Reading in a Foreign Language October 2017, Volume 29, No. 2
ISSN 1539-0578 pp. 218–244

Measuring and predicting graded reader difficulty

Trevor A. Holster
Fukuoka University
Japan

J. W. Lake
Fukuoka Jo Gakuin University
Japan

William R. Pellowe
Kindai University Fukuoka
Japan

Abstract

This study used many-faceted Rasch measurement to investigate the difficulty of graded readers using a 3-item survey. Book difficulty was compared with Kyoto Level, Yomiyasusa Level, Lexile Level, book length, mean sentence length, and mean word frequency. Word frequency and Kyoto Level were found to be ineffective in predicting students' perceptions of book difficulty. Book length was found to be highly predictive of perceived book difficulty, with the Yomiyasusa Levels predicting 68% of variance, while the Lexile measure of mean sentence length was moderately predictive, with 40% of variance explained. These results show that current headword levelling of graded readers is ineffective and that publishers' book levels do not provide useful guidance in selection of books to read. It is therefore recommended that students use book length as their primary consideration in choosing books and that reading recommendations and purchasing decisions be based on Yomiyasusa Levels rather than publishers' levels.

Keywords: extensive reading, book difficulty, readability, Rasch analysis, many-faceted Rasch measurement

Although the distinction between extensive reading (ER) and intensive reading (IR) is at least a century old (Palmer, 1917), the question of how to persuade students to engage in ER continues to provoke debate. Day and Bamford (2002) produced a list of principles for teaching ER and Prowse (2002) independently arrived at similar conclusions. The essence of these principles is that students should engage in pleasurable, fluent reading of many easy, interesting books of the students' own choosing, with teachers providing guidance and support, but not assigning tasks that overtly focus on specific details of the text rather than general intra-subjective and inter-subjective meaningfulness. These principles of ER require that students can choose from a large selection of graded readers: books made accessible to beginners by the deliberate simplification

of syntactic and semantic features. Of course, as reading proficiency increases, more difficult books will become accessible so students must continually monitor their own reading levels in order to keep reading books of appropriate difficulty. Although publishers assign graded readers to levels, there is no common scale of difficulty even between two different series of graded readers by a single publisher. This means that two books rated as Level 1 cannot be assumed to be of similar difficulty unless they are from the same graded reader series. Additionally, given the emphasis in ER on reading for its own sake (Day & Bamford, 2002), publishers' claimed levels of book difficulty will only be useful if they are predictive of students' perceptions of book difficulty, but Claridge's (2012) investigation of four publishers' specification of graded reader levels found that they rarely or never surveyed students. This leaves important questions about how students can choose appropriate level graded readers under-researched because students' perception of a book's difficulty will be a major factor in its suitability for ER, but publishers have largely ignored this.

Background

Second language reading monitoring systems

Two systems developed in Japan attempted to address the problem of matching students to appropriate level books: the Yomiyasusa Levels (YL) (Furukawa, 2014a), and the Kyoto Scale (MReader, 2016a). In the YL (Furukawa, 2014a), books were rated on a 100 point scale, from 0.0 to 9.9. Ratings largely reflected the word count of the books, but also took account of factors such as illustrations and text styles, with review and adjustment every two years following feedback from teachers and students. The YL are freely accessible, and are presented both as a general level for popular book series (Furukawa, 2014a) and as tables listing the levels of individual books (Furukawa, 2014b). This allows teachers or students to target reading at books of a specified level.

The Kyoto Scale was developed largely through the headword counts (i.e., semantic level) provided by publishing companies, but adjusted with reference to the YL. It comprised 10 levels, ranging from Starter to Level 9 (MReader, 2016a). The Kyoto Levels were the basis of the MReader online monitoring system (MReader, 2016b), developed from the earlier Moodle Reader package (Robb & Kano, 2013). Access to the free MReader system was made available to administrators of ER programs and provided multiple-choice tests for each book title, with a time limit to compel students to read books cover-to-cover before starting a test, rather than skimming and scanning to find specific information while taking a test. Based on a cursory review of 49 MReader tests, testing the recall of specific details of a story was common, meaning that general comprehension of a story would often be insufficient to answer quiz questions. Instead, MReader quizzes included items requiring students to memorize specific details of the characters and events. MReader was designed to provide progress reports to administrators and teachers in the form of word counts of the books that each student passed tests on. Limits on the level of test accessible to individual students and the time interval between tests were also implemented, preventing students from reaching mandated word targets by reading a small number of long books or by skimming and scanning short, simple books to complete many tests in a single session.

The Kyoto Scale's reliance on publishers' headword levels also raises concerns. Wan-a-rom (2008) investigated the vocabulary of different graded reader series and found that, especially at lower levels, the books diverged considerably from the publishers' lists, in part because even low level graded readers included idiosyncratic vocabulary that was essential to the story. Thus, two graded readers of the same headword level from the same publisher may be substantively different in difficulty but still be assigned to the same level on the Kyoto Scale. Further to this, Claridge (2012) showed large discrepancies between different publishers in how headword levels were used in constraining and rating the level of graded readers, raising concerns over whether the levels from different publishers can be used to assign books to a common scale, an assumption that is implicit in the Kyoto Scale.

IR versus ER

MReader's emphasis on testing recall of specific details raises concerns about whether it promotes ER or IR, as defined by Palmer:

> Reading may be intensive and extensive. In the former case each sentence is subjected to a careful scrutiny, and the more interesting may be paraphrased, translated, or learnt by heart. In the latter case book after book will be read through without giving more than a superficial and passing attention to the lexicological units of which it is composed. (Palmer, 1917, p. 205)

Palmer's conception of ER assumed fluent reading of texts simple enough that conscious attention to formal language features was unnecessary. Yamashita's (2015) review argued that ER is most appropriately associated with Carver's (1993) notion of *rauding* which involves fluent reading of easily comprehensible texts, in contrast to *learning* and *memorizing*, both of which involve much slower reading with conscious attention to details of the text. Reading to learn and reading to memorize are incompatible with ER (Yamashita, 2015). Further support for the importance of rauding was provided by Yamashita and Shiotsu's (2017) finding that listening comprehension, a key component of Carver's (1993) model of rauding, was the strongest predictor of second language (L2) reading comprehension, with the implication that L2 reading programs should emphasize the importance of activities that encourage rauding. Given MReader's emphasis on memorization of details rather than engagement in the rauding behavior that Yamashita (2015) saw as definitive of ER, MReader is more appropriately viewed as an IR monitoring system, or as a blend of IR and ER (Day, 2015). This departure from the traditional definition of ER is further evidenced by Robb's (2002) rejection of Day and Bamford's (2002) ER principles on the grounds that students read "to satisfy a course requirement" (Robb, 2002, p. 146). However, regardless of whether we view MReader as a monitoring system for ER or for IR, gathering evidence as to the effectiveness of the Kyoto Scale in matching students' reading level to book difficulty is still necessary, and research into first language (L1) reading provides potentially useful tools for this purpose.

First language readability

L1 reading researchers have adopted more technically sophisticated methodology than used in the development of the YL and Kyoto Scale, as demonstrated by Stenner, Burdick, Sanford, and Burdick's (2007) review of readability formulas. L1 readability formulas estimate readability by combining a syntactic component such as sentence length with a semantic component such as word length or word frequency. Two of the most common of these are the Flesch Reading Ease and the Flesch-Kincaid Grade Level (Stenner, et al., 2007) which are available within the Microsoft Word word-processing software. However, readability scales based on average grade level do not map the reading ability of individual students to the readability of specific books because students of the same age can vary enormously in their reading ability. Without mapping both the difficulty of individual books and the ability of individual persons onto a shared measurement scale, we cannot make detailed predictions about which books are accessible to which students.

The Lexile Framework (Stenner, 1999; Stenner, et al., 2007) achieved this calibration by combining word frequency from a 600-million word corpus with sentence length, these respectively serving as proxies for the semantic load and the syntactic load of texts. The difficulty of reading texts was then calibrated against test items using Rasch analysis (Wright & Stone, 1979) which provided equal-interval measures of both person ability and test item difficulty in log-odds units, or *logits*. In the Lexile framework, logits were transformed to a more user-friendly scale called *Lexile units*, giving a difference of 1000 units between basal primers and an electronic encyclopedia, respectively anchored at Lexile levels of 200 and 1200. This allowed person ability to be measured through a comprehension test and book difficulty to be measured through computerized analysis of textual features, with both person ability and text difficulty reported in Lexile units. Most importantly, the Lexile Framework provided for prediction of the comprehension level of new texts by individual persons. When a person had the same Lexile level as a book, 75% comprehension was predicted, but only 50% comprehension was predicted when the text difficulty exceeded the person ability by 250 Lexile units and 90% comprehension was predicted when person ability exceeded book difficulty by 250 Lexile units.

Rasch measurement

Although the Rasch measurement used in the development of the Lexile Framework may be unfamiliar to classroom practitioners accustomed to reporting test scores as raw percentages, the Rasch model is conceptually very simple (Bond & Fox, 2015; Engelhard, 2013). Furthermore, the intended use of MReader to criterion reference person ability to the difficulty of books (MReader, 2016b) relied on the same assumptions of invariance that underpin the Rasch model. Engelhard (2013) provided a conceptual overview of the assumptions of invariant measurement. On this view, sample independence is a prerequisite for measurement, meaning that the relative reading ability of persons does not change depending on which books are read and the relative difficulty of books does not change depending on who reads the books. Rasch measures are relative, not absolute, so the probability of a person being able to read a book will increase as books become easier relative to that person's ability. What is viewed as invariant, however, is that a high-ability person is always predicted to have a higher probability of success than a low-

ability person. Similarly, a difficult book is modeled to be more difficult for all persons than an easy book.

The Rasch model is not deterministic, but rather *stochastic*, or probabilistic, meaning that it assumes that sometimes low-ability persons succeed on difficult tasks and high-ability persons fail on easy tasks, resulting in *unexpected responses* to test items. McNamara (1996) provided an accessible introduction to the stochastic nature of Rasch measurement, which has the counter-intuitive implication that measurement is only possible when some unexpected responses are observed. This is because calculation of the probabilities of success and failure are only possible when observed responses do not follow perfectly predictable, or deterministic, patterns. The assumption underlying classical test theory (CTT), where perfectly deterministic patterns are seen as a theoretical ideal, is conceptually incompatible with stochastic models such as the Rasch model (Bond & Fox, 2015; Engelhard, 2013; McNamara, 1996; Wright & Stone, 1979). Instead, stochastic analyses rely on comparing patterns of responses of persons to items across a large dataset that necessarily includes unexpected responses. Stochastic datasets allow the probabilities of each person succeeding on each item to be calculated as odds-ratios; the *expected response* of each person to each item is defined as the probability of success of that person on that item. In the stochastic Rasch model, low-ability persons will have lower probabilities of success on any task than high-ability persons, but the observed responses will never perfectly match the expected responses.

Rasch data-model fit

The difference between an observed response and the expected response to an item is known as the *score residual*. The distribution of the squared standardized residuals across a dataset is expected to follow a chi-square distribution (Linacre, 2014). This provides for quality-control statistics known as *fit statistics*, showing the degree to which the observed data meet the stochastic assumptions of invariant measurement. Excessively noisy data are manifested as *misfit*, indicating distortion of the measurement scale and degradation of measurement. Overly deterministic data are manifested as *overfit*, indicating redundancy in the data, with the implication that measurement is muted. Just as in physical measurement, which is seen as an exemplar of measurement, psychometric data will never perfectly match an idealized model so Rasch analysis allows researchers to determine whether the quality of measurement is adequate for the purpose at hand.

Although the developers of the Kyoto Scale and YL did not conduct any psychometric analysis on the quality of measurement of their scales, the requirement of sample-independent measurement is clear when the nature of the measurement problem is considered. MReader flags students who read many of the same books as potential cheats, so the expectation is that every student will read a unique sample of books. The sampling of books will not be random, however, because low-ability students are constrained to read easy books while high-ability students will read more difficult books. Thus, all students are expected to score highly on the MReader tests. In the case of the Lexile Framework, students who consistently read at an appropriate level will average about 75% on comprehension tests, regardless of whether they are of high-ability or low-ability. Under both systems, raw percentage scores cannot be used to estimate either book difficulty or person ability because the sampling pattern is designed to ensure that all students

receive very high raw percentage scores. However, the sample-independent invariant measurement provided by Rasch analysis allows person ability and item difficulty to be calculated from datasets such as this, provided there is sufficient connectivity within the dataset. As well as providing powerful tools for quality control, the Rasch model thus provides the practical benefit of allowing the analysis of the very sparse, non-randomly distributed datasets expected in research into the difficulty of reading texts.

Dataset connectivity and database design

Unfortunately, MReader was designed only to provide summaries of the percentage of correct responses to each quiz item rather than the matrix of item responses needed for item analysis under either CTT (Brown, 2005) or Rasch analysis (Linacre, 1994; Wright & Stone, 1979), precluding measurement of the difficulty of books or the ability of persons. Further to this, items for different books were contributed by numerous volunteer item writers and detailed test specifications were not developed to ensure that the quizzes for different books contained items representing equivalent content (Robb, Campbell, & Bateson, 2014). Different books were not linked by common items or by items written to equivalent specifications, so items are nested within disjoint subsets of data for each book without the connectivity required to measure book difficulty, as distinct from item difficulty. Although it would be possible to rewrite the MReader database software to provide the matrix of item responses, correcting the disjoint subset problem would require an enormous amount of work to develop detailed item specifications and review and rewrite every single quiz in the item bank. The years of effort that this would entail made it preferable to start with a clean-sheet design rather than trying to rewrite MReader for a purpose it was not designed for. Thus, although the Lexile system and MReader system both test students' comprehension of texts, the Lexile system was specifically designed to provide measurement of book difficulty and student ability on a common scale, but MReader was implemented in a manner that precludes this.

Objectives and research questions

Despite the technical sophistication of the research used to develop the Lexile Framework, results are reported in terms simple enough that program administrators, classroom teachers, and parents can make instructional decisions by matching students' reading levels to book difficulty. The Lexile text analyzer (Lexile, 2016) is also provided as a free online tool providing Lexile levels for text samples of up to 1000 words. This provides a potential alternative to the Kyoto Scale and YL for the estimation of the difficulty of L2 graded readers, with the advantage that, in addition to an overall Lexile level for each text, estimates of syntactic complexity and semantic complexity are reported as mean sentence length and mean log word frequency, respectively. However, the Lexile framework was developed for L1 readers, raising questions about its validity for measuring the difficulty of graded readers for L2 readers. This research therefore aimed to compare the effectiveness of the Kyoto Scale, the YL, and the Lexile Framework in predicting student self-reports of graded reader difficulty. This necessitated development of an online ER monitoring system based on the existing MOARS audience response system (Pellowe, 2016) to gather student self-report ratings of graded reader difficulty, validation of the resulting instrument using many-faceted Rasch measurement (MFRM) (Linacre, 1994), and then comparison of the different measures of difficulty. Two research questions were posed:

RQ1. Can students' self-report ratings provide valid estimates of book difficulty?

RQ2. Does word count, semantic level, or syntactic level provide the best prediction of students' perceived book difficulty?

Method

Participants

Participants were students in compulsory reading classes at two Japanese women's universities between April 2012 and March 2015. Scores from the TOEFL IP, based on the superseded pencil and paper form of the TOEFL (ETS, 2008), were available for students at one institution comprising the majority of the participants, with mean scores of approximately 450 and 90% of students falling between TOEFL 400 and 500. The majority of participants were therefore of novice level proficiency, with insufficient English ability to handle simple everyday tasks. The assertion that it is a "fact of life" (Robb, 2002, p. 147) that Asian students cannot be relied on to take responsibility for their own learning was not supported by student behavior during the development and piloting of the ER monitoring system in 2011, with the majority of students completing both in-class and homework tasks with commendable motivation and enthusiasm. Day and Bamford's (2002) ER principles were therefore adopted as the conceptual basis of the ER program. Although ER was assigned as homework, teachers were asked not to set word goals or use ER as part of formal grades, but to treat it as recommended but non-compulsory independent study and to praise participation rather than punishing non-participation. Because multiple ratings are essential for reliable measurement of both book difficulty and student ability, students who reported reading fewer than five books were excluded from the analysis, as were books with fewer than three reviews. This led to a recursive trimming of the dataset until a core of 668 students and 1016 books were retained of the original 810 students and 1383 books.

Survey instrument

Although the MReader system was considered for the ER program, it was not adopted because of concerns that the comprehension quiz format encouraged IR rather than ER, following similar concerns to those later raised by Yamashita (2015) and Day (2015). Instead, a 6-item survey was written and piloted in pencil-and-paper form in 2011, and converted to an on-line format for operational use in 2012. The survey was primarily intended as a formative tool to remind students of the principles of ER, with three items intended to remind students that they should seek books they find personally interesting (the interest dimension), and three to remind them to read easy books (the difficulty dimension). A secondary consideration in the instrument design was to determine appropriate level graded readers for library purchases, with the use of common items across all books providing the data connectivity needed for analysis using MFRM. The survey items are shown in Appendix A, including the response options and associated rating scales used in the analysis. Items 2, 3, and 4 addressed the difficulty dimension, with Items 2 and 4 having reversed polarity, so *Very often* dictionary use for Item 2 and *Very difficult* book for Item 4 indicated lower person ability relative to book difficulty, while reading *Very quickly* for

Item 3 indicated higher person ability. It should be noted that, as the research questions of this study addressed book difficulty, the results section is limited to analysis of responses to Items 2, 3, and 4, the items which address the difficulty dimension, and no analysis is included of the three items comprising the interest dimension.

Procedure

In the first or second week of class, teachers distributed a handout giving a brief explanation of ER and its purpose, plus a pencil-and-paper version of the survey, all presented in English as classes were conducted entirely in English. As homework, students were asked to find two or three potentially interesting graded readers in the library, and to complete the survey and bring it to class the following week. In the next class, teachers distributed instructions and log-in information for the ER monitoring system, demonstrated how to complete the on-line version of the survey (also presented in English), and offered students the opportunity to enter their survey data using mobile phones if desired. Students who did not enter the first week's data in class were asked to enter it for homework. From the third week onwards, teachers were provided with a weekly report slip for each student showing the cumulative number of books read by that student along with the number read by the 25th and 75th percentiles of students. This was intended to remind students that they were expected to engage in ER outside of class and to give them feedback on their relative effort, but teachers were asked not to attempt to compel reluctant students to complete surveys by threatening grade penalties. However, students' grades and classroom management were completely determined by classroom teachers and there was no way to independently check how teachers implemented ER or to compel them to follow the recommended procedures. Therefore, teachers were trusted to adapt the procedures to the needs of their own classes as they saw appropriate.

Design

Data collection and facets model specification. The MOARS ER module (Pellowe, 2016), an internet based open-source audience response system, was used to collect data for MFRM analysis using the Facets software package (Linacre, 1994, 2010a). Classroom teachers will be implicitly familiar with two-faceted assessments, where persons respond to items (i.e., test questions). The Rasch model is derived from the simple insight that the probability of success increases with greater person ability and decreases with greater item difficulty. In Rasch analysis, the convention is to use the term *difficulty* for the property exhibited by items and *ability* for the property exhibited by persons, but this convention does not imply any particular theory about why some items result in increased or decreased probabilities of success, or why some persons have higher or lower probabilities of success. Thus, ability and difficulty are simply conventional terms that express the positions of items and persons on the same latent trait, with ability representing a positive facet reflecting an increased probability of success, and difficulty representing a negative facet reflecting a reduced probability of success.

For this study, however, the two facets of persons and items were insufficient to model the interactions leading to responses. Each response reflected the interaction of a *Reader* (i.e., person) and an *Item*, with regard to a *Book*, at a specified *Time*. Therefore, a four-faceted model was specified, where the response to each survey item was modeled to reflect the interaction of 1)

Readers (ability), 2) *Books* (difficulty), 3) *Time* (ability gain), and 4) *Items* (difficulty). For this analysis, *Readers* and *Time* were modelled as positive facets, meaning that the probability of endorsing an item increased for persons with greater ability or after more time. Books and Items were modelled as negative facets, meaning that greater difficulty of either books or survey items reduced the probability of endorsement. As all facets are measured in equal-interval logits that represent positions on the same difficulty parameter, *Readers*, *Books*, *Time*, and *Items* can be mapped onto a common measurement scale. The probability of endorsement of an item can be conceptually expressed as:

$$P = f(R + T - B - D) \quad (1)$$

Where:

P = probability of endorsement
R = reading ability of the person
T = time
B = difficulty of the book
D = difficulty of the survey item

Analysis was conducted using the Masters partial credit model, where the scale intervals for different items are not assumed to be equal, in contrast to the assumptions of Likert type scales where all items are assumed to represent a common scale, as described in detail by Linacre (2016b). Thus, for different items, the difference between adjacent raw responses is not assumed to represent equal differences in person ability or book difficulty. In addition to logit measures of each person, book, and survey item, Facets provided detailed reports including reliability indices, data-model fit, and tables of unexpected responses, allowing diagnosis ranging from the global functioning of the instrument to interactions between individual persons, books, and items.

Book sampling. Once logit measures of books were established, up to 10 popular books from different levels of 14 graded reader series were purchased, electronically scanned, and edited to remove review sections, glosses, and other elements that did not form part of the core reading text. Ideally, all of the 1016 measured texts would have been scanned and analyzed, but destructive scanning of books was necessary to meet the project deadlines and the research budget only allowed for the purchase of approximately 300 books. The most popular books from each graded reader series were identified from data gathered in the first six months of the project. When possible, 10 books from each level of each series were purchased, but relatively few reviews of higher level books were recorded, resulting in fewer than 10 books being included for some levels of some graded reader series. However, including some of these books was desirable to provide a greater range of book difficulty, providing increased variance within the dataset, and thus higher quality measurement. Ultimately, 309 samples of text were analyzed, with Microsoft Word used to obtain word counts, average sentence length, and readability statistics for each text. Kyoto Scale levels were obtained from the MReader website (MReader, 2016a), YL from Furukawa (2014b), and the free Lexile analyzer (Lexile, 2016) was used to estimate the Lexile level of each book. SPSS version 19 was then used to calculate correlations between the various estimates of difficulty: Lexile Measures, Lexile Words per Sentence (LWPS), Lexile Word Frequency (LWF), Kyoto Scale, YL, Word Count, Words per Sentence (WPS), Characters per Word (CPW), Flesch Reading Ease, and Flesch-Kincaid Reading Level.

Results and Discussion

Measurement rulers

The Facets software package provided detailed reports ranging from the global functioning of the survey down to detailed analysis of individual persons, books, and items. Rasch analyses typically begin with examining the global functioning of the test or survey instrument, followed by detailed analyses specific to the research questions of interest. The facets map provides a graphical illustration of the measurement rulers, with all measurement facets mapped to a shared logit scale. Figure 1 shows the facets map, with the logit scale on the left and the response scales for the three difficulty items on the right, the three interest items having been excluded as irrelevant to the research questions of this study. Logits are, by definition, equal-interval measures of ability-difficulty but it can be seen that the steps in the raw response scale do not represent equal intervals, precluding the use of raw scores as measures. Following conventional Rasch practice, person ability, i.e., the *Reader* facet, was non-centered, meaning that the other three facets were anchored to mean values of 0.00 logits. The mid-point of the raw rating scale, where there is a 50% probability of endorsing responses of either 1 or 2, is thus anchored to 0.00 logits. Average reader ability is much higher on the scale than average book difficulty, meaning that average responses to items were above the mid-point of the scale, consistent with students selecting easy books to read, as intended. The *Time* facet shows changes in average ability by the number of books read, in bands of 10 books. Following Engelhard's (2009) guideline of 0.30 logits as the threshold for a substantive effect size, students showed a small gain of 0.34 logits after reading 30 books and a substantively significant gain of 0.59 logits after reading 80 or more books. However, of the 668 students, the median number of books read was 25 and only 19 students read 80 books or more. It is probable that students with high book counts had different motivational orientations than those with low book counts, limiting the conclusions that can be made concerning reading gains. The research questions of this study are therefore limited to the measurement of book difficulty, with learning gains constituting a confounding variable for that purpose. The inclusion of *Time* as a facet is therefore necessary only to eliminate the effect of this confounding variable on the measures of *Readers* and *Books*.

Also of interest in Figure 1 is that Item 3, *How quickly did you read this book?*, was the most difficult item by a substantive degree, while Item 4, *Was this book difficult?*, was the easiest. In this context, difficulty means the likelihood of endorsing the response options in the scale, so, although students reported that they did not consider the books difficult, they were much less likely to report that they could read them quickly. A possible explanation of this was provided by Shiotsu and Weir (2007) who found syntactic knowledge to be a slightly better predictor of reading comprehension than vocabulary, so students may have struggled with fluent syntactic parsing while not being challenged by the semantic content of the books. Although peripheral to the research questions of this report, the finding that students reported reading quickly to be most difficult is supportive of the need for emphasizing Carvers' (1993) notion of rauding in ER programs rather than reading to learn or reading to memorize.

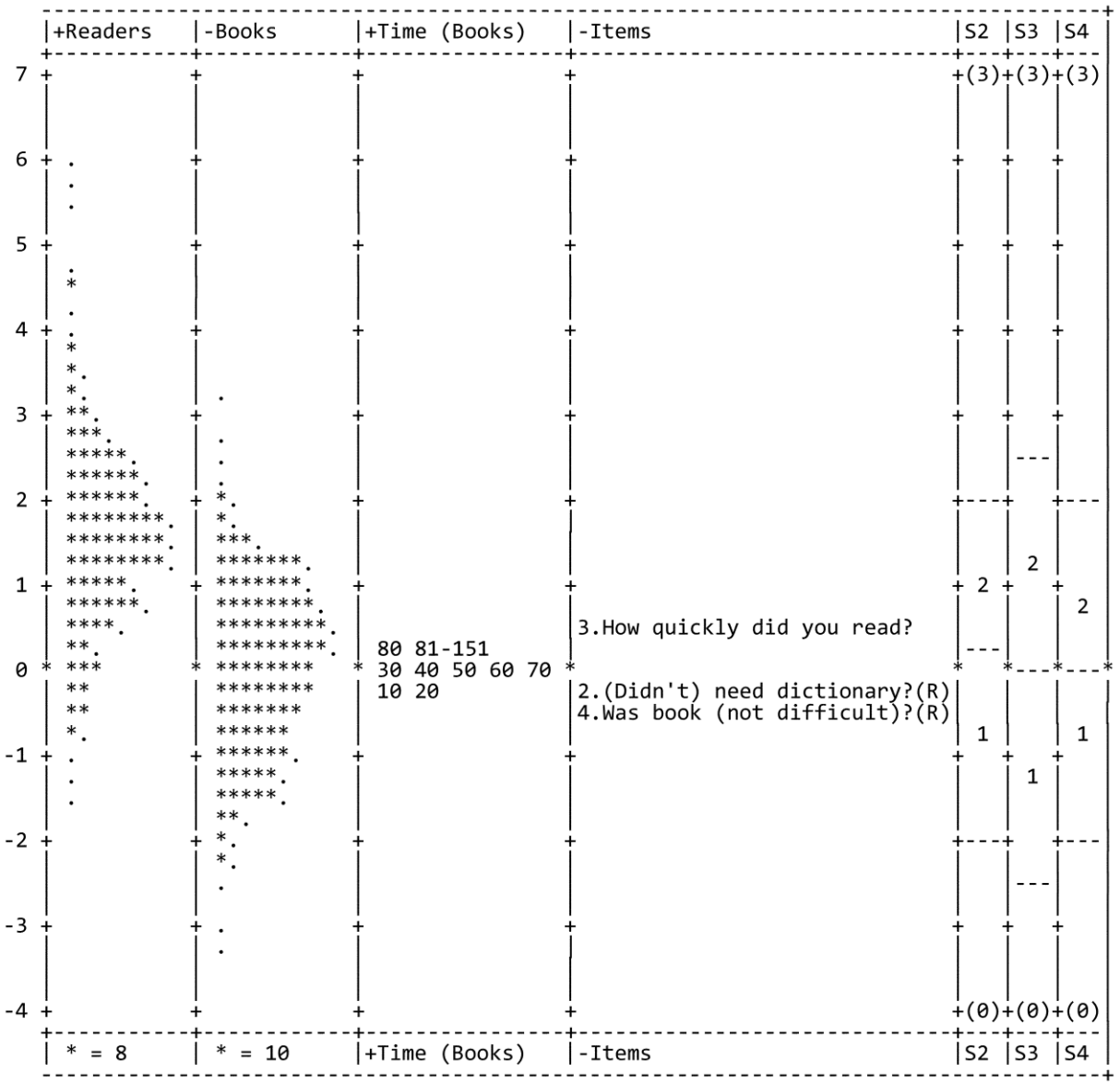

Figure 1. Facets map showing reader ability, book difficulty, learning gains, and survey item difficulty mapped onto a common logit measurement scale.

Figure B1, shown in Appendix B, confirms the unsuitability of raw survey responses as measures of book difficulty. Logit measures for each book are shown on the vertical axis, with the average raw response for each book shown on the horizontal axis in the left-hand panel. An average raw response of 2.0 can be seen to map to a range of logit measures exceeding 2.0 logits, equal to a probability of endorsement increasing from 27% to 73%. This represents approximately half the range of book difficulty, so this level of raw response could indicate a very difficult book or a book that is easier than average. Raw responses cannot provide reliable rank-ordering of book difficulty, let alone the interval level measurement required for this study. The right-hand panel shows the fair-average score: the average raw response that would be expected if all persons had rated all books. Each distinct fair-average score maps to a single logit measure, meaning that the fair-average score can rank-order book difficulty. However, while the transformation is

approximately linear up to a fair-average score of 2.5, it is clearly non-linear above this level. Although fair-average scores provide ordinal level measurement, the generation of these is equally as computationally complex as generating logit measures which provide the interval level measures needed to address RQ2. This study, therefore, used logit measures rather than raw responses.

Research question 1: Validity

Reliability of measurement. RQ1 concerned the validity of the measurement of book difficulty. Following Bachman (2000), validity concerns the interpretation of scores and their consequences, not the instrument itself. Scores might provide for valid interpretation for one purpose but not for a different purpose, so the instrument itself is not valid in any universal manner. Rather, each user of an instrument must gather evidence that the interpretation of scores made in that instance is valid. As discussed in detail by Holster and Lake (2016), standard tools used in Rasch-based validity arguments are analyses of reliability of separation, unidimensionality, and data-model fit. Facets returns a reliability of separation coefficient for all facets. Person reliability is analogous to Cronbach's alpha in classical analysis, reported as a coefficient ranging from .00 to 1.00. However, Rasch reliability is most easily interpreted as a separation index (Linacre, 2010b), with a minimum value of 0.00 and no upper limit. The separation index tells us how many statistically significant levels of ability or difficulty were observed in the sample. Table 1 provides summary statistics for all four facets, with the lowest separation index of 3.13 for Books providing a conservative estimate that the range of book difficulty exceeds the measurement error sufficiently that books in the middle of the range of difficulty are statistically separated from the most difficult and the easiest. The chi-square statistics provide a test of the fixed effect hypothesis that all elements of the facet share the same measure, with all facets showing an extremely high degree of confidence that the highest and lowest measures are different.

Table 1. *Reliability of separation*

		Readers	*Books*	*Time*	*Items*
Reliability:		.96	.91	.96	1.00
Separation:		4.62	3.13	4.91	46.89
Strata:		6.50	4.50	6.88	62.86
Chi-square:	Fixed	20090.90	16034.30	642.80	4456.10
	df	667	1015	8	2
	Sig.	.00	.00	.00	.00
	Random	620.10	921.80	7.70	2.00
	df	666	1014	7	1
	Sig.	.90	.98	.36	.16

Data-model fit. Although reliability is a prerequisite to addressing the research questions of this study, adequate data-model fit is a fundamental assumption of the Rasch model because data-model misfit indicates distortions in the measurement scale. The Rasch model describes an idealization of measurement. Although no psychometric dataset will show perfect fit to the Rasch model, this indicates that the measuring instrument is imperfect, not that the logit scale is flawed. The practical issue is whether the distortions are severe enough that productive measurement is compromised. To this end, Rasch fit statistics provide diagnosis of both the magnitude of the distortion and its source.

Rasch fit statistics are most commonly analyzed in the form of the mean-square statistic, as summarized in Table 2, provided as both infit and outfit values. The infit statistic is information weighted, so provides an important indication of the effect of misfit on measurement, while the outfit statistic is unweighted, so provides diagnostic information about the effect of outlying responses. The mean-square statistic is constrained to an average value of approximately 1.00, with a lower bound of 0.00 and no upper bound. Given the stochastic nature of the Rasch model, some degree of unpredictability is expected in the response set and a mean-square value of 1.00 indicates precisely the expected amount of unpredictability, while values greater than one indicate noisy data, or *misfit*, and values less than one indicate more predictability than expected, or *overfit*. Linacre (2014) advised that mean-square values greater than 1.50 warrant investigation, while those greater than 2.00 are unproductive for measurement. From Table 2, we can see that the facets of *Time* and *Items* were acceptably fitting, with standard deviations less than 0.10 for both infit and outfit, but that *Books* and *Readers* had very large ranges of fit. Thus, some books and some persons were much more predictable than expected, while others were much less predictable.

Table 2. *Data-model fit*

		Readers	*Books*	*Time*	*Items*
Count	Mean	92.80	61.00	6884.30	20653.00
	SD	65.00	61.80	6080.90	0.00
Infit Mean-square:	Mean	1.00	0.99	1.01	1.01
	SD	0.39	0.35	0.04	0.05
Outfit Mean-square:	Mean	1.02	1.00	1.01	1.04
	SD	0.52	0.43	0.06	0.07

Table 3 shows the measurement report for survey items. All three items showed excellent data-model fit, with maximum outfit of 1.11 for Item 2 and maximum infit of 1.05 for Item 3. Item 4 is slightly overfitting, meaning that responses to this item were marginally more predictable than expected. Fundamental to the Rasch model is the assumption that all items discriminate equally, meaning that the item characteristic curves (ICCs) for all items are expected to follow parallel trajectories (see Holster & Lake, 2016 for an illustration of the consequences of abandoning this assumption). Figure 2 shows the modelled and observed ICCs for the three items, with all three closely following the predicted values for response values above 1. Below this level, the upper and lower 95% confidence intervals become extremely large due to very few responses being recorded at this level. These responses reflect books that students reported as extremely difficult, and thus that contribute little information towards measurement. They will accordingly have more effect on the unweighted outfit mean-square statistic than on the information weighted infit mean-square statistic, which is the crucial indication of the quality of measurement. Similarly, the greater difficulty of Item 3 is reflected in wider confidence intervals as the ICC nears the maximum response of 3, due to relatively few responses at this level.

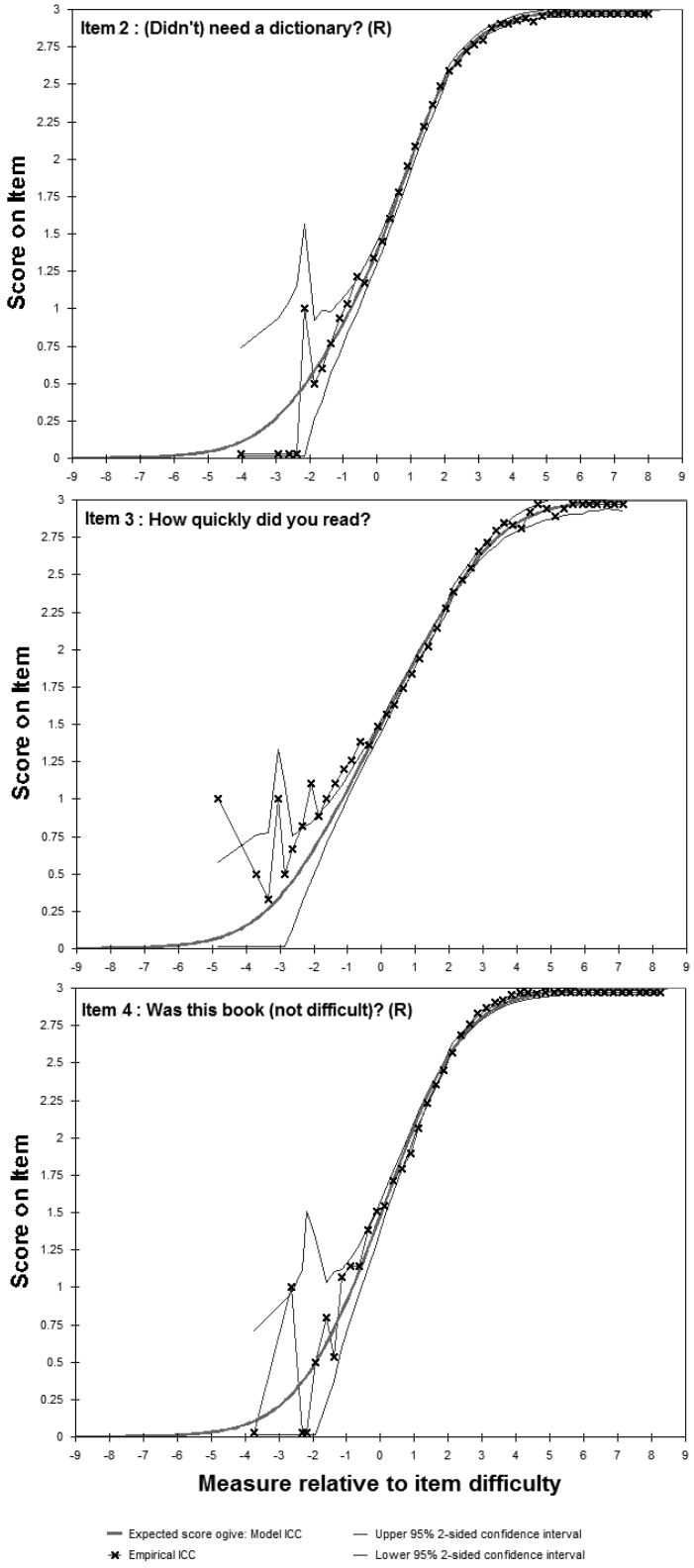

Figure 2. Item characteristic curves for the three difficulty items.

Table 3. *Item measurement report*

		Average		Logit		MnSq Fit			Correlation		
Count	Score	Obs	Fair	Measure	*SE*	In	Out	Dscr	PtM	Exp	Item
20653	50174	2.43	2.40	-0.17	0.01	1.03	1.11	0.97	.56	.58	2
20653	42194	2.04	1.90	0.62	0.01	1.05	1.06	0.95	.59	.61	3
20653	52666	2.55	2.54	-0.46	0.01	0.94	0.95	1.08	.57	.54	4
20653	48344.7	2.34	2.28	0.00	0.01	1.01	1.04		.58	Mean	
	4466.6	0.22	0.27	0.46	0.00	0.05	0.07		.01	*SD* (Pop)	
	5470.4	0.26	0.34	0.56	0.00	0.06	0.09		.01	*SD* (Samp)	

Pop RMSE 0.01 True *SD* 0.46 Separation 38.28 Strata 51.38 Rel. 1.00
Sample RMSE 0.01 True *SD* 0.56 Separation 46.89 Strata 62.86 Rel. 1.00
Fixed (all same) chi-square: 4456.1 *df* 2 significance .00
Random (normal) chi-square: 2.0 *df* 1 significance .16

Discrimination and point-measure correlations. Overall, Table 3 shows Item 4 to be slightly more discriminating than Items 2 and 3, but the magnitude of this is far below any level of concern. The point-measure correlation is analogous to the CTT notion of item discrimination, where equal discrimination is not an assumption and higher discrimination typically indicates a better functioning item (Brown, 2005). Table 3 shows all three items to have excellent point-measure correlations, the lowest being Item 2, with .56. Comparison of the mean-square fit statistics and point-measure correlations show that Item 3 has highest point-measure correlation (i.e., the highest CTT discrimination), at .59, but the lowest Rasch discrimination and worst mean-square infit, at 0.95 and 1.05, respectively. This illustrates that the Rasch and CTT assumptions about what constitutes an effective item are fundamentally different and that correlation-derived CTT indices do not provide the same information as Rasch data-model fit and discrimination. Table 3 also shows the expected point-measure correlation, namely the value predicted if the data perfectly matched Rasch model assumptions. Interestingly, Item 4, which is slightly overfitting, has the lowest expected point-measure correlation, at .54, but the observed correlation is fractionally higher, at .57. Conversely, Items 2 and 3, which are both slightly misfitting, have slightly lower observed point-measure correlations than expected.

Construct definition. Reconsidering the content of the items, Item 4, *Was this book difficult?*, aimed to directly address holistic book difficulty without consideration of sub-components; Item 2, *Did you need a dictionary?*, was intended to address the semantic component of book difficulty, based on the hypothesis that greater semantic difficulty of a book would be manifested in greater dictionary use; and Item 3, *How quickly did you read this book?*, was written with the intention of addressing the syntactic component of difficulty, based on the hypothesis that greater syntactic complexity would be manifested in slower syntactic parsing. The results presented in Table 3 are consistent with all three items contributing to measurement of a single latent trait of students' perception of book difficulty. Item 4 addressed this directly and showed slight overfit, or redundancy, indicating that responses to this item slightly over-predicted the patterns of the dataset. Items 2 and 3 addressed the latent trait less directly and showed slight misfit, indicating slightly noisy measurement and slight under-prediction of the overall patterns of the dataset. However, all three items showed data-model fit that was comfortably within acceptable limits and the fit statistics overall were consistent with a well-functioning instrument.

Dimensionality. Another fundamental assumption of Rasch measurement is that of unidimensionality. Facets reported 42.6% of variance explained by the Rasch trait, comfortably exceeding the 20% minimum recommended by Reckase (1979), a necessary, but not sufficient, condition for unidimensionality. Dimensionality was therefore investigated by cross-plotting of item subsets, following Linacre (2016a). With three survey items, there are three possible subsets of two items, so logit measures of book difficulty were calculated for each subset and cross-plotted, as shown in Figure 3. Each panel in Figure 3 compares the effect of two items on book difficulty measures. Although, as shown in Figure 1 and Table 3, Item 3 is substantively more difficult to endorse than Items 2 and 4, the Rasch assumption of sample independence means that the relative positions of books (and persons) on the latent trait should not be affected by use of different subsets of items. Figure 3 confirms this, with the scatterplots for all three comparisons closely following a linear trendline, consistent with a unidimensional survey instrument.

Table 4 shows inter-subset correlations. The lowest raw correlation of .93 was observed from the substitution of Items 2 and 3, with a raw correlation of .94 observed for the substitution of Items 3 and 4, and .96 for the substitution of Items 2 and 4. Raw correlations are attenuated by measurement error, making disattenuated values preferable as estimates of the unidimensionality of the instrument. Disattenuation was achieved by dividing the raw correlation between each pair of forms by the geometric mean of the reliability of the two forms, following Wang and Chen (2004). After disattenuation, all three sub-form comparisons returned correlations exceeding 1.00, indicating that the different sub-forms provided invariant measurement within the limits of measurement error. Further to this, Item 4, *Was this book difficult?*, found to be slightly overfitting in Table 3, showed the most consistent performance in the sub-form comparison in Table 4 and Figure 3. The substitution of Item 4 for other items had the smallest effect on book measurement, while substituting Item 3 for other items had the largest effect. Although these differences are not large enough to raise concern, they provide further evidence of the construct validity of the instrument as a measure of students' perception of the difficulty of books because Item 4, which directly addressed difficulty, showed the highest agreement with the instrument overall. Item 3, intended to address reading speed, but with potentially ambiguous wording, showed slightly lower agreement with the overall results.

Table 4. *Correlations between instrument sub-forms*

Test Forms	Changed Items	Mean Reliability	Raw Correlation	Disattenuated Correlation
2-3 vs. 2-4	4-3	.85	.94	> 1.00
2-3 vs. 3-4	4-2	.87	.96	> 1.00
2-4 vs. 3-4	3-2	.87	.93	> 1.00

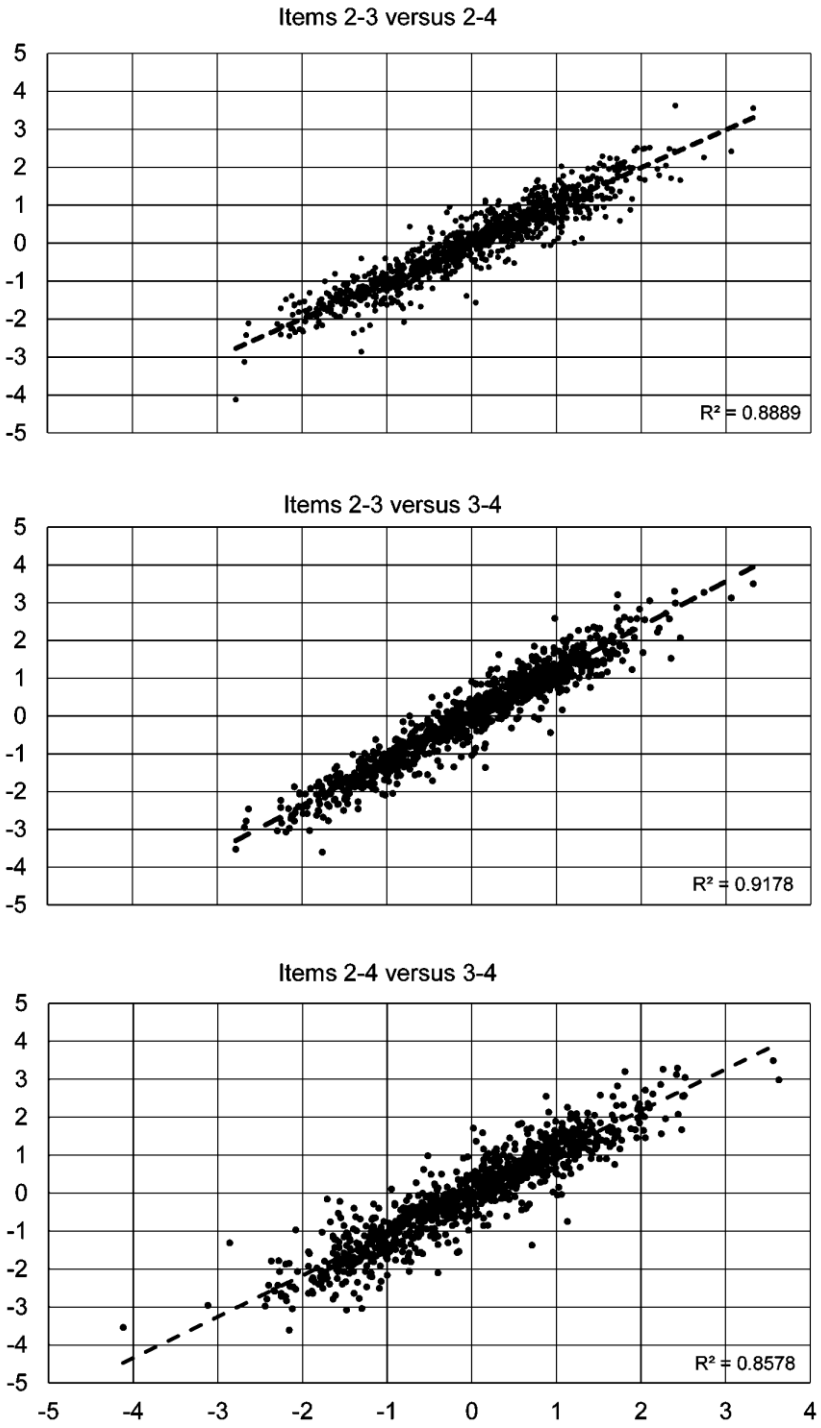

Figure 3. Cross-plots of item sub-groups.

Measurement of reader ability and book difficulty.

Overall, all three items performed within the expectations of the Rasch model: the instrument exceeded the minimum requirement for variance explained by the Rasch trait; there was no evidence of sub-dimensions of items in the cross-plotting analysis; all three items showed

excellent data-model fit; the reliability of separation of book measures was sufficient to separate easy books from difficult books with a high level of confidence. The evidence therefore supports the claim that the survey instrument allows for valid inferences about students' perceptions of book difficulty.

However, Table 2 shows that, although the facets of *Items* and *Time* showed excellent data-model fit, the measurement of *Readers* and *Books* was much noisier. This required investigation because the research questions of this study were primarily concerned with the measurement of book difficulty. Measurement reports for *Persons*, *Books*, and *Time* were also provided by Facets, but cannot be reprinted here due to space considerations. Readers who wish to examine the complete measurement reports will be provided the complete Facets output upon request, or the complete dataset and Facets control file for anyone who wishes to conduct their own Facets analysis of the data.

The source of the misfit of books and persons is suggested by the count of responses. As discussed earlier, in order to retain as much data as possible, data were included from persons with five reviews and books with three, respectively 15 responses and nine responses. The mean count of responses by persons was 92.80, indicating that the average Reader read about 31 books, while the mean for books was 61.00 responses, indicating about 20 reviews, with respective standard deviations of 65.00 and 61.80. This skewed distribution of data makes the mean-square fit statistics susceptible to a few outlying responses, as shown in Figure 4, which maps the outfit mean-square values for books against the number of survey responses. Most books with more than 100 responses have mean-square values within the acceptable range of 0.50 to 1.50; but books with fewer than 50 responses have a much larger range of values, many having values below 0.50, indicating overfit to the model. Counterintuitively, overfitting responses, indicating surprisingly predictable data, cause other responses to appear relatively less predictable and thus to misfit because the mean-square statistic is constrained to an average value of about 1.00 (see Holster & Lake, 2016 for an illustration of this phenomenon). In this case, the excessive range of fit statistics arises from a very sparse dataset. If more responses were available, many of the highly overfitting books would appear less predictable and this would reduce the relative misfit of other books. Definitive conclusions about data-model fit would require a much larger dataset, but Figure 4 suggests that books with more than 100 responses show adequate fit for the low-stakes purposes of this research. It was therefore judged preferable to retain as many reviews as possible despite the noisy data, rather than continue trimming data in order to obtain better fit statistics, avoiding Davidson's concern regarding "statistical determinism" (2000, p. 615).

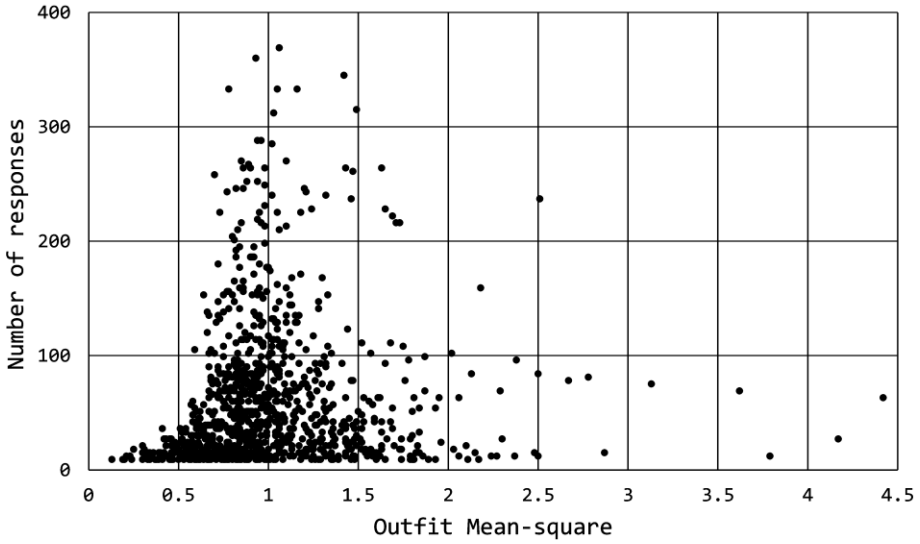

Figure 4. Outfit mean-square values for books by the number of responses

The answer to RQ1 is therefore that students' self-report ratings provided valid measurement of book difficulty, meaning that a higher logit measure of difficulty can be confidently interpreted to indicate a book that students will report as more difficult to read. As validity is not generalizable beyond the context of a particular study, this needs to be qualified with a caution that achieving adequate data-model fit for high-stakes decisions would require a much larger number of responses than was available for this study.

Research question 2: Predictors of book difficulty

Variance explained by text features. RQ2 concerned the relationship between word count, vocabulary level, and sentence length to the perceived difficulty of books. Up to 10 popular fiction books from different levels of 14 graded reader series were purchased and scanned to provide text samples for analysis, comprising 309 books totaling 721,752 running words. Microsoft Word was used to obtain statistics such as word counts, sentence counts, mean sentence length, mean word length, Flesch reading ease, and Flesch-Kinkaid grade level. The free Lexile analyzer tool (Lexile, 2016) was used to provide Lexile estimates from the first 1000 words of each text, along with sentence length and log word frequency data.

Figure 5 shows the percentage of variance explained by the different estimations of text difficulty, calculated from the squared raw correlation between variables. Contrary to expectations, the YL provided the best estimates of perceived book difficulty, with 68% variance explained, followed by the raw word count with 61%. The YL largely represented word count, so it is unsurprising that these two measures show high agreement. However, semantic difficulty explained an unexpectedly low amount of variance, shown by the LWF, Kyoto Scale, and CPW measures, with 6%, 13%, and 2% variance explained, respectively. Also unexpected was that Lexile measures, which combine both word frequency (semantic difficulty) and LWPS (syntactic difficulty), accounted for less variance than WPS alone, respectively 34% versus 40%. One potential explanation for this is that the Lexile framework was developed for native-speaker readers so the Lexile corpus may not be representative of the English texts that students

encountered in their high-school study. Additionally, concerns arise about the effectiveness of the semantic simplification used by publishers. Wan-a-rom (2008) and Claridge (2012) reported serious discrepancies between books and publishers' lists, and between different publishers, so it is possible that semantic difficulty is an important consideration but that the simplification algorithms used by publishers were ineffective.

Another consideration is demonstrated by Figure 6, showing publishers' headword levels compared with book length. The easiest books were very short, with word counts of fewer than 500 running words, but students also read many longer books, with running word totals of 5000 words or more, indicating an order of magnitude difference in word counts. There was much less variance in the semantic level, shown by the headword level, however, with relatively few books having headword levels greater than 500, an extremely low level. Thus, the correlation between book difficulty and semantic level is inevitably low because students did not read books with higher levels of semantic difficulty. This indicated a potential limitation of this study, so a stronger relationship between semantic level and book difficulty might be observed in future studies with higher ability students. This limitation, however, does not affect the finding of this report that publishers' headword levels were ineffective predictors of students' perceptions of book difficulty.

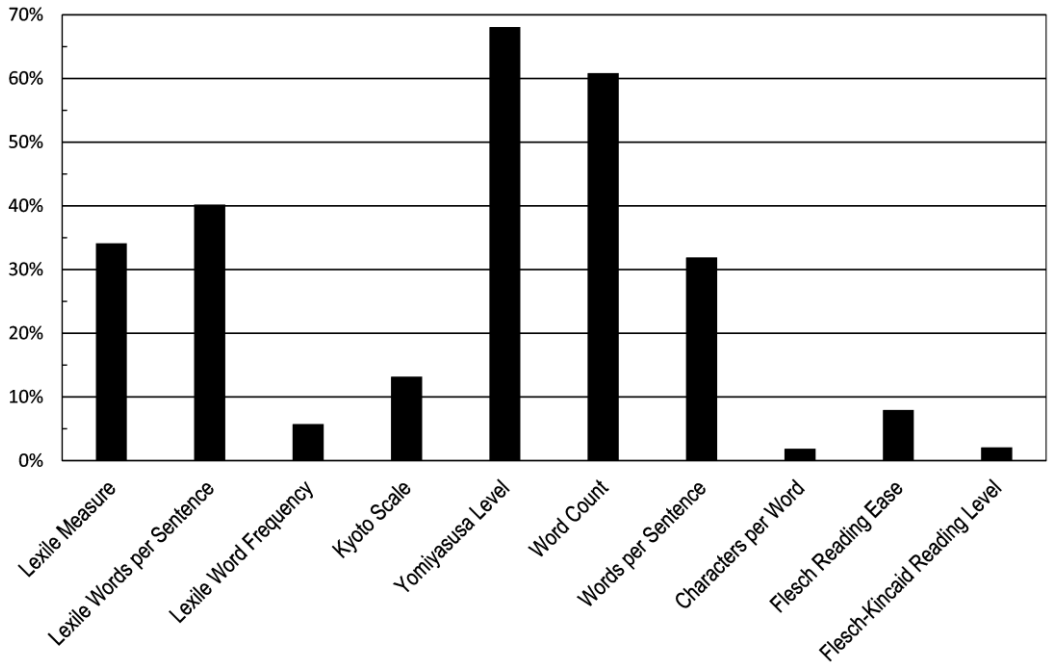

Figure 5. Variance explained by different estimates of text difficulty. Lexile results are grouped on the left, followed by Kyoto Scale and YL, with Microsoft Word derived results grouped on the right.

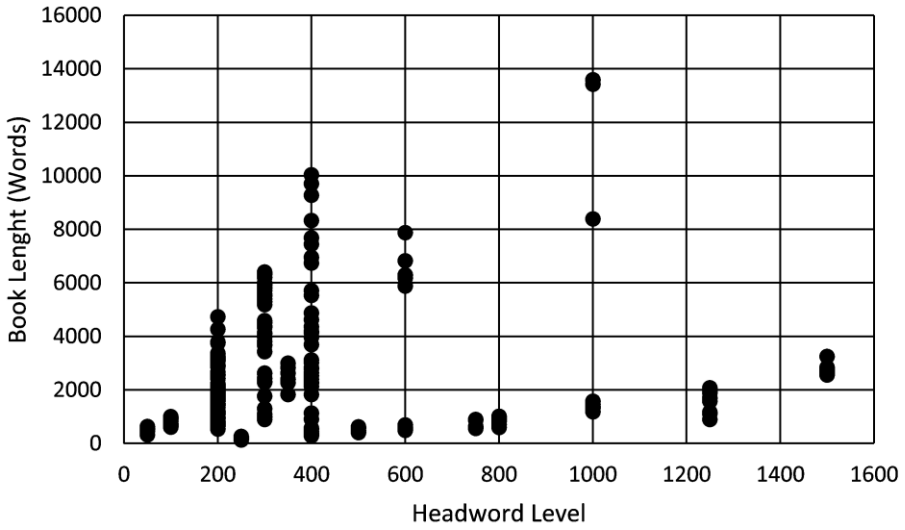

Figure 6. Headword level versus book length.

Reading speed versus reading time. A final issue of concern is the possible ambiguity of Item 3, *How quickly did you read this book?*, which raises the concern that survey responses were manifestations of book length rather than text difficulty and whether Item 3 measured a different construct to the other two items.

As discussed in detail by Holster and Lake (2016), the crucial assumption of the Rasch model is of parallel ICCs, meaning that the probabilities of success for items of different difficulty follow parallel trajectories. The Rasch model also assumes parallel person characteristic curves (PCCs), meaning that the probabilities of success of persons of different ability follow parallel trajectories. Item 3 was slightly misfitting and slightly less consistent than the other two items, but not to a degree that raised concern. The intended purpose of the survey instrument was to measure students' perception of book difficulty, hence the use of self-report measures. Item 4, *Was this book difficult?*, addressed this holistically and slightly overfitted Rasch model expectations, indicating slight dependency between this item and overall results. The investigation of RQ1 supported the unidimensionality of the survey instrument, so the performance of Item 4 is very strong evidence that the survey did indeed measure students' perception of book difficulty, not of how long it took to read the books. There is no reason to believe that students mistook Items 2 and 4 as referring to book length, so the evidence strongly contradicts the hypothesis that the survey was unintentionally measuring reading time. Therefore, the RQ2 finding that book length was the strongest predictor of the perceived difficulty of books did not arise through ambiguity concerning Item 3, but rather that students perceived longer books to be more difficult.

Thus, a tentative answer to RQ2 is possible: students' perception of graded reader difficulty was largely predicted by the word count of the text, with mean sentence length the only textual feature that provided useful predictions of difficulty, consistent with Shiotsu and Weir's (2007) finding that syntactic knowledge was a better predictor of reading comprehension than vocabulary knowledge. For the purpose of advising students on choosing appropriate books to read, the YL proved to be the most useful by a large margin. However, syntactic difficulty, indicated by the WPS indices reported by the Lexile analyzer and by Microsoft Word also

proved to have some predictive value. Given that all the books analyzed in this study were simplified to make them more accessible to novice L2 readers, this is consistent with publishers having constrained syntactic difficulty as well as semantic difficulty, but with inconsistent results. Because this study mixed the data from multiple publishers and book series, it is not clear whether the modest predictive value of syntactic difficulty is because some publishers were less effective at simplifying texts than others or whether there are inherent limitations to the algorithms used by all the publishers. Because of the budget constraints on this research, only a few graded readers could be purchased from each series, so detailed investigation of the relationship between syntactic difficulty and book difficulty could not be pursued further in the current study.

Discussion

Conclusion

This study used MFRM analysis of student ratings of graded reader difficulty to investigate two research questions: whether students' self-report ratings provided valid measurement of perceived book difficulty; and whether word count, vocabulary level, and sentence length provided useful predictions of perceived book difficulty. The use of MFRM allowed four facets to be measured: reading ability of persons, difficulty of books, learning gains by time, and the difficulty of rubric items. Measurement reliability of all four facets exceeded .90, corresponding to separation reliability exceeding 3.0, sufficient to confidently conclude that the easiest books were less difficult than the most difficult books. An assumption of Rasch measurement is adequate data-model fit, and the analysis of fit statistics indicated that the survey questions showed excellent data-model fit, although the extremely sparse data associated with less popular books provided noisy measurement of book difficulty. However, books that had a large number of responses showed acceptable data-model fit so the resulting measures of book difficulty were judged to provide acceptable psychometric properties for the low-stakes purposes of this research. Investigation of the second research question returned the unexpected result that semantic level, indicated by vocabulary frequency, was only weakly predictive of the perceived difficulty of books, and thus that publishers' headword levels and the Kyoto Scale did not provide useful predictions of students' perceptions of book difficulty. The YL, based largely on the word count of books, proved to be highly predictive of perceived book difficulty, so it is recommended that teachers and students choose books based on the YL rather than the Kyoto Scale or publishers' headword levels. Further surprising results were that Lexile levels, integrating both semantic difficulty and syntactic difficulty, were only modestly predictive of perceived book difficulty and that syntactic difficulty, measured by the LWPS measure, proved to be a better predictor of difficulty than the Lexile level itself. These findings suggest that the current practice of levelling graded readers by headword levels derived from native-speaker corpora is largely ineffective.

Limitations and future directions

The major limitations of this study arose from the sampling of books and persons. Although data were collected on more than 1000 books, many books received very few reviews, resulting in a very noisy dataset. This was partially addressed by restricting the textual analysis to more

popular books, but the quality of measurement would have been improved if more reviews were available for the less popular books. With the completion of the funded research project, it was considered preferable to publish results rather than delay publication in the hope that further data might be forthcoming. However, even if an opportunity to continue with data gathering does arise in the future, sampling bias against the unpopular books may continue to be a problem. The current study used students from two Japanese women's universities, where students were enthusiastic about, or at least accepting of, foreign language study. However, students at other Japanese universities may have very different motivational orientations, as reflected by Robb's (2002) observation that students are motivated solely by the instrumental desire to satisfy course requirements, a view supported by the low level of cooperation reported by Robb and Kano (2013). In such cases, resorting to systems such as MReader to compel students to engage in IR is understandable. However, as discussed in the results section, unidimensionality is an assumption of both Rasch analysis and CTT, and this applies to the sampling of persons as well as to items. The sample-independent measurement provided by the Rasch model (Engelhard, 2013) assumes that persons are drawn from the same population, an assumption that is potentially violated by mixing groups of students with different motivational orientations.

Results from one Japanese university cannot be assumed to generalize to other Japanese universities, let alone to non-Japanese contexts. This raises a major difficulty for the future of the current research because books that were less popular among the students in this study may prove to be more popular among other populations of students, leading to person-book and person-item interactions and differential functioning of the measurement instrument due to multi-dimensionality arising from mixing students with different motivational orientations. If a substantive level of differential functioning was found to exist, it would have profound implications for the levelling of graded readers because the perceived difficulty of books would not be invariant for different populations of students; yet, it is the perceived difficulty of books that students use in deciding whether or not a particular book is appropriate. This would make it fundamentally impossible for publishers to level graded readers in any universal manner because a levelling system based on research on one population of students would not suit students from a different population. Given that MReader and the YL were both developed within Japan, differential functioning would also mean that any ranking of book difficulty derived from those projects could not be assumed to generalize to other populations of students. Future research should therefore prioritize the investigation of differential functioning among students drawn from different populations in order to determine whether generalizable rankings of book difficulty are possible.

Acknowledgements

This research was funded by a grant-in-aid for scientific research from the Japan Society for the Promotion of Science and the Japanese Ministry of Education, Culture, Sports, Science and Technology, *kakenhi* grant 25370643.

References

Bachman, L. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, *17*, 1–42. doi: 10.1177/026553220001700101

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

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment* (New ed.). New York: McGraw-Hill.

Carver, R. P. (1993). Merging the simple view of reading with rauding theory. *Journal of Literacy Research*, *25*, 439–455. doi: 10.1080/10862969309547829

Claridge, G. (2012). Graded readers: How the publishers make the grade. *Reading in a Foreign Language*, *24*, 106–119.

Davidson, F. (2000). The language tester's statistical toolbox. *System*, *28*, 605–617. doi: 10.1016/S0346-251X(00)00041-5

Day, R. R. (2015). Extending extensive reading. *Reading in a Foreign Language*, *27*, 294–301.

Day, R. R., & Bamford, J. (2002). Top ten principles for teaching extensive reading. *Reading in a Foreign Language*, *14*, 136–141.

Engelhard, G. (2009). Using item response theory and model-data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement*, *69*, 585–602. doi: 10.1177/0013164408323240

Engelhard, G. (2013). *Invariant measurement*. New York: Routledge.

ETS. (2008). The TOEFL® Test - Test of English as a Foreign Language™. Retrieved 28 March, 2008, from http://tinyurl.com/zocgc

Furukawa, A. (2014a). Yomiyasusa levels, reading levels for Japanese students. Retrieved 10 March, 2016, from http://www.seg.co.jp/sss/YL/

Furukawa, A. (2014b). YL tables. Retrieved March 10, 2016, from http://www.seg.co.jp/sss/YL/YL_tables.html

Holster, T. A., & Lake, J. (2016). Guessing and the Rasch model. *Language Assessment Quarterly*, *13*, 124–141. doi: 10.1080/15434303.2016.1160096

Lexile. (2016). The Lexile Framework for reading. Retrieved March 10, 2016, from https://lexile.com

Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). Chicago: MESA Press.

Linacre, J. M. (2010a). *Facets* (Version 3.67.0). Retrieved from http://www.winsteps.com/facets.htm

Linacre, J. M. (2010b). Reliability and separation of measures. Retrieved 24 September, 2010, from http://www.winsteps.com/winman/index.htm?reliability.htm

Linacre, J. M. (2014). Misfit diagnosis: infit outfit mean-square standardized. Retrieved 22 August, 2014, from http://www.winsteps.com/winman/misfitdiagnosis.htm

Linacre, J. M. (2016a). Dimensionality investigation - an example. Retrieved 25 October, 2016, from http://www.winsteps.com/winman/multidimensionality.htm

Linacre, J. M. (2016b). Partial credit model. Retrieved March 10, 2016, from http://www.winsteps.com/winman/partialcreditmodel.htm

McNamara, T. F. (1996). *Measuring second language performance*. Harlow: Pearson Education.

MReader. (2016a). The Kyoto Scale. Retrieved 10 March, 2016, from http://mreader.org/mreaderadmin/s/html/Kyoto_Scale.html

MReader. (2016b). Extensive Reading: The fun way to learn English! Retrieved 10 March, 2016,

from http://mreader.org/
Palmer, H. E. (1917). *The scientific study and teaching of languages*. Edinburgh, UK: The Riverside Press.
Pellowe, W. R. (2016). MOARS feature overview. Retrieved 10 January, 2016, from http://moars.com/moars-features/4-moars-feature-overview.html
Prowse, P. (2002). Top ten principles for teaching extensive reading: A response. *Reading in a Foreign Language*, *14*, 142–145.
Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*, 207–230. doi: 10.2307/1164671
Robb, T. (2002). Extensive reading in an Asian context - an alternative view. *Reading in a Foreign Language*, *14*, 146–147.
Robb, T., Campbell, A., & Bateson, G. (2014). *MoodleReader/MReader progress and user sharing*. Paper presented at the JALT2014, Tsukuba, Japan.
Robb, T., & Kano, M. (2013). Effective extensive reading outside the classroom: A large-scale experiment. *Reading in a Foreign Language*, *25*, 234–247.
Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, *24*, 99–128. doi: 10.1177/0265532207071513
Stenner, A. J. (1999). *Instructional uses of the Lexile framework*. Durham, NC: MetaMetrics Inc.
Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2007). *The Lexile framework for reading technical report*. Durham, NC: MetaMetrics Inc.
Wan-a-rom, U. (2008). Comparing the vocabulary of different graded-reading schemes. *Reading in a Foreign Language*, *20*, 43–69.
Wang, W.-C., & Chen, H.-C. (2004). The standardized mean difference within the framework of item response theory. *Educational and Psychological Measurement*, *64*, 201–223. doi: 10.1177/0013164403261049
Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: 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*, 168–181.
Yamashita, J., & Shiotsu, T. (2017). Comprehension and knowledge components that predict L2 reading: A latent-trait approach. *Applied Linguistics*, *38*, 43–67. doi: 10.1093/applin/amu079

Appendix A

Survey items

Item	Dimension	Question	Response	Code
1	Enjoyment	Did you enjoy this book?	A Lot	3
			A Little	2
			Not Much	1
			Not at all	0
2	Difficulty	Did you need a dictionary?	Very often	0
			Sometimes	1
			Not often	2
			Never	3

3	Difficulty	How quickly did you read this book?	Very quickly	3
			Quickly	2
			A little slowly	1
			Very slowly	0
4	Difficulty	Was this book difficult?	Very difficult	0
			Quite difficult	1
			A little difficult	2
			Not difficult	3
5	Enjoyment	How much of the book did you read?	100%	3
			More than 50%	2
			Less than 50%	1
			Less than 10%	0
6	Enjoyment	Your general opinion of this book.	Very good	3
			Average	2
			Below average	1
			Very poor	0

Appendix B

Logit measures versus raw scores

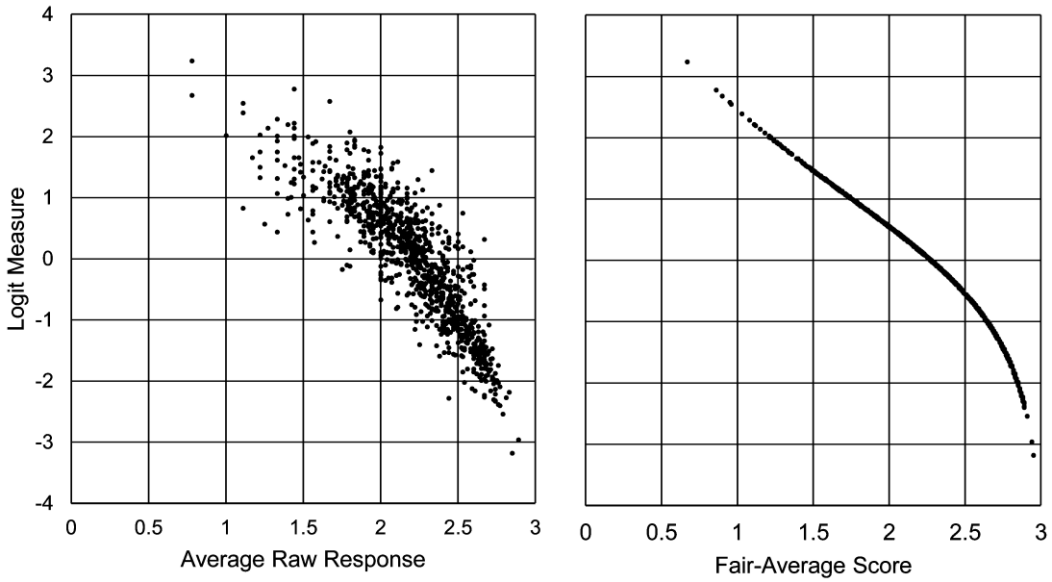

Figure B1. Logit measures of book difficulty versus average survey responses. The left-hand panel shows the average raw response. The right-hand panel shows the fair-average score.

About the Authors

Trevor A. Holster is an English instructor at Fukuoka University. He has a Master of Applied Linguistics degree from the University of Southern Queensland. His research interests include reading instruction, extensive reading, vocabulary assessment, and formative assessment. E-mail: holster@fukuoka-u.ac.jp

J. W. Lake, PhD, is a lecturer at Fukuoka Jo Gakuin University. He has taught at universities in Japan for many years. His research interests include language learning motivation, L2 reading and vocabulary development, language assessment, and positive psychology. E-mail: jlake@fukujo.ac.jp

William R. Pellowe is an associate professor at Kindai University's Fukuoka campus. He earned his MA in TEFL with distinction from the University of Birmingham. His research interests include vocabulary, CALL and student response systems. E-mail: pellowe@fuk.kindai.ac.jp