Europe's Journal of Psychology
ejop.psychopen.eu | 1841-0413

Research Reports

Rasch Measurement in Language Research: Creating the Foreign Language Classroom Anxiety Inventory

Miranda J. Walker*[a], Panayiotis Panayides[a]

[a] Lyceum of Polemidia, Limassol, Cyprus.

Abstract

The purpose of this study was to construct a new scale for measuring foreign language classroom anxiety (FLCA). It begun with the creation of an extended item pool generated by qualitative methods. Subsequent Rasch and semantic analyses led to the final 18-item Foreign Language Classroom Anxiety Inventory (FLCAI). In comparison with the Foreign Language Classroom Anxiety Scale (FLCAS), the FLCAI demonstrated more convincing evidence of unidimensionality and the optimal 5-point Likert scale functioned better. The FLCAI, while 55% the length of the FLCAS, thus more practical for classroom practitioners to administer and analyse, maintains its psychometric properties and covers a wider range on the construct continuum thus improving the degree of validity of the instrument. Finally, test anxiety was shown to be a component of FLCA.

Keywords: foreign language classroom anxiety, Rasch measurement, unidimensionality, reliability

Europe's Journal of Psychology, 2014, Vol. 10(4), 613–636, doi:10.5964/ejop.v10i4.782

Received: 2014-04-03. Accepted: 2014-06-13. Published (VoR): 2014-11-28.

Handling Editor: Maciej Karwowski, Academy of Special Education, Warsaw, Poland

*Corresponding author at: Anthemidos 12, 4007, Limassol, Cyprus. E-mail: mirandajanewalker@gmail.com

Foreign language anxiety (FLA) is a key issue to be addressed by language teachers as it can cause students to become less receptive to language input (Krashen, 1981), and thus slow down the language learning process. Furthermore, it can negatively affect student motivation (Liu, 2012; Liu & Huang, 2011) as well as achievement (Horwitz, 1986; MacIntyre & Gardner, 1991; Mahmood & Iqbal, 2010; Yan & Horwitz, 2008). Debate, initiated by Sparks and Ganschow (1991), continues over correlation and causation, in other words whether anxiety causes poor performance or poor performance causes anxiety (Horwitz, 2000; MacIntyre & Gregersen, 2012). Either way, as anxiety can have debilitating effects on foreign language learning, identifying students with high levels of foreign language anxiety is important (Horwitz, Horwitz, & Cope, 1986). By identifying such students, the teacher can take steps to help them cope with this. If the teacher is unaware that their students are suffering from FLA, they may perceive students' behaviour as lack of motivation, abilities and / or poor attitude. In fact Aida (1994) suggests that students learn more effectively when teachers take necessary measures to help them overcome their FLA.

Literature Review

Anxiety, according to Spielberger (1983), is the subjective feeling of tension, apprehension, nervousness, and worry associated with an arousal of the automatic nervous system. (Horwitz, Horwitz, & Cope, 1986, p. 125). He defines state anxiety as an ‘unpleasant emotional state or condition’ and trait anxiety as a ‘relatively stable individual difference in anxiety-proneness as a personality trait’ (p. 1). Horwitz et al. (1986) suggest that clinically ‘the subjective feelings, psycho-physiological symptoms, and behavioural responses of the anxious foreign language learner are essentially the same as for any specific anxiety’ (p. 29). Notwithstanding, Horwitz et al. (1986), as pioneers in the field, have played a significant role in facilitating an understanding of FLA by providing the following definition: FLA is ‘a distinct complex of self-perceptions, beliefs, feelings and behaviors related to classroom language learning which arise from the uniqueness of the language learning process’ (p. 128). Horwitz et al. (1986) further suggest that FLA is related to communication anxiety, fear of negative evaluation and test anxiety. Horwitz and Young (1991) do however inform us that in the literature there are two approaches to language anxiety. In one it is viewed as ‘a manifestation of other more general types of anxiety’ (p. 1). In the other it is considered as ‘a distinctive form of anxiety expressed in the response to language learning’ (ibid). Aida (1994) suggests that students with a fear of negative evaluation might become passive in the classroom and that, in extreme cases, the students may even consider skipping lessons in order to avoid anxiety situations, and thus they are left behind. Gender comparisons of FLA have been varied. Park and French (2013) report significantly higher anxiety levels in females than males whereas Matsuda and Gobel (2004) found no significant differences.

Along with their world renowned definition of FLA, Horwitz et al. (1986) designed the Foreign Language Classroom Anxiety Scale (FLCAS). This 33 item, five category Likert scale was designed for use with university students almost 30 years ago but remains a popular instrument. Chan and Wu (2004) note that ‘due to the scale’s success on construct validation and reliability, the FLCAS has been widely adopted by many researchers to explore learners’ foreign language anxiety’ (p. 292). The majority of these studies has also been with university students. Park and French (2013) suggest that as this anxiety scale has been widely used around the world, psychometric evidence has been established. They state that the internal consistency of the FLCAS was high in many cited studies. However, they add that the latent factor structures differed across studies citing, among others, Aida (1994), Horwitz (1986) and Tóth (2008). Most recently, Panayides and Walker (2013) showed through Rasch measurement, that the scale is unidimensional, and that test anxiety is indeed a component of FLCA. Nonetheless, they brought into question the extremely high reliability (internal consistency) suggesting possible flaws in the scale.

Teachers and students are direct stakeholders in language teaching and learning. Both must work in unison in order to achieve maximum results. Indeed research has shown that anxiety can ‘be changed and shaped through teacher intervention in learning’ (Robinson, 2002, p. 8). This reiterates the need for teachers to assess their students’ degree of anxiety. As teachers’ perceptions of their students’ language anxiety are not always congruent with that of the students (Levine, 2003), self-report instruments measuring student anxiety are vital tools for the educator. Over recent years the literature has begun to embrace a more dynamic, multifaceted relationship between anxiety, motivation, self-efficacy and other language learning variables such as learner characteristics and teaching styles (Liu, 2012; Liu & Huang, 2011). Such advances in research, as well as the fact that time and settings are not constant, infer a need to re-evaluate and perhaps refine even widely-accepted instruments such as the FLCAS. Besides, even ‘the most accepted working hypotheses themselves may need revising’ (Spielmann & Radnofsky, 2001, p. 261).

Validity of Scales

‘Validity always refers to the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores’ (Messick, 1993, p. 13). This is because ‘test responses are a function not only of the items, tasks, or stimulus conditions but of the *persons* responding and the *context* of measurement’ (Messick, 1993, p. 15). For this reason, as Yun and Ulrich (2002) stress, the appropriateness of an instrument should be established through validity investigations prior to its use in new situations or new attributes of population. Furthermore, Luyt (2012) advocates that such a process determines whether a test ‘requires revision or a new instrument might better be developed’ (p. 297).

Panayides and Walker (2013) conducted a study of the psychometric properties of the FLCAS on a Cypriot senior high school population (16-18 years old). They verified that test anxiety was a component of FLA, which had previously been disputed (Aida, 1994; Cheng, Horwitz, & Schallert, 1999; Matsuda & Gobel, 2004). Their analysis also revealed flaws in the original FLCAS. They found the reliability of the FLCAS to be very high, in accordance with most primary studies on the FLCAS. They suggested that such high reliability is undesirable in psychometric scales since it can lower their degree of validity. Panayides and Walker’s (2013) investigation revealed two reasons for such a high reliability. First, the items covered a rather narrow range on the construct continuum. Such construct underrepresentation threatens validity (Messick, 1993) and jeopardizes the precision of person estimates (Panayides & Walker, 2013). Second, the scale includes many parallel items. The researchers believe that the use of parallel items should be avoided as:

- they may give a false sense of a high degree of reliability
- the validity of the instrument is compromised
- for basic research, very high reliabilities are not necessary (Nunnally, 1978)
- while Ray (1988) supports the use of parallel items in order to investigate the consistency in an individual’s response pattern, such practice can be successfully avoided by using Item Response Theory (IRT) models and person fit statistics.

One approach which excels in collecting a wide range of validity evidence of an instrument is Rasch measurement.

Rasch Measurement

The most important difference perhaps, and at the same time the strength, of the Rasch models over other IRT models is its philosophy. Rasch measurement is a mathematical framework of ideal measurement, against which test and scale developers can assess their data (Bond & Fox, 2001, 2007). Any departure from the models’ requirements constitutes parting from useful measurement. Other IRT models are statistical models aimed at incorporating all characteristics observed in the data without any regard as to whether they contribute to the measurement process. Panayides, Robinson, and Tymms (2010) argue that the difference is between measurement and modelling. If the aim is to describe the data at hand, trying to model all of their characteristics, then IRT models are preferable. On the other hand if the aim is to construct good measures then the scale or test items should be constrained to the principles of measurement, and this can be achieved only through the use of Rasch measurement.

Assessing Unidimensionality, the Rasch Approach

Scales such as the FLCAS, 'where single scores are used to position individuals on a latent trait continuum should be unidimensional' (Panayides & Walker, 2013, p. 496). Thus, before such scales are used, their unidimensionality must be established as an important component in the investigation of their degree of validity.

Factor analysis (FA) is the most widely used method for assessing the dimensionality of data. However, FA has a major drawback. As with all other statistical analyses, it operates on interval-level data when in fact, the scale scores are ordinal by nature. This makes results of such analysis disputable. Rasch measurement transforms ordinal scores into an interval-level logit scale thus producing more reliable results (Wright & Masters, 1982). Schumacker and Linacre (1996) state that:

> Factor analysis is confused by ordinal variables and highly correlated factors. Rasch analysis excels at constructing linearity out of ordinality and at aiding the identification of the core construct inside a fog of collinearity. (p. 470)

Following the Rasch calibrations, principal components analysis (PCA) of the standardised residuals can be performed. This method has been shown to be more effective at identifying multidimensionality than factor analysis of the raw data (Linacre, 1998). With regard to the critical value of the eigenvalue, above which the factor extracted can be considered a different dimension, Linacre (2005) argues that any value smaller than 2 indicates the strength of less than two items thus the implied dimension has little strength in the data. He suggests that, in a test with a reasonable length, if a secondary dimension has eigenvalue less than 3, the test is most probably unidimensional.

This Study

The FLCAS is irrefutably a well-established instrument which has been used productively for so long that it is inevitable that the new scale will be met by some with scepticism and that comparisons may be made between the two instruments. This is welcomed as it will help establish the validity of the scale in other similar settings. This study follows on from Panayides and Walker (2013) who suggested further research be carried out so as to refine the FLCAS. Rasch measurement was used in this study to construct a new foreign language classroom anxiety scale that is less lengthy, covers the construct of FLA more adequately and maintains the reliability, while at the same time enhances the validity of the instrument.

The following recommendations made by Panayides and Walker (2013) were adhered to. First, five items were removed from the original FLCAS due to poor fit to the Rasch model, and a careful semantic analysis of the remaining 28 items was conducted in order to remove parallel or repetitive items. Second, new items were added so as to achieve a wider coverage of the construct and to improve item targeting. Finally, the category labels were changed in order to facilitate possible collapsing should the need arise.

The research questions guiding this study were:

1. Is the new 5-point rating scale psychometrically optimal?
2. Does the new scale provide reliable person measures?
3. Do the scale items define a single construct?
4. How does the new scale compare with the original FLCAS?

Despite the fact that Panayides and Walker (2013) demonstrated test anxiety to be a component of FLCA through Rasch analyses, previous inconsistencies in research findings (Aida, 1994; Horwitz et al., 1986; MacIntyre & Gardner, 1989) as well as the significance of test anxiety in the literature, indicated a need to explore this again in the new scale.

Method

The Creation of the New Scale

In keeping with AERA, APA, & NCME (1999) the following were documented: (1) the procedures used to develop, review, try out and select items from the item pool; (2) the model used for evaluating psychometric properties of items and data used for item selection (3) the IRT used in the test development, the item response model and evidence of model fit.

Gehlbach and Brinkworth (2011) suggest that 'a literature review and focus group-interview data can be synthesized into a comprehensive list to facilitate the development of items' (p. 380). This study sought teacher and student input and in doing so supports Pasquale's (2011) stance that 'to ignore such beliefs handicaps the language teaching and learning process at nearly every step of the way' (p. 97). To this end, two focus groups were held in January 2013 with experienced language teachers.

The creation of the new scale began with an oversized pool of items. These included 22 items from the original 33 items of the FLCAS (Horwitz et al., 1986) and 17 newly created items. The original 33 items were reduced to 22 through a multi-step process. First, the five items found by Panayides and Walker (2013) to be misfitting the Rasch model in a similar sample of Cypriot high school students were removed. They found the scale to be unidimensional once these had been removed. Second a further six items were removed including four parallel items and two which were not considered relevant to classroom anxiety. The two deemed unrelated for the classroom setting were: Item 14, I would not be nervous speaking the foreign language with native speakers and Item 32, I would probably feel comfortable around native speakers of the foreign language.

Following this, additional items were generated with the intention of improving coverage of the construct. In accordance with the substantive approach (Messick, 1993), items were included in the pool on the basis of judged relevance. The researchers based their decisions on an in-depth study of the literature, an informal discussion with EFL students, two focus group discussions with language teachers and three personal interviews with recently retired teachers of senior high school EFL.

EFL students from ten mixed ability classes were asked informally by their teacher, one of the researchers in this study, to state what makes them anxious in the EFL classroom. They were told that their answers would help her address their anxieties better and also facilitate research. All ideas raised were taken down in writing by the researchers.

In January 2013 two focus group discussions were held. Rodriguez, Schwartz, Lahman, & Geist (2011) contend that focus groups are 'a powerful qualitative research method which, especially when designed to be culturally responsive, facilitate collection of rich and authentic data' (p. 400). One of the groups comprised of 12 teachers and the other of 37 teachers. The participants, who had between one and 33 years teaching experience, were asked to share their beliefs concerning what makes students anxious. They were asked to note down their ideas without discussion a few minutes prior to the main group discussions. These notes were collected to prevent changes being made during the discussion. Ideas raised during the ensuing discussion were noted by one of the two researchers, while the other coordinated the discussion.

Finally three recently retired teachers of senior high school EFL, with an average of 37 years teaching experience, one of whom a trained psychologist, were consulted. They were asked by email to write down what they believed causes EFL students classroom anxiety. Following this, personal interviews were held between them and one of the researchers. The purpose of these was to expand on ideas raised in the written comments. Details of the data collected from the student and teacher discussions and interviews can be found in Appendix A, in frequency order.

The new items stemming from qualitative data collected for the extended item pool were written directly in Greek as the data from which they were generated was Greek. This is also the language of the participants. Items designed as a result of the literature review were written in English and translated into Greek. The Greek version of the original FLCAS, translated by Panayides and Walker (2013) was also used.

The resulting 39 item scale was administered, in April 2013, to 212 students in three senior high schools in Limassol, Cyprus, all of whom had been studying English for a minimum of seven years. In compliance with Messick's (1993) guidelines, the item responses were obtained and analyzed. The final scale items were subsequently selected from the pool. The selection process involved, in line with AERA et al. (1999), looking into two different sets of criteria (a) semantic and (b) statistical – measurement.

The addressing of semantic issues included the removal of parallel items and those items which had been shown to contain ambiguities. An example of ambiguity can be found in Item 15 of the original FLCAS (Horwitz et al., 1986) 'I get upset when I don't understand what the teacher is correcting'. Some stated that their level of English was higher than that of the lesson and so they very rarely made mistakes and wondered whether they should be answering hypothetically. It is customary for children in Cyprus to take private English tuition, from as young as six years of age, which progresses at a faster rate than school lessons do. Such confusion could lead to unreliable results and thus it was considered prudent to remove this item.

For the examination of statistical – measurement criteria, the Rasch Rating Scale Model (Andrich, 1978) was used which is appropriate for analyses of Likert scales. The investigation involved analysing the estimated measures of the items in order to ensure a wide coverage of the construct, the point measure correlations and the item fit statistics. Five items with outfit of greater than 2.0 and with both infit and outfit greater than 1.5 were removed. Finally 22 items were kept. These items included 10 from the original FLCAS (Horwitz et al., 1986) and 12 newly designed items. Eight items concerned test anxiety, five items communication anxiety, four fear of negative evaluation and five general anxiety items. Interestingly this break down reflects the emphasis given to each aspect of FLA in the group discussions, informal discussions with students and interviews held within the setting during the development stage of the extended item pool. The 22-item scale was administered to 285 high school students (16-18 years old) in October 2013.

The Likert Scale — Great emphasis and consideration was placed on the number and labels of the Likert scale to be used. In their validation study of the FLCAS, Panayides and Walker (2013) documented concerns regarding the Likert scale for the population under study. The main problem encountered was that the probability curves of the middle three categories, namely 'disagree', 'neither agree nor disagree' and 'agree', failed to peak (be the most probable choice) for sufficiently large ranges on the construct continuum. They could not however be collapsed for semantic reasons. Since this is a valid concern it has been addressed in designing the modified scale by changing the labels. The importance of semantics is that should any subsequent problems in the Likert categories be diagnosed and collapsing of categories become necessary, it will be possible to do so. As such participants are asked to indicate 'How often do the following statements apply to you?' by selecting between never, rarely, sometimes, frequently and always. This maintains a starting point of five categories.

Assessing Unidimensionality — Many of the methods offered by the Rasch models were used for the assessment of the scale dimensionality. First, PCA of the standardised residuals was used as suggested by Linacre (1998). Second, the point measure correlations were observed for the possibility of unacceptable values (negative or close to zero) and third, the well-known mean square statistics, infit and outfit. Various values are suggested in the literature for cut-off values including the most popular for scales are 1.4 (Wright, Linacre, Gustafson, & Martin-Lof, 1994), 1.5 (Bond & Fox, 2001, 2007) and 1.6 (Curtis, 2004). There is a simple explanation as to why higher values are considered adequate as cut-off values for scales.

> For high stakes multiple-choice tests the items are highly controlled, carefully constructed and piloted and the examinees respond in a highly controlled environment. Questionnaires are usually less carefully constructed and there is less control over how respondents behave. Observational instruments usually have even less control (or even no control) of how respondents behave. Therefore, less control → more off-dimensional behaviour → worse fit expected. (Linacre, personal communication, March 7, 2007)

The primary purpose of conducting a test or administering a scale is to measure the ability or position on the latent trait continuum of people. One needs measures that are good enough for the purpose. Rough measures are useful for the purpose of assessing personality traits and thus the fit criteria can be much more relaxed. Consequently 1.5 was chosen as the cut-off value for this study. In deciding whether an item should be removed from the scale or replaced, 'one should use the suggested cut-off scores as a guide, and then rely on his/her professional judgement and intuition to reach the best possible decision' (Panayides, 2009, p. 134).

Next the items were divided into two groups, test anxiety items and the rest. Person measures were obtained from the two groups of items separately. The two sets of person measures were compared using the correlation coefficient and by performing t-tests for differences in the measures from the two different calibrations as suggested by Smith (2002). A 95% confidence interval for the t-values would be approximately between -2 and 2. Values outside this range indicate significant differences between the person measures. Hence, any percentage of t-values outside this range of less than 5% indicates that the two item groups give statistically equivalent person measures and the two groups of items can be considered as measuring the same construct.

The next step was to once again divide the items into two groups, the items from the original FLCAS and newly created items. The same method of t-tests for differences between the two sets of person measures (one from each item group) was used. In both cases where the t-test method was employed, the item estimates used were the ones obtained from the final calibration of the 18-item scale (shown in Table 2 in the results section).

Finally, for the investigation of invariance, persons were divided by gender and item estimates were obtained separately for the two groups. Again the two sets of item estimates were compared using a scatter plot and the calculation of the correlation coefficient. Statistically equivalent item measures were obtained which supports the property of invariance. In other words the construct measured by the scale has the same meaning across the two groups of people.

Reliability — Three reliability indices were used in assessing the reliability of the scale: Person Reliability, which shows how well the instrument can distinguish persons; Person Separation (G_p), which takes values from zero to infinity and indicates the spread of person measures in standard error units and Strata, given by the formula $[(4G_p+1)/3]$. Strata determines the number of statistically distinct levels (separated by at least three errors of measurement) of person abilities that the items have distinguished.

Item Reliability was also used as it is useful in showing how well the items are discriminated by the sample of respondents. Wright and Masters (1982) emphasise that good item separation is a necessary condition for effective measurement.

Rasch Diagnostics for the Optimal Number of Likert Scale Categories — Optimizing a rating scale is 'fine-tuning' to try to squeeze the last ounce of performance out of a test (Linacre, 1997, para. 1). Linacre (2002) and Bond and Fox (2001, 2007) describe the Rasch measurement diagnostics for evaluating the effectiveness of the number of categories of a Likert scale. These diagnostics facilitate the investigation of the extent to which respondents can clearly identify the ordered nature of the rating scale response options, and whether they can accurately distinguish the difference between each category.

Categories with very low frequencies (Linacre suggests lower than 10) do not provide enough observations for estimating stable threshold values. Such categories should be removed or collapsed with adjacent categories, provided that the semantics permit such collapsing. The average of the measures of all persons in the sample who choose successive categories should increase monotonically. This indicates that those with higher position on the construct continuum endorse higher categories. Likewise, the thresholds or step calibrations should also increase monotonically, otherwise they are considered disordered. In addition, the range of each category (the distance between successive thresholds) should not be too wide in order to avoid large gaps in the variable (not more than 5.0 logits), or too narrow to show distinction between categories (not lower than 1.0 for the case of a 5-point Likert scale). Step disordering and very narrow distances between thresholds 'can indicate that a category represents too narrow a segment of the latent variable or correspond to a concept that is poorly defined in the minds of the respondents' (Linacre, 2002, p. 98). Finally, the outfit statistic provides another useful tool in assessing the effectiveness of the categories. Values of outfit greater than 2.0 indicate more misinformation than information, that is, the category introduces noise into the measurement process.

Estimation Method — WINSTEPS (Linacre, 2005) was used for the analysis of the data. The estimation method used with this software is Joint Maximum Likelihood Estimation (JMLE) in preference to Conditional Maximum Likelihood Estimation or Marginal Maximum Likelihood Estimation. Linacre (2005) explains that JMLE is preferred 'because of its flexibility with missing data. It also does not assume a person distribution' (p. 11). He also clarifies that any estimation bias is not a real concern as, except in rare cases with short tests or small samples.

Results

The First Calibration of the 22-Item Scale

The first calibration (22 items and 285 persons) showed very satisfactory reliability indices: Person Reliability = .93, Person Separation = 3.70 and Strata = 5.27. However, this calibration also revealed that some items had almost identical statistics.

When constructing new scales a good range of item difficulty estimates is needed (that is, a good spread of the items on the construct continuum) so as to attain a high degree of reliability. Empirically, when two or more items are functioning almost exactly the same, then any one of these is as good as the other. When this happens, one can look at these items qualitatively, asking the question: are any of the items more valuable than the rest? And then retain the item or items that are deemed more valuable and remove the others.

Table 1 shows four groups of statistically very similar items. Group A has two items, Group B three, Group C four and Group D two. Four items were removed: one from each of Groups B and D, Items 4 and 1 respectively and two from Group C, Items 12 and 16.

Table 1

Groups of Statistically Similar Items

Item group	Item number	Item difficulty	Pt-measure correlation	Wording
A	2	-0.57	.74	**In tests, I worry that I won't understand the vocabulary in the text.**
	11	-0.58	.74	**Essays make me nervous.**
B	8	-0.07	.72	I start to panic when I have to speak without preparation in the language class.
	19	-0.10	.73	I always feel that the other students speak the foreign language better than I do.
	4	-0.11	.74	I keep thinking that the other students are better at languages than I am.
C	16	0.27	.70	I am afraid that my language teacher is ready to correct every mistake I make.
	17	0.26	.72	**When we have an oral dialogue, I worry that I might not be able to understand what the other person is saying.**
	13	0.22	.70	It embarrasses me to volunteer answers in my language class.
	12	0.20	.68	**I worry that I might not understand the instructions in a listening test.**
D	1	0.55	.63	I tremble when I know that I am going to be called on in language class.
	7	0.57	.67	**I get anxious when the test has a listening component.**

Note. Newly generated items are in bold.

Semantic Justifications for the Removal of Items — 'I tremble when I am going to be called on in language class' (Item 1) was removed because this concept is an extension of 'it embarrasses me to volunteer answers in my language class' (Item 13). Furthermore the adjective 'tremble' was considered by the researchers to be so specific that even very anxious students may not choose it. Items 4 and 19 were found not only to be statistically almost identical, but their meaning was also parallel. Being nervous about listening tests in general (Item 7) could well include worrying about not understanding the instructions in listening tests (Item 12). It was therefore considered that the latter could be removed. The fear of having every mistake corrected by the teacher (Item 16) caused some undesired ambiguity. Some participants asked if they should answer hypothetically since their level of linguistic competence was higher than the difficulty level of the class. Others wished to respond theoretically as they had never experienced this.

Calibrations of the 18-Item Scale

The remaining 18 items were used for the second calibration. Initially, a person fit analysis revealed a few misfitting students. Seven of those (approximately 2.5% of the sample) were considered badly misfitting with infit and/or outfit values greater than 2.5. The responses of these students were considered to be distorting the measurement process and were removed. The remaining dataset, 18 items and 278 students, was used for the final calibration.

Table 2 shows the item statistics, in difficulty order, from the most difficult to endorse to the easiest. The item difficulties range from -1.41 to 1.40. This covers a range of 2.81 logits on the construct continuum.

Table 2

Item Statistics in Measure Order

Item	Difficulty	Error	Infit	Outfit	R
18	1.40	0.10	1.09	1.17	.60
21	1.00	0.09	0.95	0.78	.68
7	0.66	0.09	1.02	0.99	.68
14	0.56	0.09	1.29	1.09	.67
22	0.47	0.09	0.74	0.75	.74
17	0.34	0.08	0.84	0.77	.75
13	0.25	0.08	0.98	1.03	.72
5	0.10	0.08	0.79	0.86	.75
8	-0.01	0.08	0.96	0.95	.73
19	-0.05	0.08	0.93	0.89	.75
9	-0.13	0.08	1.08	1.07	.72
20	-0.25	0.08	0.91	0.90	.76
3	-0.40	0.08	0.82	0.86	.77
10	-0.45	0.08	1.06	1.11	.74
2	-0.55	0.08	0.87	0.95	.76
11	-0.60	0.08	0.95	0.91	.77
6	-0.93	0.08	1.31	1.36	.71
15	-1.41	0.08	1.50	1.51	.70

Note. **R** (final column) shows the point measure correlations.

Reliability — The Person Reliability was found to be .93, the Person Separation 3.57 and Strata 5.09. Also Item Reliability was .98. Despite the removal of four items from the scale, the reliability indices remained essentially unaffected, thus maintaining the high degree of reliability of the scale.

Dimensionality of the Scale — Various analyses were performed for assessing whether the scale is unidimensional. First the point measure correlations (shown in Table 2) of all items were positive and highly significant, ranging from .60 to .77.

PCA of standardised residuals — Table 3 shows the results of the PCA of the standardised residuals.

Table 3

Standardised Residual Variance in Eigenvalue Units

	Empirical		Modelled
Variance component	**Eigenvalue**	**%**	**%**
Raw variance explained by measures	**29.2**	**61.9**	**62.3**
Raw variance explained by persons	23.0	48.7	49.0
Raw variance explained by items	6.2	13.2	13.3
Raw unexplained variance	**18.0**	**38.1**	**37.7**
Unexplained variance in 1st factor	2.2	4.6	
Total raw variance in observations	**47.2**	**100.0**	**100.0**

To assess the strength of the measurement dimension one looks at the variance explained by the measures. In this case it is 61.9% of the total variance in the data. The first factor had an eigenvalue of 2.2, the strength of less than three items. Also the first factor explains just 4.6% of the total variation and only 12.1% of the unexplained variance. Furthermore, the ratio of the variance explained by the measures to the variance explained by the 1st factor is 13.3:1. The evidence collected from the PCA of the standardized residuals suggests that no second dimension is present in the data meaning that the scale is unidimensional.

Item Fit — With the exception of Item 15, 'I worry about my grade in English', all other items were well below the infit and outfit cut-off values. Item 15 with item difficulty estimate of -1.41 had marginal values (infit = 1.50 and outfit = 1.51). Further investigation showed that this item was the easiest in the scale and its marginal fit was caused by very few unexpected high scores by person with low positions on the construct continuum. For example, persons with entries 101 and 146 had estimates of -2.22 and -3.06 and scored 5 and 4 on the item respectively. Both persons were positioned much lower than the item on the construct continuum and their probabilities of scoring 5 and 4 respectively were below 2%. Just these two unexpected responses raised the infit and outfit from 1.45 and 1.46 to 1.50 and 1.51 respectively. As this item was considered important semantically, since it was the only item asking about students' overall assessment, and was the easiest in the scale thus widening the construct coverage, it was kept in the scale.

Property of Invariance — The dataset was divided into two subgroups. One contained the 18 items and 87 male students and the other 18 items and 194 female students. Four students did not state their gender. The correlation between the two sets of item difficulty estimates, yielded from the two separate calibrations, was .967. Furthermore, Figure 1 shows the scatter plot of these item estimates together with a 95% confidence interval.

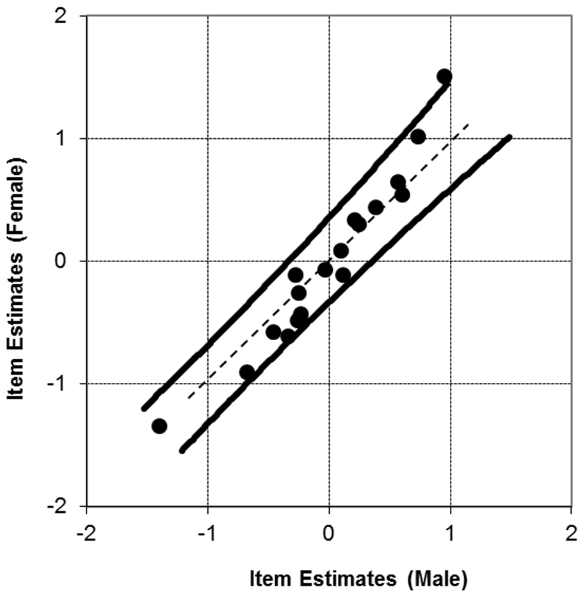


Figure 1. Scatter plot of item estimates from two separate calibrations (Male - Female).

Only one of the 18 items estimates fell slightly outside the confidence interval. This, together with the very high correlation, are a strong indication that invariance holds, meaning the construct indeed has the same meaning among the two groups of students.

Is Test Anxiety a Component of FLCA? — The aim of creating a scale is to obtain estimates of persons' level of anxiety or their positions on the FLCA continuum. In order to investigate whether test anxiety is a component of FLCA the items of the scale were divided into two groups. The first group contained the seven test anxiety items (Items 2, 6, 7, 10, 14, 18 and 20) and the second the remaining 11 items (Items 3, 5, 8, 9, 11, 13, 15, 17, 19, 21 and 22). Person measures were estimated for the two groups of items separately. The correlation of the two sets of measures was highly significant at .876. Perhaps more importantly, significant differences between measures were found only in 12 cases (four t-values below -2.0 and eight t-values above 2.0) and this represents 4.3% of the sample (12 out of 278 students). Since this percentage is below 5% one can infer that the person estimates from the two separate item groups are statistically equivalent and that the two item groups indeed measure the same construct. Figure 2 shows the distribution of the t-values which approximates very well a standard normal (Mean = 0 and standard deviation = 1) with a mean of 0.06 and standard deviation 1.053.

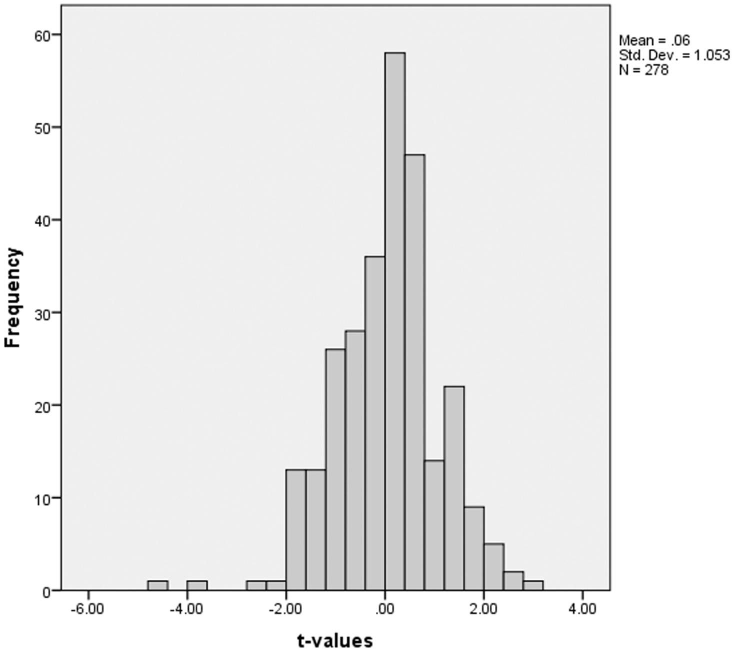


Figure 2. Distribution of t-values (1).

FLCAS Items vs New Items — Finally the items were once again divided into two groups. The first group contained the seven FLCAS items (6, 8, 13, 18, 19, 21, 22) and the second the 11 new items (2, 3, 5, 7, 9, 10, 11, 14, 15, 17, 20). Person measures were estimated for these groups of items separately. The correlation of the two measures was highly significant at .869. Significant differences between person measures were found in 11 cases (four t-values below -2.0 and seven t-values above 2.0) which represents 4.0% of the sample. This method also gave statistically equivalent person measures supporting again the unidimensional structure of the scale. Figure 3 shows the distribution of the t-values with a mean of 0.008 and a standard deviation of 1.026.

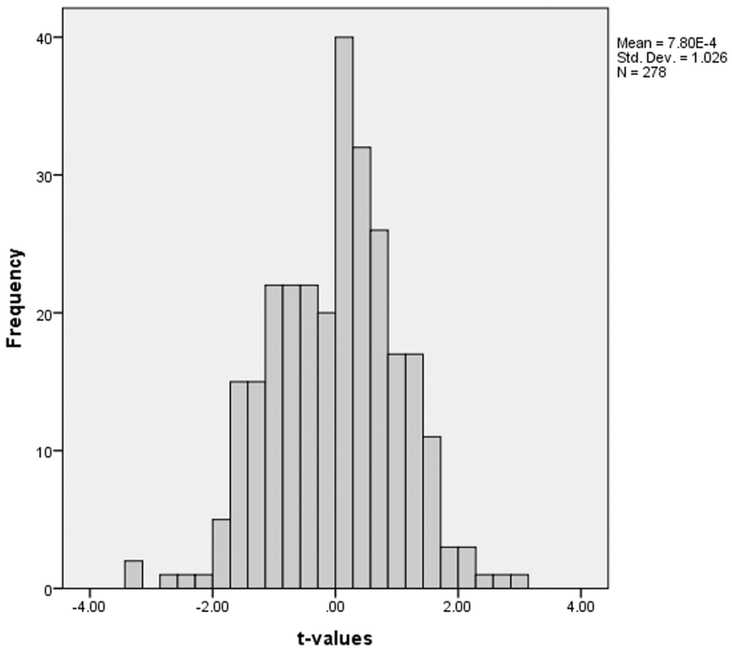


Figure 3. Distribution of t-values (2).

Category Functioning — Table 4 shows the Rasch diagnostics for the investigation of the functioning of the Likert scale categories.

Table 4

Rasch Diagnostics for the Category Functioning

Category	Label	Observed count	Average measure	Infit	Outfit	Thresholds
1	Never	1608	-2.62	0.88	0.92	None
2	Rarely	1297	-1.29	1.03	0.93	-1.73
3	Sometimes	1000	-0.35	0.99	1.03	-0.61
4	Frequently	539	0.43	1.09	1.07	0.67
5	Always	288	1.64	1.14	1.16	1.66

All categories had high observed frequencies, the average person measure corresponding to the categories monotonically increases, infit and outfit values are very close to the expected value of 1.00 and the thresholds also monotonically increase. Finally, the ranges between successive thresholds (1.12, 1.28 and 0.99) are large enough to show distinction between categories. The satisfactory ranges between successive thresholds can be seen in Figure 4 which shows the category probabilities.

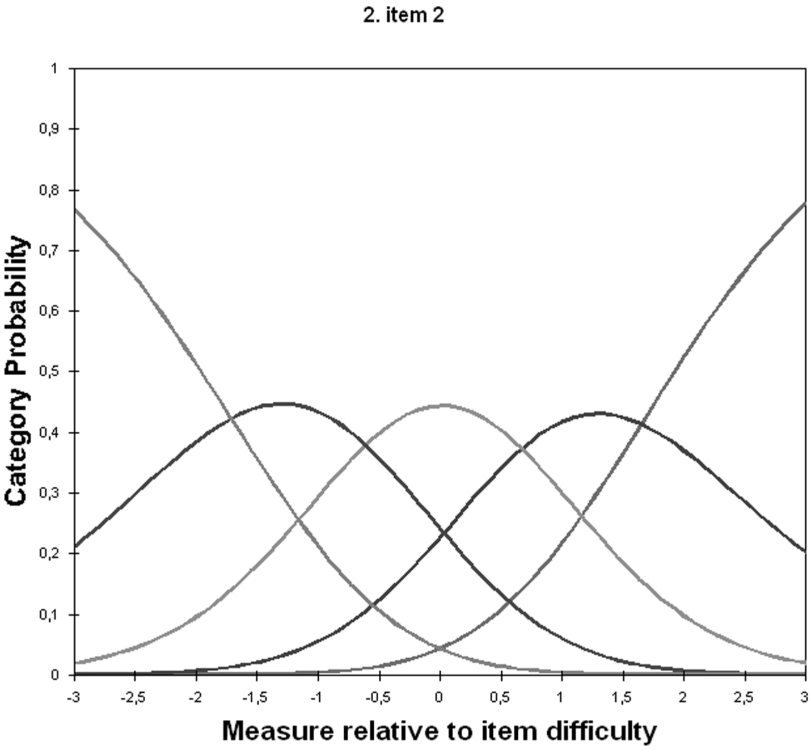
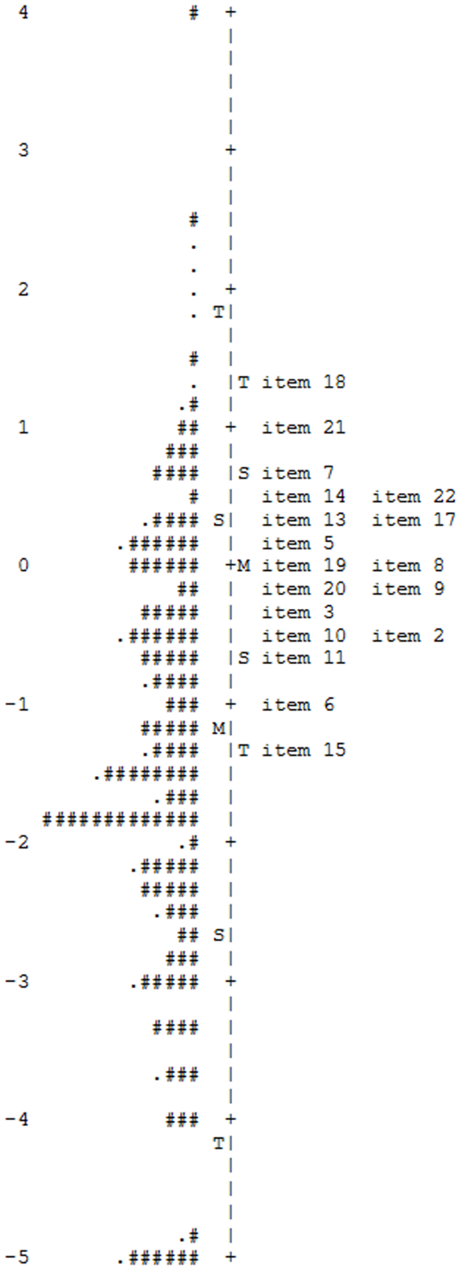


Figure 4. Category probabilities.

Each probability curve (corresponding to each of the five categories) peaks for a significant range along the continuum and this shows that each category is the most probable for the corresponding range. Since these basic criteria suggested by Linacre (2002) are met, the 5-point Likert scale can be considered optimal.

Item Targeting — Figure 5 shows the person item-map.

The items are well targeted for persons of just below the persons' mean FLCA to about one and two thirds standard deviations above the mean (from -1.41 to 1.40 logits) covering a range of 2.81 logits. The person measures range from -4.03 to 6.03.



```
 4              #  +
                   |
                   |
                   |
                   |
                   |
 3                 +
                   |
                   |
                #  |
                .  |
                .  |
 2              .  +
                . T|
                   |
                #  |
                .  |T item 18
               .#  |
 1             ##  +  item 21
              ###  |
             ####  |S item 7
                #  |  item 14  item 22
            .#### S|  item 13  item 17
          .######  |  item 5
 0         ######  +M item 19  item 8
               ##  |  item 20  item 9
            #####  |  item 3
          .######  |  item 10  item 2
            #####  |S item 11
            .####  |
-1            ###  +  item 6
            ##### M|
            .####  |T item 15
        .########  |
             .###  |
    #############  |
-2             .#  +
           .#####  |
            #####  |
             .###  |
               ## S|
              ###  |
-3         .#####  +
                   |
             ####  |
                   |
             .###  |
                   |
-4            ###  +
                  T|
                   |
                   |
                   |
               .#  |
-5        .######  +
```

Figure 5. Person-item map (Each '#' represents two persons).

Comparisons Between the FLCAI and the FLCAS

Table 5 shows comparisons between the main psychometric features of the new scale presented in this study, the Foreign Language Classroom Anxiety Inventory (FLCAI), and the equivalent FLCAS features from the 2013 study by Panayides and Walker (2013). Both instruments were administered to similar-sized samples from the same population (Cypriot high school students of age 16-18) and data were analysed through Rasch measurement thus facilitating comparisons.

Table 5

Comparisons between the FLCAI and FLCAS

Description	FLCAS	FLCAI
Scale length	33 items	18 items
Sample size	304	285
Sample composition	High school students	High school students
Age of participants	16 – 18	16 – 18
Reliability		
Person Reliability	.93	.93
Person Separation	3.64	3.57
PCA of standardised residuals		
Variance by measures	51.6%	61.9%
First factor extracted		
Eigenvalue	2.5	2.2
% of total variance	4.4%	4.6%
Ratio (Measures:1st factor)	11.9:1	13.3:1
Construct coverage	1.44 logits	2.81 logits
Category functioning	Marginally optimal	Optimal

The FLCAI (18 items) is shorter than the FLCAS (33 items) by 45%. Nevertheless, Person Reliability is the same and Person Separation is negligibly lower.

The strength of the measurement dimension is arguably better in the FLCAI as the variance explained by the measures is 61.9% of the total variance as opposed to 51.6% in the FLCAS. Furthermore, the eigenvalue of the first factor extracted by PCA of the standardised residuals is 2.2 in the FLCAI and slightly higher at 2.5 in the FLCAS. Also, the ratio of variance explained by the measures to the variance explained by the first factor is higher for the FLCAI (13.3:1) than for the FLCAS (11.9:1). The above evidence shows that core construct measured by the FLCAI has more strength in the data than that measured by the FLCAS, thus there is more convincing evidence of unidimensionality in the new scale.

Two further points are noteworthy in the comparisons made in Table 5. First there is a wider construct coverage for the 18 items of FLCAI (2.81 logits) than for the 33 items of FLCAS (1.44 logits). Second, the five categories of the Likert scale in the FLCAI function better than those in the FLCAS.

Conclusions

The purpose of this study was to design an appropriate instrument for measuring foreign language classroom anxiety in high schools. Prerequisites were it being psychometrically successful, that is, having a high degree of validity and reliability and of course being an appropriate instrument for the intended population. The FLCAS (Horwitz et al., 1986) has been efficaciously used for almost 3 decades however validity is not time and location independent. A study of its psychometric properties in senior high schools in Cyprus (Panayides & Walker, 2013) drew attention to the need for modifications. Such revisions included an examination of its length, which has been suggested to have inflated reliability (Panayides & Walker, 2013).

The new scale was created in a multi-step fashion. It began with an examination of the original scale. This was followed by the creation of an enlarged item pool which was generated by an extensive examination of the literature, informal discussions with language teachers and students, three interviews with highly experienced English language teachers, as well as focus group discussions with language teachers. Concerns raised in the literature, such as those of Sparks and Patton (2013) that the FLCAS 'is likely to be measuring individual differences in students' language learning skills and / or self-perceptions about their language skills rather than anxiety unique to L2 learning' (p. 870), were reflected upon in the selection of items. The enlarged item pool (39 items) was administered to 212 16 – 18 year old students in early 2013. This was then shortened again through a multi-step process which included Rasch analyses and thorough semantic examination. This resulted in a 22 item scale which was administered to 285 16 – 18 year old students in October 2013.

Finally fine tuning was performed in order to create the most appropriate, efficient and functional scale whilst maintaining reliability. This included careful reanalysis of both the statistic and the semantic properties of the remaining items. Four further items were deemed unnecessary for statistical and semantic reasons. The resulting 18-item form of the scale was used for the final analyses.

Rasch measurement was judged to be the most appropriate method of yielding answers to the following research questions:

Is the New 5-Point Rating Scale Psychometrically Optimal?

The effectiveness of the scale was investigated through the Rasch diagnostics suggested by Linacre (2002) and Bond and Fox (2001, 2007). All categories had high observed frequencies and thus an adequate number for stable estimates. Both the average person measure corresponding to each category and the threshold estimates monotonically increased and the ranges between successive thresholds were satisfactory rendering each category the most probable choice in an adequate range. Finally the outfit values of the categories were all close to the expected value of 1.00. All the evidence collected supports the hypothesis that categories function satisfactorily and thus five can be considered the optimal number of categories for the FLCAI.

Does the New Scale Provide Reliable Person Measures?

All the indices calculated reveal a high degree of reliability. The Person Reliability was .93 indicating that the scale can distinguish individuals very well. The Person Separation was 3.57 and Strata 5.09 confirming the good separation of persons along the FLCA continuum. Finally, the Item Reliability was .98 indicating that the items are well discriminated by the sample of respondents and this is a necessary condition for effective measurement (Wright & Masters, 1982).

Do the Scale Items Define a Single Construct?

A variety of evidence was collected to support the unidimensional structure of the scale. First, all point measure correlations were positive and highly significant (.60 to .77). Second, PCA of the standardised residuals revealed a strong measurement dimension explaining approximately 62% of the variation in the data. Furthermore, there was no significant second dimension present in the data (its eigenvalue was only 2.2 explaining only 4.6% of the total variation). Also the measurement dimension was 13.3 times stronger than that second dimension. Third, all items fit the Rasch model well. Only one item (Item 15) was marginal but it was not removed from the scale for two reasons; its misfit was caused by very few students responding unexpectedly to it and it was the easiest item in the scale, widening the coverage of the FLCA construct.

Fourth, two separate calibrations were performed on the 18 items from two distinct groups, male and female students. The item difficulties obtained from the two groups were statistically equivalent, showing that the property of invariance holds for this scale. That is, the construct has the same meaning across the two different groups of people.

Fifth, in agreement with Panayides and Walker (2013), the researchers showed that test anxiety is indeed a component of FLCA by separating the items into two groups, one containing the seven test anxiety items and the other the 11 remaining items. Person estimates were obtained from the two groups and these estimates were shown to be statistically equivalent, strengthening the belief that this scale indeed measures one construct and is unidimensional.

Finally, it was shown that the seven FLCAS items included in FLCAI measure the same construct as the 11 newly created items. Since the FLCAS is a well-established scale which has been validated many times, this result not only strengthens the hypothesis that FLCAI is unidimensional, but also that the dimension measured by the inventory is indeed FLCA.

How Does the New Scale Compare With the Original FLCAS?

Panayides and Walker's study (2013) facilitated direct comparisons between the FLCAS and FLCAI as data in both studies were analysed through the use of Rasch measurement with a time lapse of just one year. Panayides and Walker (2013) suggested that the very high reliability of the FLCAS was caused by the length of the scale, the use of parallel items and the narrow construct coverage. This study confirmed this since the new scale has just 55% the length of the original scale, containing 18 items rather than the 33, without losing any of its psychometric properties. In fact, the FLCAI is psychometrically superior to the FLCAS for the following reasons. First its degree of reliability is not lowered by the downsizing of the scale. Thus, a new shorter scale has been created without any loss in the degree of reliability from the original. The shortening of the instrument was beneficial as teachers and 'most researchers are constrained by another real-world factor: survey length' (Matthews, Kath, & Barnes-Farrell, 2010, p. 76). Second, it covers a much wider range of FLCA (2.81 logits as opposed to 1.44 logits). Third the categories function better and fourth, the unidimensional structure of the scale is more convincing with a much higher strength in the main dimension measured by it.

Despite achieving a wider spread of item difficulties and wider construct coverage than that in the FLCAS, the items of the FLCAI are well targeted for students with higher levels of FLCA. The mean item difficulty is, as always in Rasch analyses, zero whereas the mean person measure is well below at -1.40 logits. They suggest that the targeting of the items of the FLCAS can be explained by the fact that senior high school students in Cyprus have been studying English since early primary school, and they are therefore familiar with the language. Consequently their anxiety levels are well below that of high school or university beginner level students of English, who have been the participant population in many other studies in the literature. This study aligns itself with Panayides and Walker's (2013) stance, and suggests that should the FLCAI be used in a different setting, with less experienced students of English, or another foreign language, the mean item difficulty would probably be much closer to the mean person measure.

Limitations

The results reported here are very convincing and the researchers believe that the FLCAI will also prove worthy and appropriate for other EFL and foreign language learner populations where the students have been studying

English for more than 8-10 years. Notwithstanding, the validity of the scale cannot be taken for granted for any other population, just as it could not for the FLCAS. The FLCAI is a highly appropriate instrument for measuring FLCA among the Cypriot high school population.

Finally, even though every effort was made to provide an accurate equivalent for these items, slight semantic differences cannot be ruled out between the English and Greek versions of the FLCAS and the FLCAI.

Concluding Remarks

For the creation of the FLCAI Rasch analyses were complemented by semantic analyses. This was deemed necessary as this inventory is intended not only for use in further educational research but also, and perhaps more importantly, it is intended to be used by language teachers at the beginning of each academic year to assess their students' anxiety level. It was therefore of paramount importance to design an instrument that is as time economic as possible during both administration and analysis, without compromising validity, reliability or usefulness. It is also recommended that when teachers administer this instrument in the future they not only consider their students' total anxiety score but also what they find most anxiety provoking. This will lead to better student support.

The FLCAI was shown here to have a high degree of validity and reliability. However, further studies of the validity and appropriateness of the scale in diverse settings such as among university students and different countries are encouraged. Given the thoroughness of the methodology used, and the strengths of Rasch measurement, it is believed that the FLCAI will prove to be a valuable instrument maintaining its psychometric properties in other settings.

Funding

The authors have no funding to report.

Competing Interests

The authors have declared that no competing interests exist.

Acknowledgments

The authors have no support to report.

References

Aida, Y. (1994). Examination of Horwitz, Horwitz, and Cope's construct of foreign language anxiety: The case of students of Japanese. *The Modern Language Journal, 78*(2), 155-168. doi:10.1111/j.1540-4781.1994.tb02026.x

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME] (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573. doi:10.1007/BF02293814

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the social sciences*. Mahwah, NJ: Lawrence Erlbaum.

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

Chan, D., & Wu, G. (2004). A study of foreign language anxiety of EFL elementary school students in Taipei County. *Journal of National Taipei Teachers College, 17*(2), 287-320.

Cheng, Y.-s., Horwitz, E. K., & Schallert, D. L. (1999). Language anxiety: Differentiating writing and speaking components. *Language Learning, 49*(3), 417-446. doi:10.1111/0023-8333.00095

Curtis, D. D. (2004). Person misfit in attitude surveys: Influences, impacts and implications. *International Education Journal, 5*(2), 125-144.

Gehlbach, H., & Brinkworth, M. E. (2011). Measure twice, cut down error: A process for enhancing the validity of survey scales. *Review of General Psychology, 15*(4), 380-387. doi:10.1037/a0025704

Horwitz, E. K. (1986). Preliminary evidence for the reliability and validity of a foreign language anxiety scale. *TESOL Quarterly, 20*(3), 559-562. doi:10.2307/3586302

Horwitz, E. K. (2000). It ain’t over ’til it’s over: On foreign language anxiety, first language deficits, and the confounding of variables. *The Modern Language Journal, 84*, 256-259. doi:10.1111/0026-7902.00067

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal, 70*(2), 125-132. doi:10.1111/j.1540-4781.1986.tb05256.x

Horwitz, E. K., & Young, D. J. (1991). *Language anxiety: From theory and research to classroom implications*. Englewood Cliffs, NJ: Prentice Hall.

Krashen, S. D. (1981). *Second language acquisition and second language learning.* Retrieved from http://www.sdkrashen.com/content/books/sl_acquisition_and_learning.pdf

Levine, G. S. (2003). Student and instructor beliefs and attitudes about target language use, first language use, and anxiety: Report of a questionnaire study. *The Modern Language Journal, 87*(3), 343-364. doi:10.1111/1540-4781.00194

Linacre, J. M. (1997). *Guidelines for rating scales and Andrich thresholds* (MESA Research Note #2). Retrieved from http://www.rasch.org.rn2.htm

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement, 2*(3), 266-283.

Linacre, J. M. (2002). Understanding Rasch measurement: Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

Linacre, J. M. (2005). WINSTEPS Rasch measurement computer program (Version 3.65) [Computer software]. Chicago, IL: Winsteps.com.

Liu, H.-j. (2012). Understanding EFL undergraduate anxiety in relation to motivation, autonomy, and language proficiency. *Electronic Journal of Foreign Language Teaching, 9*(1), 123-139.

Liu, M., & Huang, W. (2011). An exploration of foreign language anxiety and English learning motivation. *Education Research International, 2011*, Article 493167. doi:10.1155/2011/493167

Luyt, R. (2012). A framework for mixing methods in quantitative measurement development, validation, and revision: A case study. *Journal of Mixed Methods Research, 6*(4), 294-316. doi:10.1177/1558689811427912

MacIntyre, P. D., & Gardner, R. C. (1989). Anxiety and second-language learning: Toward a theoretical clarification. *Language Learning, 39*(2), 251-275. doi:10.1111/j.1467-1770.1989.tb00423.x

MacIntyre, P. D., & Gardner, R. C. (1991). Investigating language class anxiety using the focused essay technique. *The Modern Language Journal, 75*(3), 296-304. doi:10.1111/j.1540-4781.1991.tb05358.x

MacIntyre, P. D., & Gregersen, T. (2012). Affect: The role of language anxiety and other emotions in language learning. In S. Mercer, S. Ryan, & M. Williams (Eds.), *Psychology for language learning: Insights from research, theory and practice* (pp. 103-118). New York, NY: Palgrave Macmillan.

Mahmood, A., & Iqbal, S. (2010). Difference of student anxiety level towards English as a foreign language subject and their academic achievement. *International Journal of Academic Research, 2*(6, Pt. 1), 199-203.

Matthews, R. A., Kath, L. M., & Barnes-Farrell, J. L. (2010). A short, valid, predictive measure of work–family conflict: Item selection and scale validation. *Journal of Occupational Health Psychology, 15*(1), 75-90. doi:10.1037/a0017443

Matsuda, S., & Gobel, P. (2004). Anxiety and predictors of performance in the foreign language classroom. *System, 32*(1), 21-36. doi:10.1016/j.system.2003.08.002

Messick, S. (1993). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104), Phoenix, AZ: American Council on Education and The Oryx Press.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Panayides, P. (2009). *Exploring the reasons for aberrant response patterns in classroom maths tests* (Doctoral Dissertation, Durham University, Durham, United Kingdom). Retrieved from http://etheses.dur.ac.uk/2042/

Panayides, P., Robinson, C., & Tymms, P. (2010). The assessment revolution that has passed England by: Rasch measurement. *British Educational Research Journal, 36*(4), 611-626. doi:10.1080/01411920903018182

Panayides, P., & Walker, M. J. (2013). Evaluating the psychometric properties of the Foreign Language Classroom Anxiety Scale for Cypriot senior high school EFL students: The Rasch measurement approach. *Europe's Journal of Psychology, 9*(3), 493-516. doi:10.5964/ejop.v9i3.611

Park, G., & French, B. (2013). Gender differences in the Foreign Language Classroom Anxiety Scale. *System, 41*(2), 462-471. doi:10.1016/j.system.2013.04.001

Pasquale, M. (2011). Folk beliefs about second language learning and teaching. *AILA Review, 24*, 88-99. doi:10.1075/aila.24.07pas

Ray, J. J. (1988). Semantic overlap between scale items may be a good thing: Reply to Smedslund. *Scandinavian Journal of Psychology, 29*, 145-147. doi:10.1111/j.1467-9450.1988.tb00784.x

Robinson, P. (Ed.). (2002). *Individual differences and instructed language learning.* Amsterdam, The Netherlands: John Benjamins.

Rodriguez, K. L., Schwartz, J. L., Lahman, M. K. E., & Geist, M. R. (2011). Culturally responsive focus groups: Reframing the research experience to focus on participants. *International Journal of Qualitative Methods, 10*(4), 400-417.

Schumacker, R. E., & Linacre, J. M. (1996). Factor analysis and Rasch analysis. *Rasch Measurement Transactions, 9*(4), 470. Retrieved from http://www.rasch.org/rmt/rmt94k.htm

Smith, E. V., Jr. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal components analysis of residuals. *Journal of Applied Measurement, 3*(2), 205-231.

Sparks, R. L., & Ganschow, L. (1991). Foreign language learning differences: Affective or native language aptitude differences? *The Modern Language Journal, 75*(1), 3-16. doi:10.1111/j.1540-4781.1991.tb01076.x

Sparks, R. L., & Patton, J. (2013). Relationship of L1 skills and L2 aptitude to L2 anxiety on the Foreign Language Classroom Anxiety Scale. *Language Learning, 63*(4), 870-895. doi:10.1111/lang.12025

Spielberger, C. D. (1983). *Manual for the stait-trait anxiety inventory (STAI-Form Y).* Palo Alto, CA: Consulting Psychologists Press.

Spielmann, G., & Radnofsky, M. L. (2001). Learning language under tension: New directions from a qualitative study. *The Modern Language Journal, 85*(2), 259-278. doi:10.1111/0026-7902.00108

Tóth, Z. (2008). A foreign language anxiety scale for Hungarian learners of English. *Working Papers in Language Pedagogy, 2*, 55-77.

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 from http://www.rasch.org/rmt/rmt83b.htm

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

Yan, J. X., & Horwitz, E. K. (2008). Learners' perceptions of how anxiety interacts with personal and instructional factors to influence their achievement in English: A qualitative analysis of EFL learners in China. *Language Learning, 58*(1), 151-183. doi:10.1111/j.1467-9922.2007.00437.x

Yun, J., & Ulrich, D. A. (2002). Estimating measurement validity: A tutorial. *Adapted Physical Activity Quarterly, 19*, 32-47.

Appendix

Appendix A

The data collected from the discussions with students and teachers, and the three interviews was subjected manually to theme analysis. The complete list of ideas and their frequencies can be seen below. These had a direct impact on the designing of items for inclusion in the extended item pool. The items in **bold** feature in the final version of the scale.

Tests	**19**
Oral participation	**10**
Grades	**7**
Being mocked if they make serious mistakes	**5**
Peer pressure / other students' reactions	**5**
Unclear instructions for listening tests	4
Essays / creative writing	**4**
Fear of making mistakes / reproduce incorrectly	**4**
Unclear instructions for written tests	3
Oral test	3
Homework – **unfamiliar tasks**, too much	3
Worry that they are not understood correctly	2
Teacher's reaction if they make a mistake	2
Teacher criticism	2
Talking about mistakes	2
Unknown vocabulary	2
Unfamiliar tasks in tests	**2**
Being interrupted whilst giving an answer	1
Threats from the teacher related to their grades	1
Working individually	1
Lack of self-confidence / sense of trust	1
Difficult concepts in the curriculum	1
Not knowing what is required	1
Listening tests	**1**
Amount of studying required for the lesson	1
When the teacher is unfriendly	1
When the teacher does not show understanding	1
Teacher not using exercises which could relax them in the lesson	1
Not being given encouragement	1
Fear of expressing their opinions	1
Teacher centred teaching	1
Technology	1
Sharing ideas	1
Their pronunciation	**1**
Fear that they haven't understood the teacher correctly	**1**
Not being well-prepared	1
Too much grammatical theory	1
Written exercises in the classroom	1
Double period	1
Lack of knowledge about the teacher	1

Appendix B: The Foreign Language Classroom Anxiety Inventory

The numbering on this inventory is that used for the 22-item scale. This is so as to avoid any potential confusion in references made to items throughout the study. New items are in **bold.**

Instructions: Below you will find a list of statements related to the feelings a student may experience during foreign language lessons. Please respond with √ or X in the box which best reflects how often these statement apply to you.

	Never	Rarely	Sometimes	Often	Always
2. In tests, I worry that I won't understand the vocabulary in the texts.					
3. I get nervous when the tasks are unfamiliar to me.					
5. I worry that my English teacher might ask me something that I won't understand.					
6. I am usually at ease during tests in my language class.					
7. I get anxious when the test has a listening component.					
8. I start to panic when I have to speak without preparation in language class.					
9. I am afraid I may mispronounce a word in front of the class.					
10. Thoughts of doing poorly interfere with my concentration in tests.					
11. Essays make me nervous.					
13. It embarrasses me to volunteer answers in my language class.					
14. During important tests I am so tense that it upsets my stomach.					
15. I worry about my grade.					
17. When we have an oral dialogue, I worry that I might not be able to understand what the other person is saying.					
18. The more I study for a language test, the more confused I get.					
19. I always feel that the other students speak the foreign language better than I do.					
20. During tests I find myself worrying about the consequences of failing.					
21. Language class moves so quickly I worry about getting left behind.					
22. I get nervous when I don't understand every word the language teacher says.					

About the Authors

Miranda Jane Walker holds a BA in Hispanic Studies and Modern Greek (King's College, University of London) a BA in English Language and Literature (University of Cyprus) and an MA in Education Leadership and Management (Open University, UK). She is currently an EdD candidate at the Open University, UK. She teaches Spanish in Secondary Education in Limassol, Cyprus. Her research interests include teacher and student motivation and anxiety in the foreign language classroom as well as educational leadership and management.

Panayiotis Panayides holds a BSc in Statistics with Mathematics (Queen Mary College, University of London), an MSc in Educational Testing (Middlesex University, UK) and a PhD in Educational Measurement (University of Durham, UK). He is an assistant headmaster and head of the Mathematics department in Secondary Education in Limassol, Cyprus. His research interests include educational and psychological measurement as well as research into Mathematics education.