Koda and Yamashita's (2019) reading to learn framework.

Given the aims of the reading quizzes, PMs may wish to consider question types beyond textually explicit and literal comprehension questions, though these would likely require some degree of higher-order thinking skill and/or background knowledge and so should be considered with a great deal of caution. In order to better facilitate text-meaning building, allowing students to access the reading while completing the quiz may also be beneficial (Day & Park, 2005), though a potential negative washback effect would be that students do not deem it necessary to complete the reading ahead of time in order to complete the quiz. This potentially leaves students with less input for use later on in the lesson as output, which is a detriment to knowledge refinement and an impairment of students' ability to meet lesson aims.

Considering the arguments against MC quizzes more generally, PMs may instead wish to move away from using this format for reading quizzes altogether. However, all other formats carry additional disadvantages, mostly related to ease of administration and scoring, which is important in a course at the scale and level of unification as EDC. In such a large-scale course with routine reading quizzes made standard across all faculties and proficiency levels, cheating is and will remain a concern. While the results here indicate that resequencing the key and distractors for each MC item does not affect students' ability to identify the correct answer, it is logistically unreasonable in the present context to create and administer enough different quiz versions to protect against cheating. Currently, two versions of MC quizzes likely remain the most viable method of encouraging students to complete the homework reading with the ultimate aim of recycling that content to practice the target language forms in class.

In addition to the principles currently observed by PMs when creating and revising MC quiz questions, the following additional principles should be considered in future semesters to further meet reading quiz aims. Quiz questions should: (a) have lower readability targets than the associated readings, (b) avoid synonymy or other aspects that do not specifically target literal reading comprehension, and (c) be textually explicit based on both the higher and lower level readings. In additions, PMs should consider better standardizing the procedure for quiz administration, for example setting a standard time limit for all students to follow rather than providing instructors with a range of three to four minutes. Regardless of the reading quiz being administered to students in any given lesson, its format and item construction should help students to interact with and derive meaning from the text.

REFERENCES

Alderson, J. C. (2000). *Assessing reading*. Cambridge, UK: Cambridge University Press.

Bailey, K. M. & Curtis, A. (2015). *Learning about language assessment: dilemmas, decisions, and directions* (2nd ed.). Boston, MA: National Geographic Learning.

Brereton, P., Lesley, J., Schaefer, M. Y., & Young, D. (2018a). *What do you think: Interactive skills for effective discussion, Book I* (8th ed.). Tokyo, Japan: DTP Publishing.

Brereton, P., Lesley, J., Schaefer, M. Y., & Young, D. (2018b). *What do you think: Interactive skills for effective discussion, Book II* (9th ed.). Tokyo, Japan: DTP Publishing.

Brereton, P., Lesley, J., Schaefer, M. Y., & Young, D. (2018c). *What do you think: Interactive skills for effective discussion, Book III* (9th ed.). Tokyo, Japan: DTP Publishing.

Brereton, P., Lesley, J., Schaefer, M. Y., & Young, D. (2018d). *What do you think: Interactive skills for effective discussion, Book IV* (9th ed.). Tokyo, Japan: DTP Publishing.

Brinton, D. M., & Snow, M. A. (2017). The evolving architecture of content-based instruction.

In M. A. Snow & D. M. Brinton (Eds.), *The content-based classroom: New perspectives on integrating language and content* (2nd ed.) (pp. 2-20). Ann Arbor, MI: University of Michigan Press.

Day, R. R. & Park, J. (2005). Developing reading comprehension questions. *Reading in a Foreign Language, 17*(1), 60-73.

Koda, K. & Yamashita, J. (2019). *Reading to learn* in a foreign language: An integrated approach to FL instruction and assessment. In K. Koda & J. Yamashita (Eds.), *Reading to learn in a foreign language: An integrated approach to foreign language instruction and assessment* (pp. 3-8). Oxford, UK: Routledge.

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R* (2nd ed.). New York, NY: Routledge.

Linacre, M. (2018). Winsteps® Rasch measurement computer program. Beaverton, OR: Winsteps.com.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge, UK: Cambridge University Press.

Oh, S. (2001). Two types of input modification and EFL reading comprehension: Simplification versus elaboration. *TESOL Quarterly, 35*(1), 69-96. doi:10.2307/3587860

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danmarks Paedigogiske Institut.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing, 23*(4), 441-474. doi:10.1191/0265532206lt337oa

Trochim, W. M., Donnelly, J. P., & Arora, K. (2016). *Research methods: The essential knowledge base.* Delhi, India: Cengage Learning.

Tukey, J. W. (1977). *Exploratory data analysis.* Boston, MA: Addison-Wesley.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer-Verlag.

Yano, Y., Long, M. H., & Ross, S. (1994). The effects of simplified and elaborated texts on foreign language reading comprehension. *Language Learning, 44*(2), 189-219. doi:10.1111/j.1467-1770.1994.tb01100.x

Young, D. (2016). Textbook revision in the EDC context: Readability and topic interest. *New Directions in Teaching and Learning English Discussion, 4*, 295-302.