生駒経済論叢　第9巻第2号　2011年11月

A Systematic Approach to Validating a Likert-scale Instrument Used to Measure Change across Multiple Occasions with Multiple Groups in an EFL setting

Lance Burrows

Abstract This paper focuses on heightening researchers' awareness of the problems that could occur if instruments are not properly validated, which could ultimately end in obscured results. It will demonstrate a systematic approach to validating a Likert-scale instrument used to measure change across multiple occasions with multiple groups in an EFL context. It will detail many of the possible pitfalls that researchers may encounter when processing their data and solutions to overcoming these problems. This will be facilitated by utilizing the Rasch Rating Scale Model to expel confounding effects from the data and, thereby, present a more accurate representation of the actual changes that occur in a study.

Key words instrument validation, Rasch model, Likert-scale instrumentation
September 21, 2011 accepted

要旨　このことをふまえ，本論文は上記のような長期的研究を行う研究者にとって測定手段を有効化することがいかに重要であるかを実証する。具体的には，EFLで複数の被験者グループを対象にした長期的研究において用いるリッカート尺度方式による測定手段の有効化の方法を，ラッシュ評定尺度モデルを使って例示する。またデータ分析の際多くの研究者が直面するであろう問題点やその解決方法も紹介する。

キーワード　測定手段の有効化，ラッシュモデル，リッカート尺度方式の測定手段
原稿受理日　2011年9月21日

Introduction

Generally-speaking, cross-sectional studies seem to outnumber longitudinal studies in most academic journals. This can be said for journals in the English as a Foreign Language field, as well. Certainly, cross-sectional studies offer insight into constructs and the relationships that may exist between them, but they often are limited in providing practical guidance regarding how knowledge can be used to improve pedagogy in foreign language classrooms over an extended period of time, such as a school semester or academic year. To this end, longitudinal studies that chart changes over time are absolutely necessary.

This may be of particular importance when researching topics in EFL education. Language education, like all kinds of education, requires a process of learning and often a considerable amount of time. These types of longitudinal studies, that take into account this emphasis on change over time, are essential in facilitating research that will further the knowledge base of the EFL field.

Creating these types of Likert-scale instruments to be used on multiple occasions and with multiple groups can present its own set of difficulties. Literature reviews must be completed, a theoretical framework for the instrument must be established, the instrument must be developed, pilot studies must be conducted, and improvements to the instrument must be made before even attempting to use the instrument in a formal study. And unfortunately, even after hurdling all of these obstacles the researcher may still not be guaranteed accurate results.

Judging change in Likert-scale measures over time presents a multitude of possible obstacles, as well. If researchers are not careful, they can be led down the wrong path by confounds that may distort actual changes, making it unclear whether observed changes were caused by the intervention or some other uncontrolled design effect. Some of the possible problems that could occur are problems that influence evaluative study control groups (e.g., treatment diffusion) or statistical procedures (e.g., unreliability of measurement instruments).

To combat these problems, a systematic approach to validating the instrument must be executed. When evaluating changes in people over time, items on the Likert-scale instrument and the rating scale must be stabilized. This can be accomplished by following a series of steps set out by Wolfe and Chiu (1999), with the aid of the Rasch Rating Scale Model.

Problem

Data used to demonstrate the above systematic approach was taken from a questionnaire that was designed to assess the beliefs of 322 Japanese university students in regards to extensive reading practices and to determine to what degree they believed practicing extensive reading could help improve their reading comprehension. To this end, only one version of the questionnaire was developed and the same questionnaire items were used over three rounds of testing.

The questionnaire was given to the participants three times over the course of the study which lasted one academic year. The first was given at the beginning of the school year, the second was given just before the end of the first semester, and the third was given at the end of the second semester. The study lasted for 40 weeks.

The questionnaire was given to four groups: a control group, a group that was practicing extensive reading, a group that was studying reading strategies, and a group that was both practicing extensive reading and studying reading strategies. The 17 items on the questionnaire (See Appendices A for a completed version of the questionnaire) were adapted from Day and Bamford's (2002) Guidelines for Extensive Reading.

The questionnaire required participants to provide judgments based on a 6-point Likert scale, ranging from 1 (*Strongly disagree*), 2 (*Disagree*), 3 (*Slightly disagree*), 4 (*Slightly agree*), 5 (*Agree*), and 6 (*Strongly agree*)(see Appendix B for the full questionnaire). The various items are phrased in a way to elicit responses based on how useful participants consider extensive reading to improvements in their overall reading comprehension. One example from the questionnaire, item 7, highlights the discouraged use of dictionaries in the extensive reading method, "In order to improve my reading comprehension, it is better not to stop to check a dictionary if I find an unknown word while I am reading."

Method

The systematic approach to validating this instrument will take the following four steps（Wolfe and Chiu, 1999):

1）evaluate the rating scale and item invariance with separate analyses data from each occasion,

2）create common category threshold calibrations,

3）create corrected item calibrations for occasion 1 by anchoring the rating scale, and

4）create a corrected item calibration for the remaining occasions by anchoring the rating scale（Wolfe and Chiu, 1999, p. 72).

By following these procedures, the data will be rid of confounds that could mislead the researcher.

The above four steps are conducted by utilizing the Rasch Rating Scale Model. All of the above procedures use the Rasch analysis and, therefore, a brief explanation of the model is warranted.

Rasch Analysis

In the above study, raw scores were obtained, however, these scores are fundamentally difficult to compare across groups and time. Rasch analysis was utilized to assess validity and reliability of surveys and tests in this study, as well as, to create true interval-scale measures from the raw scores obtained.

Rasch analysis is performed by attempting to fit a data set to an *a priori* model. Gross misfitting between the *a priori* model and real world data is considered to signify poor definition of constructs, statistical bias, and/or an error in measurement. This concept of *fit* is one of the keys to Rasch analysis.

Fit analyis

The software package used for this study, Winsteps (Linacre, 2009) analyzes data and outputs various statistics for determining fit for both items and participants in the study. Participants who misfit the model might not have answered the items on the test truthfully or seriously. It might also indicate that factors are creating a sub-population within the group of participants in the study with different measurement attributes. Misfitting items can indicate a bias in the items, a deviation from unidimensionality of the construct, and/ or a redundancy in the items.

There are two types of misfit, outfit and infit, commonly represented as Outfit mean square (Outfit MNSQ) and Infit mean square (Infit MNSQ), respectively. Outfit represents the degree of unexpected responses within a group of test takers or set of items. In Rasch analysis, 1.00 is the expected value by which all values can be compared and levels of outfit and infit can be determined. For item Outfit MNSQ, a high value would indicate an item that is being answered often by low-level test takers but missed by high-level test takers.

Debate continues to rage over the appropriate values of fit. A number of researchers offer guidelines to represent fit limits. McNamara (1996) claims that MNSQ values between .77 and 1.30 are acceptable, while Wright, Linacre, Gustafson, and Martin-Lof(1994)claim that depending on the context of the test, acceptable values can fluctuate. For example, on a high-stakes test, the appropriate range would be 0.8–1.2, but for a lower stakes test, the acceptable range might be 0.4–1.2. For the purposes of this study, a low-stakes test, the more lenient guidelines, 0.7–1.3 were utilized for the dichotomous tests and 0.6–1.4 were used for the Likert-scale questionnaires.

Reliability Analyis

In addition to showing how well items fit a model, Rasch analysis also provides statistics signifying item reliability. Simply stated, reliability refers to how accurately a test can be replicated to provide similar results to the current test. For example, item reliability predicts how well the results for a particular test can be reproduced with a different sample of test takers but the same set of items. Person reliability refers to the degree to which the same sample of test takers could reproduce their obtained results from a unique set of items measuring the same construct. In addition, standard error statistics for items can be obtained through Rasch analysis as well. These error statistics can help researchers pinpoint problematic individual items or persons that might be distorting the reliability data.

Principal Components Analysis of Rasch Residuals

In addition to the fit statistics, Rasch analysis includes an evaluation of Rasch residuals. This principal components analysis(PCA)pertains to a fit analysis that searches for systematic variance outside of that determined by item difficulty. In a well-functioning model, the residuals consist of random noise only. It is when a significant grouping of items can be explained by these residuals that a problem might be present. These residuals confirm or falsify the hypothesis that the construct under investigation is unidimensional or not. It is a sensitive test of dimensionality that indicates to what degree additional dimensions to the construct distort the measurement.

First, where a Likert scale is used, the functioning of the scale is checked using the criteria suggested by Linacre (2007). If necessary, the scale is adjusted until it meets those criteria. Second, item functioning is reported with a particular emphasis on how well the items fit the Rasch model. The fit criteria used in this study are 0.6–1.4 for the Likert-scale questionnaires (Wright, Linacre, Gustafson, & Martin-Lof, 1994). Third, the dimensionality of the items hypothesized to measure the same construct is investigated with a Rasch PCA of item residuals analysis.

Conducting the Validation Process

Rating Scale Calibrations

In an attempt to create meaningful measures using Likert-scale questionnaire items across multiple occasions, several guidelines must be followed (Linacre, 1999).

1. There must be a minimum of 10 observations in each category.
2. The rating scale distribution should form a normal distribution and should, therefore, be peaked.
3. The average category measures should increase with the rating scale measures.
4. The outfit mean square statistics should be between .8 and 1.4.
5. The category thresholds should increase along with the rating scale categories.
6. The category threshold calibrations should be between 1.4 and 5.0 logits apart.

These criteria were applied to the rating scale of the perceived utility of extensive reading questionnaire. However, in regards to criterion 6 above, the minimum acceptable separation value, 1.4, refers to a scale with only 3 categories. As a 6-category scale was used in the current study, a more complete set of acceptable separation values was necessary.

Values in Table X for 3-, 4-, and 5-point scales can be found in Wolfe and Smith (2007, p. 210). However, because the value for a 6-point scale was not included, it was extrapolated from the sequence using the following equation:

$$\text{minimum separation } (j) = \text{minimum separation } (j-1) - [\mathit{ln}\ (j-1) - \mathit{ln}\ (j-2)].$$

Therefore, the minimum separation for six categories is $(j=6)$ is $.81-[1.61-1.39]$, which simplifies to $.81-.22$, which equals .59. If we analyze the minimum separation values more closely, we find a steadily decreasing series of values (i.e., differences are .29, .22, and .18). These more accurate values for the number of categories were then used to investigate the minimum separation of each scale in the perceived utility of extensive reading questionnaire.

Table 1.　Category Separation Series

Category (*j*)	*ln* (*j*)	*ln* (*j*) −*ln* (*j*−1)	Minimum separation (logits)	Minimum separation (CHIPS)
3	1.10		1.40	6.37
4	1.39	0.29	1.10	5.00
5	1.61	0.22	0.81	3.69
6	1.79	0.18	0.59	2.68

Note. The minimum separation (logits) and minimum separation (CHIPS) for categories 3, 4, and 5 are from Wolfe and Smith (2007, p. 210).

The category structure of the six original categories for the perceived utility of extensive reading questionnaire (See Table 2 for the uncorrected rating scale calibrations, standard errors, fit statistics, and standardized differences for the perceived utility of extensive reading questionnaire [Times 1, 2, and 3]) was examined using the above criteria(Linacre, 1999; Wolfe & Smith, 2007). On all three occasions, Category 1 had the lowest observed count, but was still well above the minimum of ten, so the first criterion was met. In addition, the counts increase for each category and peak near the middle at category 4, therefore meeting the second and third criteria. According to the above criteria, the outfit mean square statistics should be between .8 and 1.4. This is clearly the case for all the outfit ratings, except for two, Category 1 on Times 2 and 3. Due to this violation, criterion 4 was not met. The category thresholds increase along with the rating scale categories, so criterion 5 was met. In regards to the last criterion, all the category threshold calibrations should be at least. 59 and no more than 5.0. All of the thresholds meet this criterion except for the threshold between categories 1 and 2 on Times 1 and 2. In these instances, there is a separation value of 0.49 and 0.50, respectively.

In addition to the problems related to the above criteria, inspection of the standardized differences also revealed problems with the invariance of the categories across time (See Table 2 for the uncorrected rating scale calibrations, standard errors, fit statistics, and standardized differences for the perceived utility of extensive reading questionnaire[Times 1, 2, and 3]). Standardized differences values less than 2.00 indicate invariance across time and represent a stable category. However, all of the categories have at least one standardized value that is over

2.00; therefore, none of the categories behaved in a stable fashion across the three rounds of testing.

In regards to the misfitting category 1, the unacceptable threshold levels between categories 1 and 2, and the unfavorable standardized differences, Wolfe and Chiu (1999) recommend a series of steps that free the data of confounding factors. To combat these problems, scaling methods are used to place measures from different administrations of a measurement instrument onto a common underlying scale. Following the recommendations outlined by Wolfe and Chiu (1999), the data for all three occasions, using the 6-point scale, were stacked in a single data set, maintaining the item identity across the three occasions but regarding each person as unique at each time period. Therefore, for this data set, there were essentially 966 participants, instead of the actual 322 (See Table 3 for a summary of the stacked data of the category structure for the 6-point rating scale for the perceived utility of extensive reading questionnaire [Times 1, 2, and 3]).

Table 2. Uncorrected Rating Scale Calibrations, Standard Errors, Fit Statistics, and Standardized Differences for the Perceived Utility of Extensive Reading Questionnaire

Category	Observed count (%)			Average Measure			Category Threshold (S.E.)			Mean Square Outfit[a]			Standardized Differences[b]	
	T 1	T 2	T 3	T 1	T 2	T 3	T 1	T 2	T 3	T 1	T 2	T 3	T 1−2	T 1−3
1	209(4 %)	100(2 %)	87(2 %)	−2.78	−3.01	−3.21	—(—)	—(—)	— (—)	1.33	2.07	1.69	—	—
2	551(10%)	332(6 %)	357(7 %)	−1.35	−1.55	−1.67	−1.36(.07)	−1.59(.11)	−1.84(.11)	1.02	1.19	1.16	1.77	3.68
3	1,185(22%)	985(18%)	1,109(20%)	−0.42	−0.53	−0.56	−0.87(.04)	−1.09(.06)	−1.15(.05)	0.81	0.87	0.86	3.05	4.38
4	1,496(27%)	1,667(30%)	1,823(33%)	0.39	0.44	0.52	0.00(.03)	−0.10(.04)	−0.07(.03)	0.94	0.82	0.86	2.00	1.67
5	1,263(23%)	1,482(27%)	1,340(24%)	1.35	1.56	1.69	0.76(.03)	0.99(.03)	1.18(.03)	0.96	0.94	1.00	−5.48	−10.00
6	770(14%)	908(17%)	758(14%)	2.84	3.14	3.25	1.47(.04)	1.79(.04)	1.88(.04)	1.01	0.91	0.92	−5.61	−7.19

[a] Mean square outfit indices less than .8 and greater than 1.4 are considered to indicate rating scale misfit.
[b] Absolute standardized differences less than 2.00 are condsidered to indicate rating scale invariance across time.

By stacking the data in this way, an "average" underlying rating scale was determined and was used to describe the data over the three rounds of testing. These calibrations were then intended to be used as rating scale anchor values in the analyses of the data for all three occasions. However, despite this intervention, the unacceptably small threshold level, 0.35, between categories 1 and 2 persisted. Therefore, categories 1 (*Strongly disagree*) and 2 (*Disagree*) were collapsed and renamed (*Disagree*) (See Table X for a summary of the stacked data of the category structure for the collapsed 5-point rating scale for the perceived utility of extensive reading questionnaire [Times 1, 2, and 3]). It must be noted that once the collapse from a 6-point scale to a 5-point scale was conducted, the new minimum separation criterion for the thresholds was 0.81. After collapsing categories 1 and 2, some minor problems persisted. Although the separation value for categories 2 to 3 is only 0.74 and categories 4 to 5 is only 0.78, the values were considered close enough to the minimum. 0.81, to reject the notion of further collapsing the category scale.

Table 3. Summary of the Stacked Data of the Category Structure for the 6-point Rating Scale for the Perceived Utility of Extensive Reading Questionnaire for Times 1, 2, and 3

	Count (%)	Infit MNSQ	Outfit MNSQ	Structure Calibration	Category Measure
1 Strongly disagree	423 (3%)	1.40	1.44	None	(−2.85)
2 Disagree	1,194 (7%)	1.11	1.12	−1.40	−1.45
3 Slightly disagree	3,301 (20%)	0.89	0.88	−1.05	−0.50
4 Slightly agree	4,964 (30%)	0.95	0.90	−0.08	0.40
5 Agree	4,131 (25%)	0.97	0.96	0.89	1.46
6 Strongly agree	2,408 (15%)	0.92	0.94	1.63	(3.00)

By performing this collapse, the outfit values for category 1 fell to 1.19 for Time 1, 1.55 for Time 2, and 1.23 for Time 3. Although the value for Time 2 was still slightly misfitting, it was considered minimal enough to continue the analysis. It was only after following this series of steps that the category data was considered accurate and the calibrations were then used as rating scale anchor values in the analyses of the data for all three occasions.

Table 4. Stacked Data of the Category Structure for a Collapsed 5-point Rating Scale for the Perceived Utility of Extensive Reading Questionnaire for Times 1, 2, and 3

	Count (%)	Infit MNSQ	Outfit MNSQ	Structure Calibration	Category Measure
1 Disagree	1,617 (10%)	1.23	1.21	None	(−2.63)
2 Slightly disagree	3,301 (20%)	0.92	0.93	−1.26	−1.08
3 Slightly agree	4,964 (30%)	0.95	0.94	−0.52	−0.01
4 Agree	4,131 (25%)	0.99	1.01	0.50	1.08
5 Strongly agree	2,408 (15%)	0.89	0.94	1.28	(2.64)

Item Calibrations

The analysis of the 17 items on the perceived utility of extensive reading questionnaire administered at Times 1, 2, and 3 (See Table 5 for the uncorrected item calibrations, standard errors, fit statistics, and standardized differences [Times 1, 2, and 3]) revealed that only one item, ER11, misfit the Rasch model using the 0.6–1.4 outfit MNSQ criterion for Likert-scale questionnaires (Wright, Linacre, Gustafson, & Martin-Lof, 1994). At Time 1, although the outfit MNSQ was not optimal, 1.38, it was not misfitting. However, at Times 2 and 3, the outfit MNSQ scores clearly misfit with fit statistics of 1.61 and 1.70, respectively.

Because item ER11 misfit the Rasch model, a brief explanation of the rationale for originally including it in the questionnaire is warranted. Although the philosophy expressed in item ER11 does not adhere to the basic tenets of extensive reading outlined by Day and Bamford (2002), many participants strongly endorsed the item, "I can improve my reading comprehension by writing a short summary of what I have read, after I finish reading." Some educators who use extensive reading believe in assigning homework or additional assignments to improve student accountability. Therefore, item ER11 was originally included in the questionnaire. However, from an extensive reading *purist* standpoint, these types of extra activities detract from student motivation to read and therefore undermine the value of extensive reading. The premise behind this item that learners must do something in addition to simply reading (e.g., writing a short summary) to become better readers is antithetical to Day and Bamford's basic guidelines and the results were recoded during the analysis process. Even so, the item still misfit. In the end, due to its loose connection with the basic tenets of extensive reading and its misfit, item ER11 was deleted from the data set.

After deleting item ER11, further investigation of the items revealed little change in the measurement, standard error, and the outfit MNSQ over the three occasions (See Table X for the uncorrected item calibrations, standard errors, fit statistics, and standardized differences [Times 1, 2, and 3]). The low standard errors (.06–.07) for Times 1 and 2, and (0.06) for Time 3 indicated that the item difficulty estimates were reasonably precise. The item separation reliabilities for

Table 5. Uncorrected Item Calibrations, Standard Errors, Fit Statistics, and Standardized Differences for the Perceived Utility of Extensive Reading Questionnaire (Times 1, 2, and 3)

Item	Measurement			Standard Error			Standardized Mean Square Outfit[a]			Standardized Differences[b]	
	T 1	T 2	T 3	T 1	T 2	T 3	T 1	T 2	T 3	T 1 - 2	T 1 - 3
ER01	−0.41	−0.38	−0.28	0.06	0.06	0.06	1.20	0.88	0.85	−0.35	1.53
ER02	−0.09	−0.14	−0.26	0.06	0.06	0.06	1.02	0.78	0.66	0.59	2.00
ER03	−0.53	−0.53	−0.60	0.06	0.06	0.06	1.05	0.73	0.76	0.00	0.82
ER04	1.09	0.31	0.39	0.06	0.06	0.06	1.49	1.13	1.07	9.18	8.24
ER05	1.26	1.26	1.25	0.07	0.06	0.06	1.16	1.30	1.39	0.00	0.11
ER06	−1.35	−1.05	−0.85	0.07	0.07	0.06	1.04	1.08	0.89	−3.03	−5.43
ER07	0.30	0.16	0.27	0.06	0.06	0.06	1.21	0.98	1.04	1.65	0.35
ER08	0.30	0.05	0.07	0.06	0.06	0.06	0.98	0.87	0.72	2.94	2.71
ER09	−0.49	−0.54	−0.50	0.06	0.06	0.06	0.93	0.75	0.79	0.59	0.12
ER10	0.12	0.34	0.26	0.06	0.06	0.06	1.15	0.96	0.92	−2.59	−1.65
ER11		DELETED			DELETED			DELETED		DELETED	
ER12	−0.05	0.04	0.02	0.06	0.06	0.06	1.18	0.91	0.87	−0.12	−0.35
ER13	0.50	0.56	0.49	0.06	0.06	0.06	0.91	0.98	0.83	−0.71	0.12
ER14	0.88	0.73	0.64	0.06	0.06	0.06	0.93	0.88	0.89	1.76	2.82
ER15	−1.06	−0.67	−0.69	0.07	0.06	0.06	0.69	0.58	0.59	−4.24	−4.02
ER16	−0.65	−0.59	−0.54	0.06	0.06	0.06	0.93	0.74	0.77	−0.71	1.29
ER17	0.18	0.46	0.34	0.06	0.06	0.06	1.28	1.16	1.07	−3.29	−1.88
Mean	0.00	0.00	0.00	0.06	0.06	0.06	1.07	0.92	0.88		
(SD)	0.71	0.59	0.55	0.00	0.00	0.00	0.18	0.18	0.19		

Note. ER=perceived utility of extensive reading questionnaire item; [a] Absolute standardized mean square fit indices greater than 2.00 are considered large enough to indicate item misfit; [b] Absolute standardized differences less than 2.00 are considered small enough to indicate item invariance across time; SD=standard deviation; Item ER11 was deleted.

the three occasions were .99 for all three occasions, indicating that these distributions of item parameters contain enough variability to create distinct strata of item difficulties. Item separation indices for the three occasions were, 11.16, 9.48, and 9.08, respectively. The robustness of the items is underscored by the relatively low standard deviation of the standardized mean square outfit statistics that are well below the expected value of 1.00.

However, some problems did surface after analyzing the standardized differences of the items. The standardized differences of the items illustrate the extent to which calibrations for individual items exhibit statistically significant change across measurement occasions. The stability of two parameter estimates that are obtained on different occasions is evaluated by examining the standardized difference between the two estimates. Because Time 1 was considered to be the base from which participant beliefs could later be compared, invariance was evaluated by comparing Time 2 to Time 1, and Time 3 to Time 1. Some of the item parameters (38% between Time 1 and Time 2 and 38% from Time 1 to Time 3) display parameter instability over time. Overall, eight of the seventeen items, ER02, ER04, ER06, ER08, ER10, ER14, ER15, and ER17, reveal parameter instability on at least one of the time transitions.

Because eight of the items are not behaving in a stable fashion over the three occasions, it is imperative to anchor the items displaying appropriate levels of invariance over the three occasions from Time 1 to Times 2 and 3. With these items anchored, accurate comparisons of the person ability estimates can be better made over the three occasions. Therefore, the measurement values for Time 1 for ER01, ER03, ER05, ER07, ER09, ER12, ER13, and ER16, were anchored to analyze the data in Times 2, and 3.

One concern that arises when anchoring values for items is the possibly negative impact that anchoring can have on the data. To safeguard against this, random displacement values were calculated. Wright and Douglas (1976) claim that random displacement values of less than 0.5 logits are unlikely to have much impact in a test instrument. On all three occasions, the anchored values were less than 0.5 logits. On time 2, the values ranged from −0.14 to 0.09, and on time 3, the values ranged from −0.10 to 0.11. These small displacement values indicate that the anchored item difficulty estimates and the estimates that would have resulted from freely estimating the items differed only slightly.

Summary of Corrections to the Data Set

Due to a misfitting item (ER11 for Times 2 and 3), a misfitting category on the rating scale (category 1 for Times 2 and 3), unacceptable levels of invariance across time for both the items and the categories, the following corrections were made to the data set:

1. The data from all three Times was stacked and an"average"rating scale was determined. These values were then used as anchor values for the rating scale for all three times.
2. Categories 1 and 2 were collapsed.
3. Item ER11 was deleted.
4. Items demonstrating sufficient levels of invariance were identified and the values for those items on Time 1 were used as anchors when analyzing the data from Times 2 and 3.

The focus of this paper was to provide a step-by-step explanation of validating a Likert-scale instrument used on multiple occasions and with multiple groups in the EFL field. To this end, the data taken from a questionnaire given to 322 Japanese university students based on the perceived utility of extensive reading was utilized. This data was analyzed using the Rasch Rating Scale Model. Problems that occurred (i.e., instability in the rating scale and items) were highlighted for the reader. Solutions to these problems were laid out. By following the procedures laid out by Wolfe and Chiu (1999), the above data was rid of confounding factors. Once that the items and the rating scales were stabilized, the changes in person measures were believed to be more accurate and to reflect the actual changes that occurred.

APPENDIX A

GUIDELINES FOR EXTENSIVE READING (ENGLISH VERSION)

(adapted from Day & Bamford, 2002, pp. 137-140)

What is extensive reading ?

1. Learners read in and out of class as much as possible. (on avg. 100,000 to 200,000 words per year)
2. Learners choose their own books based on their own purpose and objectives from a large variety of topics and genres.
3. Learners are allowed to choose the books that they want to read and are able to stop in the middle of reading, if they find the book to be uninteresting.
4. The purpose of reading is usually related to pleasure, information, and general understanding (not just for learning English).
5. Reading, alone, is its own reward, so no reading comprehension questions or homework should be assigned after reading.
6. Learners should read at a level that they can understand the basic gist of the material without using a dictionary (unknown words should include less than 5% of the text).
7. Learners should be given the opportunity to read quietly when, where, and at whatever pace they want.
8. Reading should be relatively fast (at least 100 words per minute).
9. In order to improve the benefits of extensive reading for students, teachers should explain the basics of extensive reading to the students and monitor their reading.
10. The teacher should act as a role-model, reading in class along with the students.

APPENDIX B

PERCEIVED UTILITY OF EXTENSIVE READING QUESTIONNAIRE

Perceived Utility of Extensive Reading Questionnaire

What is Extensive Reading？

1. Reading books in English where less than 5% of the words in the book are unknown words
2. Reading extensively（reading 100,000–200,000 words per year）（textbooks in the six years of junior and senior high school have a total of 30,000 words in them, altogether）
3. Reading fast（more than 100 words per minute）
4. Reading for meaning
5. Choosing and reading the books that the reader is interested in

Example of Extensive Reading Level

The Amazon Rain Forest is the largest rain forest in the world. It is 10,000,000 years old and many different kinds of plants and animals live here. The forest is important for the world’ s weather and wildlife, but it is disappearing fast.

In regards to improving reading comprehension, how do you feel about the following survey items？ To what degree do you agree that the items below help to improve your reading comprehension？ Please answer by referring to the following scale（1～6）below.

1	2	3	4	5	6
Strongly disagree	**Disagree**	**Slightly disagree**	**Slightly agree**	**Agree**	**Strongly Agree**

I can improve my reading comprehension by . . .

1	reading many pages in easy books.	1 2 3 4 5 6
2	trying to read faster rather than slower.	1 2 3 4 5 6
3	reading books that I have chosen myself.	1 2 3 4 5 6
4	reading silently.	1 2 3 4 5 6
5	reading while focusing on unknown words or grammar in a text.	1 2 3 4 5 6
6	reading for enjoyment.	1 2 3 4 5 6
7	stopping to check a dictionary if I do not understand the meaning of a word while I am reading.	1 2 3 4 5 6
8	reading without a dictionary.	1 2 3 4 5 6
9	choosing books to read that I like from a large selection.	1 2 3 4 5 6
10	trying to translate into Japanese everything that I read.	1 2 3 4 5 6
11	writing a short summary of what I have read, after I finish reading.	1 2 3 4 5 6
12	waiting until I have finished reading to check the dictionary for unknown words that I encountered while reading.	1 2 3 4 5 6
13	reading many books at my current proficiency level rather than listening to my teacher explain grammar.	1 2 3 4 5 6
14	reading many easy books rather than listening to my teacher explain new vocabulary words.	1 2 3 4 5 6
15	trying to guess the meaning of unknown words from the reading.	1 2 3 4 5 6
16	trying to understand English as English instead of translating English into Japanese	1 2 3 4 5 6
17	looking up only those unknown words in the dictionary that I have encountered several times in my reading.	1 2 3 4 5 6

Note. Adapted from Bamford & Day (2003). Extensive Reading Activities for Teaching Language. Item 11 was deleted from the data due to poor fit to the Rasch Model and because the item concerned writing instead of reading, and was therefore theoretically illogical to the survey.

REFERENCES

〔1〕 Day, R. R., & Bamford, J.(2002). Top ten principles for teaching extensive reading. *Reading in a Foreign Language,* 14(2), 136–141.

〔2〕 Linacre, J. M. (2009). WINSTEPS Rasch measurement computer program (Version 3.68.0) [Computer software]. Chicago: Winsteps.com.

〔3〕 McNamara, T. F. (1996). Measuring second language performance. London: Longman.

〔4〕 Wolfe, E. W., & Chiu, C. W. T.(1999). Measuring change across multiple occasions using the Rasch Rating Scale Model. *Journal of Outcome Measurement,* 3, 360–381.

〔5〕 Wolfe, E. W., & Smith, E. V. Jr. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II- Validation activities. *Journal of Applied Measurement,* 8, 204–234.

〔6〕 Wright, B. D., & Douglas, G. A. (1976). Rasch item analysis by hand. *Research Memorandum No.* 21. Statistical Laboratory, Department of Education, University of Chicago.

〔7〕 Wright, B. D., Linacre, J. M., Gustafson, J. E., & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions,* 8(3), 370. *Retrieved September* 24, 2010 *from* http://rasch.org/rmt/rmt83b.htm