【投稿論文】

A Confirmatory Approach to the Contested Factor Structure of the Foreign Language Classroom Anxiety Scale

Ian ISEMONGER* and Darren LINGLEY**

Abstract

Foreign language learning anxiety has emerged as an important explanatory construct for theorizing variation in foreign language learning outcome. It is conceptualized as situation specific and the Foreign Language Classroom Anxiety Scale (FLCAS) is the most influential measure in this area. Despite this influence, the construct formulation of the instrument is a source of confusion. While the original authorship advocated a single scale, subsequent researchers have adopted multi-scale models informed either by their own exploratory factor analyses (EFAs) or their particular reading of the seminal work itself. Much more recent work has used confirmatory factor analysis (CFA), and found the single-scale model implausible. The study reported in this paper extends the use of CFA, in adjudicating the single-scale hypothesis, to the Japanese population, and results were negative for this hypothesis.

Keywords: FLCAS, Foreign Language Classroom Anxiety Scale, Anxiety, Validity, SLA

MAIN ARTICLE

Research into anxiety and foreign language learning extends back to the 1970s. As early as the mid-1970s, and with respect to second language acquisition (SLA), Krashen (1976) had recognized anxiety as an affective filter compromising the comprehensibility of the linguistic input of the learner. Work conducted by Kleinmann (1977, 1978) brought the anxiety construct into association with avoidance behavior research and, by extension, with the issue of communicative strategy use. Early instrumentation in this area borrowed from adjunct areas of the behavioral sciences. For example, Kleinmann (1978), in language learning research, employed a translated version of the Achievement Anxiety Test (Alpert & Haber, 1960) which is claimed to measure debilitating and facilitating anxiety. This measure originated from abnormal and social psychological

* Kumamoto University
** Kochi University
高知人文社会科学研究第６号（2019）

research and has undergone continuing psychometric scrutiny (Plake, Smith, & Damsteegt, 1981; Watson, 1988) and research deployment (Carden, Bryant, & Moss, 2004; Wolf & Smith, 1995) subsequent to Kleinmann's use of the measure. Initial lines of instrumentation with regard to anxiety emerged principally through general counseling and clinical interests and sought to measure trait anxiety (e.g. Taylor, 1953) or both state and trait anxiety (e.g. Spielberger, Gorsuch, & Lushene, 1970).

Also emerging in the 1970s, however, were instruments targeted at situation-specific anxiety. For example, Richardson and Suinn (1972) developed an instrument that claimed to measure math-specific anxiety (Mathematics Anxiety Rating Scale), and Spielberger et al. (1980) developed an instrument for measuring test anxiety (Test Anxiety Inventory) – see also Sarason (1978) and McCroskey (1970) for instrumentation on test anxiety and communication anxiety, respectively. In this context, where the recognition of a distinction between trait anxiety and situation-specific anxiety was seen as an important theoretical step forward, it was not too long before a situation-specific view of language learning anxiety, i.e. second or foreign language learning anxiety, emerged in the form of the Foreign Language Classroom Anxiety Scale (FLCAS; Horwitz, 1986; Horwitz, Horwitz, & Cope, 1986). In fact, a French Class Anxiety Scale comprising five items had emerged earlier as part of the Attitudes and Motivation Test Battery (Gardner, Clement, & Smythe, 1979), but failed to find traction and fully establish itself within the literature. This instrument was cited by Horwitz et al. in 1986 and its limitations were identified as part of the context for their contribution in the form of the FLCAS.

The 1970s had seen contradictory results with regard to the impact of anxiety on foreign language achievement (Gardner & MacIntyre, 1993a; Scovel, 1978, 1991), and the FLCAS, as a situation-specific instrument, was designed to address deficits in instrumentation argued to be involved in these ambivalent results (Horwitz et al., 1986; MacIntyre, 1999). The instrument had a significant impact on foreign language learning and anxiety research with a number of early studies quickly emerging which indicated, or engaged with, the deleterious effects of anxiety on performance and achievement (e.g. Aida, 1994; Cheng, Horwitz, & Schallert, 1999; Gardner & MacIntyre, 1993b; Horwitz et al., 1986; MacIntyre & Gardner, 1991; Phillips, 1992; Young, 1986, 1991). The theoretical importance of language learning anxiety was also amplified by concomitant theoretical

developments in the area of willingness to communicate (WTC) which was presumed to have a negative association with anxiety. The notion of WTC was initially advanced in the area of communication theory more generally (e.g. McCroskey, 1977, 1992; Sallinen-Kuparinen, McCroskey, & Richmond, 1991) and later found theoretical purchase within SLA (e. g. Daly, 1991; MacIntyre, Dornyei, Clement, & Noels, 1998) to explain why learning outcome can vary so significantly.

Given that the FLCAS was rapidly adopted as the tool of preference for research into anxiety, attention also naturally turned to the psychometrics of scores produced by the instrument (e.g. Aida, 1994; Cheng et al., 1999; Liu & Jackson, 2008; Matsuda & Gobel, 2004) either as the major part of a study or as a subpart of the study-usually the latter. The methods adopted in this regard have typically included Exploratory Factor Analysis (EFA) or Cronbach's α or both of these. Typical of sequential studies employing EFA, the findings were often contradictory and the latent constructs which emerged did not correspond satisfactorily across studies. This may have been because the latent structure was indeed not stable across samples, but importantly, it may also have been because the method of EFA is predisposed to such variation in factor solutions. Thus, the problem of multiple solutions to the simple structure of the FLCAS may have been an artifact of the EFA method rather than a property of the scores generated by the instrument in each study. The case has been made that many of the analytical decisions involved in the execution of an EFA are inherently subjective and vary across studies (Henson, Capraro, & Capraro, 2004; Henson & Roberts, 2006). The researcher has to decide the number of factors to extract in the solution and there is more than one method to do this (e.g., the Eigenvalue greater-than-one rule, a parallel analysis, and inspection of a scree plot). The researcher, in arriving at simple structure, also has to set a threshold "loading" for inclusion of an item in a factor in the rotated solution, and these thresholds, even if stipulated *a priori*, which they should be, can vary across studies. Thus, although recommended practice and good judgment make the determination of the best solution more than an arbitrary affair, it would be fair to say that the procedure is subject to a level of indeterminacy. EFA is not, therefore, designed to confirm construct formulation and its use thus far as the primary method for dealing with the structure of the FLCAS should be seen as having produced results which are suggestive and potentially informative, but also potentially artifactual, and not

confirmatory.

Some specific examples of contradictions which have emerged with respect to EFAs conducted on scores generated by the FLCAS include studies conducted by Aida (1994), Cheng et al. (1999) and Matsuda and Gobel (2004). In the case of the first study conducted by Aida, four factors were extracted, via a varimax (orthogonal) rotation, and the threshold of the factor loading for inclusion of an item in a factor was set at .50. According to this criterion, six items were not included in the final model. The method for determining the number of factors to extract was inspection of the sums of square loadings in a rotated matrix for having a value greater than one. In the case of the second study conducted by Cheng et al., in authorship which overlaps with the original authors of the FLCAS, two factors were extracted (varimax rotation) and, similar to the Aida study, .50 was used as the threshold for inclusion of an item in a factor, but with the additional requirement that the item not have a secondary loading higher than .20. According to this criterion, 10 items were not included in the final model. The method for determining the number of factors to extract was initial inspection of a scree plot (Cattell, 1966) and then additionally examining extractions with numbers of factors one above and one below the number suggested by the scree plot with the overall criterion for adoption being interpretability of the solution. Finally, in the case of Matsuda and Gobel, two factors were extracted (varimax rotation). The threshold factor loading for inclusion of an item in a factor is not reported explicitly, but inspection of the tabulated results of the rotated factor solution (Table 2 in the report) indicates that the threshold must have been .30. Only two items were not included in the model which is lower than the number not included in the study reported by Cheng et al., for example, but the threshold was more permissive (.30) than that of Cheng et al. (.50). The criterion used for selecting the number of factors to extract was inspection of a scree a plot. The EFA conducted by Liu and Jackson (2008) is not covered here, because the decision sequences for the EFA were under-reported and items were added to the instrument.

Notably, the findings from the above three independently conducted EFAs are different, and more importantly, this may be explained by the variation in the decision sequences adopted for the analysis in each case. The method of Confirmatory Factor Analysis (CFA) on the other hand, which has emerged in the literature later than EFA (and which may explain the decision to use EFA in the three studies covered above),

overcomes many of the difficulties with respect to determinacy evident in the method of EFA. We do not suggest that all issues of determinacy are overcome by CFA, because fit indexes for tested models are interpreted on a continuum, and while cutoffs are empirically informed (e.g. Hu & Bentler, 1999), there is a level of subjective judgment involved in interpreting the returned value for an index against a cutoff. However, CFA does afford an *a priori* approach to a hypothesized model, and the issues of determinacy are not the same as those which attend EFA which is essentially a post hoc form of analysis. Thus, CFA is more disposed to examining the psychometrics of scores generated by an instrument when 1) there is an *a priori* theoretical model underpinning the instrument, either explicitly stated by the author/s of an instrument or implicitly stated via the scoring regime for the instrument, which requires confirmation 2) when there are competing claims surrounding the componential structure or construct formulation of an instrument which require empirical arbitration, and 3) when unidimensionality of scores is an important interpretive assumption for use of the instrument (which is almost always the case and which is a matter for the routine process of normal science).

With regard to the first point, the FLCAS was presented to the wider research and practitioner community with a scoring regime advocating treatment of the instrument as a single scale (Horwitz, 1986; Horwitz et al., 1986; Personal Communication, August, 2006). This necessarily presupposes a measurement model for the instrument comprising a single latent construct which should include all 33 items. The most appropriate method for direct engagement with the empirical plausibility of this model is CFA. With regard to the second point, it is notable that the single-scale measurement model for the FLCAS has been attended by some ambivalence in the literature and there are competing claims. This is because the seminal article through which the instrument entered the literature stated that the items making up the instrument were "reflective of communication apprehension, test-anxiety, and fear of negative evaluation in the foreign language classroom" (Horwitz et al., 1986, p. 129). This has led to a perception among subsequent researchers engaging with scores generated by the instrument that there is a sub-componential measurement model for the instrument. The problem of ambivalence is further exacerbated by the fact that subsequent EFAs undertaken (e.g. Aida, 1994; Cheng et al., 1999; Matsuda & Gobel, 2004) have produced

factor solutions which were also sub-componential − though the number of, and labels ascribed to, these components vary and do not directly mirror communication apprehension, test-anxiety, and fear of negative evaluation. These studies exemplify the confusion because they necessarily imply the abandonment of the single-scale hypothesis in each case, and yet this critical implication is also discursively neglected in each case.

Horwitz (2010), in a subsequent intervention in the literature, reasserted the position that the 1986 seminal article was misinterpreted as implying multiple components when she was in fact implying that the items were merely related to communication apprehension, test-anxiety, and fear of negative evaluation and not composed of them (see discussion section below). This scholarly, rather than empirical, reprise of the case for a single non-componential scale for the FLCAS is not reconciled with Cheng et al. (1999), where there is overlapping authorship with Horwitz (2010), and where the abandonment of the single scale is discursively omitted though being an empirical implication of the results of the EFA. The reprise is also not reconciled with any of the other EFAs (Aida, 1994; Matsuda & Gobel, 2004) suggesting sub-componential simple structures.

Research conducted far more recently by Park (2012, 2014) has employed CFA, and the results obtained, represent a significant empirical intervention in the confusion surrounding the scale structure for the FLCAS. In particular, the 2012 study (N = 918) conducted by Park directly tested the single-scale hypothesis for the FLCAS in the Korean population, and the model was rejected. Other models with multiple factors were tested, and these produced better results than the one-factor model, but nonetheless, still not satisfactory. These results do not empirically support the position of the original author (Horwitz, 2010) that the instrument is underpinned by a single scale. The 2014 study by Park tested a two-factor model in a sample (N = 244) using CFA as the method; with this two-factor model having been suggested, in a separate sample (N = 217), by an EFA. According to Park the two-factor model was, in general, acceptable under the CFA analysis. However, it is possible to dispute this interpretation. For example, Park used a cutoff for the CFI of (> .90), and this cutoff is not aligned with Hu and Bentler (1999), but rather with a more relaxed threshold, also associated with Hu and Bentler in earlier work (1998), and which was adjusted by these authors to be more

stringent (> .95) in the 1999 paper. Nonetheless, the results obtained by Park, while open to interpretation with respect to being acceptable, were certainly not poor. What is more notable, however, is that these results were not obtained for the single factor model. Although it would have been empirically helpful to have also tested the single-factor model in the 2014 study, overall it is clear that Park's work does not support a single-factor model for the instrument.

Park's (2014) study prompted a response from the original author of the instrument (Horwitz, 2016), in which she appeared to argue that results which depart from a single factor model in the Korean population could be due to population variance. This is the first time, to date, that the claim for a single-factor scale has been qualified as possibly population variant, rather than population invariant. It is impossible to empirically check this claim, because the FLCAS has not undergone a CFA in the original North American population, which would be the baseline for comparison against results from CFAs in other populations. Horwitz's original claims for a single scale are premised on data presented with the introduction of the instrument, including Cronbach's α and correlation coefficients with concurrent instruments, which do not demonstrate unidimensionality. The evidence for the single-scale claim is simply insufficient in the original population, and therefore claims concerning departures from the single-scale hypothesis in non-North American populations being explained by population variance are empirically premature. Measurement invariance studies, which would be the next step after empirical confirmation of a single-factor model in the original population, have also not been conducted. In this context, the work conducted by Park (2012) remains the best evidence so far that simple structure for this instrument does not reduce to a single unidimensional scale, and that the instrument comprises more than one scale, in spite of its otherwise conception.

In this context, the purpose of the research reported in this paper was to use the method of CFA to submit the single-factor hypothesis to an *a priori* test using a Japanese translation of the FLCAS, and thus extend this line of research into the plausibility, or perhaps implausibility, of the single-scale model in further populations. An empirical approach to the dimensionality of scores generated by the FLCAS involving confirmation has epistemic authority over both the empirical (but exploratory) results from EFAs, and the non-empirical scholarly positions concerning the originally-

intended construct formulation of the instrument.

Method

The FLCAS takes the form of 33 self-report items which, according to the authorship, comprise a single scale. Each item uses a five-point Likert scale with the following semantic anchors: Strongly Agree (5), Agree (4), Neither Agree nor Disagree (3), Disagree (2) and Strongly Disagree (1).

The FLCAS was translated (with permission) into Japanese. Although work had been reported in the literature indicating that Japanese translations were already in use (e.g. Matsuda & Gobel, 2004), it was translated again so that the process could be controlled and so that back translation could become an explicit part of the translation process – thus conforming to the guidelines of the International Test Commission (ITC; International Test Commission, 2001).

Participants and Procedure

The sample was collected, for the purpose of this study, from a national university in Japan and comprised 1407 freshmen who participated under informed consent. There were 117 cases where one or more items had not been responded to by the participant. Inspection indicated no systematic loss of data and thus the approach was to remove all these cases leaving a final dataset of 1290 cases.

With respect to age, there were 29 non-responses. The age range was from 18 years to 25 years with 93 percent of respondents being 18 or 19. With respect to gender, there were 26 non-responses, and of those who responded 584 (46%) were male and 680 (54%) were female.

Results

Results are reported in terms of two analytical procedures, namely, Cronbach's α (including the 95% confidence intervals for α) and CFA. It is important to note that CFA is the analytically more powerful tool and is the essential component of the analysis.

Cronbach's α does not demonstrate unidimensionality of scores and the rationale for reporting it is essentially to replicate past statistical reporting.

Reliability Analysis

The value for Cronbach's α was .93 and this is relatively consistent with Aida (.94; 1994) and Cheng et al. (.95; 1999); but not consistent with Matsuda and Gobel (.78; 2004). It is also relatively consistent with Park (.94, 2012). The 95% confidence interval (Fan & Thompson, 2001) was very narrow with the lower bound being .928 and the upper bound being .938. Cronbach's α was derived for all 33 items on a single scale consistent with the measurement model implied by the original authors (Horwitz et al., 1986).

Confirmatory Factor Analysis

A CFA of the single-factor unidimensional model (Horwitz et al., 1986) was directly tested. Given the ambivalence in the literature around a componential structure for the instrument, the possibility of a rival model (three-factor) specified to represent communication apprehension, test-anxiety, and fear of negative evaluation was considered. Horwitz et al. (p. 127) claimed that these three performance anxieties were related to foreign language anxiety, and Horwitz (2010, p. 158) reasserted this by claiming that foreign language anxiety was related to, but not composed of, them. There has been no explicit specification, by the authors, of items in the FLCAS operationalizing the three performance anxieties although inspection of the FLCAS clearly shows items which resonate with them. However, some items are difficult to align with any of the three aspects and some are arguably aligned with more than one of them. The rival model was therefore abandoned due to indeterminacy of specification, and only the model explicitly argued for by Horwitz et al. (1986) and Horwitz (Personal Communication, August, 2006) was directly tested; this being the single-scale or one-factor model.

The model was specified so that all items would indicate a single latent factor (Foreign Language Classroom Anxiety). The loading for one indicator was fixed to a value of 1 as a metric for the scale. The model comprised 561 distinct sample moments,

66 free parameters and 495 degrees of freedom meeting the criterion of overidentification. Univariate non-normality was a significant property of the data. The critical ratio for skew exceeded a pre-selected threshold of 3.0 in all but nine items, and the critical ratio for kurtosis exceeded the same threshold of 3.0 in all but 13 items (See Table 1 below). Furthermore, the value for the critical ratio for multivariate normality (Mardia's coefficient) was also high (99.0). These properties of the data threaten the normal-theory assumptions made by most likelihood (ML) estimation, and thus results should be considered with care. Mean- and variance-adjusted weighted least squares (WLSMV) estimation was considered as an alternative estimator. However, while some empirical work has occurred on the empirical properties of this estimator (e.g. Yu, 2002), it is still considered experimental and requires further testing (Byrne, 2012). Also, the properties of the ML estimator under violation of normal-theory assumptions are well reported (e.g. Hu & Bentler, 1999). In addition, the method of bootstrapping (Bollen & Stine, 1993) was adopted in this study as a further method in the sequential analysis of the data in order to specifically assist with the non-normal properties of the data.

Table 1

Content, Means, Standard Deviations, Skew (CR), Kurtosis (CR) and Loadings (Standardized Regression Weights) for Items Comprising the FLCAS

Item	Content	Mean	SD	Skew	Kurtosis	Loading
Item 01	I never feel quite sure of myself when I am speaking in my foreign language class.	3. 95	0. 97	-11. 53	0. 49	0. 58
Item 02	I don't worry about making mistakes in language class.	3. 20	1. 17	-3. 40	-6. 93	0. 50
Item 03	I tremble when I know that I'm going to be called on in language class.	3. 02	1. 23	-1. 73	-7. 67	0. 63
Item 04	It frightens me when I don't understand what the teacher is saying in the foreign language.	2. 74	1. 19	2. 27	-7. 28	0. 60
Item 05	It wouldn't bother me at all to take more foreign language.	3. 40	1. 17	-4. 93	-5. 46	0. 33
Item 06	During language class, I find myself thinking about things that have nothing to do with the course.	3. 10	1. 15	-2. 13	-6. 57	0. 18
Item 07	I keep thinking that the other students are better at languages than I am.	3. 87	1. 15	-14. 00	1. 09	0. 37
Item 08	I am usually at ease during tests in my language class.	3. 63	1. 13	-8. 41	-3. 28	0. 53
Item 09	I start to panic when I have to speak without preparation in language class.	3. 96	1. 05	-13. 97	2. 14	0. 65
Item 10	I worry about the consequences of failing my foreign language class.	4. 04	1. 10	-17. 19	4. 88	0. 50
Item 11	I don't understand why some people get so upset over foreign language classes.	4. 01	0. 99	-12. 47	0. 91	0. 51
Item 12	In language class, I can get so nervous I forget things I know.	3. 56	1. 07	-8. 36	-2. 55	0. 57
Item 13	It embarrasses me to volunteer answers in my language class.	3. 58	1. 10	-8. 64	-2. 32	0. 57
Item 14	I would not be nervous speaking in the foreign language with native speakers.	3. 78	1. 08	-10. 93	-0. 81	0. 55
Item 15	I get upset when I don't understand what the teacher is correcting.	3. 59	1. 05	-8. 51	-2. 09	0. 59
Item 16	Even if I am well prepared for language class, I feel anxious about it.	3. 55	1. 10	-8. 00	-3. 51	0. 61
Item 17	I often feel like not going to my language class.	2. 89	1. 27	1. 47	-7. 64	0. 54
Item 18	I feel confident when I speak in foreign language class.	4. 01	0. 87	-11. 37	3. 82	0. 58
Item 19	I am afraid that my language teacher is ready to correct every mistake I make.	3. 02	1. 18	0. 03	-7. 14	0. 65
Item 20	I can feel my heart pounding when I'm going to be called on in language class.	3. 55	1. 17	-8. 01	-4. 49	0. 71
Item 21	The more I study for a language test, the more confused I get.	2. 54	1. 04	6. 45	-1. 70	0. 49
Item 22	I don't feel pressure to prepare very well for language class.	3. 35	1. 09	-4. 45	-4. 53	0. 50
Item 23	I always feel that the other students speak the language better than I do.	3. 62	1. 21	-8. 89	-3. 95	0. 37
Item 24	I feel very self-conscious about speaking the foreign language in front of other students.	3. 76	1. 04	-9. 32	-1. 77	0. 56
Item 25	Language class moves so quickly I worry about getting left behind.	3. 34	1. 21	-3. 48	-6. 82	0. 65
Item 26	I feel more tense and nervous in my language class than in my other classes.	3. 19	1. 20	-2. 24	-6. 57	0. 77
Item 27	I get nervous and confused when I am speaking in my language class.	3. 14	1. 14	-0. 89	-5. 97	0. 76
Item 28	When I'm on my way to language class, I feel very sure and relaxed.	3. 76	0. 98	-7. 81	-0. 91	0. 62
Item 29	I get nervous when I don't understand every word the language teacher says.	3. 43	1. 21	-5. 35	-6. 66	0. 51
Item 30	I feel overwhelmed by the number of rules you have to learn to speak a foreign language.	3. 13	0. 95	0. 01	0. 27	0. 41
Item 31	I am afraid that the other students will laugh at me when I speak the foreign language.	2. 94	1. 16	0. 12	-6. 24	0. 61
Item 32	I would probably feel comfortable around native speakers of the foreign language.	3. 61	1. 07	-6. 25	-3. 02	0. 36
Item 33	I get nervous when the language teacher asks questions which I haven't prepared in advance.	3. 84	1. 00	-12. 25	1. 63	0. 64

Model fit was assessed using the χ^2 which has a tendency to over-reject with large samples and in data with non-normal properties (Byrne, 2001; Hu & Bentler, 1999), and both of these conditions were arguably met in the dataset for this study. Thus the analysis was assisted with a combination of fit indexes (and associated cut-off values)

recommended by Hu and Bentler (1999). These cutoffs were empirically derived to minimize both Type I and Type II error.

The χ^2 produced a significant result (χ^2 = 3190.307, *df* = 495, p < 0.01). In the logic of CFA, the reverse logic of traditional inferential statistics for group differences on the dependent variable, this means that the model is rejected because the dimensionality of the data departs significantly from the model against which it was tested. However, as stated immediately above, χ^2 tends to over-reject in larger samples due to excessive statistical power. Thus the results for the additional procedures involving fit indexes are important and were as follows (cutoffs recommended by Hu and Bentler [1999] in parentheses): RMSEA, .065 (< .06); TLI, .816 (> .95); CFI, .828 (> .95); and SRMSR .052 (< .08). It is important to note that these indexes are interpreted on a continuum unlike the χ^2 value which is a test statistic, and which is interpreted in absolute terms using the associated probability threshold (decided in advance). The values for the TLI and CFI were well below the threshold of .95 and would have to be taken as emphatic evidence against the model. The value for the RMSEA is very close to satisfying the threshold and so this is not emphatic negative evidence, but it is not positive evidence for model fit either. The value for the SRMSR, which is essentially an expression of the residuals, was well within the threshold. The p value for the Bollen and Stine (1993) bootstrap procedure (1000 bootstrap samples), which copes with non-normally distributed scores, was . 001 indicating a significant result and rejection of the hypothesized model.

The factor loadings (Standardized Regression Weights) for each item are available for inspection in Table 1. Items 05, 06, 07 and 23 were interpreted as problematic for having a coefficient of less than .40. However, a further test of the model with these items removed did not provide any appreciable improvement in fit. The values for the fit indexes were as follows (cutoffs recommended by Hu and Bentler [1999] in parentheses): RMSEA, .063 (< .06); TLI, .854 (> .95); CFI, .864 (> .95); and SRMSR .047 (< .08). The RMSEA and the SRMSR values were still good, but the TLI and CFI values only increased by a trivial amount, and remained well below the thresholds for an acceptable value.

Discussion

Turning first to the result for Cronbach's α which was high (.93), and which is reported here for replicative purposes given that this index has been used routinely in past literature, the result was positive. However, there are analytical limitations with respect to this index. The index cannot be used to test whether the FLCAS comprises a single scale, because the index does not demonstrate unidimensionality of scores. It is an index of reliability and presumes rather than demonstrates unidimensionality of scores (see, Cortina, 1993; Green, Lissitz, & Mulaik, 1977). Its other potential limitation is that the value derived for the index is predisposed to being higher when the number of observables (i.e. items on the scale) is higher; and there are 33 items on the scale which is high by almost any standard. This limitation is particularly unhelpful if items are operationally repetitive rather than operationally extensive or diverse. The limitations of the index have invited a reappearance of its critique in more recent literature (Bentler, 2009; Green & Yang, 2009a, 2009b; Revelle & Zinbarg, 2009; Sijtsma, 2009a, 2009b). One notable feature of the result for this index, however, is that the value derived in this study is closely aligned with Aida (1994) and Cheng et al. (1999) which involved non-Japanese samples, and not at all closely aligned with Matsuda and Gobel (2004) which did involve a Japanese sample. The sample size (N=252) in the Matsuda and Gobel study was significantly smaller than in this study (N=1290), although the sample also represented freshmen at a Japanese university. The translations, which are different, is another possible area of explanation but suffice to state that these results invite further research in the future.

With reference to the CFA, despite the marginal case for the RMSEA and strong case for the SRMSR, the interpretation placed on these results, when considered in triangulation, is that there is not satisfactory empirical evidence for the single-factor model as a plausible solution to the dimensionality of the data. Hu and Bentler (1999) intended this combination of indexes to be used together, or in triangulation, meaning that a satisfactory result is required on all of them to minimize both Type I and Type II error. The results for both the CFI and the TLI were significantly below the threshold of .95 and a positive interpretation is not possible here. The RMSEA as an index is noted for its rewarding for model parsimony and this may have contributed to the relatively

better result on this index when compared with the TLI and CFI, because clearly the single-factor model for the FLCAS is the most parsimonious expression of the dimensionality of scores possible; nonetheless, the result remains only relatively better and is not meritorious. The SRMR indicates the size of residuals which were relatively small, but this by itself is insufficient for claiming plausible fit of the model to the data. Also, and beyond the indexes, the result for the Bollen and Stein (1993) bootstrap procedure, which analytically accommodated for evidence of non-normal distributions in the data, indicated rejection of the model. Finally, the value for the χ^2 and its associated probability level indicated rejection of the model, although the limitations of this test statistic, given the proclivity towards excessive statistical power and over-rejection of models, should be noted.

In discussing and situating the results from this study and their associated bearing on the issues, it is important to reiterate and elaborate the principle psychometric question which surrounds the FLCAS; and this concerns whether the instrument should be treated as a single scale or should be treated as being comprised of sub-components or sub-scales. The question exists because enough authors have treated the instrument as sub-componential, or have found it to be empirically sub-componential (usually through EFAs), to have prompted a correction or clarification by Horwitz (2010) which reasserts the original claim that the instrument was, and is, underpinned by a single scale and associated scoring regime. The confusion surrounding the issue was also enough to prompt subsequent research by Park (2012, 2014), and the 2012 paper, especially, empirically drew into question the single-scale claim for the instrument, and quite directly so, because this was the first use of CFA in the literature in engaging with this problem (to the best knowledge of the current authors). The 2012 paper also includes a direct test of the single-scale model, for which the evidence was negative.

The 2010 correction by Horwitz, where the single-scale claim was reasserted, came despite overlap in authorship between the correction itself (2010) and another paper (Cheng et al., 1999) which supported an EFA-inferred, sub-componential measurement model. It came in a research-timeline article (Horwitz, 2010) which cited the Horwitz et al. (1986) seminal paper and stated the following with respect to it:

> Often credited with introducing the construct of FLA (foreign language anxiety) as a situation-specific anxiety, Horwitz et al. [1986] discusses the

> ego-threatening nature of language learning and includes the Foreign Language Classroom Anxiety Scale (FLCAS), which has become the standard measure of language anxiety. The authors identify three related situation-specific anxieties − communication apprehension (CA), fear of negative evaluation (FNE), and test anxiety (TA) − to help language teachers and scholars understand the anxiety-provoking potential of language learning. This article has sometimes been misinterpreted to mean that FLA is composed of CA, FNE, and TA rather than as simply being related to them. (p. 158, our underline)

We argue that there is a lot resting in the semantic distinction made here between foreign language anxiety (FLA) being "composed of" and being "related to" these three performance anxieties of communication apprehension, fear of negative evaluation and test anxiety. The distinction is fundamental and requires specific elaboration in order to stand as an empirically testable claim, and many of the items in the FLCAS do in fact bear strong operational resemblances to these three performance anxieties, despite having never been explicitly hypothesized as operationalizing them. Furthermore, claims with respect to FLA as a theoretical notion should be logically distinguished from claims with respect to the measurement model for the FLCAS which is a measurement device presumed to be theoretically informed, and this distinction is somewhat lost in the clarification. Although, in principle, the measurement model for an instrument should be aligned with theory, it may not be; and while a measurement model might be theoretically informed by intent, the empirical presence of such theory in the dimensionality of the scores generated by an instrument is ultimately a matter for empirical confirmation and cannot be arbitrated via scholarly interchanges. In other words, there may be empirical misalignment with the theory purported to inform the construct formulation of an instrument. It is relatively straight forward to accumulate empirical evidence for the dimensional properties of scores generated by an instrument and, in fact, non-confirmatory evidence in the form of EFAs has already accumulated in significant measure, and this evidence suggests that scores generated by the FLCAS do not reduce to a single latent construct; and it is simply that the implications of this have not satisfactorily been absorbed by researchers in the area. This non-confirmatory evidence via EFAs invited a confirmatory approach, and recent work by Park (2012,

2014) has responded to this invitation, with the results being negative for the single-scale claim. The results obtained in this study corroborate those of Park (2012), but in the Japanese population.

However, having made the case for the necessity of a confirmatory approach, it should be cautioned that a single confirmatory study is contributive rather than definitive. Validity evidence is cumulatively acquired (Messick, 1989) and for more secure conclusions to be drawn, studies across populations and associated translations need to be conducted. We recommend that available datasets for both the original English-language form of the instrument, and for other translations into other languages, be revisited under a direct test of the single-factor hypothesis, if advocacy for scoring the test under a single scale is to be maintained by its authorship. This is especially important given the recent claims Horwitz (2016) has made, in responding to Park (2014), with respect to the possibility of the factor-structure of the instrument being population dependent. This may well be the case, but the outstanding issue is that claims for a single factor in the primary population (North American) for which the instrument was developed are absent empirical evidence, notwithstanding reported Cronbach's α values and correlation coefficients with concurrent measures.

Thus far, theoretical specification of an alternative componential measurement model for the instrument has not occurred explicitly by the authorship of the instrument, but it would have to if single-scale advocacy was abandoned; and it should be abandoned if cumulative confirmatory studies reject the single-factor hypothesis. Given the operational resonance of some items with the performance anxieties of communication apprehension, fear of negative evaluation and test anxiety, one would presume these three performance anxieties to be one avenue for an alternative, multi-factor specification of the measurement model for the FLCAS. However, as stated above, these resonances are just that, and there is no clear one-to-one mapping of items to factors that the independent researcher can securely discern. Nonetheless, in the case of the negative evidence reported in this study for the one-factor model hypothesized by the authors of the FLCAS, the outcome for the practitioner remains clear; the claim that the FLCAS should be scored on a single scale should be treated with skepticism until further empirical evidence can support this; scholarly arguments and claims to population variance notwithstanding.

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*, 155-168.

Alpert, R., & Haber, R. N. (1960). Anxiety in academic achievement situations. *Journal of Abnormal and Social Psychology*, *61*, 207-215.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, *74*(1), 137-143.

Bollen, K. A., & Stine, R. A. (1993). Bootstrapping goodness-of-fit measures in structural equation modeling. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 115-135). Newbury Park, CA: Sage.

Byrne, B. M. (2001). *Structural equation modeling with AMOS*. London: Lawrence Erlbaum Associates.

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York: Routledge.

Carden, R., Bryant, C., & Moss, R. (2004). Locus of control, test anxiety, academic procrastination, and achievement among college students. *Psychological Reports*, *95* (2), 581-582.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245-276.

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

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98-104.

Daly, J. (1991). Understanding communication apprehension: An introduction for language educators. In E. K. Horwitz & D. Young (Eds.), *Language anxiety from theory and research to classroom implications* (pp. 3-15). Hemel Hempstead: Prentice Hall International.

Fan, X., & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial. *Educational and Psychological Measurement*, *61* (4), 517-531.

Gardner, R. C., Clement, R., & Smythe, C. C. (1979). *Attitudes and motivation test battery*

(rev manual). London, Ontario: University of Western Ontario, Dpt. of Psychology.

Gardner, R. C., & MacIntyre, P. D. (1993a). On the measurement of affective variables in second language learning. *Language Learning*, *43*, 157-194.

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

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of Coefficient Alpha as an Index of Test Unidimensionality. *Educational and Psychological Measurement*, *37*, 827-838.

Green, S. B., & Yang, Y. (2009a). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, *74*(1), 121-135.

Green, S. B., & Yang, Y. (2009b). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, *74*(1), 155-167.

Henson, R. K., Capraro, R. M., & Capraro, M. M. (2004). Reporting practices and use of exploratory factor analyses in educational research journals: Errors and explanation. *Research in the Schools*, *11*(2), 61-72.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, *66*, 393-416.

Horwitz, E. K. (1986). Preliminary evidence for the reliability and validity of a foreign language anxiety scale. *TESOL Quarterly*, *20*(3), 559-564.

Horwitz, E. K. (2010). Foreign and second language anxiety. *Language Teaching*, *43*(2), 154-167.

Horwitz, E. K. (2016). Factor structure of the Foreign Language Classroom Anxiety Scale: Comment on Park (2014). *Psychological Reports*, *119*(1), 71-76.

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal*, *70*(2), 125-132.

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424-453.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1-55.

International Test Commission. (2001). *International Test Commission guidelines for test*

adaptation. London: Author.

Kleinmann, H. H. (1977). Avoidance behavior in adult second language learning. *Language Learning*, *27*, 93-101.

Kleinmann, H. H. (1978). The strategy of avoidance in adult second language acquisition. In W. Ritchie (Ed.), *Second language acquisition research: Issues and implications* (pp. 157-174). London: Academic Press.

Krashen, S. D. (1976). Formal and informal linguistic environments in language acquisition and language learning. *TESOL Quarterly*, *10*(2), 157-168.

Liu, M., & Jackson, J. (2008). An exploration of Chinese EFL learners' unwillingness to communicate and foreign language anxiety. *Modern Language Journal*, *92*(1), 71-86.

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: McGraw-Hill.

MacIntyre, P. D., Dornyei, Z., Clement, R., & Noels, K. A. (1998). Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence and affiliation. *The Modern Language Journal*, *82*, 545-562.

MacIntyre, P. D., & Gardner, R. C. (1991). Anxiety and second language learning: Toward a theoretical clarification. In E. K. Horwitz & D. Young (Eds.), *Language anxiety from theory and research to classroom implications* (pp. 41-56). Hemel Hempstead: Prentice Hall International.

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

McCroskey, J. C. (1970). Measures of communication bound anxiety. *Speech Monographs*, *37*, 269-277.

McCroskey, J. C. (1977). Oral communication apprehension: A summary of recent theory and research. *Human Communication Research*, *4*(1), 78-96.

McCroskey, J. C. (1992). Reliability and validity of the willingness to communicate scale. *Communication Quarterly*, *40*(1), 16-25.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Park, G. P. (2012). Investigation into the constructs of the FLCAS. *English Teaching*, *67*,

207-220.

Park, G. P. (2014). Factor analysis of the foreign language classroom anxiety scale in Korean learners of English as a foreign language. *Psychological Reports*, *115*(1), 261-275.

Phillips, E. M. (1992). The effects of language anxiety on students' oral test performance and attitudes. *The Modern Language Journal*, *76*, 14-26.

Plake, B. S., Smith, E. P., & Damsteegt, D. C. (1981). Validity investigation of the Achievement Anxiety Test. *Educational and Psychological Measurement*, *41*, 1215-1222.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the GLB: Comments on Sijtsma. *Psychometrika*, *74*(1), 145-154.

Richardson, F. C., & Suinn, R. M. (1972). The Mathematics Anxiety Rating Scale. *Journal of Counseling Psychology*, *19*, 551-554.

Sallinen-Kuparinen, A., McCroskey, J. C., & Richmond, V. P. (1991). Willingness to communicate, communication apprehension, introversion, and self-reported communication competence: Finnish and American comparisons. *Communication Research Reports*, *8*, 55-64.

Sarason, I. G. (1978). The test anxiety scale: Concept and research. In C. D. Spielberger & I. G. Sarason (Eds.), *Stress and anxiety* (Vol. 5, pp. 193-216). Washington, DC: Hemisphere.

Scovel, T. (1978). The effect of affect on foreign language learning: A review of the anxiety research. *Language Learning*, *28*, 129-142.

Scovel, T. (1991). The effect of affect on foreign language learning: A review of the anxiety research. In E. K. Horwitz & D. Young (Eds.), *Language anxiety from theory and research to classroom implications* (pp. 15-23). Hemel Hempstead: Prentice Hall International.

Sijtsma, K. (2009a). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107-120.

Sijtsma, K. (2009b). Reliability beyond theory and into practice. *Psychometrika*, *74*(1), 169-173.

Spielberger, C. D., Gonzalez, H. P., Taylor, C. I., Anton, E. D., Algaze, B., Ross, G. R., & Westberry, L. G. (1980). *Manual for the Test Anxiety Inventory*. Redwood City, CA: Consulting Psychologists Press.

Spielberger, C. D., Gorsuch, R., & Lushene, R. (1970). *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologist Press.

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

Watson, J. M. (1988). Achievement Anxiety Test: Dimensionality and utility. *Journal of Educational Psychology*, *80*(4), 585-591.

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*(3), 227-242.

Young, D. J. (1986). The relationship between anxiety and foreign language oral proficiency ratings. *Foreign Language Annals*, *12*, 439-448.

Young, D. J. (1991). The relationship between anxiety and foreign language oral proficiency ratings. In E. K. Horwitz & D. Young (Eds.), *Language anxiety from theory and research to classroom implications* (pp. 57-64). Hemel Hempstead: Prentice Hall International.

Yu, C. U. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. (Doctor of Philosophy), University of California, Los Angeles.

Received: 13 September 2018
Accepted: 16 November 2018

（イアン・アイズマンガー　熊本大学）
（ダレン・リングリー　高知大学）