Articles

Using Rasch Analysis to Create and Evaluate a Measurement Instrument for Foreign Language Classroom Speaking Anxiety

Matthew T. Apple
Ritsumeikan University

Despite the existence of skill-specific anxiety instruments measuring reading anxiety, writing anxiety, and listening anxiety, there is still no single measurement instrument specifically designed to measure levels of speaking anxiety. This research had two purposes. The first was to provide for classroom-based foreign language teachers and researchers an example of the advantages of Rasch model analysis, the use of which is increasing in first-language educational contexts for measurement instrument creation and validation. The second purpose was to create and evaluate an instrument for measuring foreign language speaking anxiety within the classroom in an EFL learning context, in which few native speakers of the language are available for interaction. Using data from a sample of Japanese university students (N = 172), the Rasch model identified misfitting items and examined the construct validity of a 20-item questionnaire to measure levels of Foreign Language Classroom Speaking Anxiety (FLCSA).

リーディング、ライティング、リスニングといった特定のスキルに関する不安を測定する手段はあるが、スピーキング不安のレベルを測定する手段は現在のところ存在しない。本研究の目的は二つある。一つ目は、教室で教える外国語教師や研究者にラッシュモデル分析の利点の例を示すことである。ラッシュモデルによる分析は、第一言語の教育環境において、測定手段の作成やその妥当性を高める目的でますます使用されるようになっている。本研究の二つ目の目的は、

母語話者とのやりとりの機会がほとんどない「外国語としての英語」を学ぶ教室において、外国語スピーキング不安を測る手段を作成し、評価することである。日本の大学生（*N*=172）のデータを使用し、外国語スピーキング不安（FLCSA）の度合いを測定する20項目からなる質問紙からラッシュモデルにより不適当な項目を割り出し、構成概念妥当性を検証した。

Anxiety has attracted a considerable amount of attention in SLA research (MacIntyre, 1999; Scovel, 2001; Young, 1999). Although some SLA researchers have insisted that L2 anxiety was primarily due to poor language skills in a student's L1, the so-called "linguistic coding deficit hypothesis" (Sparks & Ganschow, 1991, 1995; Sparks, Ganschow, & Patton, 1995), other researchers have persuasively argued that anxiety negatively affects cognitive processes and thus is not a result but a cause of performance (Gardner & MacIntyre, 1993; MacIntyre, 1995a, 1995b; MacIntyre & Gardner, 1991). Some researchers (e.g., Brown, Robson, & Rosenkjar, 2001; MacIntyre & Gardner, 1994) indicated the possible benefits of "facilitative" anxiety. Whether facilitating or debilitating, anxiety is a critical concern for foreign language instructors in the classroom setting.

Before dealing effectively with anxiety in the classroom, foreign language instructors need to measure levels of anxiety among their learners. Anxiety itself was originally conceived of as a stable trait variable that holds true in any social situation (Eysenck, 1970; Taylor, 1953), but quickly became categorized into two forms by Spielberger (1983): trait anxiety (constant across contexts) and state anxiety (changeable according to circumstances). Foreign language anxiety is generally considered a state-related, situation-specific anxiety that arises only within foreign language contexts (MacIntyre, 1999; 2007). As such, questionnaire instruments[1] designed to measure foreign language anxiety typically focus on anxiety as it occurs within the foreign language classroom, the most salient learning context for language learners.

Quantitatively oriented foreign language anxiety studies typically use short two-, four-, and six-item Likert-type questionnaire instruments (e.g., MacIntyre, Babin, & Clement, 1999; MacIntyre & Charos, 1996; MacIntyre, Noels, & Clément, 1997) or the 33-item Foreign Language Classroom Anxiety Scale (FLCAS) that measures a combination of communication apprehension, fear of negative evaluation, and test anxiety (e.g., Elkhafaifi, 2005; Gregersen, 2007; Gregersen & Horwitz, 2002; Horwitz, Horwitz, & Cope, 1986; Mak, 2011; Matsuda & Gobel, 2004). Skill-specific measurements of foreign language anxiety created through correlation to previous questionnaires include reading anxiety (Saito, Garza, & Horwitz, 1999), writing

anxiety (Cheng, Horwitz, & Schallert, 1999), and listening anxiety (Kimura, 2008). However, there is still no foreign language anxiety questionnaire instrument designed specifically to measure classroom speaking anxiety. SLA anxiety researchers have noted that language anxiety development seems particularly linked to the fear of speaking in front of one's peers (Cohen & Norst, 1989; MacIntyre, 1999; Price, 1991), making the lack of a foreign language speaking anxiety questionnaire all the more surprising.

Drawbacks to Traditional Statistical Methods of Measuring Anxiety

Many previous instruments measuring foreign language anxiety contained items related to various aspects of anxiety. Many researchers claim to have "validated" responses gathered from such instruments through the use of Cronbach's alpha reliability estimates, factor analyses, and correlation to data obtained from other questionnaires. However, the use of existing foreign language anxiety measurement instruments without detailed examination of the data scores generated for specific student samples may have resulted in four potentially misleading assumptions.

The first assumption is that each item in a questionnaire contributes equally to the measurement of the construct[2], regardless of whether some items are easier to agree or disagree with than others (Bond & Fox, 2007, p. 120). For example, when many instructors want to determine the level of anxiety of their students, they often assume that the "3" of one item means a "3" on another item and that by simply adding raw Likert scores, the resulting combined anxiety score will accurately represent the amount of anxiety that each student experiences. A second, related assumption is that all items share equal endorsement difficulty; that is, that all items are equally easy or difficult to agree with. However, answering "agree" to one item may actually be more difficult than answering "agree" to another item. Adding raw scores from Likert-type categories treats the items as having the same endorsement difficulty, potentially leading to erroneous interpretations of the data (Wolfe & Smith, 2007a). Thus, scores obtained from summing Likert-scale responses may misrepresent the anxiety present in the classroom.

The third assumption is that existing questionnaire instruments measure unidimensional constructs. A measurement instrument must measure a single construct before it can be deemed reliable (Thurstone, 1931; Wright, 1999); construct validity is thus a prerequisite of reliability. SLA researchers whose questionnaire instrument items have not been checked for construct unidimensionality cannot be certain that the data obtained from

the questionnaire instrument represents the psychological construct that the researcher intended the item to measure. In other words, while foreign language anxiety may indeed be a multidimensional *concept*, each individual *construct* that comprises foreign language anxiety (i.e., reading, testing, listening, speaking) should produce scores with demonstrable unidimensional properties. This is true even when using exploratory factor analysis (EFA), a procedure based on correlational analysis, because EFA is designed to find multiple constructs within data and cannot determine the construct validity for individual constructs (Wright, 1996a). Thus, prior to any analysis of questionnaire instruments that consist of multiple constructs, each individual construct should be created and evaluated independently to determine the ability of items to measure the intended construct.

A final assumption is that Likert-type categorical data are true interval data. The treatment of categorical data raw scores as interval data rather than ordinal data may help explain why results from questionnaire studies have proved difficult to reproduce across different samples (Wright, 1999). An alternative method of analysis, which transforms Likert-type categorical data from ordinal data into interval data from which a linear construct can be mathematically extracted, is the Rasch model (Rasch, 1960).

The Advantages of Rasch Model Analysis

The Rasch model is a unidimensional measurement model that uses respondents' positive endorsements (responding "agree") or negative endorsements ("disagree") of questionnaire items to calculate the relationship between the amount, or *levels*, of the construct present in each of the respondents (called "persons") and the *endorsability difficulty level* of questionnaire items, that is, the degree to which respondents find it difficult to agree or disagree with items. This relationship is expressed in log-odds, or *logits* (Embretson & Reise, 2000). Data are fit to the Rasch model by mathematically transforming raw scores on items into interval data and then by placing both persons and items on the same linear scale for comparison (Bond & Fox, 2007; Wright, 1999).

Analysis of Likert-type categorical data using the Rasch Rating Scale model (Andrich, 1978) offers several advantages over traditional analysis. The first advantage is the use of *fit* to demonstrate the quality of both persons and items measured by the hypothesized construct. By identifying misfitting person responses and items that do not contribute to the construct being measured, Rasch model analysis can assist the researcher in revising the

questionnaire instrument in order to provide a more accurate estimation of the construct (Wolfe & Smith, 2007a). A second advantage is that, whereas a typical analysis, such as Cronbach's alpha reliability estimates, only shows the consistency of person responses (Sijtsma, 2009), Rasch analysis provides reliability figures both for person responses and for questionnaire items. Additionally, Rasch model analysis uses *separation,* which shows the number of different groups within the sample and the number of different item difficulty levels (Fisher, 1992; Wright, 1996b).

A third advantage is that the Rasch model uses the concept of *measure* to indicate the level of the construct within each questionnaire respondent as well as the item *endorsability difficulty level*, in other words, the degree to which respondents find it difficult or easy to agree with items (Smith, E. V., 2001). Higher person measures indicate greater amounts of the construct present in respondents, and higher item measures indicate items that are more difficult to endorse. A fourth, related advantage is the use of an *item-person map* to display the relative levels of the construct for each person and relative endorsability difficulty level of each item on a single scale. The levels of construct for each questionnaire respondent and the endorsability difficulty level of each questionnaire item are displayed side by side on the same logit scale for easy visual comparison (Wilson, 2005, p. 96).

Finally, the use of Rasch principal components analysis (PCA) of item residuals can demonstrate the degree to which all items demonstrate coherence to a single dimension, thus satisfying the one-construct, one-dimension criterion of measurement theory (Thurstone, 1931). Rasch PCA of the item residuals provides an estimation of internal construct validity by examining not only the items that load onto the hypothesized construct but also the error left over from extracting the construct from the data (Waugh & Chapman, 2005; Wright, 1996a). The use of Rasch analysis to evaluate the data obtained from questionnaires can thus provide the researcher with a wealth of information about the quality of the questionnaire items that the researcher is using to measure latent psychological traits such as foreign language speaking anxiety in the classroom.

Purpose of This Study

In this paper I seek to demonstrate the advantages of using Rasch analysis for the creation and evaluation of a measurement instrument specifically aimed at the construct of foreign language classroom speaking anxiety. Although the sample in this study is based in Japan, it is hoped that the result-

ing questionnaire will become a beneficial measurement resource for others as well.

There were two research goals for this study:

1. To examine how SLA instructors and researchers can use Rasch model analysis to create and evaluate Likert-type category questionnaire instruments, and
2. To evaluate the degree to which the levels of foreign language classroom speaking anxiety in a typical four-skills EFL classroom in Japan can be measured with a questionnaire instrument that has been analyzed and evaluated using Rasch analysis.

Materials and Methods

Participants

There were initially 172 participants in this study. All were 1st-year students in a large private university in western Japan in six intact four-skills English classes that met twice per week for 90 minutes and were taught by native speakers of Japanese. The average class size was 30; the smallest was 25 and the largest was 38, which is representative of typical Japanese university EFL classrooms. Eighty of the participants were economics majors, 69 were engineering majors, 22 were English literature majors, and one participant did not give a major. There were 49 females, 122 males, and one participant who did not indicate a sex.

Measurement Instrument

As a sample of how researchers can use Rasch analysis to create and evaluate questionnaires, in this study I examined the Foreign Language Classroom Speaking Anxiety Scale (FLCSAS). The original FLCSAS consisted of 20 items related to typical speaking situations within a communicative English classroom in Japan (see Appendix). Items had been previously piloted with a smaller sample (N = 116) and modified prior to the analysis in this study (Apple, 2008). Items that measured participants' speaking anxiety in different communicative situations in the foreign language classroom (Items 2, 4, 6, 8, 13, 14, 16, 17, 18, 19, and 20) were based on items from the Personal Report of Communication Apprehension (PRCA-20; McCroskey, 1977, 1978). Items concerned with speaking to the teacher (Items 3, 5, 7, 10, and 15) as well as general speaking anxiety items (Items 1, 9, 11, and 12) were based on items from the FLCAS (Horwitz et al., 1986).

The questionnaire items were translated into Japanese by a native speaker of Japanese and back-translated by another native speaker of Japanese to confirm accuracy of statement. A 6-point Likert-type category scale was employed with the end points of the scale labeled (1 = "Strongly disagree" and 6 = "Strongly disagree"). Other points were not labeled, and there was no middle or neutral option.

Analysis Procedure

The questionnaire instrument was implemented during class time in the middle of the spring semester. Data were analyzed with Winsteps 3.63 software (Linacre, 2006) using the Rasch Rating Scale model for categorical data (Andrich, 1978). Rasch analysis consisted of Rasch person and item fit analysis, Rasch item-person maps, Rasch PCA of item residuals, and disattentuated person measures correlational analysis.

Person Fit and Item Fit Analysis

To determine the fit of persons and items to the construct, Rasch analysis produces both *infit* and *outfit* statistics, which have two forms: one unstandardized (*mean squares*) and one standardized (*z-scores*) (Linacre, 2002). Infit statistics are weighted to give more information about persons who are at or near item endorsability difficulty levels; that is, questionnaire respondents whose probability of endorsing items is similar to the difficulty of endorsing items, thus giving insight into the item quality. Outfit statistics are not weighted and are more easily affected by respondents who find items too easy or too difficult to endorse (i.e., statistical outliers), and thus provide less useful information about the majority of questionnaire respondents. Researchers typically pay more attention to infit in the interests of determining the quality of items as they apply to the majority of respondents (Bond & Fox, 2007, p. 57).

A mean-square fit statistic of 1.0 indicates perfect fit. Person and item responses that misfit the model may be the result of carelessness, response set answering, or item bias and may need further examination to determine whether they are contributing to the construct as intended (Wolfe & Smith, 2007b, p. 211). For this study, misfit was defined as less than 0.50 mean-squares or -3.0 *z*-scores, or greater than 1.50 mean-squares or 3.0 *z*-scores, based on the recommended criteria of Linacre (2007).

The questionnaire instrument's reliability was estimated using four statistics: (a) person reliability (to determine the consistency of person

responses), (b) person separation (to estimate the ability of the instrument to separate participants into different levels of the construct), (c) item reliability (to estimate how well the items cohered), and (d) item separation (to estimate the ability of the participants to distinguish between items measuring different levels of the construct) (Wright & Masters, 2002).

Item-Person Maps

Item maps were requested as part of item fit analysis to provide a visual representation of person and item locations on the construct. Persons are located along the linear scale based on the level of construct (i.e., the amount of construct present in individual respondents), while items are located based on their endorsement difficulty level (i.e., degree of difficulty of answering "agree" or "disagree"). A person has a 50% chance of endorsing an item located at the same level of the construct on the opposite side of the vertical line (Bond & Fox, 2007). Items located above the person are more difficult to endorse and items located below the person are easier to endorse.

PCA of Item Residuals

A Rasch PCA of item residuals was conducted to examine the unidimensionality of the construct in two ways. First, a unidimensional construct should account for at least 50% of the total variance in the data. Second, the principal contrast, the residual errors left over after the linear construct is extracted, should represent uncorrelated error with an eigenvalue of less than 3.0 and less than 10% of the variance (Linacre, 2007). If the principal contrast has a greater eigenvalue and variance than these criteria, there may be an additional, unwanted construct present in the data that the items were not meant to measure. Item residuals on the main dimension of the construct are termed "positively loading items," while items on a possible secondary dimension are termed "negatively loading items" (Wright, 1996a).

Disattentuated Person Measures Correlational Analysis

Disattentuation refers to the process by which the Rasch model takes into account the error of measurement when transforming raw, ordinal scores from questionnaire item responses into true interval measures. The transformed responses are referred to as *disattentuated person measures*, which can be used in the place of raw scores in correlational analysis. The correlational analysis of disattentuated person measures serves two functions.

The first is to verify whether the removal of items that misfit the model will adversely affect the ability of the questionnaire instrument to measure the level of the latent construct of questionnaire respondents (Smith, R. M., 2000). A strong correlation ($r > .7$) indicates that the items are measuring the same construct. The second is to provide a further verification of construct validity by correlating person measures from positively and negatively loading items in the Rasch PCA of item residuals (Smith, E. V., 2002). A strong correlation ($r > .7$) suggests construct unidimensionality and thus supports claims of construct validity.

Results

Person Fit Analysis

The data acquired from the questionnaire instrument were input into the Rasch model and Rasch person fit analysis was conducted to determine whether all participant responses fit the model's expectations. The Rasch reliability of person responses was estimated at .88, with a Rasch person separation of 2.88. Based on the fit criteria, the responses of 20 persons were found to systematically misfit the model on all questionnaire items. Examination of the misfitting person responses showed the existence of set response patterns and repeated extreme response scores, indicating possible carelessness or lack of seriousness in answering the questionnaire. Because extreme scores add measurement error and adversely affect item fit and unidimensionality of construct, the misfitting person responses were excluded from further analysis, making an adjusted *N*-size of 152. The data were input into the Rasch model again; further examination of person fit indicated no misfitting person responses. The revised Rasch person reliability was .90 and the person separation was 3.02. The Rasch person separation of above 3.0 indicated the ability of the measurement instrument to stratify person responses into at least four separate groups, or levels, across the construct.

For comparison, traditional descriptive statistics were calculated based on raw response scores (Table 1). Cronbach's alpha was calculated at $\alpha = .93$, indicating that Rasch person reliability was a more conservative estimate.

Table 1. Descriptive Statistics for Items Measuring the Foreign Language Classroom Speaking Anxiety Construct

Item	Item description	*M*	*SD*
A1	I'm worried that other students in class speak better than I do	3.29	1.57
A2	I feel nervous speaking in front of the entire class	4.03	1.51
A3	I tremble when the teacher is about to ask me a question	2.26	1.21
A4	I am reluctant to express my opinion in a group	2.50	1.25
A5	I'm worried about making mistakes when I speak with the teacher	2.73	1.46
A6	I'm worried that my partner speaks better English than I do	2.69	1.42
A7	I am reluctant to ask the teacher a question	2.57	1.48
A8	I start to panic when I speak with a classmate in a pair	2.03	1.14
A9	I dislike speaking entirely	2.09	1.27
A10	I'm worried that the teacher will think my speaking is no good	2.37	1.37
A11	I'm worried about making mistakes while speaking	2.58	1.52
A12	I feel nervous when I can't express my opinion	3.17	1.57
A13	I'm afraid my partner will laugh when I speak with a classmate in a pair	1.89	1.09
A14	I'm worried about making mistakes when I speak with a partner	2.02	1.21
A15	Answering a teacher's question in class is embarrassing	2.14	1.22
A16	Speaking in a group of classmates makes me feel self-conscious	2.26	1.27
A17	I feel tense when I have to speak with a classmate in a pair	2.02	1.10
A18	I start to panic when I have to speak in a group	1.96	1.08
A19	I'm afraid that others in a group discussion will laugh if I speak	2.95	1.62
A20	I can feel my heart pounding when it's my turn to speak in a group	1.82	1.07

Note. A Likert scale from (1) *Strongly disagree* to (6) *Strongly agree* was used.

Item Fit Analysis

After person fit analysis was concluded, Rasch item fit analysis was conducted (Table 2). The Rasch item reliability was .98, demonstrating the strong coherence of the items. The Rasch item separation was 6.41, indicating that the study participants were able to distinguish approximately six different levels of the construct as measured by the items. Item fit statistics showed that several items (Items 1, 2, 7, 13, 15, 18, and 19) were well above the ±3.0 *z*-score criterion; however, only three of these were also outside the mean-squared criteria of 0.5 and 1.5 (Items 1, 7, and 19).

Table 2. Rasch Item Analysis for Items Measuring the Foreign Language Classroom Speaking Anxiety Scale

Item	Measure	Error	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
A16	.90	.11	1.30	2.1	1.25	1.5
A13	.72	.10	.59	-3.7	.53	-3.7
A19	.57	.10	*.47*	-5.2	.43	-5.0
A14	.50	.09	.79	-1.8	.73	-2.0
A18	.47	.09	.55	-4.3	.53	-4.0
A9	.46	.09	1.27	2.0	1.29	1.9
A8	.41	.09	.91	-.7	.80	-1.5
A15	.38	.09	.66	-3.1	.63	**-3.0**
A3	.21	.09	.93	-.6	.99	.0
A17	.18	.09	.81	-1.7	.81	-1.5
A10	.03	.08	.90	-.8	.89	-.9
A4	-.07	.08	1.01	.2	1.10	.9
A7	-.07	.08	1.72	5.2	1.90	5.8
A11	-.16	.08	.94	-.5	.87	-1.1
A6	-.33	.08	1.02	.2	1.06	.6
A5	-.35	.08	1.00	.1	1.09	.8
A20	-.65	.08	1.29	2.5	1.30	2.4
A12	-.79	.08	1.00	.0	1.01	.1
A1	-.82	.08	1.52	4.1	1.66	4.9
A2	-1.61	.08	1.46	3.7	1.45	3.5

Note. MNSQ = mean-squared; ZSTD = standardized *z*-scores. Measures are in Rasch logits. ZSTD misfit is indicated by boldface; MNSQ misfit is indicated by italics. $N = 152$.

The items misfitting both sets of criteria were removed and the analysis was run again. However, there was no significant difference in person separation (before = 3.02; after = 2.98) or item separation (before = 6.41; after = 6.68), meaning in practical terms that the range of persons and items across the construct remained essentially the same. Both person and item reliability remained unchanged. To examine changes in construct validity, a correlational analysis of disattentuated person measures was conducted with person measures obtained from before removing misfitting items and from after removing the items. The two sets of person measures were strongly correlated, $r = .99$, $p < .01$, signifying that the misfitting items could be removed without adversely affecting the measurement. For the time being, misfitting items were retained for further confirmation in the PCA of residuals analysis.

Item-Person Map

An item-person map was obtained as a visual representation of person and item difficulty levels along the construct (Figure 1). The most difficult item to endorse was Item 16 ("Speaking in a group of classmates makes me feel self-conscious," Rasch item difficulty measure = .90), and the item easiest to endorse was Item 2 ("I feel nervous speaking in front of the entire class," Rasch item difficulty measure = -1.61).

To give a clearer visual picture of levels of the construct for each study participant, the item-person map was reconfigured to show individual participants along the linear scale (Figure 2). The item-person map in Figure 2 shows items on the left side and individual study participants minus the 20 misfitting persons who were removed, labeled 1 through 172, on the right side. Even though the mathematical mean of the person ability estimates was slightly lower than Item 1 ("I'm worried that other students in class speak better than I do") on the left side of the linear scale, the median of persons was the same level as Item 20 ("I can feel my heart pounding when it's my turn to speak in a group"), suggesting that Item 20 was the best item to distinguish between low and high levels of foreign language classroom speaking anxiety among the participants.

PCA of Item Residuals

Following person and item fit analyses, a Rasch PCA of item residuals was conducted for all 20 items (Table 3). Results indicated that 71.0% of the variance (eigenvalue = 49.1) was explained by the construct. The principal

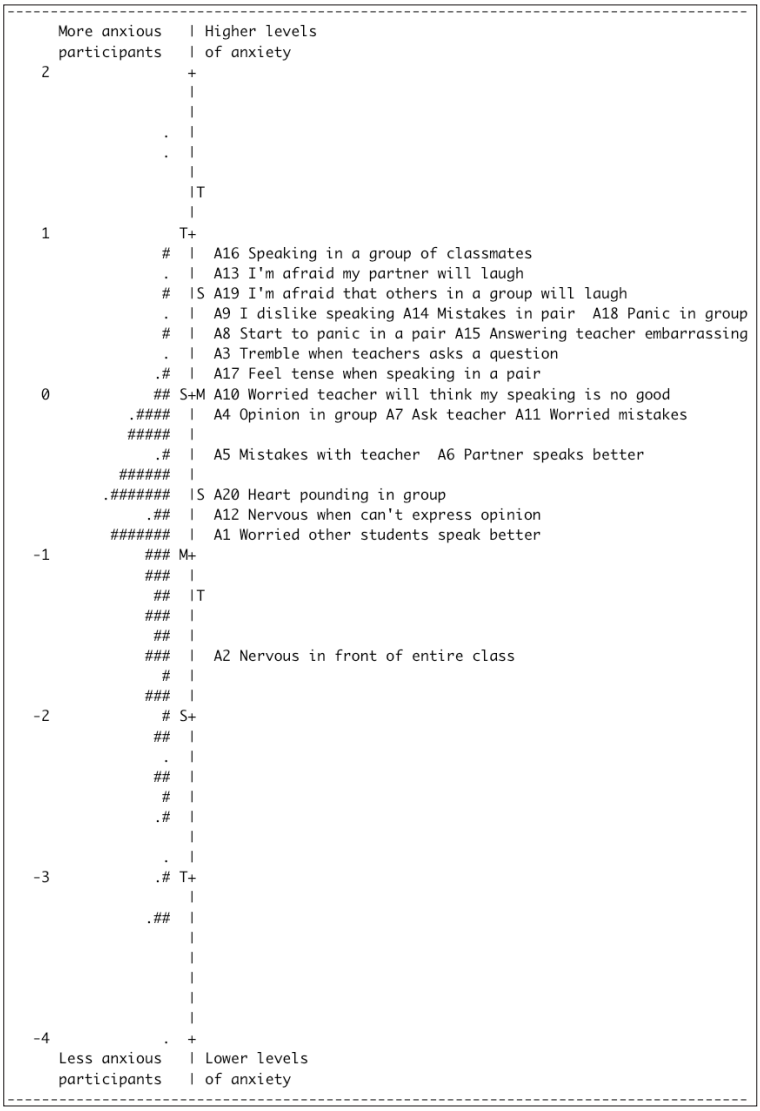

Figure 1. Item-Person Map for Items Measuring the Foreign Language Classroom Speaking Anxiety Scale, Including Misfitting Items

Note. Each # represents two participants. Items more difficult to endorse and thus indicative of higher levels of anxiety are near the top of the scale. M = mean; S = one standard deviation from the mean; T = two standard deviations from the mean.

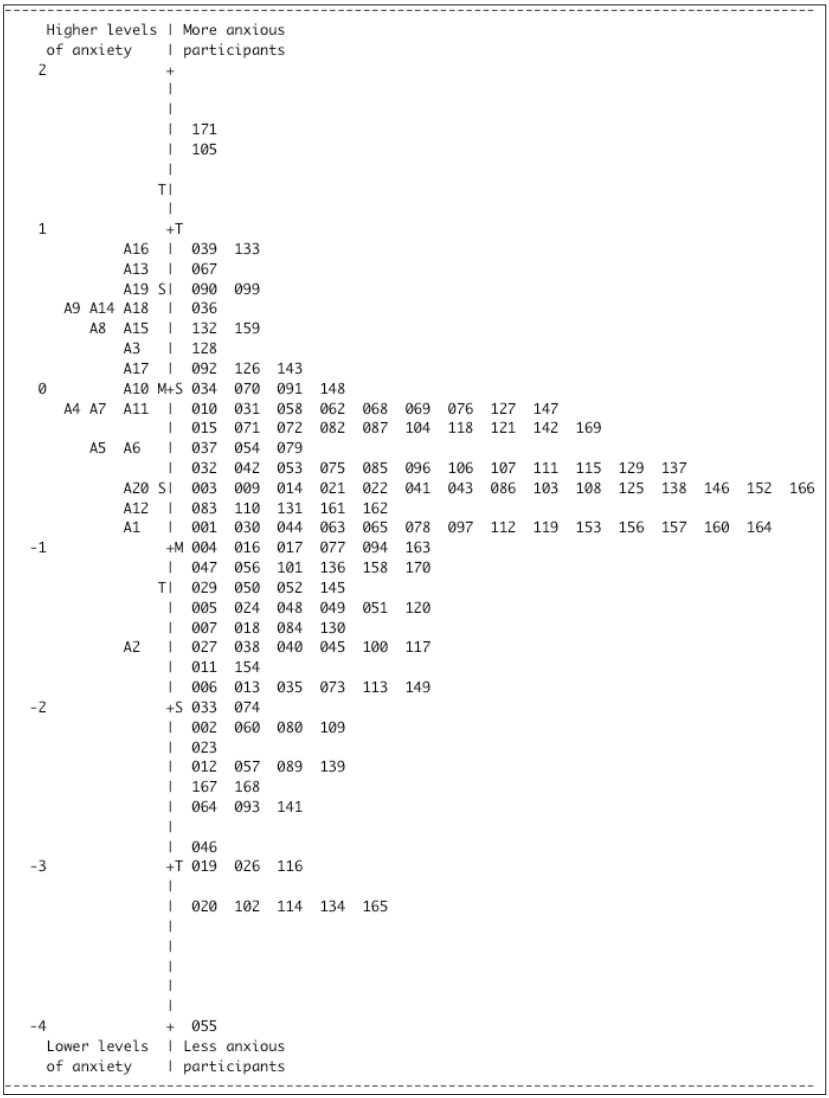

Figure 2. Item-Person Map for Persons Along The Foreign Language Classroom Speaking Anxiety Scale

Note. Study participants with higher levels of anxiety are located near the top of the scale. M = mean; S = one standard deviation from the mean; T = two standard deviations from the mean.

contrast explained 3.9% of the variance (eigenvalue = 2.7). The variance accounted for by the construct was above the minimum criterion of 50%, and the variance accounted for by the correlated residuals in the principal contrast were below the criterion of 10%, indicating that the items demonstrated strong construct validity. Disattentuated person measures from items that loaded positively and negatively onto the principal contrast were correlated, $r = .74$, $p < .01$, as an additional support of construct validity.

Table 3. Rasch Principal Components Analysis of Item Residuals of the Principal Contrast to the Foreign Language Classroom Speaking Anxiety Scale

Item	Loading	Measure	Infit MNSQ	Outfit MNSQ
A19	**.68**	.57	.47	.43
A18	**.64**	.47	.55	.53
A17	**.55**	.18	.81	.81
A13	**.52**	.72	.59	.53
A14	**.52**	.50	.79	.73
A8	.32	.41	.91	.80
A15	.28	.38	.66	.63
A11	.28	-.16	.94	.87
A16	.21	.90	1.30	1.25
A2	**-.50**	-1.61	1.48	1.45
A4	**-.40**	-.07	1.01	1.10
A7	-.35	-.07	1.72	1.90
A1	-.26	-.82	1.52	1.66
A12	-.16	-.79	1.00	1.01
A6	-.15	-.33	1.02	1.06
A9	-.13	.46	1.27	1.29
A20	-.10	-.65	1.29	1.30
A5	-.08	-.35	1.00	1.09
A3	-.07	.21	.93	.99
A10	-.01	.03	.90	.89

Note. Measures are in Rasch logits. *Positive loading* items above the line indicate the main construct identified by the Rasch model and *negative loading* items below the line indicate a possible secondary dimension to the main construct. Item loadings above .40 are in boldface. MNSQ = mean-squared.

Examination of the item loadings on the principal contrast in Table 3 revealed that five items (Items 13, 14, 17, 18, and 19) had high positive loadings above .40, and two items (Items 2 and 4) had high negative loadings above -.40. Because Item 19 ("Others in a group will laugh…") had been previously identified as a misfitting item, there was a possibility that the item was exerting undue influence on other positive loading items in the construct. Therefore, the item was temporarily removed and the PCA of item residuals was conducted again with the remaining items. However, the variance accounted for by the construct was not improved (before = 71.0%; after = 70.3%). The other misfitting items (Item 1, "Others students speak better…" and Item 7, "Reluctant to ask the teacher…") were similarly examined. The variance accounted for was improved in both instances; after the removal of Item 1, the variance accounted for was 73.6%, while with Item 7 removed, the variance accounted for was improved to 74.3%.

The PCA of item residuals indicated that three other teacher-related items (Items 3, 5, and 10) had extremely low loadings of -.07, -.08, and -.01, respectively, indicating that the items did not strongly cohere to other items measuring the construct. When all five of the teacher-related items (Items 3, 5, 7, 10, and 15) were removed, the resulting variance accounted for improved (77.0%, eigenvalue = 50.3), and the principal contrast accounted for 3.9% of the variance (eigenvalue = 2.6). The disattentuated person measures from positively and negatively loading items were correlated again and the result was stronger (r = .89) than prior to removing the teacher-related items (r = .74). Thus, Rasch analysis indicated that teacher-oriented items could be removed to improve construct validity for the study sample (Table 4).

Discussion

At first glance, the descriptive statistics in Table 1 seem to match the Rasch model analysis results; for example, Item 2 had the highest mean score (M = 4.03), and thus was the easiest with which students agreed. In the Rasch analysis, Item 2 was likewise the easiest to endorse (Rasch item difficulty measure = -1.61). However, while Item 20 had the lowest mean score in the descriptive statistics (M = 1.82) and thus was the most difficult item with which to agree, Item 20 was not the most difficult to endorse in the Rasch analysis. In fact, the results of Rasch model analysis showed Item 20 to be relatively easy to endorse (Rasch item difficulty measure = -.65). The most difficult item for students to endorse, as measured by Rasch analysis, was Item 16; in the descriptive statistics, this item was relatively difficult for students to agree with (M = 2.26). This demonstrates how a reliance on mean

Table 4. Rasch Principal Components Analysis of Item Residuals of the Principal Contrast to the Foreign Language Classroom Speaking Anxiety Scale Without Teacher-Related Items

Item	Loading	Measure	Infit MNSQ	Outfit MNSQ
A19	**.67**	.60	.47	.43
A18	**.64**	.50	.51	.48
A17	**.52**	.53	.81	.79
A13	**.52**	.76	.60	.55
A14	**.47**	.21	.75	.74
A8	.31	.44	.89	.78
A11	.30	-.15	1.02	.92
A16	.24	.95	1.40	1.28
A2	**-.51**	-1.65	1.55	1.52
A1	-.35	-.83	1.52	1.64
A4	-.34	-.06	1.07	1.20
A6	-.26	-.32	1.02	1.04
A12	-.24	-.81	1.02	1.00
A20	-.23	-.66	1.28	1.30
A9	-.11	.49	1.30	1.32

Note. Measure is in Rasch logits. *Positive loading* items above the line indicate the main construct identified by the Rasch model and *negative loading* items below the line indicate a possible secondary dimension to the main construct. Item loadings above .40 are in boldface. MNSQ = mean-squared.

scores to judge which items are the best indications of levels of a psychological variable within a participant population may be potentially misleading.

Likewise, although the questionnaire data had a high Cronbach's alpha of .93, this figure was not helpful in determining to what degree the items measured the construct, or to what degree the persons were ranged along the construct. Researchers can use Rasch analysis to take into account measurement error, item location, person location, and fit statistics to better determine the degree to which speaking anxiety levels exist for individual students as well as to determine which questionnaire items were the best indicators of speaking anxiety. Thus, the first research goal, of examining how researchers can use Rasch analysis to create and evaluate questionnaire instruments, has been demonstrated: Using Rasch analysis of the example questionnaire data in this paper, we were able to determine per-

son fit, item fit, the degree to which items contributed to the construct, the amount of variance accounted for by the items, and the degree of construct unidimensionality. The Rasch analysis results provide valuable information that a researcher could use to revise such a questionnaire instrument prior to further implementations with other learner population samples.

The second research goal, regarding the measurement of different levels of foreign language classroom speaking anxiety among the study participants, was also reached. Rasch person separation (3.02) indicated the presence of four levels of anxiety within the sample, while the Rasch item-person map indicated the location of individual persons along the construct. While the identification of three items (Items 1, 7, and 19) as misfitting the intended construct on both infit mean-squared and standardized *z*-scores provided a useful starting point for the examination of construct validity in the PCA of item residuals, an examination of Item 7 proved most useful, as subsequent removal of all five teacher-related items in the PCA of item residuals resulted in more variance accounted for by the construct. Additionally, the location of three teacher-oriented items (Items 5, 7, and 15) next to items of equal endorsement difficulty level on the item-person map in Figure 1 indicated item redundancy.

Finally, the item fit (Table 2) indicated only a .03 logit difference in item difficulty measure between a fourth teacher-oriented question, Item 3 ("I tremble when the teacher is about to ask me a question"), and Item 17 ("I feel tense when I have to speak with a classmate in a pair"). Thus, the Rasch analysis indicated that the five teacher-related items did not contribute as much to the construct as other items and could be eliminated to improve measurement of the construct.

By way of comparison, Cronbach's alpha would slightly decrease ($\alpha = .92$) if one were to remove the teacher-related items, leading to the potentially erroneous conclusion that the items should be kept. The use of Rasch model analysis in this paper demonstrates that over-reliance on Cronbach's alpha reliability estimates of internal consistency, in which more items and more participants equals better reliability, does not necessarily lead to a better questionnaire (Nunnally, 1978; Sijtsma, 2009).

Conclusion

Results from this study demonstrate that using Rasch analysis can identify items that may be inappropriate for the sample population. The results indicated that the questionnaire performed adequately as a measure of foreign

language classroom speaking anxiety. Results obtained from Rasch item fit analysis, item-person map, and Rasch PCA of item residuals revealed that the questionnaire items regarding communication with the teacher did not contribute to the intended construct. From these results it can be inferred that, for students in the sample, questionnaire items concerning direct interaction with the teacher were not as salient as other items pertaining to communication among classmates for determining speaking anxiety levels as measured by the questionnaire.

The analyses also indicated room for improvement before future implementations. The measurement instrument presented in this study as an example of how Rasch analysis results can assist researchers in creating and evaluating their own questionnaire instruments was part of a pilot study for a larger study (Apple, 2011), in which 11 of the original 20 items were retained (see Appendix). A shorter version was also used in two recent studies (Apple, Falout, & Hill, 2012, in press). In each case, items from the FLCSAS were reevaluated for the participant sample to determine the effectiveness of items to measure the theoretical construct of foreign language classroom speaking anxiety. When implementing a previously created and tested measurement instrument, researchers should keep in mind that instruments do not "have" a certain reliability; in that sense, questionnaire instruments should always be evaluated to check the fit of the items to the construct and the fit of the participants to the questionnaire. Rasch analysis provides researchers with a full set of evaluation tools to help them improve existing questionnaires and create their own if needed.

Notes

1. The terms *questionnaire instrument* and *questionnaire measurement instrument* are preferred to *questionnaire* in this paper. In the quantitative measurement tradition, any method of obtaining and evaluating responses of persons that "express their achievements, attitudes or personal points of view" is a measurement instrument (Wilson, 2005, p. 5). Instrument is thus a more precise term than questionnaire when attempting to collect, measure, and summarize participant responses in a questionnaire survey study.
2. The term *construct* is used in this paper rather than *factor* to represent a "theoretical concept that explains observable behaviors and refers to assumed latent (unobservable) characteristics of respondents" (Wolfe & Smith, 2007a, p. 106).

Acknowledgements

The author would like to thank the university language instructors who assisted in data collection and David Beglar of Temple University who gave much statistical advice regarding the Rasch model.

Matthew Apple has an MFA (University of Notre Dame) and an MEd and EdD (Temple University) and has taught English at various educational levels in Japan since 1999. His research interests include CALL and individual differences in SLA and his teaching interests include academic writing and presentation.

References

Andrich, D. (1978). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement, 38,* 665-680.

Apple, M. (2008, May). *Developing an anxiety Rasch*. Paper presented at the PanSIG 2008 Conference, Kyoto.

Apple, M. (2011). *The Big Five personality traits and foreign language speaking confidence among Japanese EFL students* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3457819)

Apple, M., Falout, J., & Hill, G. (2012). The L2 motivational selves of technical college students. In H. Terai (Ed.), *Proceedings of the 6th International Symposium on Advances in Technology Education* (pp. 189-194). Retrieved from http://www.isate2012.jp/dl/iste2012_web_proceedingsv2.pdf

Apple, M., Falout, J., & Hill, G. (in press). Exploring classroom-based constructs of EFL motivation for science and engineering students. In M. Apple, D. Da Silva, & T. Fellner (Eds.), *Foreign language motivation in Japan.* Bristol, UK: Multilingual Matters.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Brown, J. D., Robson, G., & Rosenkjar, P. R. (2001). Personality, motivation, anxiety, strategies, and language proficiency of Japanese students. In Z. Dörnyei & R. Schmidt (Eds.), *Motivation and second language acquisition* (pp. 361-391). Honolulu: University of Hawai'i.

Cheng, T.-S., Horwitz, E. K., & Schallert, D. L. (1999). Language anxiety: Differentiating writing and speaking components. *Language Learning, 49*, 417-446.

Cohen, Y., & Norst, M. J. (1989). Fear, dependence, and loss of self-esteem: Affective barriers in second language learning among adults. *RELC Journal, 20*, 61-77.

Elkhafaifi, H. (2005). Listening comprehension and anxiety in the Arabic language classroom. *The Modern Language Journal, 89,* 206-220.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Eysenck, H. J. (1970). *The structure of human personality* (3rd ed.). London: Methuen.

Fisher, W. (1992). Reliability statistics. *Rasch Measurement Transactions, 6,* 238.

Gardner, R. C., & MacIntyre, P. D. (1993). A student's contribution to second language learning. Part II: Affective variables. *Language Teaching, 26*, 1-11.

Gregersen, T. (2007). Breaking the code of silence: A study of teachers' nonverbal decoding accuracy of foreign language anxiety. *Language Teaching Research, 11,* 209-221.

Gregersen, T., & Horwitz, E. K. (2002). Language learning and perfectionism: Anxious and non-anxious language learners' responses to their own oral performances. *The Modern Language Journal, 86,* 562-570.

Horwitz, E. K., Horwitz, M. B., & Cope, J. A. (1986). Foreign language classroom anxiety. *The Modern Language Journal, 70,* 125-132.

Kimura, H. (2008). Foreign language listening anxiety: Its dimensionality and group differences. *JALT Journal, 30,* 173-195.

Linacre, J. M. (2002). What do infit, outfit, mean-square, and standardized mean? *Rasch Measurement Transactions, 16,* 878.

Linacre, J. M. (2006). Winsteps (Version 3.63) [Computer software]. Chicago: Winsteps.com

Linacre, J. M. (2007). *A user's guide to WINSTEPS: Rasch-model computer program*. Chicago: MESA.

MacIntyre, P. D. (1995a). How does anxiety affect second language learning? A reply to Sparks and Ganschow. *The Modern Language Journal, 79,* 90-99.

MacIntyre, P. D. (1995b). On seeing the forest through the trees: A rejoinder to Sparks and Ganschow. *The Modern Language Journal, 79,* 245-248.

MacIntyre, P. D. (1999). Language anxiety: A review of the research for language teachers. In D. J. Young (Ed.), *Affect in foreign language and second language learning: A practical guide to creating a low-anxiety classroom atmosphere* (pp. 24-45). Boston, MA: McGraw-Hill.

MacIntyre, P. D. (2007). Willingness to communicate in a second language: Understanding the decision to speak as a volitional process. *The Modern Language Journal, 91*, 564-576.

MacIntyre, P. D., Babin, P. A., & Clément, R. (1999). Willingness to communicate: Antecedents & consequences. *Communication Quarterly, 47*, 215-229.

MacIntyre, P. D., & Charos, C. (1996). Personality, attitudes, and affect as predictors of second language communication. *Journal of Language and Social Psychology, 15,* 3-26.

MacIntyre, P. D., & Gardner, R. C. (1991). Language anxiety: Its relation to other anxieties and to processing in native and second languages. *Language Learning, 41,* 513-534.

MacIntyre, P. D., & Gardner, R. C. (1994). The subtle effects of language anxiety on cognitive processing in the second language. *Language Learning, 44,* 283-305.

MacIntyre, P. D., Noels, K. A., & Clément, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language Learning, 47,* 265-287.

Mak, B. (2011). An exploration of speaking-in-class anxiety with Chinese ESL learners. *System, 39,* 202-214.

Matsuda, S., & Gobel, P. (2004). Anxiety and predictors of performance in the foreign language classroom. *System, 32,* 21-36.

McCroskey, J. C. (1977). Oral communication apprehension: A summary of recent theory and research. *Human Communication Research, 4,* 78-96.

McCroskey, J. C. (1978). Validity of the PRCA as an index of oral communication apprehension. *Communication Monographs, 45,* 192-203.

Nunnally, J. C. (1978). *Psychometric theory.* New York: McGraw-Hill.

Price, M. L. (1991). The subjective experience of foreign language anxiety: Interviews with highly anxious students. In E. K. Horwitz & D. J. Young (Eds.), *Language anxiety: From theory and research to classroom implications* (pp. 101-108). Englewood Cliffs, NJ: Prentice-Hall.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago.

Saito, Y., Garza, T. J., & Horwitz, E. K. (1999). Foreign language reading anxiety. *The Modern Language Journal, 83*, 202-218.

Scovel, T. (2001). *Learning new languages: A guide to second language acquisition*. Boston, MA: Heinle and Heinle.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,* 107-120.

Smith, E. V. (2001). Evidence for the reliability of measures and validity of measurement interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2,* 281-311.

Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3,* 205-231.

Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement, 1,* 199-218.

Sparks, R. L., & Ganschow, L. (1991). Foreign language learning differences: Affective or native language aptitude. *The Modern Language Journal, 75,* 2-16.

Sparks, R. L., & Ganschow, L. (1995). A strong inference approach to causal factors in foreign language learning: A response to MacIntyre. *The Modern Language Journal, 79,* 235-244.

Sparks, R. L., Ganschow, L., & Patton, J. (1995). Prediction of performance in first-year foreign language courses: Connections between native and foreign language learning. *Journal of Educational Psychology, 87,* 638-655.

Spielberger, C. D. (1983). *Manual for the State-Trait Anxiety Inventory* (Form Y). Palo Alto, CA: Consulting Psychologists.

Taylor, J. A. (1953). A personality scale of manifest anxiety. *The Journal of Abnormal and Social Psychology, 48,* 285-290.

Thurstone, L. L. (1931). Measurement of social attitudes. *The Journal of Abnormal and Social Psychology, 26,* 249-269.

Waugh, R. F., & Chapman, E. S. (2005). An analysis of dimensionality using factor analysis (true-score theory) and Rasch measurement: What is the difference? Which method is better? *Journal of Applied Measurement, 6,* 80-99.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Wolfe, E. W., & Smith, Jr., E. V. (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I – Instrument development tools. *Journal of Applied Measurement, 8,* 97-123.

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

Wright, B. D. (1996a). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling, 3,* 3-24.

Wright, B. D. (1996b). Reliability and separation. *Rasch Measurement Transactions, 9,* 472.

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 65-104). Mahwah, NJ: Erlbaum.

Wright, B. D., & Masters, G. N. (2002). Number of person or item strata. *Rasch Measurement Transactions, 16,* 888.

Young, D. J. (1999). *Affect in foreign language and second language learning: A practical guide to creating a low-anxiety classroom atmosphere.* Boston, MA: McGraw-Hill.

Appendix

The Foreign Language Classroom Speaking Anxiety Scale (FLCSAS)

I'm worried that other students in class speak better than I do.*
I feel nervous speaking in front of the entire class.*
I tremble when the teacher is about to ask me a question.
I am reluctant to express my opinion in a group.
I'm worried about making mistakes when I speak with the teacher.
I'm worried that my partner speaks better English than I do.*
I am reluctant to ask the teacher a question.
I start to panic when I speak with a classmate in a pair.*
I dislike speaking entirely.
I'm worried that the teacher will think my speaking is no good.
I'm worried about making mistakes while speaking.* †
I feel nervous when I can't express my opinion.* †
I'm afraid my partner will laugh when I speak with a classmate in a pair.*
I'm worried about making mistakes when I speak with a partner.* †
Answering a teacher's question in class is embarrassing.
Speaking in a group of classmates makes me feel self-conscious.
I feel tense when I have to speak with a classmate in a pair.*
I start to panic when I have to speak in a group.*
I'm afraid that others in a group discussion will laugh if I speak.* †
I can feel my heart pounding when it's my turn to speak in a group.* †

Notes. Items marked with an asterisk (*) represent items included in the final long version of the construct (Apple, 2011). Items marked with the symbol † represent items included in the short version (Apple, Falout, & Hill, 2012, in press).