Reading in a Foreign Language October 2022, Volume 34, No. 2
ISSN 1539-0578 pp. 249–270

Exploring Perceived Difficulty of Graded Reader Texts

Yuya Arai
Waseda University
Japan

Abstract

Although proponents of extensive reading (ER) have recommended easy reading material, book difficulty has been poorly defined and operationalized in previous studies. The present study argues for the use of perceived text difficulty for operationalizing book difficulty based on empirical findings (Holster et al., 2017), reading purposes in ER, and the importance of readers' perspectives. A total of 162 Japanese English-as-a-foreign-language university students rated the difficulty of 15 texts excerpted from graded readers (GRs). The data were analyzed by conducting a many-facet Rasch analysis (Linacre, 1989; Rasch, 1960/1980), where a rating scale model (Andrich, 1978) was tested with persons, texts, and graded readers' levels as the facets of measurement. The results revealed that perceived text difficulty could not replicate the stated difficulty level provided by the GR publisher, reinforcing the necessity of examining perceived text difficulty in ER research and practice in the second and foreign language classroom.

Keywords: perceived text difficulty, graded readers, many-facet Rasch measurement, extensive reading, pleasure reading

Second and foreign language (L2) extensive reading (ER) has emphasized the importance of easy reading material in its programs in order that learners are exposed to a large amount of reading material. Bamford and Day (2004) defined ER as "an approach to language teaching in which learners read a lot of easy material in the new language" (p. 1). Based on Krashen's (1985) hypothesis concerning the importance of input in L2 acquisition, Day and Bamford (1998) have also suggested that books for ER should be the level "well within [learners'] linguistic competence" (p. 16) represented by "*i* minus 1," where "*i*" stands for their linguistic proficiency and "minus 1" represents the levels slightly below their proficiency.

However, the jury is still out on how to operationalize such ambiguous concepts as easy reading material. This could lead to inconsistent interpretations of reading material among ER programs and reduce the effectiveness of ER suggested in previous empirical studies (e.g., Jeon & Day, 2016; Nakanishi, 2015). Some studies (e.g., Bahmani & Farvardin, 2017; Chiang, 2016; Yang et al., 2021) have suggested that the difference in book difficulty represented by "*i* plus 1" (i.e., the level slightly above learners' linguistic proficiency; see Krashen, 1985) and "*i* minus 1" can affect the degree of reading comprehension as well as motivation, anxiety, and reading attitude.

Accordingly, ER researchers should define book difficulty in an appropriate way, without which the validity of the results in such studies may be doubted.

Nevertheless, it is also true that researchers have tried to operationalize book difficulty in the context of ER. For instance, some studies examining the differences in book difficulty (e.g., Bahmani & Farvardin, 2017) used the six levels of the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001) for the operationalization of "*i* plus or minus 1," despite the possibility that the six-level CEFR may be too roughly attuned to distinguish the subtle difference between "*i* plus or minus 1." Other studies (e.g., Chiang, 2017; Yang et al., 2021) employed the publishers' stated levels of graded readers (GR) (i.e., books written mainly for L2 learners with strictly controlled vocabulary) to classify the participants into either "*i* plus 1" or "*i* minus 1" reading groups. However, it has been pointed out that developing grade levels may not always be consistent with the publishers' own word list (Wan-a-rom, 2008), which casts doubt on the reliability of the grading system.

Similarly, previous ER studies have operationalized book difficulty using lexical coverage, word families, and readability indices, each of which can have its own limitations in discussions of book difficulty. Due to these limitations, which will be identified in the related literature in the next section, the present study instead examined perceived text difficulty as a candidate for operationalization of ER material difficulty by conducting a many-facet Rasch analysis (Rasch, 1960/1980). As discussed below, examining how perceived text difficulty functions in comparison to actual book difficulty (e.g., grade levels) could encourage readers' perspectives to be included in the discussion of book difficulty in the ER programs where the importance of pleasure reading is emphasized.

Literature Review

Previously proposed approaches to defining ER book difficulty

Lexical coverage. Lexical coverage means "the percentage of known words" in the reading text (Webb & Nation, 2017, p. 280). The importance of lexical coverage as an index for book difficulty has often been emphasized in ER studies to operationalize book difficulty. For example, Day and Bamford (2002) suggested that ER materials should ensure "that learners must know at least 98% of the words" (p. 137).

To the best of the present author's knowledge, however, what such percentages really mean has not been thoroughly discussed. In light of various study contexts, studies on lexical coverage may not always support the use of this index as a criterion for ER material (Arai, 2022). In particular, lexical coverage would not be appropriate for operationalization of book difficulty for the following three reasons.

First, most studies on lexical coverage are not conducted in the ER context. For instance, although Nation (2006) found 98% lexical coverage should be important for "unassisted comprehension of written" material (p. 59), he calculated this percentage based on a corpus analysis, and not by requiring learners to read actual texts extensively. Hu and Nation (2000) also recommended 98% lexical coverage for pleasure reading, but their participants read a text that

had no more than 673 words. In addition, Herman and Leeser (2022) conducted a lexical coverage study, but they required their study participants to read only one GR text. Given the small amount of reading, the results in these studies may not easily be generalized to pleasure reading where learners are required to read more.

Second, studies on lexical coverage assumed various reading purposes, some of which would not be consistent with those in ER programs. For example, while Laufer (1989) suggested 95% lexical coverage, her participants were simply asked to underline unknown words during reading, which cannot be considered as reading but rather as searching for unknown words in text (Schmitt et al., 2011). In addition, Schmitt et al. required their participants to read academic texts and take a comprehension test, the purposes of which could thus be reading in order to respond to test items, and this may not be considered as reading for pleasure (Day & Bamford, 2002).

Third, factors other than lexical coverage may explain a large part of reading difficulty. For example, Gillis-Furutaka (2015) suggested factors such as syntactic information in the texts, readers' background knowledge, and the use of illustrations. To sum up, relying only on lexical coverage might not be enough to adequately define book difficulty.

Word families. Word families, which are frequently mentioned in GRs (including Oxford Bookworms Library series published by Oxford University Press), are comprised of a headword (e.g., *use*), inflections (e.g., *using*), and derivations (e.g., *useful*) (Webb & Nation, 2017). Such GR publishers often provide learners with the information on word families alone without referring to lexical coverage. However, recent studies argue against the use of word families because of the concept of word families' overestimation of vocabulary size, meaning that a learner who knows the word *use* cannot necessarily understand the meaning of *usability* in the reading text (e.g., McLean, 2018). For this reason, word families can be too roughly attuned to operationalize book difficulty for ER where the subtle difference between "*i* plus and minus 1" may matter.

Readability indices. There are some examples of readability indices, or reading difficulty indices, as represented by the Flesch Reading Ease (Flesh, 1948) or the Flesch-Kincaid Grade Level (Kincaid et al., 1975) in first language (L1) reading studies. However, Nation and Waring (2020) cautioned against applying L1 indices to L2 ER. This suggestion was partly supported empirically by Holster et al. (2017), who found that the Lexile Framework, an L1 reading scale taking reading and text measures into consideration, was not as predictive of L2 perceived book difficulty as expected with small percentages of variance explained.

Some L2 studies have used L2 indices. Among them is the Yomiyasusa Level, which is "a 100-point scale ... [that not only] reflected the word count of the books, but also took account of factors such as illustrations and text styles" (Holster et al., 2017, p. 219). These authors found that the Yomiyasusa Level could play an important role in predicting perceived book difficulty in L2 ER. However, it is worth noting that the Yomiyasusa levels of most GRs are already determined and fixed by laypeople who love reading extensively but are not trained as raters. Furthermore, the lack of empirical analyses and reliable procedures to establish the levels would make it difficult to interpret and use the index for defining book difficulty. To sum up, the three

options for operationalizing ER material difficulty (i.e., lexical coverage, word families, and readability indices) have some limitations.

An alternative operationalization of ER book difficulty: Perceived text difficulty

Perceived text difficulty is based on readers' perceptions of text difficulty. One reason for applying the concept in the present study was that using perceived text difficulty to define book difficulty for ER was empirically supported by Holster et al. (2017). They conducted a pioneering study that revealed that "students' self-report ratings provided valid measurement of book difficulty" (p. 236). The participants were 810 female Japanese university students on an ER course, among whom 668 students read 1,016 books in total and rated book difficulty based on a four-point Likert scale. The data were analyzed by applying the partial credit model in the many-facet Rasch analysis (Wright & Masters, 1982; see the next section for the explanation of Rasch measurement) because Holster et al. also wanted to include other categories than perceived book difficulty (e.g., time taken to complete the reading) based on different rating scales. It was found that perceived book difficulty generally corresponded to the grade levels developed by the GR publisher. They also concluded that book length was one of the important factors that were predictive of perceived book difficulty.

Although Holster et al.'s (2017) findings were important in that perceived book difficulty could be a good predictor of actual book difficulty in ER settings, the finding that book length played an important role in learner perceptions may suggest that the participants rated their books as difficult due to other length-related factors, such as fatigue, anxiety, avoidance, or time pressure. For example, some participants might have rated a lengthy book "difficult" just because they became fatigued by the idea of finishing it even before opening it. For the purpose of examining perceived text difficulty, therefore, controlling the effects of book length may contribute to furthering our understanding of how perceived text difficulty will function similarly or differently from the stated difficulty levels provided by GR publishers.

Another reason for examining perceived text difficulty is its relationship to one of the important reading purposes in ER: reading for pleasure. Day and Bamford (2002) advocated pleasure reading as a key principle for teaching ER, which makes the term pleasure reading interchangeable with ER (Arai, 2022). Proponents of ER have emphasized the importance of pleasure experiences as a means of motivating learners to read extensively (Nation & Waring, 2020).

Given that pleasure reading is a subjective experience where readers feel pleasure during reading, it is compatible with the concept of perceived text difficulty, which is also subjective. For example, Arai (2022) examined this compatibility between pleasure experiences and perceptions of text difficulty from the perspectives of the flow theory (Csikszentmihalyi, 1975), partly because the concept of pleasure could be operationalized by flow experiences (Yamashita, 2015). Because the flow theory emphasizes the importance of the relationship between the perceived challenges of a task and the perceived skills of a person engaging with the task, which can be a factor in generating flow, perceived text difficulty could play an important role in pleasure reading (Arai, 2022).

Finally, the discussion of perceived text difficulty could emphasize the importance of assuming readers' perspectives in (extensive) reading studies. As Alderson (2000) suggested, readers and reading material are the two fundamental variables in reading studies. However, most book difficulty indices frequently mentioned in previous ER studies (e.g., grade levels, lexical coverage, word families, and some readability indices) do not always reflect readers' perspectives. This might give ER practitioners and researchers the impression that book difficulty is rather fixed despite the possibility that readers may have different perceptions of the difficulty of a book. Therefore, perceived text difficulty should be examined along with other variables to explore a better way of assessing book difficulty.

Many-facet Rasch measurement

Following Holster et al. (2017), the present study employed a many-facet Rasch analysis to examine the functioning of perceived text difficulty. The Rasch measurement (Rasch, 1960/1980) offers an estimate of "a simple mathematical relationship between ability and difficulty [expressed] as the probability of a certain response" (McNamara, 1996, pp. 152–153). Later, Linacre (1989) further developed the many-facet Rasch measurement to enable more facets (i.e., variables) than item difficulty and person's ability to be analyzed. Unlike item response theory (IRT) models, the Rasch analysis has the property of measurement invariance, which enables item difficulty measures to be free from a given set of test takers (e.g., Bond & Fox, 2015; Eckes, 2015).

Although the Rasch analysis, as well as IRT, has not been common in previous ER studies, the present study used this method mainly because these characteristics of many-facet Rasch measurement would encourage discussions of perceived text difficulty to be based on person-free item difficulty estimations, thus creating some implications for the development of GRs (for more detailed discussion of measurement invariance, see Sick, 2008). Furthermore, employing this method in the present study could change raw scores into logit (log odds unit) values on an interval scale, suggesting that it would be possible to identify not only whether, but also by how much, each grade level is different in terms of difficulty. This would make it easier to discuss the relationships among GR texts and the minute difference between "*i* plus and minus 1." Therefore, by employing the many facet-Rasch analysis, this study addressed the following three research questions concerning how perceived text difficulty would function compared with grade levels:

1. How difficult are GR texts in the higher grade levels compared with those in the lower ones in terms of perceived difficulty?
2. How consistent are the texts in the same grade levels in terms of perceived difficulty?
3. How distinguishable are the grade levels from one another in terms of perceived difficulty?

Method

Participants

The participants consisted of 162 Japanese undergraduate students from five universities in Japan, all of whom were non-English majors taking mandatory English-as-a-foreign-language (EFL) courses to fulfill

graduation requirements. University students were targeted in this study mainly because results and discussion would be compatible with previous studies that assumed the same population (e.g., Holster et al., 2017). Information on the participants' gender was not collected. Their English proficiency levels ranged from the A2 to B2 CEFR based on proficiency tests taken within two years prior to their participation in the study, including the Eiken Test in Practical English Proficiency (Eiken; Eiken Foundation of Japan, https://www.eiken.or.jp/eiken/en/); Global Test of English Communication for Students (GTEC; Benesse Corporation, https://www.benesse.co.jp/gtec/en/); the Test of English for International Communication (TOEIC®; Educational Testing Service, https://www.ets.org/toeic); the Test of English as a Foreign Language Internet Based Testing (TOEFL® iBT; Educational Testing Service, https://www.ets.org/toefl/test-takers); and self-evaluation of their proficiency reported by those who had not taken any of the above-mentioned tests (n = 16).

Materials

Texts from GRs. Fifteen first-page texts were excerpted from the beginning parts of GRs from five grade levels (i.e., three texts from each grade level). The GRs were Oxford Bookworms Library series published by Oxford University Press. Although the grade level of the series ranges from Starter to Level 6, the present study relabeled them as Level 1 to 7 for the convenience of analysis. This study only used the texts from Level 1 to Level 5. The exclusion of Levels 6 and 7 was based on Holster et al. (2017), whose participants read fewer books in higher grade levels than those in lower ones. Table 1 shows the texts used in the present study. These books were sampled because the texts were around median lengths in each grade level so that the texts were representative of their grade levels. The books were also selected in light of the genres: No more than three books of the same genre were included so as to minimize bias toward particular book genres. Instead of sampling pages randomly, the first pages from the beginning were excerpted because it would be more natural to assume that such books are read from the beginning, and because excerpts from the middle of the story would have resulted in difficulties in the understanding of the stories' plots. Although the present study employed only the first pages of each GR due to the time constraint, excerpting the same number of pages contributed to minimizing the differences in book length among the texts to some extent. This was an attempt to control the effects of book length discussed in the literature review section.

The sampled texts were then randomly assigned to three booklets (Booklets 1 to 3). Two copies of each text were assigned to two of the three booklets (i.e., 10 texts in each booklet). The texts were randomly ordered in the booklets for the purpose of counterbalancing, and to satisfy local independence, an assumption of the Rasch analysis whereby a response on one item should not influence that on another (Eckes, 2015, p. 27). Consequently, booklets always enabled the participants to read two texts from each grade level, satisfying another requirement of the Rasch analysis concerning data connectedness between participant and text facets (Eckes, 2015; McNamara, 1996).

Table 1
Summary of Texts Used in the Many-Facet Rasch Analysis (Oxford University Press, 2019)

No.	Title	Grade level	Book length	Text length[a]	Genre[b]
1	*Sing to Win*	1	1,400	64	H
2	*Police TV*	1	1,500	51	CM
3	*Starman*	1	1,600	65	FH
4	*The Coldest Place on Earth*	2	5,500	232	T
5	*Titanic (Factfiles)*	2	5,600	130	NF
6	*The Bridge and Other Love Stories*	2	5,605	188	H
7	*The Year of Sharing*	3	6,390	191	FH
8	*Ear-rings from Frankfurt*	3	6,422	134	TA
9	*The Jungle Book*	3	6,510	108	C
10	*The Picture of Dorian Gray*	4	10,245	237	FH
11	*'Who, Sir? Me, Sir?'*	4	10,295	137	H
12	*A Christmas Carol*	4	10,385	103	C
13	*20,000 Leagues under the Sea*	5	15,748	227	TA
14	*The Big Sleep*	5	15,960	286	CM
15	*The Scarlet Letter*	5	15,965	226	C

Notes. [a]text length excerpted from the first pages of the 15 GRs.
[b]C = Classics; CM = Crime and Mystery; FH = Fantasy and Horror; H = Human Interest; TA = Thriller and Adventure; T = True Stories; NF = Non-Fiction.

Questionnaire on perceived text difficulty. Every time they read each text in the randomly distributed booklets, the participants were required to rate the difficulty on a five-point Likert scale: (1) "*i* minus 2 or below," (2) "*i* minus 1," (3) "*i*," (4) "*i* plus 1," and (5) "*i* plus 2 or above." The scale, also used by Arai (2022), was employed in the present study partly because each category should be on an interval scale, which is compatible with the Rasch analysis changing raw scores into equal-interval logit values. The participants had some chances to practice rating in advance of the questionnaire administration.

Data Collection

The participants were convenience sampled. The present author asked for the help of four instructors in charge of EFL classes, where the students willing to take part in this study were recruited. The instructors informed the students of their right to anonymity within the study and the voluntary nature of study participation. After that, the instructors distributed the three types of booklets randomly in order to minimize the order effects. The students spent thirty minutes completing the questionnaire. After its administration, the questionnaire was then mailed back to the present author.

In addition to the participants who were contacted indirectly by the instructors, the present author contacted four students directly, to whom the questionnaire was distributed and collected via email in accordance with the same procedure described above.

Analysis

As discussed in the literature review section, the data obtained were analyzed by means of many-facet Rasch analysis (Rasch, 1960/1980). The rating scale model (Andrich, 1978) with persons, texts, and grade levels as the facets of measurement was employed. To this end, Minifac, a demonstration version of Facets software, was used for analysis (http://www.winsteps.com/minifac.htm).

Before the main analysis, global model fit and other fit statistics were examined as an important assumption of Rasch measurement. These statistics show the degree to which the data of each facet are consistent with the expected responses generated by the mathematical model.

It is also noted that Rasch measurement, as discussed in the literature review, estimates both participants' ability and text difficulty. Although this "ability" variable usually represents participants' ability measured by a reading comprehension assessment, the present study did not measure reading comprehension (see also the limitation section). Therefore, "ability" in this study is considered to be "agreeability" (Bond & Fox, 2015), or the degree of agreement, with the questionnaire's categories ranging from 1 (i.e., least agreeable) to 5 (i.e., most agreeable). For example, if a participant frequently chose categories of smaller numbers (i.e., less agreeable in terms of difficulty), this meant that they perceived the texts to be easy because they could be assumed to have had a higher "ability." In contrast, if another participant tended to choose categories such as 4 or 5 representing the difficulty of the texts (i.e., more agreeable in terms of difficulty), it could be said that they had a lower "ability." For convenience, both agreeability and "ability" are used interchangeably in the present study.

In addition, it should be pointed out that the present study had the text facet nested within the grade level facet. In order to resolve the disconnectedness across subsets in the dataset due to the nested structure between these two facets, two Minifac runs were conducted by adding group anchoring to the initial three-facet model, following Linacre (2012). In Run 1, the grade level difficulty measures were anchored to the group average in order to obtain the text difficulty measures (Research Question 2). In contrast, in Run 2, the text measures were anchored in order to obtain the grade level difficulty measures (Research Question 3).

Results

As for global model fit, it was found that the percentage of unexpected responses was 0.3%, lower than the proposed criterion of 1% (Eckes, 2015), meaning that the global model fit was satisfactory.

The results of fit statistics for texts suggested that, among the 15 texts excerpted from GRs, Texts 4 and 9 (see also Appendix A) were considered as misfitting (i.e., poorly fitting the model) because of the values of infit and outfit standardized *z*-scores, an indicator of fit anomalies, given the criteria used in the previous literature (e.g., Aryadoust et al., 2021; Bond & Fox, 2015; Knoch & McNamara, 2015). This could suggest that some participants had a higher ability but perceived the texts to be more difficult than expected by the model, or vice versa (Knoch & McNamara, 2015). Because the reasons for such a misfit should be examined in more detail

instead of eliminating the misfitting data from further analysis (Bond & Fox, 2015), a possible reason will be examined with Research Question 1. As for fit statistics for grade levels (ranging from 1 to 5), Level 1 had a small value of outfit standardized *z*-score, suggesting the probability of overfit (i.e., being too predictable to be true; see also Appendix B). As with the case of the misfitting texts, a possible reason will be discussed later in relation to Research Question 1. Finally, the results of fit statistics for participants ($N = 162$), suggested that eight participants were misfitting, while 16 participants were overfitting, as suggested by their infit or outfit standardized *z*-scores. As discussed below, one of the reasons for this result could be the imbalance between the participants' ability and text difficulty.

RQ1: How difficult are GR texts in the higher grade levels compared with those in the lower ones in terms of perceived difficulty? To address this research question, Table 2 presents the descriptive and frequency statistics for the participants' perceived text difficulty. Each item of data was normally distributed. The frequency tables revealed that small numbers of participants chose Category 5 ("*i* plus 2 or above") for each text, suggesting that the texts seemed not to be so difficult for them. As for the descriptive statistics, it was found that Text 14 (from Level 5) and Text 7 (from Level 3) had the highest mean values ($M = 3.1$), while Text 13 (from Level 5) had the smallest one ($M = 2.0$), suggesting that the former two texts were perceived as the most difficult, while the latter was perceived as the least difficult. It was remarkable that two texts from the same grade level (i.e., Level 5) were perceived as most difficult and easiest. Therefore, the ranking of the mean difficulty value across two texts for each grade level suggested that perceived difficulty was not related to the ordering of the original grade levels provided by the GR publisher, and that texts from higher grade levels were not necessarily perceived as more difficult. The ranking also appeared not to be systematically related to other features including text length in light of the findings that shorter texts (e.g., Texts 1, 2, and 3) were located in the middle of the ranking, while longer texts (e.g., Texts 10 and 13) were not necessarily perceived as difficult.

Overall, the descriptive results were consistent with those obtained in the many-facet Rasch analysis. Figure 1 maps the logit values (in the "Measr" columns) of the three measurement facets on the same scale ranging from 1 to 5 (in the "Scale" columns): "Participants," "Texts" (from 1 to 15), and "Grade Levels" (from 1 to 5). The left map was outputted from Run 1 (i.e., to examine text difficulty), while the right one being from Run 2 (i.e., to examine grade difficulty). The "Participants" column represents the participants' "ability" in the form of "." (i.e., one participant) or "*" (i.e., two participants). The higher the value of measure a participant obtained, the higher the "ability" they were assumed to have. Meanwhile, lower values of measures in the "Texts" and "Grade Levels" columns correspond to easier texts or grade levels. For example, the left map showed that Text 13 was considered to be easiest.

Table 2
Frequency Tables and Descriptive Statistics (N = 162)

				Frequency					Descriptive statistics			
Text	GL	TL	*n*	1	2	3	4	5	*M*	*SD*	Min	Max
14	5	286	100	4	18	46	30	2	3.1	0.86	1	5
7	3	191	101	5	13	55	25	3	3.1	0.82	1	5
6	2	188	106	9	8	61	27	1	3.0	0.86	1	5
15	5	226	113	13	20	47	32	1	2.9	0.99	1	5
4	2	232	102	15	26	30	26	5	2.8	1.13	1	5
1	1	64	108	12	23	51	21	1	2.8	0.94	1	5
9	3	108	113	16	26	48	20	3	2.7	1.03	1	5
2	1	51	112	24	22	49	16	1	2.5	1.03	1	5
3	1	65	102	18	32	34	17	1	2.5	1.01	1	5
10	4	237	106	22	28	37	18	1	2.5	1.05	1	5
12	4	103	102	19	33	37	12	1	2.4	0.96	1	5
5	2	130	112	20	44	36	12	0	2.4	0.90	1	4
8	3	134	108	31	32	36	8	1	2.2	0.99	1	5
11	4	137	114	40	38	27	8	1	2.1	0.98	1	5
13	5	227	108	40	34	29	4	1	2.0	0.93	1	5

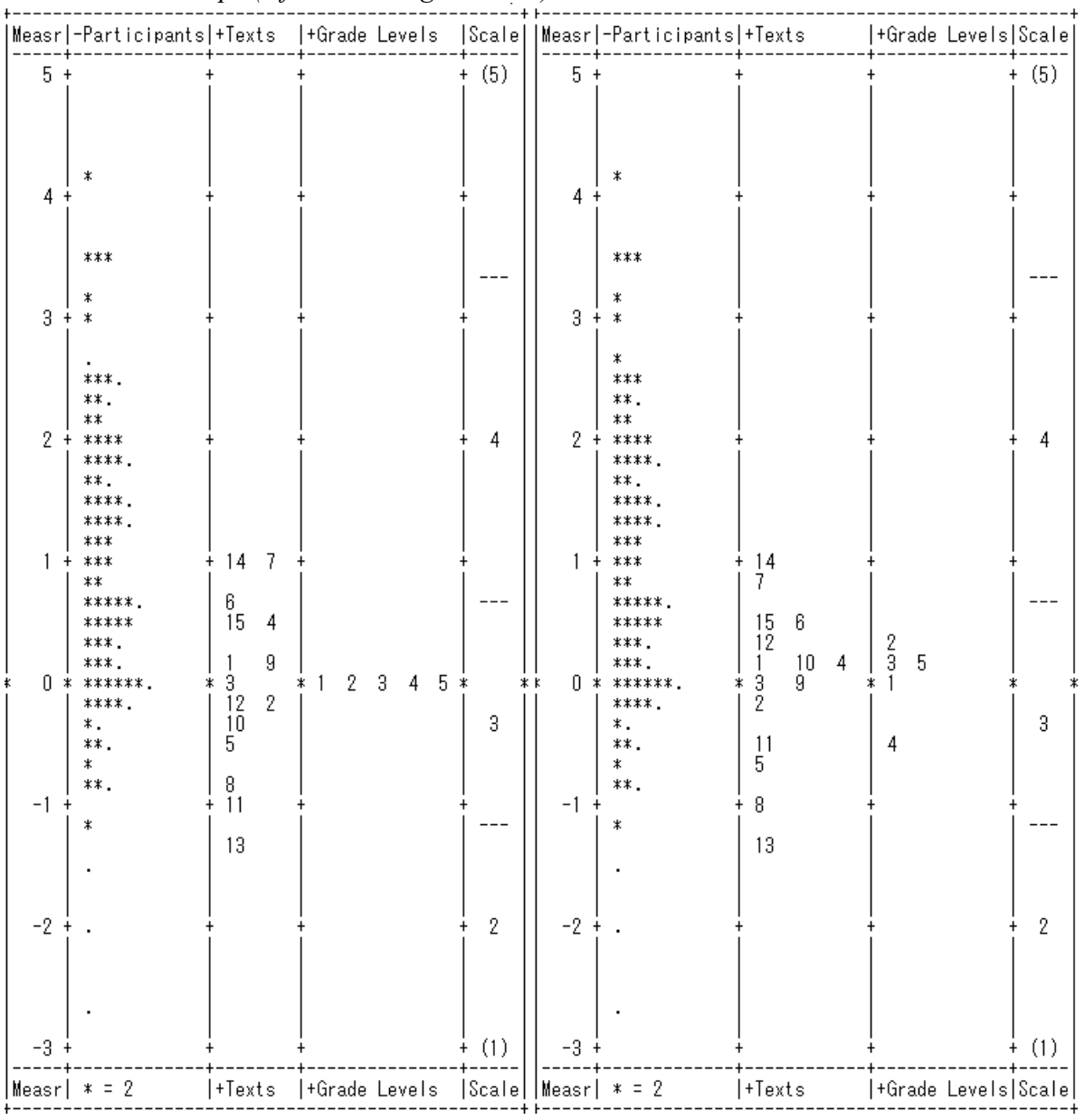
Notes. Text = text number (ranging from 1 to 15); GL = grade level (ranging from 1 to 5); TL = text length, or the word count for the text except used in this study; *n* = participant sample size for each text; *M* = mean; *SD* = standard deviation.

The relationship between ability and difficulty shown in Figure 1 was consistent with the pattern observed in the descriptive statistics. Among others, the distribution of the participants in the "Participants" columns was positively skewed. Higher values of the Participants measures meant higher ability (i.e., lower agreeability). In contrast, the left map (from Run 1) showed that the Texts measures were not severely skewed. Because logit values made it possible to compare the participants and text difficulty on the same scale, these results mean that the participants may have generally considered the texts easy. This interpretation would also be supported by the mean logit value of the participant facet, which was 0.99, while the average logit values of the other two facets were 0.00.

Being on an equal-interval logit scale, the map also shows the degree of differences in perceived difficulty among texts and grade levels. In the "Texts" column on the left map (from Run 1), for example, Texts 7 and 14 had almost the same and largest logit values, meaning that both of them were equally rated as the texts that were the most difficult. Meanwhile, the "Grade Levels" column on the right map from Run 2 showed that Level 4 had the smallest value of measures and could be considered as easiest.

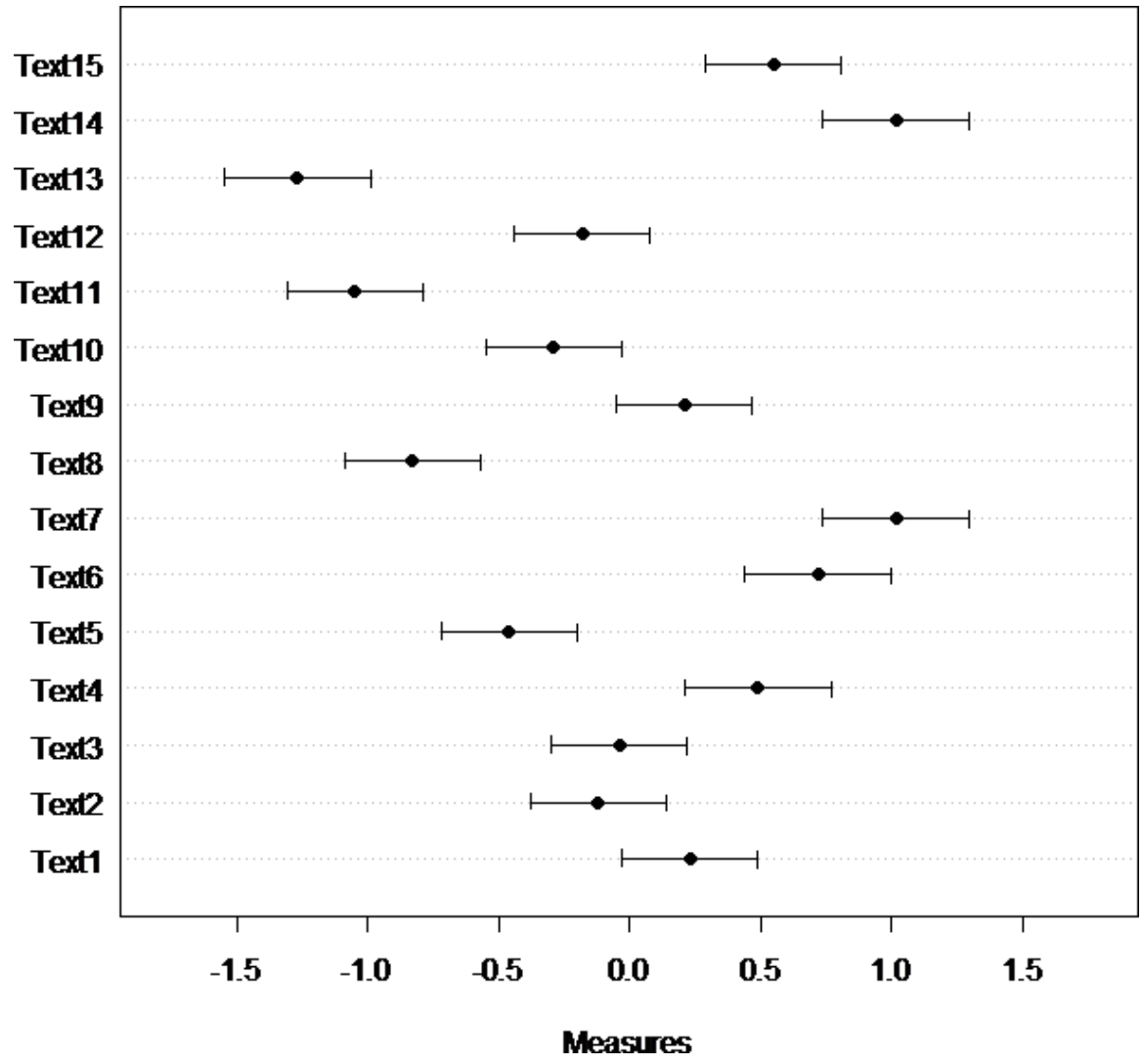
Figure 1

Rasch variable maps (left: Run 1; right: Run 2)



In order to further examine text difficulty, Figure 2 visualizes the result of text measurement obtained from Run 1 with the information on 95% confidence intervals (for the detailed information, see Appendix A). Although being from different grade levels, Texts 7 (Level 3) and 14 (Level 5) had the largest logit values and could be considered as the most difficult. Interestingly, Texts 1 and 9 had almost the same logit values despite being from different grade levels (Levels 1 and 3).

Figure 2
Text measures and 95% confidence intervals



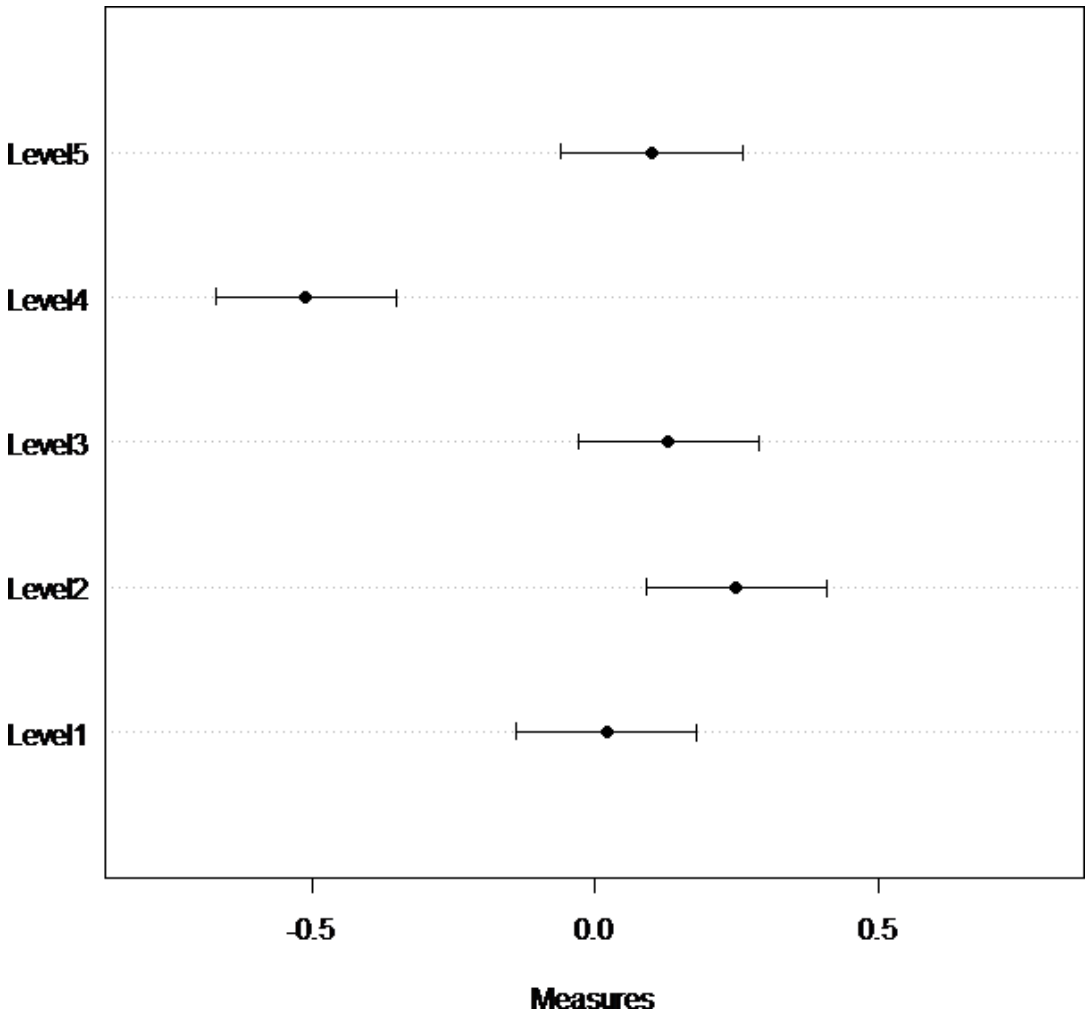
In addition, as noted above, the values of infit mean square suggested that Texts 4 and 9 were considered as misfitting. A possible reason could lie in the text length. As shown in Table 1, Text 4 had more words (232) than the other texts from Level 2, which were Texts 5 (130 words) and 6 (188 words). It is also worth noting that the number of words in Text 4 were more than those in the next grade level (i.e., Level 3). In contrast, Text 9 had fewer words (108 words) than the other texts from Level 3 (i.e., Text 7: 191 words; Text 8: 134 words) and was even shorter than

those from Level 2. This being the case, text length may explain why the difficulty of some texts was not perceived as predicted by the Rasch model.

Similarly, Figure 3 visualizes the result of grade level measurement obtained from Run 2 with the information on 95% confidence intervals (for the detailed information, see Appendix B). The value of Level 2 was largest, meaning that it was the most difficult level of all, while that of Level 4 was smallest, meaning that it was the easiest level of all. In addition, Levels 1, 3, and 5 had almost the same values, suggesting that they had the same difficulty.

Figure 3

Grade Level Measures and 95% Confidence Intervals



Furthermore, as discussed earlier, Level 1 was considered as overfitting. A possible reason could lie in the functions of five categories used for rating text difficulty (i.e., from "*i* minus 2 or below" to "*i* plus 2 or above"). The Rasch analysis estimated thresholds, or boundaries, between two adjacent categories, which could be an index of how distinctive each category was. The

results showed that the threshold value between Categories 1 (i.e., "*i* minus 2 or below") and 2 (i.e., "*i* minus 1") was -2.37 while that between Categories 2 and 3 (i.e., "*i*") was -1.50. The difference between the two thresholds was therefore 0.87, which was considered as too close in light of Eckes' (2015) criterion. This means that Category 2 was less distinctive than expected and should have been combined with the other categories. Because this category was related to the easiness of texts, it could be suggested that the participants could not discriminate the difficulty of lower grade levels, especially because of their high "ability."

To sum up, neither text difficulty nor grade levels perceived by the participants in this study could replicate the original difficulty provided by the GR publisher. Although text length may not be able to explain the gap between perceived and actual GR difficulty, it may contribute to the discussion as to why some texts were not perceived to be as difficult as expected. The results also revealed the imbalance between the participants' ability and text difficulty.

RQ2: How consistent are the texts in the same grade levels in terms of perceived difficulty? This research question was concerned with examining the within-grade consistency in text difficulty. To this end, the results of separation statistics should be examined because separation can indicate the degree to which levels of ability or difficulty are "reliably different" (Linacre, 2020, p. 351). This means that higher values of separation statistics indicate the large number of "statistically distinguishable levels of performance" (p. 352). Separation statistics was also represented as the concrete number of such levels in the form of reliability strata value.

The separation index of text measures showed that the value of reliability was .96 and considered as high and well-separated. The strata value was 6.95, outnumbering the original number of grade levels (i.e., five) provided by the GR publisher. In other words, texts were more distinguishable from one another than the original grade levels. Therefore, it could be pointed out that some texts from the same grade levels had unexpected inconsistencies in terms of perceived difficulty.

For the purpose of examining the within-grade inconsistencies further, Figure 2, based on Run 1, displays text measures and 95% confidence intervals represented by the upper and lower limits of the measures with standard errors of measurement taken into consideration (see also Appendix A). If there is no overlap among the texts in terms of the upper and lower limits of the measures, it could be said that texts were statistically significantly different from the perspective of perceived text difficulty.

If the original grade levels could have been replicated, all the texts from one grade level would have overlapped with one another because they were considered to be the same difficulty. It was true that all of the texts from Level 1 (i.e., Texts 1 to 3) overlapped with one another, meaning that they were not statistically significantly different. A possible reason could be related to the results of category statistics and mean logit values of facets discussed above. In other words, the participants might have considered Level 1 texts to be equally easy and could not be distinguished from one another.

However, as for the Levels 2 (Texts 4 to 6), 4 (Texts 10 to 12), and 5 (Texts 13 to 15), only two of the three texts overlapped with each other. Furthermore, any of the texts from Level 3 (Texts 7

to 9) did not overlap with one another. Therefore, the results suggested that, for all grade levels except Level 1, the text difficulty was statistically significantly inconsistent within the grade level. This meant that, from the perspective of perceived text difficulty, texts from a grade level may not have the same perceived difficulty level. On the contrary, the original grade levels may not be valid in light of the findings that Text 1 (Level 1) was not statistically significantly different from Text 15 (Level 5), meaning that they could not be differentiated in terms of perceived text difficulty. Taken together, these findings suggest that the texts from the same grade levels may not necessarily have had the same perceived text difficulty.

RQ3: How distinguishable are the grade levels from one another in terms of perceived difficulty? Instead of the within-grade difference discussed in the second research question, the third one was related to across-grade differences from the perspective of perceived text difficulty.

First, the separation index of grade levels showed that the value of reliability was .91, indicating relatively good separation. However, the value of strata was 4.67 and lower than the number of grade levels (i.e., five) provided by the GR publisher.

In more detail, Figure 3, based on Run 2, displayed text measures and 95% confidence intervals represented by the upper and lower limits of the measures with standard errors of measurement taken into consideration (see also Appendix B). If the original grade levels could have been replicated, all the grade levels would not have overlapped with one another, or at least a trend of gradual increase in perceived difficulty with the level increase would have been observed. However, the results revealed that only Level 4 was statistically significantly easier than the other grade levels, although the differences were small, and that the other grade levels overlapped with one another, suggesting no statistically significant differences. Surprisingly, this means that the difficulty between Levels 1 and 5 may not differ in terms of perceived text difficulty, concurring with the findings in Research Question 2, showing that the grade levels may not function as originally developed by the publisher. To sum up, it could be pointed out that each grade level was generally not different from one another despite their original five grade levels.

Discussion and conclusion

Addressing the necessity of examining readers' perspectives of text difficulty, the present study examined 162 Japanese EFL university students' perceptions of 15 GR excerpts, by conducting a many-facet Rasch analysis in order to address the three research questions concerning how perceived text difficulty functioned.

As for the first research question concerning the relationship between perceived text difficulty and actual grade levels, it was found that the ranking of grade levels that had originally been assumed was not confirmed, suggesting that perceived text difficulty functioned differently from the difficulty specified by the GR publisher. This being the case, readers may not always perceive the difficulty of GR texts as GR publishers do, and thus not enjoy the benefits of the grade system laddering of the reading material, making it difficult to strike a balance between their (perceived) linguistic proficiency and (perceived) text difficulty, an important factor generating flow (Csikszentmihalyi, 1975) and pleasure reading experiences (Yamashita, 2015).

Interestingly, Claridge (2009) pointed out that GR publishers tend not to take account of language learners' opinions about ER materials. In light of the gap between perceived and GR text difficulty, the results of this study concurred with Claridge's suggestion that when it comes to developing and grading the reading material for ER, GR publishers should include perspectives of readers who have various backgrounds and changing proficiencies. To this end, as Wan-a-rom (2008) pointed out, GR publishers may also have to provide more empirical information on the development of GRs.

As for the second and third research questions concerning the within-grade and across-grade differences in terms of text difficulty, the participants' perceptions did not support the stated difficulty provided by the GR publisher both within and across the grade levels. From the perspective of the use of GRs in the classroom, on the one hand, these findings may suggest that students as well as practitioners cannot rely on the grade system alone when choosing GRs. As for the research on GRs, on the other hand, these findings may appear to be inconsistent with those in Holster et al. (2017) in that there was a difference between perceived and GR text difficulty when text length was controlled. As hypothesized in the literature review section, this may be because book length in Holster et al. included other confounding variables such as readers' fatigue and negative reading attitude, which can make the readers consider lengthy books to be difficult even without reading the first pages. At the same time, however, the finding that Texts 4 and 9 were misfitting partly because they were longer or shorter than other texts from the same grade revealed that text length should be examined further to understand why some texts are perceived as being more difficult than others. Therefore, further studies controlling for text length by, for example, requiring each text to have the same number of words should be encouraged.

More importantly, however, the findings that text length can play an important role in perceived GR difficulty may be inconsistent with the current text grading system whereby GRs are developed with strictly controlled vocabulary and grammar, which is "the single defining feature of a graded reader, distinguishing it from other books" (Nation & Waring, 2020, p. 19). Therefore, the onus might be on GR publishers to explain the relationship between text length and the way grade levels are developed.

In light of the imbalance between readers' ability and text difficulty identified in this study, some may think that these findings were not surprising because it is natural for readers with higher ability to consider most GR texts to be easy. Importantly, however, these results suggest that "good" and "poor" readers perceive text difficulty differently, and that "good" readers might consider most texts as sufficiently easy to encourage "poor" readers to read them extensively. In other words, practitioners as well as researchers may not be able to truly understand how students perceive the difficulty of the reading material. The situation where there might be such a difference between experts (i.e., teachers) and novices (i.e., students) can result in some practical issues when it comes to using GRs in the reading classroom, including teachers' inappropriate support for students' choice of what they want to read. As emphasized in the present study, this can be considered as a reason for the importance of including readers' perspectives in the discussion of text difficulty because the existing grading system cannot always be fixed and appropriate for all learners.

The present study had at least three limitations that should be addressed in future studies. First, the Rasch analysis revealed that there was a discrepancy between the participants' ability and text difficulty, which may have been caused by the lack of more difficult texts. The results suggested that the present study should have employed four, instead of five, categories in the rating scale. Many participants considered the texts to be easy in general, making it difficult to distinguish between "*i* minus 2 or below" and "*i* minus 1" texts, despite the fact that the participants did not take the same proficiency test and may not be grouped homogeneously. Further studies should therefore be encouraged to add more difficult texts and participants with a wider range of language proficiency and individual differences in consideration of the rating scale. Second, unlike Holster et al. (2017), this study could not be considered as an ER program in that the amount of text exposure was quite small, and that the participants did not take an ER course. This makes it difficult to generalize the results of perceived one-page text difficulty to the discussion of ER programs, although one of the aims of the present study was to control the effects of book length, which Holster et al. considered to be an important factor in reader perceptions. In particular, given the fact that the present study employed only three texts from each grade level, future studies are encouraged to examine more texts in order to draw firmer conclusions. Such studies should also take into consideration the text length to be analyzed and other factors in GRs, including the presence of illustrations that would affect perceived text difficulty. Using GRs from one publisher also makes it difficult to generalize the results to other publishers' GRs, leading to the necessity of examining other book series from other publishers in the future. Finally, as with Arai (2022), the present study did not measure the degrees of the participants' reading comprehension. Further studies are expected to examine the relationship between reading comprehension and perceptions.

Despite these, the present study offers the following three implications for further studies. First, this study has emphasized that further discussions should be encouraged concerning how to define and operationalize book difficulty in ER. The situation whereby most ER studies have not discussed their own definitions of book difficulty causes various interpretations of book difficulty, leading to varied ER definitions and inconsistent ER implementation across the classrooms (Arai, 2019). Clarifying the method, and its reasons, of operationalizing book difficulty may therefore contribute to the discussion concerning how to define book difficulty for ER in the long run. Second, perceived text difficulty should be examined and explored in more detail because the concept is compatible with pleasure reading, which is one of the most important reading purposes in ER. Third, it is also important to examine what can affect readers' perceptions of text difficulty. While the present study has focused mainly on text length, there can be other factors that are text-related (e.g., topic familiarity, title, and illustrations), reader-related (e.g., comprehension, motivation, and prior ER experience), and program-related (e.g., time-pressure, classroom environment, and course assignment) in addition to text length on which readers' perceptions are based. The present study would be an avenue to further studies that examine such factors in readers' perceptions.

Furthermore, as for the implications for using GRs in the classroom, the present study has emphasized that it is important for practitioners not to consider GRs difficulty as fixed because students may perceive a higher GR as easy and vice versa. As Lee (2022) suggested, book recommendation can be personalized with consideration of students' preferred book difficulty, and these may be different from one student to another. In the ER classroom, where students are

required to read a large quantity of reading material, including GRs for pleasure, the teachers' guidance in choosing a book to read could be an important factor in successful ER practice (Arai, 2022). Therefore, practitioners could support students' extended reading for pleasure by taking factors other than language proficiency into consideration including individual differences in perceived text difficulty.

In conclusion, the present study examining Japanese EFL university students' perceptions of GR text difficulty found that perceived text difficulty could not replicate the grade levels provided by the GR publisher, thus reinforcing the necessity of exploring the uniqueness of perceived difficulty in further studies and in the classroom.

Acknowledgement

The author thanks his supervisor Professor Yasuyo Sawaki for her mentorship throughout the research. His special thanks also go to anonymous reviewers for their insightful comments.

References

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

Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika, 43*(4), 561–573. https://link.springer.com/article/10.1007/BF02293814

Arai, Y. (2019). Extensive reading definitions, effectiveness, and issues concerning practice in the EFL classroom: Japanese teacher trainees' perceptions. *Journal of Extensive Reading, 7*, 15–32. https://jalt-publications.org/content/index.php/jer/article/view/476

Arai, Y. (2022). Perceived book difficulty and pleasure experiences as flow in extensive reading. *Reading in a Foreign Language, 34*(1), 1–23. https://nflrc.hawaii.edu/rfl/item/542

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive view of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Bahmani, R., & Farvardin, M. T. (2017). Effects of different text difficulty levels on EFL learners' foreign language reading anxiety and reading comprehension. *Reading in a Foreign Language, 29*(2), 185–202. https://nflrc.hawaii.edu/rfl/item/375

Bamford, J., & Day, R. R. (2004). *Extensive reading activities for teaching language.* Cambridge University Press.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Chiang, M.-H. (2016). Effects of varying text difficulty levels on second language (L2) reading attitudes and reading comprehension. *Journal of Research in Reading, 39*(4), 448–468. https://doi.org/10.1111/14679817.12049

Claridge, G. (2009). Teachers' perspectives on what makes a good graded reader. *New Zealand Studies in Applied Linguistics, 15*(1), 13–25. https://www.alanz.org.nz/wp-content/uploads/2018/11/NZSAL-Journal-151-2009.pdf

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* https://rm.coe.int/16802fc1bf

Csikszentmihalyi, M. (1975). *Beyond boredom and anxiety: Experiencing flow in work and play.* Jossey-Bass.

Day, R. R., & Bamford, J. (1998). *Extensive reading in the second language classroom.* Cambridge University Press.

Day, R. R., & Bamford, J. (2002). Top ten principles for teaching extensive reading. *Reading in a Foreign Language*, *14*(2), 136–141. https://nflrc.hawaii.edu/rfl/item/61

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Peter Lang.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3)*,* 221–233. https://doi.org/10.1037/h0057532

Gillis-Furutaka, A. (2015). Graded reader readability: Some overlooked aspects. *Journal of Extensive Reading, 3*, 1–19. https://jalt-publications.org/content/index.php/jer/article/view/7

Herman, E., & Leeser, M. J. (2022). The relationship between lexical coverage and type of reading comprehension in beginning L2 Spanish learners. *The Modern Language Journal, 106*(1), 284–305. https://doi.org/10.1111/modl.12761

Holster, T. A., Lake, J. W., & Pellowe, W. R. (2017). Measuring and predicting graded reader difficulty. *Reading in a Foreign Language, 29*(2), 218–244. https://nflrc.hawaii.edu/rfl/item/377

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430. https://nflrc.hawaii.edu/rfl/item/43

Jeon, E. Y., & Day, R. R. (2016). The effectiveness of ER on reading proficiency: A meta-analysis. *Reading in a Foreign Language*, *28*(2), 246–265. https://nflrc.hawaii.edu/rfl/item/354

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel.* Research Branch.

Knoch, U., & McNamara, T. (2015). Rasch analysis. In L. Plonsky (Ed.) *Advancing quantitative methods in second language research* (pp. 275–304). Routledge.

Krashen, S. D. (1985). *The input hypothesis: Issues and implications.* Longman.

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316–323). Multilingual Matters.

Lee, J. S. Y. (2022). An editable learner model for text recommendation for language learning. *ReCALL, 34*(1), 51–65. https://doi.org/10.1017/S0958344021000197

Linacre, J. M. (1989). *Many-facet Rasch measurement.* MESA Press.

Linacre, J. M. (2012). *Many-facet Rasch measurement: Facets tutorial 4—anchoring*. https://www.winsteps.com/a/ftutorial4.pdf

Linacre, J. M. (2020). *A user's guide to FACETS Rasch-model computer programs: Program manual (ver. 3.83.3)*. Retrieved August 3, 2020, from https://www.winsteps.com/a/Facets-Manual.pdf

McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics, 39*(6), 823–845. https://doi.org/10.1093/applin/amw050

McNamara, T. F. (1996). *Measuring second language performance.* Longman.

Nakanishi, T. (2015). A meta-analysis of extensive reading research. *TESOL Quarterly, 49*(1), 6–37. https://doi.org/10.1002/tesq.157

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*(1), 59–82. https://www.lextutor.ca/cover/papers/nation_2006.pdf

Nation, I. S. P., & Waring, R. (2020). *Teaching extensive reading in another language.* Routledge.

Oxford University Press (2019). *Oxford graded readers: A catalogue (Japanese version)*. Retrieved April 12, 2020, from https://www.oupjapan.co.jp/download/materials/detail/oup_cat_jp_readers.shtml

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* University of Chicago Press. (Original work published in 1960)

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Sick, J. (2008). Rasch measurement in language education Part 2: Measurement scales and invariance. *Shiken: JALT Testing & Evaluation SIG Newsletter, 12*(2). 26–31. https://hosted.jalt.org/test/sic_2.htm

Wan-a-rom, U. (2008). Comparing the vocabulary of different graded-reading schemes. *Reading in a Foreign Language, 20*(1), 43–69. https://nflrc.hawaii.edu/rfl/item/167

Webb, S., & Nation, P. (2017). *How vocabulary is learned.* Oxford University Press.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* MESA Press.

Yamashita, J. (2015). In search of the nature of extensive reading in L2: Cognitive, affective, and pedagogical perspectives. *Reading in a Foreign Language, 27*(1), 168–181. https://nflrc.hawaii.edu/rfl/item/324

Yang, Y.-H., Chun, H.-C., & Tseng, W.-T. (2021). Text difficulty in extensive reading: Reading comprehension and reading motivation. *Reading in a Foreign Language, 33*(1), 78–102. https://nflrc.hawaii.edu/rfl/item/526

Appendices

Appendix A

Measurement Results for the Text Facet

Text[a]	Measure	*SE*	MS_W	t_W	MS_U	t_U	Total Count	Lower Limit[b]	Upper Limit[b]
7	1.02	.14	.97	-.1	.96	-.2	101	.74	1.3
14	1.02	.14	1.23	1.5	1.19	1.3	100	.74	1.3
6	.72	.14	.91	-.6	.93	-.4	106	.44	1.0
15	.55	.13	.85	-1.1	.85	-1.1	113	.29	.81
4	.49	.14	1.46	3.0	1.43	2.7	102	.21	.77
1	.23	.13	.83	-1.3	.81	-1.4	108	-.03	.49
9	.21	.13	1.51	3.4	1.58	3.7	113	-.05	.47
3	-.04	.13	.90	-.7	.90	-.7	102	-.30	.22
2	-.12	.13	.82	-1.4	.80	-1.5	112	-.38	.14
12	-.18	.13	.85	-1.1	.88	-.8	102	-.44	.08
10	-.29	.13	1.03	.2	.99	.0	106	-.55	-.03
5	-.46	.13	.92	-.6	1.03	.2	112	-.72	-.20
8	-.83	.13	.62	-.3	.61	-.3	108	-1.09	-.57
11	-1.05	.13	.83	-1.4	.88	-.7	114	-1.31	-.79
13	-1.27	.14	.94	-.4	.91	-.5	108	-1.55	-.99

Notes. SE = standard error of measurement; MS_W = mean-square infit statistic; t_W = standardized infit statistic; MS_U = mean-square outfit statistic; t_U = standardized outfit statistic

[a]Texts are displayed in order of the measure values.

[b]Lower and upper limits were calculated based on standard errors of measurement to provide the information on 95% confidence intervals.

Appendix B

Measurement Results for the Grade Level Facet

Grade Level[a]	Measure	*SE*	MS_W	t_W	MS_U	t_U	Total Count	Lower Limit[b]	Upper Limit[b]
2	.25	.08	1.1	1.2	1.13	1.5	320	.09	.41
3	.13	.08	1.05	.6	1.07	.8	322	-.03	.29
5	.1	.08	1.0	.0	.98	-.2	321	-.06	.26
1	.02	.08	.85	-2.0	.84	-2.1	322	-.14	.18
4	-.51	.08	.90	-1.2	.92	.9	322	-.67	-.35

Notes. *SE* = standard error of measurement; MS_W = mean-square infit statistic; t_W = standardized infit statistic; MS_U = mean-square outfit statistic; t_U = standardized outfit statistic

[a]Grade levels are displayed in order of the measure values.

[b]Lower and upper limits were calculated based on standard errors of measurement to provide the information on 95% confidence intervals.

About the Author

Yuya Arai is a PhD student at Waseda University, Japan. His research interests include extensive reading and language assessment. His recent work appeared in *Reading in a Foreign Language* and co-authored work in E.H. Jeon and Y. In'nami (Eds), *Understanding L2 proficiency: Theoretical and meta-analytic investigations* (John Benjamins).
Email: yuyason-gfi.w4-7@ruri.waseda.jp