Reading in a Foreign Language
ISSN 1539-0578

November 2025, Volume 37, No. 1
pp. 1–19
https://doi.org/10.64152/10125/67521

Developing a checklist to support learners' choice of graded readers in extensive reading

Yuya Arai
Chiba University
Japan

Abstract

Many extensive reading (ER) programs encourage second/foreign language learners to choose reading materials by themselves. However, given the difficulty of choosing appropriate reading materials, it is necessary to provide support. Hypothesizing the importance of learners' subjectivity in their choice of reading materials, this study aimed to develop a checklist that measured learners' perceptions of text difficulty in order to support their text choice. First, a multivariate generalizability theory analysis was conducted to determine the number of graded reader texts included in the checklist. Then, 18 Japanese female university students learning English as a foreign language used the checklist during the ER program. The change in their perceptions of text difficulty in the ER program was summarized descriptively, while an open-ended questionnaire was analyzed by conducting reflexive thematic analysis. The results suggested the effectiveness of the checklist in helping students choose appropriate texts, encouraging future research and practice to support learners' choice of ER materials.

***Keywords*:** extensive reading, graded readers, learners' choice of reading materials, perceived text difficulty, generalizability theory

Introduction

An extensive reading (ER) program in the second and foreign language (L2) classroom encourages language learners to read a large number of easy reading materials (Day & Bamford, 1998). Day and Bamford (2002) proposed ten principles for teaching ER, some of which concern reading materials employed in ER programs. Among others, their third principle states that "[l]earners choose what they want to read" (p. 137), meaning that ER programs should allow learners to read self-selected materials. This principle has been widely applied in many ER programs (Arai & Takizawa, 2025; Day, 2015; Puripunyavanich & Waring, 2024) and regarded as an ideal ER practice (Robb & Ewert, 2024). Zhou (2024) suggested that the importance of this principle has been supported by research findings that learners tend to enjoy their self-selection of reading materials in ER. Nation and Waring (2020) also pointed out that allowing learners to choose books by themselves can be "by far the best method" (p. 12) to encourage them to read at the right level.

Therefore, learners in the ER classroom are expected to choose appropriate reading materials by themselves, especially in terms of text difficulty, because ER emphasizes the importance of easy reading materials, as defined above. However, it is not always easy for them to choose reading materials appropriately (Lyon & Barr, 2019; Ramonda, 2020). It is thus fair to say that their choice requires some external support from others, such as teachers and peer learners, but ER research has still suffered from a lack of empirically well-established methods to appropriately support learners' choice of reading materials.

In order to address this issue, the present study aimed to develop and implement a checklist for learners' choice of graded readers (GRs) in an ER program. GRs are books written for L2 learners with strictly controlled vocabulary and grammar. Developing and implementing the checklist would provide some insights for further research and practice of ER in terms of the importance of support for learners' choice and its potential effects on language learning.

Literature Review

Importance and Difficulty of Learners' Choice of Reading Materials in ER

ER researchers have suggested that ER programs should encourage learners to choose freely what they want to read especially from the perspective of motivational and affective aspects of language learning. For example, Mori (2015) pointed out that choice is critical for learners in an ER program to become intrinsically motivated. Stoller (2015) also suggested, "Giving students choices in what they read can empower students and lead to more student engagement in reading" (p. 152). Empirically, Arai (2022b) examined flow experiences in an ER program. Flow is a pleasurable experience allowing learners to be intrinsically motivated and totally absorbed in the books they are reading (see Csikszentmihalyi, 1975/2000). His study suggested that the freedom of reading material choice in ER would help the learners find their most suitable book difficulty, leading to flow experiences.

Despite the importance of learners' choice of reading materials in ER (Day & Bamford, 1998, 2002; de Burgh-Hirabe & Feryok, 2013; Nation & Waring, 2020; Stoller, 2015), its difficulty has also been pointed out. For example, Lyon and Barr (2019) found in their ER program that it was not always easy for learners to choose appropriate reading materials by themselves, emphasizing the need for some guidance in their self-selection. Learners tend to choose difficult, unenjoyable reading materials because they tend to consider easy texts childish (Takase, 2009) or boring (Lien, 2010). If learners fail to choose appropriate reading materials, they can be discouraged from reading for pleasure (Lee & Ro, 2020; Leung, 2002). In such a case, Day and Bamford's (2002) principle of learners' self-selected reading materials is incompatible with another principle (e.g., reading for pleasure).

Importance and Issues of Supported Choice

Given the difficulty learners have in choosing reading materials in ER, teachers may need to provide appropriate support for learners' self-selection (Katz & Assor, 2007). As Day and

Bamford (2002) suggested, ER programs should provide a wide range of reading materials from which learners choose. Therefore, a lack of external support can make it difficult for learners to choose appropriate texts from a large number of books (Iyengar & Lepper, 2000; Patall et al., 2008). Furthermore, learners from some specific cultures (e.g., East Asia) are likely to be motivated by more supported choice (Iyengar & Lepper, 1999), as Mori (2015) suggested the potential of choice that is "made for them [=learners] by trusted authority figures or peers" (p. 133), including teachers. As a case in point, Sun (2020) found that both teacher- and learner-selected reading materials were effective for Chinese EFL learners who trusted their teachers. Chang and Renandya (2017) also suggested that some teacher guidance is necessary, especially for beginner-level language learners in Asian contexts.

At the same time, however, it is worth discussing the extent to which teachers should allow learners' choice of reading materials. In some cases, teachers do not allow learners to choose texts by themselves. For example, Tan (2018) suggested that teachers are generally unwilling to let learners have total control of book choice "for fear that they do not make good choices" (p. 74). Robb and Ewert (2024) also pointed out that teachers tend to face difficulty in allowing the freedom of learners' choice of reading materials and trusting their self-selected reading. Instead of freedom of reading material choice, some ER practitioners use teacher-selected class readers, which might be preferable to learner-selected reading materials for some practical reasons. For example, with teacher-selected class readers, in-class activities can easily be designed because learners read the same books (see also Davis, 1995; Ramonda, 2020). Ramonda found in his ER program that the learners generally responded positively to the use of class readers. In addition, class readers could play an important role in contexts where a limited number of reading materials is available (Macalister, 2015; Robb, 2022). Meanwhile, it should also be noted that teacher-selected texts and/or the use of class readers are not always appropriate materials for individual learners. Donald (2018) pointed out the teacher's dilemma concerning the degree of their intervention: teachers' "pre-selection of texts may ensure learners read at a suitable level but at the expense of choosing readings that are of personal interest" (p. 245). Claridge (2011) also suggested that teachers are likely to prefer difficult reading materials despite the importance of easy reading materials in ER. Given the difficulty of deciding to what extent learners can be given control over reading material choice, it is no wonder that some researchers (e.g., Waring & McLean, 2015) regard the freedom of reading material choice as an optional element of ER programs.

Examples of Support for Reading Material Choice and Their Limitations

The ER literature has suggested some practical examples of support for learners' choice of reading materials, including the need to conduct a survey to understand learners' preferences and book recommendations among learners (Bamford & Day, 2004). Another kind of support for learners' choice is the use of a placement test to assess linguistic proficiency at the beginning of an ER program. As a case in point, the Edinburgh Project on Extensive Reading Placement/Progress Tests (EPER PPTs) is widely recognized by ER practitioners and researchers as a test assessing general English proficiency before an ER program as a placement test and as a progress test measuring the effects of ER programs (Nakanishi, 2015). However, because the project has been terminated, the whole set of the EPER PPTs is currently not available (Sakurai, 2020). Furthermore, the EPER PPTs are cloze tests assessing general English proficiency. It is

thus debatable whether and to what extent the tests could make a unique contribution to understanding reading proficiency that is necessary for ER and inform learners' choice.

Some ER researchers have also proposed other ways to place learners into appropriate levels of reading materials, mainly in terms of vocabulary (Mermelstein, 2015; Wan-a-rom, 2010). Among others, Mermelstein proposed a method for making a placement decision in ER programs named the reading level measurement method (RLMM). In this method, learners read GR texts and underline words whose meanings they do not know. The RLMM encourages them to choose grade levels whose texts contain less than two percent unknown words (i.e., 98 percent lexical coverage). However, one important issue with the RLMM is that underlining unknown words during reading does not simulate reading in ER programs and unnecessarily makes learners pay more attention to linguistic forms than to meaning, the latter being an indispensable purpose of ER (Day & Bamford, 2002).

Utilizing Learner Perceptions of Text Difficulty for Text Choice

While those kinds of support may be effective in guiding learners' choice of reading materials mainly in terms of linguistic proficiency and text difficulty, it should be noted that text appropriateness depends on individual learners and is usually beyond such linguistic factors. Donald (2018) found that learners chose reading materials based on criteria other than linguistic difficulty, including book covers. Empirically, Holster et al. (2017) investigated the relationship between learner-perceived difficulty of GRs and some lexical indices, such as headwords. They found that such lexical indices provided by GR publishers were not a good predictor of learners' perceptions of text difficulty and concluded that they "do not provide useful guidance in [the] selection of books to read" (p. 218).

It is worth noting that reading material choice is a subjective act on the part of learners, highlighting the need to consider learners' subjective perceptions and experiences when ER practitioners support their self-selected ER. Interestingly, even if a text is difficult in terms of its linguistic profile, it can contribute to pleasure reading when a learner perceives it as pleasurable. For example, Ro's (2018) focus group study found that a student called Sam tended to choose difficult books in an ER program but still experienced pleasure because he was happy to become a "good student" (p. 784) who attempted to read difficult books. This example suggests that the meaning of pleasure reading could differ among individual learners, and that, for Sam, choosing difficult reading materials was important for his own reading pleasure. Similarly, de Burgh-Hirabe and Feryok (2013) examined nine students' interviews and written reflection data and found that while a student named Nick preferred easy reading materials, another student named Alan was willing to read challenging books in their ER program. Because ER is a pedagogical activity that emphasizes the importance of pleasure reading (Yamashita, 2015), it is necessary to discuss how learners' choice of reading materials can be supported from the perspective of learners' perceptions.

To this end, this study hypothesized that learners' perceptions of text difficulty could play an important role in their choice of reading materials. Some previous studies (Arai, 2022a, 2022b, 2023; Holster et al., 2017) have examined perceived text difficulty, which "is based on readers' perceptions of text difficulty" (Arai, 2022a, p. 252), and found that perceived text difficulty is

not always the same as objective text difficulty, such as grade levels. Perceived text difficulty is also conceptually compatible with the experience of pleasure and the act of reading material choice in that all of them concern learners' subjectivity (Arai, 2022b). Perceived text difficulty could thus provide unique insights into learners' text choice compared to linguistic information on reading materials, and support based on perceived text difficulty could help individual learners choose books that are appropriate for them. It is true that some researchers have attempted to include learner-perceived text difficulty in the assessment of appropriate text levels, such as the Extensive Reading Foundation's Online Self-Placement Test (https://erfpt.ealps.shinshu-u.ac.jp/; see Brierley et al., 2012). However, little empirical research has been conducted on the effectiveness of a checklist that measures only learner-perceived text difficulty in helping learners choose appropriate texts.

The Current Study

To summarize, while ER emphasizes learners' choice of reading materials, learners are not always good at choosing appropriate reading materials by themselves, necessitating effective support for their choice. Given the lack of a well-established method of such support, this study aimed to develop a checklist that would allow learners to understand reading material difficulty before an ER program so that they can choose reading materials effectively during the ER program.

An important issue when developing such a checklist that diagnoses appropriate text levels is reliability. In order to control potential sources of measurement error that negatively affect the reliability of the checklist, the number of texts included in the checklist should be considered. In general, the more texts there are in the checklist, the more reliable it will become (Liao, 2023). However, it is necessary to strike a balance between reliability and practicality (Bachman & Palmer, 1996). In order to determine how many texts should be used to reliably measure text difficulty, this study used generalizability theory (G theory; see Grabowski & Lin, 2019; Sawaki & Xi, 2019; Shavelson & Webb, 1991) and estimated the reliability of the checklist. G theory will be discussed in more detail later in the Multivariate Generalizability Theory section.

After developing the checklist, it is also necessary to investigate whether the checklist really contributed to learners' appropriate text choice. Therefore, this study addressed the following two research questions (RQs) concerning the development of the checklist (RQ1) and its implementation in an ER program (RQ2):

1. How many texts should be used in the checklist to reliably measure text difficulty perceived by the learners participating in an ER program?

2. To what extent is the checklist effective in helping learners choose appropriate reading texts?

In order to address these RQs, the present study employed a two-phase sequential research design: checklist development (Phase 1) and its implementation in an ER program (Phase 2).

Phase 1: Checklist Development

Methods

The Dataset. This study used data collected in previous studies (Arai, 2022a, 2023). The participants in the dataset were 255 Japanese students in total (157 university students [aged 18 to 22] and 98 high school students [aged 16 to 17]) who spoke Japanese as their first language (L1) and were learning English as a foreign language (EFL). Their English proficiency corresponded to the A2 to B2 levels on the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001), measured by large-scale English proficiency tests or based on their teachers' judgments.

The dataset concerned the students' rating of the difficulty of 15 GRs published by Oxford University Press in the Oxford Bookworms Library (OBL) series. They were randomly sampled from Starter (the easiest grade level in the series, corresponding to the Pre-A1 and A1 CEFR levels) to Level 4 (the third most difficult grade level in the series, corresponding to the B1 and B2 CEFR levels). Three GRs were randomly sampled from each of the grade levels (i.e., three GRs*five grade levels = 15 texts). The texts used in the dataset were extracted from the first pages of each GR (for more detailed information on the texts, see Arai, 2023).

The students in the dataset read ten GR texts in total (i.e., two out of three texts randomly selected from each grade level). They rated the difficulty of each text 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", where "i" represented their own linguistic level based on Krashen's (1985) comprehensible input hypothesis (for more detailed information on the scale and how it was explained to the participants, see Arai, 2022a, 2023). Before the rating, the students were informed of the rating scheme and provided with an opportunity to practice rating text difficulty.

The dataset was used for the checklist development in this study because it included the data on Levels 3 (B1 CEFR level) and 4. Among the five grade levels, the information on these two grade levels was utilized for the checklist because the checklist was intended for an ER program in this study (see Phase 2), which employed GR texts only from Levels 2 (A2 to B1 CEFR levels) to 5 (B2 CEFR level). This made it unnecessary to include grade levels other than Levels 3 and 4. This is because if the participants felt Level 3 texts were too difficult, they were encouraged to choose Level 2 texts; if Level 4 texts were too easy for them, Level 5 texts could be recommended. The inclusion of GR texts only from Levels 3 and 4 also made the checklist concise and practical for placement purposes in Phase 2.

As discussed so far, the dataset used for the checklist development was obtained from other published studies (Arai, 2022a, 2023). However, the present study differed from the precursor studies because the research purpose of this study was to develop the checklist by conducting a G theory analysis, whereas the main focus of the precursor studies was to understand how learners perceived text difficulty by conducting a Rasch analysis. In addition to the difference in research purposes and analytic methods, the study contexts also differed. While the present study was conducted in the ER context to examine the effectiveness of the checklist, the precursor studies were not in an ER program but simply examined perceived text difficulty.

Multivariate Generalizability Theory. In developing the checklist, it is necessary to investigate how many GR texts there should be both per grade level and in the entire checklist in order to make a reliable suggestion for reading material choice (RQ1). To estimate the reliability, this study conducted a multivariate G theory analysis (Brennan, 2001a; Cronbach et al., 1972). The software mGENOVA (Version 2.1; Brennan, 2001b) was used to conduct the analysis.

G theory was employed in this study because traditional reliability indices, including Cronbach's alpha, do not allow the calculation of reliability estimates that adequately reflect the complex measurement design of the present checklist. The reliability of this checklist is affected by multiple sources of rating variation (called *facets of measurement*, or *facets* for short), such as the grade levels and the difficulty of individual texts. However, traditional reliability indices can take account of only one facet at a time (e.g., text difficulty only). In contrast, G theory's strength lies in specifying multiple facets that could affect rating variation, including, in the present study's case, GR texts and their grade levels.

Moreover, this study used *multivariate* G theory because it has an advantage over *univariate* G theory in "model[ing] multiple measures *simultaneously*" (Grabowski & Lin, 2019, p. 56, emphasis original). In this study, Level 3 and Level 4 texts should be modeled together but distinguished from each other in order to properly estimate the reliability of both each grade level and the entire checklist. Furthermore, because the target grade levels (i.e., Levels 3 and 4) were selected purposefully, grade levels should be treated as a *fixed* facet, where a target grade level is conceptualized as not exchangeable with another grade level. In contrast, GR texts could be treated as a *random* facet because, for the purpose of the checklist, they are randomly selected among texts at the same grade levels. Multivariate G theory can specify one random facet (GR texts) and one fixed facet (grade levels) simultaneously. In G theory terms, this study employed a crossed design, notated as a $p^{\bullet}$ x $t^{\bullet}$ design, where participants (p) were crossed with (x) GR texts (t), and the superscript filled circles $\bullet$ mean that both participants and GR texts were crossed with grade levels (a fixed facet). In other words, the design means that (1) all participants (called *objects of measurement* in the G theory terminology) read two of the three texts at each grade level (p x t); (2) participants read texts from both grade levels ($p^{\bullet}$); and that (3) texts were sampled from both grade levels ($t^{\bullet}$).

Another merit of using G theory is that, while traditional reliability indices (e.g., Cronbach's alpha) only concern the reliability for relative scale score interpretations and decisions (i.e., how learners rank-ordered texts in terms of difficulty), G theory can calculate the *index of dependability* (phi, or Φ), which can be used for absolute scale score interpretations and decisions (i.e., to what extent learners perceived texts as difficult, in addition to how learners rank-ordered texts in terms of difficulty). Suppose that four learners rated two texts' difficulty on a six-point Likert scale (see Table 1). In Table 1, the rank-ordering of the texts is exactly the same among the learners (i.e., Text 1 < Text 2). However, the rank-ordering information alone does not tell us to what extent each text was perceived as difficult (i.e., mean difficulty ratings for each text). For the purpose of the checklist in this study, both the relative rank-ordering of texts in terms of difficulty and the absolute difficulty of individual texts are quite important. Therefore, in this study, it was necessary to calculate a reliability coefficient taking account of both for the purpose of absolute scale score interpretations and decisions (*dependability* in G theory terminology), which is possible by conducting a G theory analysis.

Table 1. *Example of text difficulty rating results*

Student No.	Text 1	Text 2
1	1	3
2	2	4
3	3	5
4	4	6
Mean	2.5	4.5

Note. Although this example is on a six-point Likert scale, the present study used a five-point one (see the Dataset section above).

G theory had two analytical steps. First, a generalizability study (G study) provided baseline information on the reliability of the checklist. Then, based on the G study results, decision studies (D studies) were conducted to simulate the change in reliability levels with different numbers of GR texts in the checklist. The results of D studies would thus help determine how many GR texts there should be per grade level as well as in the entire checklist.

Results and Discussion. The results of descriptive statistics showed that the means and standard deviations for Levels 3 and 4 were 2.8 and 1.08, and 3.0 and 1.06, respectively (see Appendix A for the results of the G study). Based on the results, D studies were conducted to model dependability while changing the number of GR texts included in the checklist. Table 2 presents the results of the phi coefficients for the D studies simulating different numbers of GR texts. For example, as the second row of Table 2 shows, when the checklist included two texts each for Levels 3 and 4, the phi (Φ) coefficients for Levels 3 and 4 were 0.57 and 0.40 respectively, and the composite phi coefficient (i.e., the dependability of the entire checklist combining two Level 3 texts and two Level 4 texts) was 0.62. Overall, the results showed that Level 4 texts had relatively low dependability even when the number of texts increased.

Table 2. *Estimated phi (Φ) coefficients for the D studies*

# of Level 3 texts[a]	Φ	Error	# of Level 4 texts[a]	Φ	Error	Composite Φ[a]
2	0.57	0.35	2	0.40	0.43	0.62
3	0.66	0.24	3	0.50	0.29	0.71
4	0.73	0.18	4	0.57	0.21	0.77
5	0.77	0.14	5	0.63	0.17	0.80
6	0.80	0.12	6	0.67	0.14	0.83
3	0.66	0.24	4	0.57	0.21	0.76
3	0.66	0.24	5	0.63	0.17	0.77
3	0.66	0.24	6	0.67	0.14	0.79

Notes. Error = absolute error variance.
[a]Texts at each grade level were treated as a random facet and exchangeable within the grade levels.
[b]The two grade levels were equally weighted to compute composite phi coefficients.

Shavelson and Webb (1991) suggested that phi coefficients of about .80 are "reasonable levels of estimated generalizability" (p. 101). However, this rule of thumb may not necessarily be strictly applied to the development of the checklist in this study in light of the low-stakes, classroom-

based use of the checklist in the ER program. In addition, practicality should be considered in implementing the checklist, meaning that a checklist with too many GR texts would take too much time for learners' self-evaluation of their choice of reading materials and might unintentionally affect their perceptions of text difficulty. Based on the results, this study decided to include a total of nine texts (three from Level 3 and six from Level 4) because this is the number of texts that allowed almost the same levels of dependability between the two grade levels ($\Phi = 0.66$ for Level 3 texts, $\Phi = 0.67$ for Level 4 texts). Furthermore, the composite phi coefficient in this case would be 0.79. The relatively low dependability for Level 4 texts will be further discussed in the General Discussion section.

Phase 2: Checklist Implementation

Methods

Study Context and Participants. An ER program was conducted in a semester-long EFL academic writing course at a private women's university in Japan. The course, which met for 105-minute in-person sessions once a week for 13 weeks, aimed at academic paragraph and essay writing in English, within which the ER program was incorporated because reading has a close relationship with writing. Second-year students took this course to fulfill graduation requirements. Their L1 (Japanese) was used as the medium of instruction.

The course introduction and orientation were conducted in the first week, including an explanation of the importance and procedure of ER. From Weeks 2 to 13, the ER program was implemented for 15 minutes at the beginning of each session. In each ER activity, the students were asked to choose an OBL GR text from Levels 2 to 5 and read it individually for pleasure. They did not have to finish reading within 15 minutes but were instead instructed to do so outside the classroom before the next class. Appendix B shows further information on the characteristics of the ER program and their correspondence with Day and Bamford's (2002) ten principles for teaching ER.

The participants were in total 36 female students (aged 19 and 20) from two intact classes. Among them, 18 students were provided with the checklist before their participation in the ER program from Week 2 (checklist group), while the other 18 students were not (non-checklist group). The participants' scores on the Test of English for International Communication (TOEIC®; Educational Testing Service, https://www.ets.org/toeic) corresponded to the B1 level of the CEFR.

Data Collection and Analyses. Before the study participation, the participants were informed of ethical considerations, including the voluntary nature of study participation, confidentiality, the fact that there was no relationship between the study participation and course grades, and the right to withdraw from the study at any time. They were also told that this study was approved by the research ethics committee at the author's university at the time of this study. Before the ER program, only the checklist group was provided with the checklist developed in Phase 1 and was asked to rate the text difficulty outside the classroom as homework. The checklist comprised nine OBL GR texts randomly selected from each grade level (i.e., three from Level 3 and six

from Level 4). Table 3 summarizes the characteristics of the GR texts included in the checklist. Three versions of the checklist were prepared and randomly distributed to the participants so that the order of presentation of the texts was counterbalanced. The procedure of rating text difficulty was the same as in Phase 1.

Table 3. *Characteristics of graded reader texts included in the checklist in Phase 2*[a]

Text_GL	Title	CEFR level	Book length[b]	Flesch Reading Ease	Flesch-Kincaid Grade Level	Text length[b]	Genre
1_GL3	*The Call of the Wild*	B1	10,965	93.9	2.9	101	C
2_GL3	*'Who, Sir? Me, Sir?'*	B1	10,295	64.5	8.0	137	H
3_GL3	*The Picture of Dorian Gray*	B1	10,245	88.8	2.8	237	FH
4_GL4	*The African Queen*	B1-B2	15,250	70.0	8.0	254	H
5_GL4	*The Eagle of the Ninth*	B1-B2	14,950	61.7	9.1	90	TA
6_GL4	*We Didn't Mean to Go to Sea*	B1-B2	14,860	87.2	3.2	123	FH
7_GL4	*The Scarlet Letter*	B1-B2	15,965	79.2	6.5	226	C
8_GL4	*The Big Sleep*	B1-B2	15,960	87.8	3.2	267	CM
9_GL4	*Eat, Pray, Love*	B1-B2	18,805	74.3	6.4	287	T

Notes. GL = grade level, C = classics, H = human interest, FH = fantasy and horror, TA = thriller and adventure, CM = crime and mystery, T = true stories.
[a]Titles, grade levels, CEFR levels, book lengths, and genres were reported in Oxford University Press (2019).
[b]Text length stands for the word counts of the excerpted first pages.

In order to investigate the effectiveness of the checklist, the participants in both the checklist and non-checklist groups were asked to rate the difficulty of GR texts that they chose every week after the 15-minute reading period with the same rating procedure and scale in Phase 1. In addition to the scale of perceived text difficulty, the participants completed two open-ended questions concerning (1) reasons for their text choice and (2) the effects of their text choice on reading experiences. The participants completed the scale and the questions every time after the 15-minute reading. Due to space limitations, this study only reported the results in the first four weeks of the ER program (i.e., one-third of the 12-week ER program). The quantitative results of the scale were summarized descriptively, while the qualitative data from the two open-ended questions were analyzed using reflexive thematic analysis (Braun & Clarke, 2006, 2022). Reflexive thematic analysis was used because its strength lies in the qualitative, cross-case exploration of both semantic and latent meanings of the data, which could inform the potential patterns behind the effects of the checklist on the participants' text choice. After familiarizing himself with the data, the present author coded them inductively and manually by utilizing his

expertise in ER research. Based on the codes, the present author developed themes by clustering codes and exploring potential meaning patterns. Because, in reflexive thematic analysis, coding by a single coder is common (Braun & Clarke, 2022), this study did not aim at inter-coder consensus during the qualitative analysis. The following two themes were developed, which will be discussed in the next section in more detail:

1. Worries about linguistic difficulty
2. Falling behind in finding their comfort zone

Results and Discussion. Table 4 shows the frequency statistics for the perceived difficulty of GRs that the participants chose in the first four weeks of the ER program. Despite the small sample size, there were two noticeable results. First, in Week 2 (i.e., the first week of the ER program), the majority of the students in the non-checklist group perceived their GR texts to exceed their proficiency (n = 9; 52.9%); this can be compared to those in the checklist group (n = 2; 11.8%). This result could be explained by their responses to the open-ended questions analyzed by reflexive thematic analysis. One of the two themes developed in this study was "Worries about linguistic difficulty". This theme suggested that the students in the non-checklist group tended to have difficulty with the linguistic characteristics of their texts. For example, a student commented, "I couldn't comprehend this text very well because there were some important words whose meanings I didn't know." Another student in the non-checklist group, who rated her text as "i plus 1", said, "The vocabulary [in the text] was difficult for me, so I became sleepy. I think this [grade] level was too difficult. I'm not good at reading in English, so I don't know which level to choose", suggesting the need for some kind of support for her text choice and reading. Meanwhile, the students in the checklist group were less likely to refer to their texts' linguistic difficulty but instead mentioned the stories they read. One student put it, "The story was interesting and easy to comprehend because I had read the text in Japanese." Overall, the results suggested that while the students in the non-checklist group generally had difficulty finding texts that were linguistically appropriate for them, those in the checklist group were able to concentrate on the stories that they chose.

A second notable result was that more students in the checklist group perceived their GR texts as matching their proficiency in Weeks 2 (n = 11; 64.7%) and 3 (n = 12; 80.0%) than in the non-checklist group (Week 2: n = 6; 35.3%, Week 3: n = 10; 58.8%). Meanwhile, it is interesting that more students in the non-checklist group were able to choose "i" texts in Weeks 4 (n = 6; 42.9%) and 5 (n = 9; 64.3%) than in the checklist group (Week 4: n = 4; 30.8%, Week 5: n = 6; 46.2%). A possible reason why the number of students in the checklist group who were able to choose "i" texts was smaller than that of those in the non-checklist group in Weeks 4 and 5 is that they started to try another grade level while understanding their comfort zone (i.e., the grade levels corresponding to their proficiency), whereas the students in the non-checklist group fell behind those in the checklist group and gradually found their "comfort zones" in Weeks 4 and 5. As evidence supporting this view, Theme 2 suggested that the students in the non-checklist group fell behind those in the checklist group in the search for their comfort zone. The qualitative results show that some students in the checklist group intentionally did not choose "i" texts, as a student suggested in Week 4, "Because I know I can read Level 3 texts, I thought I could read this [i.e., a Level 4 text]. It was more difficult [than Level 3 texts], but I was able to understand the story." Another student in the checklist group decided to choose a GR text whose difficulty was lower than her proficiency in Week 5 because "the topic seemed interesting to me. I was

thinking that I was able to read this level, and I actually was." Other students in the checklist group started to feel that the grade levels that they had continued to choose became easy. For example, one student pointed out that while two Level 3 texts in Weeks 2 and 3 had matched her proficiency, she felt another Level 3 text was easier to comprehend in Week 4 than the previous two texts, suggesting the possibility of her readiness to move on to the next grade level (Level 4). In contrast, the students in the non-checklist group came to know their comfort zone around Weeks 4 and 5. A student said, "I tried a Level 4 text at first, but it was difficult for me. I chose this level [Level 3] for the first time today, and it turned out that this is the level I should read." Another student suggested in Week 5, "I should have read this level earlier. I was able to comprehend this level, maybe because there were few difficult words."

Table 4. *Number of answers to each category in the scale of perceived text difficulty*

Week[a]	Group	*n*	(1)	(2)	(3)	(4)	(5)
2	C	17	0 (0%)	3 (17.6%)	11 (64.7%)	2 (11.8%)	1 (5.9%)
	NC	17	1 (5.9%)	1 (5.9%)	6 (35.3%)	9 (52.9%)	0 (0%)
3	C	15	0 (0%)	0 (0%)	12 (80.0%)	2 (13.3%)	1 (6.7%)
	NC	17	1 (5.9%)	2 (11.8%)	10 (58.8%)	3 (17.6%)	1 (5.9%)
4	C	13	0 (0%)	3 (23.1%)	4 (30.8%)	5 (38.5%)	1 (7.7%)
	NC	14	0 (0%)	3 (21.4%)	6 (42.9%)	5 (35.7%)	0 (0%)
5	C	13	0 (0%)	2 (15.4%)	6 (46.2%)	5 (38.5%)	0 (0%)
	NC	14	0 (0%)	0 (0%)	9 (64.3%)	5 (35.7%)	0 (0%)

Notes. (1) = "i minus 2 or below," (2) = "i minus 1," (3) = "i," (4) = "i plus 1," (5) = "i plus 2 or above," C = checklist group, NC = non-checklist group.
[a]Week 1 was for the course orientation without ER activities.

Taken together, it could be said that the checklist was useful for the students to understand their comfort zone from the beginning of the ER program. Meanwhile, the students in the non-checklist group tended to fall behind and take some time to find their comfort zone.

General Discussion and Conclusion

This study aimed to develop a checklist that can be useful for learners to understand GR text difficulty in an ER program. RQ1 concerned the development of a reliable checklist. The results in Phase 1 suggested the inclusion of three texts from Level 3 ($\Phi = 0.66$) and six texts from Level 4 ($\Phi = 0.67$) in the checklist (Composite $\Phi = 0.79$). It was also found that the dependability of Level 4 texts did not improve drastically even when the number of texts increased, for which there were two possible reasons. First, the participants in the original dataset

may have provided relatively unreliable perceptions of Level 4 text difficulty for some reason, including their reading proficiency not being sufficient for them to perceive the text difficulty accurately. Another reason could lie in the internal consistency of the text difficulty within Level 4. As previous studies (Arai, 2022a, 2023) have suggested, the perceived difficulty of texts within the same grade levels can differ. Therefore, in order to make the checklist more effective for learners' choices of appropriate reading materials, the grading system should be more strictly designed and applied when GR publishers develop reading materials. As Claridge (2011) pointed out, GR publishers tend not to consider reader perspectives when developing reading materials, which aligns with the present study's suggestion of the importance of learners' perceptions of text difficulty. Wan-a-rom (2008) also suggested that GR publishers provide little information on the development of GRs and their grading systems. The procedure of GR development should thus be more transparent so that researchers can examine the validity and reliability of the grading system and the quality of GRs.

In Phase 2, the checklist was implemented before an ER program to examine the checklist's effects on learners' choice of appropriate GR texts. This study operationalized "i" as the appropriate text difficulty in light of previous studies providing inconsistent results as to whether learners should read "i plus 1" or "i minus 1" texts in ER (Arai, 2022b; Bahmani & Farvardin, 2017; Chiang, 2016; Yang et al., 2021). The results suggested that the checklist helped the students in the checklist group choose appropriate texts from the beginning, while those in the non-checklist group took some time to find appropriate texts. The findings were consistent with previous studies suggesting the need for some support for learners' choice (Katz & Assor, 2007) because learners tend to have difficulty choosing appropriate texts in ER (Lyon & Barr, 2019; Ramonda, 2020). The results of the open-ended questions also suggested that learner perceptions of text difficulty were affected not only by linguistic features but also by the stories that they read, concurring with previous studies emphasizing various factors affecting perceived text difficulty (Arai, 2022b) and learners' choice of reading materials (Donald, 2018). The finding that perceived text difficulty was not only affected by linguistic features suggested that some placement techniques, such as the RLMM (Mermelstein, 2015), have a limitation in terms of their exclusive attention to linguistic features. Support for learners' choice of reading materials in terms of perceived text difficulty would thus simulate the actual reading experience in ER programs more precisely than such methods.

This study has some limitations in its contextual specificity. In Phase 1, the dataset comprised Japanese EFL students (Arai, 2022a, 2023), meaning that caution is required in the generalization of the results to other study contexts. Future attempts to develop a similar checklist are thus encouraged in various ER settings. The low phi (Φ) coefficients also point to the need for cautious interpretations of the results. In addition, this study only used OBW GR texts, necessitating further research using a similar checklist containing other publishers' GRs. In Phase 2, this study employed intact classes to examine the effectiveness of the checklist, suggesting that the results should be interpreted considering the contextually specific nature. The lack of experimental control of the research design requires caution when concluding that it was the checklist that made the difference between the checklist and non-checklist groups in terms of participants' text choice practices and perceived text difficulty.

Despite these limitations, this study has aimed to demonstrate the potential of the checklist and perceived text difficulty in the ER classroom. Furthermore, this study could be an avenue to future research and practice in the following three ways. First, future research should be conducted to examine perceived text difficulty and its importance in ER in light of its conceptual compatibility with learner psychology in text choice and the subjective experience of pleasure reading. Given the limited number of studies on perceived text difficulty (Arai, 2022a, 2022b, 2023; Holster et al., 2017), further studies are needed to understand how perceived text difficulty can be useful for learners' choice of appropriate reading materials and support by ER practitioners. Second, this study could lead to further discussions on how learners' choice of reading materials can be supported. In light of the importance (Day & Bamford, 1998, 2002; de Burgh-Hirabe & Feryok, 2013; Nation & Waring, 2020; Stoller, 2015) and difficulty of text choice (Lyon & Barr, 2019), some kind of learner support is necessary (Chang & Renandya, 2017). The few attempts to provide such support in the ER literature, including the EPER PPTs and RLMM, focus exclusively on linguistic aspects of GRs, and little has been discussed on how to support text choice, especially in terms of learner psychology. Finally, further research is needed to investigate the quality of GRs not only in terms of their linguistic features but also from readers' perspectives (Claridge, 2011). The present study has also emphasized the importance of further disclosure of information on the grading systems from GR publishers in order to encourage more studies on them in the future (Wan-a-rom, 2008). For practice, it is important that ER practitioners do not consider the grade level as the absolute and sole criterion for learners' choice of reading materials but instead provide support for individual learners by considering their perceptions.

In conclusion, this study aimed to develop and implement a checklist that can support learners' choice of reading materials in an ER program in terms of perceived text difficulty. The effectiveness of the checklist was supported in light of the descriptive change in learners' perceptions of reading material difficulty in the ER program and the results of reflexive thematic analysis of an open-ended questionnaire. Future research and practice are encouraged to understand how learners' choice of reading materials can be supported and how perceived text difficulty plays a role in ER programs.

Acknowledgment

I sincerely thank Professor Yasuyo Sawaki for her comments on earlier versions of this paper.

References

Arai, Y. (2022a). Exploring perceived difficulty of graded reader texts. *Reading in a Foreign Language*, *34*(2), 249–270. https://nflrc.hawaii.edu/rfl/item/563

Arai, Y. (2022b). 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

Arai, Y. (2023). Comparability between graded readers and the Common Test in Japan in terms of text difficulty perceived by learners. *Language Assessment Quarterly*, *20*(3), 277–295. https://doi.org/10.1080/15434303.2023.2237497

Arai, Y., & Takizawa, K. (2025). Day and Bamford's (2002) ten principles for teaching extensive reading revisited: A methodological synthesis of research practice. *Reading in a Foreign Language*, *37*(1), 1–23. https://nflrc.hawaii.edu/rfl/item/629

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.

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.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2022). *Thematic analysis: A practical guide*. Sage.

Brennan, R. L. (2001a). *Generalizability theory*. Springer-Verlag.

Brennan, R. L. (2001b). mGENOVA (Version 2.1) [Computer software]. The University of Iowa. https://education.uiowa.edu/casma/computer-programs

Brierley, M., Lemmer, R., & Adachi, K. (2012). The Extensive Reading Foundation's online self-placement test. *The 2012 Pan-SIG Proceedings*, 22–30. https://pansig.org/publications/2012/pansig2012proceedings.pdf

Chang, A. C.-S., & Renandya, W. A. (2017). Current practice of extensive reading in Asia: Teachers' perceptions. *The Reading Matrix*, *17*(1), 40–58. https://www.readingmatrix.com/files/16-4vj138u3.pdf

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/10.1111/1467-9817.12049

Claridge, G. (2011). *What makes a good graded reader: Engaging with graded readers in the context of extensive reading in L2* [Unpublished doctoral dissertation]. Victoria University of Wellington.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. https://rm.coe.int/16802fc1bf

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Wiley.

Csikszentmihalyi, M. (2000). *Beyond boredom and anxiety: Experiencing flow in work and play*. Jossey-Bass. (Original work published in 1975)

Davis, C. (1995). Extensive reading: An expensive extravagance? *ELT Journal*, *49*(4), 329–336. https://doi.org/10.1093/elt/49.4.329

Day, R. R. (2015). Extending extensive reading. *Reading in a Foreign Language*, *27*(2), 294–301. https://nflrc.hawaii.edu/rfl/item/332

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

de Burgh-Hirabe, R., & Feryok, A. (2013). A model of motivation for extensive reading in Japanese as a foreign language. *Reading in a Foreign Language*, *25*(1), 72–93. https://nflrc.hawaii.edu/rfl/item/271

Donald, S. (2018). Reading to learn or teaching to read: The roles of learners and teachers in an extensive reading programme. *Extensive Reading World Congress Proceedings*, *4*, 245–252. https://jalt-publications.org/content/index.php/jer/issue/view/6

Grabowski, K. C., & Lin, R. (2019). Multivariate generalizability theory in language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment Volume I: Fundamental techniques* (pp. 54–80). Routledge.

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

Iyengar, S. S., & Lepper, M. R. (1999). Rethinking the value of choice: A cultural perspective on intrinsic motivation. *Journal of Personality and Social Psychology*, *76*(3), 349–366. https://doi.org/10.1037/0022-3514.76.3.349

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, *79*(6), 995–1006. https://doi.org/10.1037//0022-3514.79.6.995

Katz, I., & Assor, A. (2007). When choice motivates and when it does not. *Educational Psychology Review*, *19*, 429–442. https://doi.org/10.1007/s10648-006-9027-y

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

Lee, J. M., & Ro, E. (2020). "I actually picked up a physics textbook:" Complexities of the Freedom Principle in extensive reading. *English Teaching*, *75*(1), 3–23. https://doi.org/10.15858/engtea.75.1.202003.3

Leung, C. Y. (2002). Extensive reading and language learning: A diary study of a beginning learner of Japanese. *Reading in a Foreign Language*, *14*(1), 66–81. https://nflrc.hawaii.edu/rfl/item/55

Liao, R. J. T. (2023). The use of generalizability theory in investigating the score dependability of classroom-based L2 reading assessment. *Language Testing*, *40*(1), 86–106. https://doi.org/10.1177/02655322211070840

Lien, H.-Y. (2010). EFL college learners' perceptions of self-selected materials for extensive reading. *The English Teacher*, *39*, 194–204. https://www.melta.org.my/journals/TET/downloads/tet39_01_10.pdf

Lyon, P., & Barr, A. (2019). An investigation into the criteria students use when selecting graded readers. *Studies in Self-Access Learning Journal*, *10*(3), 258–281. https://sisaljournal.org/archives/sep19/lyon_barr/

Macalister, J. (2015). Guidelines or commandments? Reconsidering core principles in extensive reading. *Reading in a Foreign Language*, *27*(1), 122–128. https://nflrc.hawaii.edu/rfl/item/318

Mermelstein, A. D. (2015). Reading level placement and assessment for ESL/EFL learners: The reading level measurement method. *ORTESOL Journal*, *32*, 44–55. https://ortesol.wildapricot.org/Journal2015

Mori, S. (2015). If you build it, they will come: From a "Field of Dreams" to a more realistic view of extensive reading in an EFL context. *Reading in a Foreign Language*, *27*(1), 129–135. https://nflrc.hawaii.edu/rfl/item/319

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

Patall, E. A., Cooper, H., & Robinson, J. C. (2008). The effects of choice on intrinsic motivation and related outcomes: A meta-analysis of research findings. *Psychological Bulletin*, *134*(2), 270–300. https://doi.org/10.1037/0033-2909.134.2.270

Puripunyavanich, M., & Waring, R. (2024). Investigating the implementation of extensive reading in four Asian countries. *Reading in a Foreign Language*, *36*(1), 1–34. https://nflrc.hawaii.edu/rfl/item/611

Ramonda, K. (2020). Extensive reading and class readers: The case for no choice. *ELT Journal*, *74*(3), 277–286. https://doi.org/10.1093/elt/ccaa017

Ro, E. (2018). Understanding reading motivation from EAP students' categorical work in a focus group. *TESOL Quarterly*, *52*(4), 772–797. https://doi.org/10.1002/tesq.426

Robb, T. N. (2022). Encouraging schools to adopt extensive reading: How do we get there? *Reading in a Foreign Language*, *34*(1), 184–194. https://nflrc.hawaii.edu/rfl/item/551

Robb, T. N., & Ewert, D. (2024). Classroom-based extensive reading: A review of recent research. *Language Teaching*, 57(3), 295–324. https://doi.org/10.1017/S0261444823000319

Sakurai, N. (2020). Exploring a placement test for extensive reading programs. *ACTA Humanistica Et Scientifica Universitatis Sangio Kyotiensis*, *53*, 53–70. http://hdl.handle.net/10965/00010409

Sawaki, Y., & Xi, X. (2019). Univariate generalizability theory in language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment Volume I: Fundamental techniques* (pp. 30–53). Routledge.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.

Stoller, F. L. (2015). Viewing extensive reading from different vantage points. *Reading in a Foreign Language*, *27*(1), 152–159. https://nflrc.hawaii.edu/rfl/item/322

Sun, X. (2020). An exploration of students' and teachers' perceptions of a two-year extensive reading program in a Chinese secondary school. *The Reading Matrix*, *20*(1), 201–219. https://www.readingmatrix.com/files/22-a0623w45.pdf

Takase, A. (2009). The effects of different types of extensive reading materials on reading amount, attitude, and motivation. In A. Cirocki (Ed.), *Extensive reading in English language teaching* (pp. 451–465). Lincom.

Tan, Y. T. (2018). Effective strategies to get teenagers hooked on reading. *Extensive Reading World Congress Proceedings*, *4*, 73–78. https://jalt-publications.org/content/index.php/jer/issue/view/6

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

Wan-a-rom, U. (2010). Graded reading vocabulary-size tests for placement in extensive reading. *The Reading Matrix*, *10*(1), 15–44. https://www.readingmatrix.com/articles/april_2010/wan_a_rom.pdf

Waring, R., & McLean, S. (2015). Exploration of the core and variable dimension of extensive reading research and pedagogy. *Reading in a Foreign Language*, *27*(1), 160–167. https://nflrc.hawaii.edu/rfl/item/323

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

Zhou, J. (2024). *Extensive reading*. Cambridge University Press.

Appendix A

Estimated Variance Components[a] for the G Study

Source of variance	Level 3		Level 4	
	Variance	% of variance explained	Variance	% of variance explained
p	0.46596	39.8%	0.28634	25.0%
t	0.00723	0.6%	0.04981	4.4%
pt, e	0.69866	59.6%	0.80706	70.6%
Total	1.17185	100.0%	1.14321	100.0%

Notes. *p* = participants, *t* = GR texts, *e* = error.

[a]To estimate reliability, G theory employs an ANOVA-like measurement model to produce variance per different sources, including the target construct (i.e., object of measurement) and each type of systematic error in the measurement (i.e., facets of measurement). This table suggests G study variance components for the three sources of variance (i.e., participants [*p*], GR texts [*t*], and interaction of the two and error [*pt, e*]) obtained when one text from each grade level was included in the checklist. The variance components for the participant-text interaction were the largest among the three for both Levels 3 (59.6%) and 4 (70.6%).

Appendix B

Correspondence between the Characteristics of the ER Program and Day and Bamford's (2002) Ten Principles for Teaching ER

Characteristics of the ER program in this study	Corresponding principles by Day and Bamford (2002, pp. 137–140)
• Students were instructed to read easy texts. They were informed that the text difficulty should be neither too difficult nor too easy, and that slow reading can be a manifestation of reading difficulty. This study examined their perceptions of text difficulty to understand whether they read easy texts, but reading	• Applying Principle 1 (Easy texts) Not applying Principle 7 (Faster reading speed)

speed was not measured.	
• Students were instructed to choose texts freely from OBL GR texts (from Levels 2 to 5). Narrowing down the number of texts available was intended to reduce the possibility that they would choose too difficult/easy texts.	• Not applying Principle 2 (Various kinds of texts) Applying Principle 3 (Self-selected reading)
• Students were instructed to read as much as possible, individually, and silently. Students did not have to finish reading the entire text within 15 minutes but needed to do so before next class, meaning that they were required to read 12 GRs during the semester.	• Applying Principle 4 (Reading amount) Applying Principle 8 (Individual and silent reading)
• Students were informed that they did not have to comprehend the text perfectly, and that general comprehension would suffice. They were supposed to read for pleasure. Their pleasure experiences were measured as flow experiences.	• Applying Principle 5 (Reading purposes)
• There were neither reading comprehension questions nor post-reading activities after reading. ER was not a part of course grades.	• Applying Principle 6 (Reading for its own sake)
• Students were expected to choose different texts in each session.	• No corresponding principles
• Dictionary use was not prohibited.	• No corresponding principles
• The teacher explained all of these rules along with the merits and the procedure of the ER program.	• Applying Principle 9 (Teacher guide/orientation)
• The teacher did not read together with the students.	• Not applying Principle 10 (Teacher as a reader model)

About the Author

Yuya Arai, Ph.D., is an assistant professor at Chiba University, Japan. His current research interests include L2 extensive reading, language assessment, learner psychology, and research synthesis. His recent studies were published in *Language Assessment Quarterly*, *Reading in a Foreign Language*, and *Research Methods in Applied Linguistics*. Email: yuyaarai1@gmail.com; ORCID: https://orcid.org/0000-0001-7922-6616.