Current Psychology (2022) 41:2250–2263
https://doi.org/10.1007/s12144-020-00610-w

Psychometric properties of TAS, TAI, FAT test anxiety scales 6 in Chinese university students: a Bifactor IRT study

Fengfeng Xu[1] · Yan Cai[1] · Dongbo Tu[1]

Published online: 23 April 2020

Abstract

In this study, the psychometric properties of three commonly used rating scales of test anxiety were examined, including the test anxiety inventory (TAI), the test anxiety scale (TAS) and the Friedman-Bendas Test Anxiety Scale (FAT). Under the framework of item response theory (IRT), the Bifactor multi-dimensional item response model was employed to compare the psychometric properties of the three scales. Results showed that the Bifactor structures were suitable for the three scales, which were then used in the subsequent Bifactor multidimensional item response theory analysis. Although the three commonly used TA scales were likely to measure the same underlying construct—test anxiety, they had very different psychometric properties. The findings of the Bifactor Multi-IRT provided suggestions for determining which scale to use in a given study design: the TAI and the FAT evaluated information at greatly overlapping ranges; however, the TAI, performing a litter better at the same levels of severity of TA, may be a good choice when we recruit those with various levels of TA severity to ensure a high precision. What's more, FAT may be a good choice for measuring those with moderate TA severity. Meanwhile, the TAS provided more information at the lower level of TA symptomatology, which was to say, TAS was more suitable for epidemiological TA studies and for measuring those with lower TA severity.

Keywords Item response theory · Bifactor model · Psychometric properties · Test anxiety · Dimensionality · Factor analysis

Introduction

Test anxiety (TA), first described in the psychological literature by Mandler and Sarason (1952), was characterized by a heightened state of anxiety that occurs before or during tests (Sommer and Arendasy 2015), and it also be described as a series of physiological and behavioral responses with specific performances that accompany concerns that the test may fail or result in poor performance (Zeidner 1998). TA is a serious and pervasive problem among students (Bodas and Ollendick 2005; Ergene 2003), and students with TA will feel nervous, fear and worry in the evaluation situation (Spielberger et al. 1979; Spielberger and Vagg 1995). Researches that correlate TA with academic achievement suggest that high levels of TA are associated with lower levels of learning and performance (Sub and Prabha 2003). At all levels of education, students who often feel test-anxious perform poorly on standardized tests (Everson, Millsap, & Rodriguez, 1991a, b) and receive poorer grades (Chapell et al. 2005), which is mainly due to that anxiety and other test-taking deficiencies interfere with their performance either directly or indirectly (Efklides et al. 1997, 1999; Lowe et al. 2008; Metallidou and Vlachou 2007). Accordingly, it is extremely critical to have an accurate assessment and diagnosis of those with TA and provide timely treatment. Measuring TA using self-report scales has become a common method over the past several decades. A number of different self-report scales have been used in previous study, including the test anxiety inventory (TAI; Spielberger 1980), the test anxiety scale (TAS; Sarason 1978), the Friedman-Bendas Test Anxiety Scale (FAT; Friedman and Bendas-Jacob 1997), and the state—trait anxiety inventory (STAI; Marteau and Bekker 1992). In spite of some differences concerning items numbers, severity of the symptom, time period, and so forth, each scale measures the similar general construct—TA (Friedman and Bendas-Jacob 1997; Sarason 1978; Spielberger 1980; Umegaki and Todo 2017). In the past, psychometric properties of most self-reporting scales have been assessed by classical test theory (CTT), which focused on reliability, validity, and norms, etc. (Hunsley and Mash 2007, 2008). Moreover, validity and reliability are two

✉ Dongbo Tu
tudongbo@aliyun.com

[1] School of Psychology, Jiangxi normal university, Nanchang, China

important characteristics of measurement instruments (Devellis 2005). Reliability captures the consistency of scores obtained from applications of the instrument, and commonly used index of reliabilities are test-retest reliability, split-half coefficient, Cronbach's alpha. Validity consists of a complex set of criteria including convergent validity, divergent validity, and factorial validity used to judge the extent to which inferences, based on scores derived from the application of an instrument, are warranted. Norm is more of a reference system for evaluating the position of the test score in the team, that is, the index used to evaluate the test score. For convenience, we had summarized them into an overview table in Appendix 1. What's more, knowledge about the range of severity evaluated by an instrument is critically important for tailoring measurements to solve specific questions and to solve them in specific settings (Embretson and Reise 2000, 2012; Olino et al. 2012). To achieve this goal it is likely to be achieved by applying the approaches of item response theory (IRT).

In terms of TA scales, existing researches of TA scales mainly based on CTT and focused on: (1) Analyzing the psychometric properties of TA scales cross different culture (Manavipour et al. 2013; Mowbray et al. 2015; Lowe et al. 2011a, b, Sebastian et al. 2012). Bi (2002) first translated the FAT into China. He pointed the FAT of Chinese version had good reliability (> .85; see Table 1), and convergent validity between the scale and Spielberg TAI was .84 for boys and.82 for girls; the Greek version of Spielberger (TAI; 1980) self-report measure of test anxiety was verified the well-established two-factor structure for the TAI (Dimitra et al. 2011); Raju et al. (2010) translated Sarason's Test Anxiety Scale into an Ethiopian language and pointed the results of confirmatory factor analyses with extraction of four factors. The Ethiopian version of the Test Anxiety Scale as a whole could be considered reliable and useful for Ethiopian students. (2) Revising the scale and developing a short form. Taylor and Deane (2002) pointed that a 5-item short form produced optimal reliability (> .80; see Table 1) and validity, and a balance of items from the Worry and Emotionality subscales of the TAI. The 5-item short form of the TAI shows promise, particularly for contexts in which time demands preclude the use of longer versions; a brief version of the FRIEDBEN Test Anxiety Scale (B-FTAS) was investigated, which had the unique strength of measuring test anxiety using a contemporary biopsychosocial model. Exploratory and confirmatory factor analyses identified a 3-factor, brief, 12-item test anxiety assessment consistent with a biopsychosocial model including social, cognitive, and physiological factors. Results provide sufficient evidence for internal reliability (> .80; see Table 1) and validity of this brief measure of test anxiety (Dave et al. 2013). (3) Using TA Scale to conduct related research. The study of Yazici (2017) revealed that competitive and cooperative learning styles had positive, low-level and significant relationship with the TAS' emotionality sub-dimension, and the same relationship was observed between the competitive learning style and the worry sub-dimension. Lori and Lori and Richard (1998) found that there were no significant differences among age groups with respect to test anxiety. And poor study behavior was related to higher levels of test anxiety, and better study behavior was related to lower levels of test anxiety. Multiple regression analysis also revealed that test anxiety, gender, age, and ethnicity were all statistically significant predictors of study behavior; Everson et al. (1991a, b) pointed the invariance of the traditional two-factor structure for both males and females, and the psychometric properties of TAI had acceptable reliability (>. 60; see Table 1); Other study pointed that researchers should be careful when drawing conclusions based on original TAI norms, especially in the case of female undergraduates (Szafranski et al. 2012). With the literatures, most of all the existing researches on TA scales were based on the framework of CTT. However, CTT methods cannot offer specific information on the severity of TA symptomatology with respect to different trait levels. In addition, unidimensionality is an important assumption in IRT, and it is difficult to be satisfied for the most scales. If the unidimensional model is applied to estimate the item parameters of multi-dimensional instruments, it is likely to yield inaccuracy in parameters estimation. Third, although plenty of instruments are available, the agreement between them is less than optimal and no scale can be considered as a gold standard (Umegaki and Todo 2017). Therefore, it may be difficult for researchers and clinicians to choose an optimal instrument when assessing for TA. To address this gap, new approaches to analyzing multi-dimensional structure scales are essential and should be applied to reanalyze the TA scales. Above all, CTT alone is not sufficient to illustrate the ability of a measure to accurately assess the severity of various symptoms. Item response theory (IRT) is a new psychometric theory, which is developed on the basis of overcoming the limitations of CTT. IRT methods are based on probabilities of individual response options and estimate TA independently of the selection of test items, and provide estimates about the position on the latent trait (theta level; i.e., test anxiety) where each item or inventory provides the most information (Olino et al. 2012).

This study aims to address the issues by (1) investigating the structures of some commonly-used scales and (2) simultaneously comparing their psychometric properties under the framework of a Bifactor multi-dimensional structure approach of IRT. To fairly compare the psychometric properties for the three scales, the TA scales used here include the TAI, TAS, and FAT. The reasons why these were chosen for this study are as follows: (1) the three instruments are widely used in several fields of psychology studies. The TAI and TAS are widely used in research and practical settings and have particular application to the assessment and treatment of TA in student populations (Song and Zhang 2008; Zhu et al. 2019). The FAT also is applied to research its validation and standardization (Fereshteh et al. 2012). (2) Some critical evidence has indicated

Table 1 Previous studies on psychometric properties of TAI, TAS and FAT

Scales	Author(year)(country)	Version	Items	Sample	Factor	Reliabilities
TAI	Spielberger (1980) (USA)	English	20	College and high school students	Two factors 1.worry(TAI-W) 2.emotionality (TAI-E)	α > .92
	Song and Zhang (1987) (China)	Chinese	20	355 college students	Two factors 1.worry(TAI-W) 2.emotionality (TAI-E)	Test-retest reliability = .70
	Dimitra et al. (2011) (Greek)	Greek	20	231 undergraduate students	Two factors 1.worry(TAI-W) 2.emotionality (TAI-E)	α1 = .81 α2 = .94.
	Everson et al. 1991a, b(USA)	English	20	501college freshmen	Two factors 1.worry(TAI-W) 2.emotionality (TAI-E)	α1,2males = .74 α1,2females = .64.
S5-TAI	Taylor and Deane (2002) (New Zealand & Australia)	English	5	333 undergraduate psychology students	Single factor	α = .87
TAS	Sarason (1978) (USA)	English	37	College students	Single factor	Test-retest reliability > .80
	Wang (2001)(China)	Chinese	37	345 college students	Single factor	α = .64
	Lori& Lori and Richard (1998) (USA)	English	37	1441college students (community)	Single factor	–
FAT	Friedman and Bendas-Jacob 1997 (USA)	English	23	2294 junior high and 1422 high school students	Three factors 1.social derogation 2.cognitive obstruction 3.tensenses	α1 = .86 α2 = .85 α3 = .81
	Bi (2002) (China)	Chinese	23	3858 senior high school students	Three factors 1.social derogation 2.cognitive obstruction 3.tenseness	α1 = .91 α2 = .86 α3 = .85
B-FTAS	Dave et al. (2013) (USA)	English	12	1463 high school students	Three factors 1.social derogation 2. cognitive obstruction 3.physiological tenseness	α1 = .88 α2 = .86 α3 = .81

TAI Test Anxiety Inventory, *S5-TAI* short version of 5 items Test Anxiety Inventory, *TAS* Test Anxiety Scale, *FAT* Friedman-Bendas Test Anxiety Scale, *B-FTAS* brief version of the Friedman-Bendas Test Anxiety Scale

that the three scales have high reliability and validity. For example, Song and Zhang (1987) pointed out that the total and each subscale of the TAI had a good Cronbach's alpha (α = .90; α1 = .80; α2 = .84) and high scale construct validity. Wang (2001) suggested the TAS had good Cronbach's reliability (α > .60) and high concurrent validity. Bi (2002) found each subscale of the FAT had good internal consistency (range between 0.85 and 0.91) and high construct validity. (3) The same scoring methods ensured that psychometric properties of three TA instruments could be compared fairly (the higher the score, the more serious the test anxiety; Bi 2002; Newman 1996; Song and Zhang 1987; Wang 2001). With our best knowledge, there is no research that compares the psychometric properties of different TA scales under the framework of IRT. To address this issue and take full advantage of IRT, the study aims to compare the psychometric properties of three commonly-used TA scales in Chinese university students. This study is expected to provide suggestions for selecting and applying the most optimal and precise measures for researchers with different study purposes (Umegaki and Todo 2017). For instance, the scale may be designed to be used in studies where it can provide the most information at the lower TA severity level; or it may be useful for assessing changes in TA severity in treatment studies where it can more precisely measure the mean of TA severity. It may also be designed to obtain information about a clinical diagnosis for the best assessment at the higher TA severity level. Furthermore, a multi-dimensional approach—(the Bifactor multi-dimensional item response theory model) is first used here to analyze and compare three widely used TA scales, which is expected to derive more appropriate parameters estimation of items and individuals than unidimensional approaches. This article might play a significant role in the selection, development and revision of TA measures.

Method

Sample

A total of 790 university students from China were recruited. Participants were mainly from two general universities of Jiangxi province. The age of participants were range from

18 to 23 with mean of 19.40 (SD = 1.51). The proportions of male and female participants were 57.2% and 42.8%, respectively. In terms of region, of 57.5% students were from the countryside and 42.5% students were from cities.

Procedure

Data were collected across multiple sessions ranging in size from 10 to 30 participants. Three TA instruments were administered before participants' academic examination. Participants also provided demographic information, including age, gender, class level (freshman, sophomore, junior or senior) and region (city or countryside) prior to completing the questionnaires. All participants in the study agreed to participate and were informed about the purpose of this research. Furthermore, the study was conducted anonymously, and no information that could identify individuals was collected.

Measures

TAI (Spielberger 1980; Chinese Version: Song and Zhang 1987) The Test Anxiety Inventory (TAI) is a self-report inventory designed to measure test anxiety (TA) as a situation-specific personality trait. The TAI consists of 20 items, with a 4 point Likert-type scale ranging from 1 (rarely or none of the time) to 4 (always). The TAI provides a measure of total TA (TAI-T) as well as measures of two TA components of emotionality (E) and worry (W). Emotionality refers to perceived autonomic reactions (physiological arousal) evoked by evaluative stress (Spielberger and Vagg 1995), whereas worry refers to cognitive concerns about the consequences of failure (Morris and Liebert 1969). Worry tends to be associated with performance decrements on cognitive and intellectual tasks, but emotionality is not (Hembree 1988; Hong 1998; Spielberger et al. 1979). The Chinese version of TAI was first tested by Song and Zhang (1987) and the Cronbach's alpha of the total and subscales are .90, .80, and .84 in Chinese university students. The inventories describe phenomena associated with TA. For example, *I feel confident and relaxed when I take the exam. At the exam, I was upset.* Of the 20 items, one is positive statement. Furthermore, The TAI has been used extensively, and the manual indicates that "most high school and college students complete the inventory in 8 to 10 minutes" (Spielberger et al. 1980).

TAS (Sarason 1978; Chinese Version: Wang 2001) The TAS is a unidimensional self-report scale. The TAS is comprised by 37 statements, and each item asks for a yes or no answer. The Chinese version of TAS was first tested by Wang (2001), and the test-retest reliability for university students was .62, and the Cronbach's alpha is .64. The statements reflect common symptoms of TA—such as, *when a major exam is coming, I always think of others smarter than me. If I was to attend a large exam, I would be very anxious before starting.* Of the 37 items, 5 are positive statements. Newman (1996) suggested that 12 points or below of TAS total score indicated that the TA was a low level; 12 to 20 points are moderate, 20 and above were higher levels.

FAT (Friedman and Bendas-Jacob 1997; Chinese Version: Bi 2002) The Friedman-Bendas Test Anxiety scale (FAT), contains 23 items with a Likert 5-point scale ranging from 1 (not at all) to 5 (completely suitable). The FAT measured three subscales: Social Derogation (worries of being socially belittled and deprecated by significant others following failure on a test), Cognitive Obstruction (poor concentration, failure to recall, difficulties in effective problem solving, before or during a test), and Tenseness (bodily and emotional discomfort) (Friedman and Bendas-Jacob 1997). The Chinese version of FAT (Bi 2002), Cronbach's alpha of the total and subscales are .81, .91, .86, .85. The items correspond to the two other scales. For example, *even if I'm well prepared, I will be nervous before the exam. If the test is not good, I am worried that the teacher wills torment me.* Of the 23 items, 5 are positive statements.

Analysis

Description of Total Scores and Reliability

First of all, the total scores of each scale and the correlations and reliability of each scale based on CTT were reported.

Factor Analysis

As for the Bifactor model, Holzinger and Swineford (1937) pointed the Bifactor model refers to a general-specific model. A Bifactor measurement model allows all items to load onto a common general dimension of psychopathology in addition to any specific symptom domains or "group" factors. The Bifactor model assumes that: (1) there is a general factor (for example, a general ability factor) that can explain the common variation of all topics; (2) there are multiple local specific factors (for example, special ability factors), After controlling the effects of general factors, each special factor can additionally explain the common variation of some topics (Chen et al. 2006). If a multi-dimensional test consists of p topics $x1$, $x2$, …, xp measured A general factor G and a special factor F1, F2, …, Fn, then the titlex_i can be expressed as (Ye and Wen 2012):

$$x_\mathrm{i} = a_i G + \sum_{j=1}^{n} bijFj + \delta i, i = 1, 2, \ldots \mathrm{p}.$$

Where a_i is the load of the topic x_i on the global factor G, b_{ij} is the load of the topicx_i on the local factor F_j, and $\delta_i$is the test error of the topic x_i. It is generally assumed that general factors, special factors, and errors are not related (Chen et al.

2012). The Bifactor model integrates the unidimensional and multi-dimensionality of multi-dimensional tests, and can simultaneously test the common effects and unique effects of each dimension. The loading pattern and factor structure of the Bifactor model, consisting of nine items and three specific factors, is shown as an example in Fig. 1.

The confirmatory factor analysis (CFA) was carried out to investigate the structure of three scales in Chinese university students. Three types of structure were considered in this study, which included unidimensional structure, the initial multi-dimensional structure of each scale and their initial multi-dimensional structure with Bifactor structure. The comparative fitted index (CFI), the incremental fitted index of Tucker and Lewis (TLI) and the root mean square error of approximation (RMSEA) were employed to investigate whether the proposed structures fitted the data well.

If all above three structures were not fitted the data well, the structure of the scale needed to be re-explored and re-confirmed. In this situation, the exploratory factor analysis (EFA) and CFA with Bifactor structure were both used to investigate the structure of scale with two randomly split-half data, respectively.

The above statistical analyses were conducted by SPSS (23.0) and MPLUS (7.4).

Item Response Theory Analysis

Three commonly-used polytocous multi-dimensional model, including the multi-dimensional Generalized Partial Credit Model (mGPCM; Muraki 1992), the multi-dimensional Graded Response Model (mGRM; Samejima 1969) and the multi-dimensional Ratings Scale Model (mRSM; Muraki 1992), were used to analyze the data. Three test level model-fit criteria, including the Akaike's information criterion (AIC, Akaike 1974), Bayes information criterion (BIC) and negative 2 times log likelihood (−2*Log-Lik), were employed to select a more suitable IRT model for the data. The less value of the three criteria represented the better of the mode -fitted.

IRT statistical analyses were conducted using R (Version 3.1.2; http://www.R-project.org/) and the R packages psych (Version1.5.1; http://CRAN.R-project.org/package_psych).

Results

Description of Total Scores and Reliability

Table 1 documented the descriptive statistics, internal consistency and mutual correlations of summed scores of the different scale based on classical test theory (CTT).

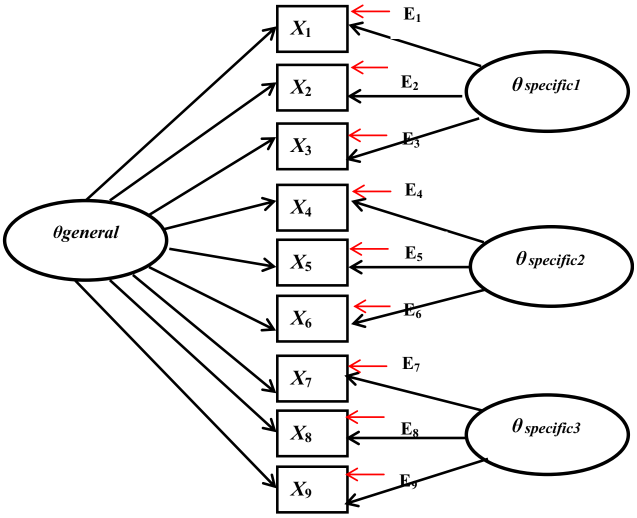

Fig. 1 A Bifactor model with three specific factors

Factor Analysis

First, with all items loading on only one dimension, the unidimensionality of each scales using CFA have been tested. Unexpectedly, no scale showed a good fit. This result indicated that FAT, TAI, and TAS were not efficiently a unidimensionality measure.

As the one-factor CFA (*structure A*) did not provide a close fit, and then the initial structure (structure B) of each scale via CFA has been verified. After that, the Bifactor CFA of each initial scale (structure C) has been fitted to investigate whether the Bifactor structure can fit better. Results were displayed in Table 2.

Table 2 showed the TAI fitted the Bifactor CFA of initial scale (*structure C*) well. However, the FAT and TAS did not. Therefore, a Bifactor EFA (*structure D*) for FAT and TAS have been performed to find a better fitted Bifactor structure. It was found that both the FAT and the TAS showed a good fit with four-special-factor explaining 56% and 50% of the variance, respectively. Then, their corresponding Bifactor structure was further confirmed by Bifactor CFA (*structure E*). Results in Table 2 indicated that the RMSEAs were less than 0.05, and the CFI and TLI were approximation 0.9 for FAT and TAS, which showed that the *structure E* (i.e., Bifactor structure with four-special-factor) was moderately fitted by both FAT and TAS.

Overall, the TAI fitted two-Bifactor structure very well, while the FAT and TAS moderately fitted four-Bifactor structure well. More details can be found in Table 3. As can be seen from the whole, the new scale structures based on Bifactor model are well-fitting.

Table 2 Descriptive statistics, internal consistency and mutual correlations of summed scores of the different scale

Scale	Scores					Cronbach's α	Spearman-Brown Half Coefficient	Correlations	
	N	*Min*	*Max*	*M*	*SD*			FAT	TAS
FAT	790	9	77	35.796	11.490	.822	.804	1	–
TAS	790	1	36	14.917	5.899	.796	.749	.531**	1
TAI	790	0	57	17.114	8.961	.896	.866	.502**	.605**

FAT Friedman-Bendas test Anxiety Scale, *TAS* Test Anxiety Scale, *TAI* Test Anxiety Inventory

** Indicates significant at the level of 0.1 at two-sided test

Item Response Theory Analysis

IRT Model Comparison and Selection

Three multi-dimensional IRT models with Bifactor structure were used for IRT analysis and the results of model-fit indexes were documented in Table 4. As shown in Table 5, the multi-dimensional GRM had the smallest values of Akaike information criterion (AIC), Bayes information criterion (BIC), and −2*Log-Lik in all three scales, which indicated that the mGRM fit the data of three scales best. Therefore, mGRM was chosen to estimated item parameters of three scales and analyzed their psychometric properties.

Psychometric properties for three scales

From the factor analysis, it was showed that the TAI fitted two-Bifactor structure very well, the FAT and TAS moderately fitted four-Bifactor structure, and all of the three scales extracted a general factor- that was test anxiety. Besides, the correlations of test scores among three scales ranged from 0.5 ($p < .01$) to 0.6 ($p < .01$), which showed that the three scales measured the similar latent trait—test anxiety. Based on the general factor-test anxiety, the psychometric properties of different self-report TA scales were further investigated.

One of the advantages of IRT is that it can provide the corresponding measurement accuracy for each subject. First, test information for each scale was to be calculated. Test information (TI) is the inverse of a squared standard error of measurement (*SE*), that is to say $SE(\theta_\alpha) = \frac{1}{\sqrt{I(\theta_\alpha)}}$.

Test information (TI) is an important index of measurement precision in IRT. Because test information increases with an increase of scale length, test information was divided by each scale's length obtained the average test information that denoted test information per item and enabled comparison of measurement precision among scales with different lengths. The average test information curves of three scales were shown in Fig. 2. Among the three scales, it showed an advantage test information for the FAT and TAI over TAS, and the TAI's advantage test information was the highest from the −1 to +3 (i.e., $-1 < \theta < +3$) of the standardized θ scale. For almost all other areas (i.e., $-3 < \theta < -1$), the FAT's advantage test information was the highest among the three scales. Conversely, the TAS's advantage test information was almost always lower than that of the other scales. These indicated that the TAI assessed information well for various degrees of TA severity.

What the study does is a comparison between the three scales, so it is necessary to compare the strengths and weaknesses of the measurement performance at a certain point or interval between the two on the θ scale. It is also necessary to examine which test has the best accuracy at the specified point or interval, and how efficient it is compared to other tests. This makes it easy to determine which test to choose is the best

Table 3 Fit index of each scale to test how the structure derived from A to E

Structure	FAT			TAI			TAS		
	TLI	*CFI*	*RMSEA*	*TLI*	*CFI*	*RMSEA*	*TLI*	*CFI*	*RMSEA*
A	0.73	0.76	0.09	0.88	0.89	0.07	0.71	0.73	0.05
B	0.78	0.80	0.08	0.92	0.93	0.05	0.71	0.73	0.05
C	0.73	0.77	0.09	**0.94**	**0.95**	**0.04**	0.60	0.64	0.06
D	0.94	0.97	0.04	–	–	–	0.87	0.89	0.03
E	**0.87**	**0.89**	**0.05**	–	–	–	**0.87**	**0.90**	**0.04**

FAT Friedman-Bendas Test Anxiety Scale, *TAS* Test Anxiety Scale, *TAI* Test Anxiety Inventory, *Structure A* one-factor structure, *Structure B* initial multidimansioanl structure, *Structure C* initial multidimansioanl Bifactor structure, *Structure D* Bifactor EFA with four-special-facotr structure, *Structure E* Bifactor CFA with four-special-factor structure, *CFI* comparative fitted index, *TLI* incremental fitted index of Tucker and Lewis, *RMSEA* root mean square error of approximation

Table 4 Factor loading for FAT, TAS and TAI

Item	G	S1	S2	S3	S4
FAT-1	0.467	**0.651**			
FAT-2	0.570	**0.123**			
FAT-3	0.522		**0.603**		
FAT-4	−0.202	**−0.404**			
FAT-5	0.653				**0.467**
FAT-6	0.693	**0.287**			
FAT-7	0.675				**−0.207**
FAT-8	−0.289			**0.472**	
FAT-9	0.520		**0.415**		
FAT-10	0.762				**0.494**
FAT-11	0.592				**−0.144**
FAT-12	0.624	**−0.194**			
FAT-13	0.719				**0.469**
FAT-14	−0.235			**0.618**	
FAT-15	0.761				**0.565**
FAT-16	0.490		**−0.186**		
FAT-17	0.588			**−0.239**	
FAT-18	−0.300			**0.530**	
FAT-19	0.668				**0.297**
FAT-20	−0.217			**0.624**	
FAT-21	0.570		**−0.319**		
FAT-22	0.714		**0.084**		
FAT-23	0.494				**−0.079**
TAS-1	0.210	**0.017**			
TAS-2	0.259		**−0.162**		
TAS-3	−0.107		**0.229**		
TAS-4	0.255				**−0.028**
TAS-5	0.171	**−0.107**			
TAS-6	0.210				**0.112**
TAS-7	0.244	**0.058**			
TAS-8	0.261			**0.028**	
TAS-9	0.041			**0.020**	
TAS-10	0.165	**0.058**			
TAS-11	0.209	**−0.013**			
TAS-12	−0.144			**0.183**	
TAS-13	0.199	**−0.056**			
TAS-14	0.263	**0.080**			
TAS-15	0.210			**0.034**	
TAS-16	0.268				**−0.073**
TAS-17	0.176	**−0.042**			
TAS-18	0.217		**0.096**		
TAS-19	0.213			**0.012**	
TAS-20	0.117		**0.081**		
TAS-21	0.210				**0.017**
TAS-22	0.114				**0.056**
TAS-23	0.256		**0.039**		
TAS-24	0.082		**0.166**		
TAS-25	−0.014		**0.151**		
TAS-26	−0.075		**0.198**		

Table 4 (continued)

Item	G	S1	S2	S3	S4
TAS-27	0.023		**0.144**		
TAS-28	0.222			**0.008**	
TAS-29	0.094		**0.067**		
TAS-30	0.243			**−0.048**	
TAS-31	0.207	**0.116**			
TAS-32	0.166				**0.159**
TAS-33	−0.037				**−0.059**
TAS-34	0.061			**0.014**	
TAS-35	0.005	**−0.163**			
TAS-36	0.200			**0.086**	
TAS-37	0.101				**0.329**
TAI-1	−0.084	**0.083**			
TAI-2	0.488	**0.137**			
TAI-3	0.628	**0.287**			
TAI-4	1.055	**0.763**			
TAI-5	0.889	**0.556**			
TAI-6	0.847	**0.470**			
TAI-7	1.000	**0.649**			
TAI-8	0.806	**0.328**			
TAI-9	0.516	**0.043**			
TAI-10	0.515	**−0.084**			
TAI-11	0.411		**−0.076**		
TAI-12	0.486		**−0.062**		
TAI-13	0.404		**−0.094**		
TAI-14	0.490		**−0.093**		
TAI-15	0.558		**0.101**		
TAI-16	0.630		**0.552**		
TAI-17	0.512		**−0.005**		
TAI-18	0.474		**−0.191**		
TAI-19	0.486		**0.245**		
TAI − 20	0.488		**0.254**		

All of the factor loading are significant. *FAT* Friedman-Bendas Test Anxiety Scale, *TAS* Test Anxiety Scale, *TAI* Test Anxiety Inventory. *G* general factor, *S* special factor

decision. The ratio of the test information functions at the specified trait level $\theta = \theta_0$ is called the relative efficiency between the two tests.

$$RE(\theta) = I_A(\theta)/I_B(\theta)$$

RE(θ) is relative efficiency, $I_A(\theta)$ and $I_B(\theta)$ are the test information functions on tests A and B, respectively.

Next, given that the three scales measure test anxiety as a whole, relative efficiency curves were plotted of the three scales (see Fig. 3). The relative efficiency of the TAI compared to the FAT was likely to be greater than 0.2 from approximately −3 to +3 (i.e., $-3 < \theta < +3$) of the standardized θ scale. That is, FAT can only achieve TAI test strength by extending 0.2 times on the basis of the original number of items. As the test information of TAI was a bit higher than FAT. This means that, when comparing the TAI with the FAT, the TAI have higher discrimination between students with test anxiety around or above the average, while the FAT have a little higher discrimination between students with test anxiety below the average. Furthermore, the relative efficiency of the TAI compared to the TAS was higher than 2 from −3 to −1 (i.e., $-3 < \theta < +1$) of the standardized θ scale and + 1 to +3 (i.e., $+1 < \theta < +3$) of the standardized θ scale. As far as the test function is concerned, the TAI is 100% stronger than the TAS, and the TAS test items need to be doubled on the original basis to achieve the TAI test strength. Because the TAI's test information was more than four

Table 5 The compare of mGRM, mGPCM, and mRSM

	mGRM			mGPCM			mRSM		
	AIC	BIC	−2*Log-Lik	AIC	BIC	−2*Log-Lik	AIC	BIC	-2*Log-Lik
FAT	**44,883.19**	**45,527.93**	**44,607.19**	45,355.51	4600.25	45,079.52	45,396.52	45,732.91	45,252.52
TAI	**29,156.05**	**29,623.26**	**28,956.05**	29,271.78	29,738.98	29,067.78	29,351.12	29,640.78	29,227.15
TAS	**32,374.42**	**32,894.02**	**32,135.42**	32,375.42	32,849.02	32,153.42	32,375.39	32,893.98	32,153.39

FAT Friedman-Bendas Test Anxiety Scale, *TAS* Test Anxiety Scale, *TAI* Test Anxiety Inventory, *AIC* Akaike information criterion, *BIC* Bayes information criterion, −2*Log-Lik = negative 2 times log likelihood, *mGPCM* multi-dimensional Generalized Partial Credit Model, *mGRM* multi-dimensional Graded Response Model, *mRSM* multi-dimensional Ratings Scale Model

times as much as the TAS, this means that, when comparing the TAI with the TAS, the TAS provides more information for the students who have test anxiety. In addition, the relative efficiency of the FAT compared to the TAS was greater than 4 when the θ was lower than approximately −1 and greater than +1 (i.e., $\theta < -1$, $\theta > +1$). Although the item of TAS was about more than twice as long as the FAT, when comparing the TAS with the FAT, the FAT provided more information only for the students whose TA severity (θ) were less than −1 and more than +1 (i.e., $\theta < -1$, $\theta > +1$).

Above all, the relative efficiency curve shows that TAI provides the most test information in the entire interval, and in the whole θ level, The test function of TAI and FAT is not very different. Besides, the test function of TAI and FAT is much stronger than TAS.

Finally, the standard error of measurement (SEM) and marginal reliability were calculated via $\mathrm{SE}(\theta_{\alpha})$. As the formula showed the larger test information for a θ is, the smaller the standard error of a scale for the θ is, and at the same time the measurement will be more accurate, and more reliable (high reliability).

In Fig. 4, the curve reflects that for FAT, when θ exceeds −1 (i.e., $\theta > -1$), the marginal reliability of the test is higher than 0.8, which means that FAT had a good reliability for the participants whose θ were more than −1 (i.e., $\theta > -1$). With regard to TAI, the accuracy of the whole scale is high and the change of curve is relatively flat. That is to say, for a standardized θ scale greater than −1.5 (i.e., $\theta > -1.5$), TAI is a good choice because it has a higher reliability (edge reliability >0.8) (Fig. 5). As shown in Fig. 6, the TAS has good reliability for a participant whose normalized θ scale is between −1 and + 2 (i.e., $-1 < \theta < +2$). In general, in terms of measured marginal reliability, the FAT and TAI in the three scales not only have higher test reliability, but also ensure the relative accuracy of the test at both ends. The accuracy of the TAS test is less optimistic than the other two.

Conclusions and Discussion

Using a Bifactor approach with a large sample of Chinese university students, the current study

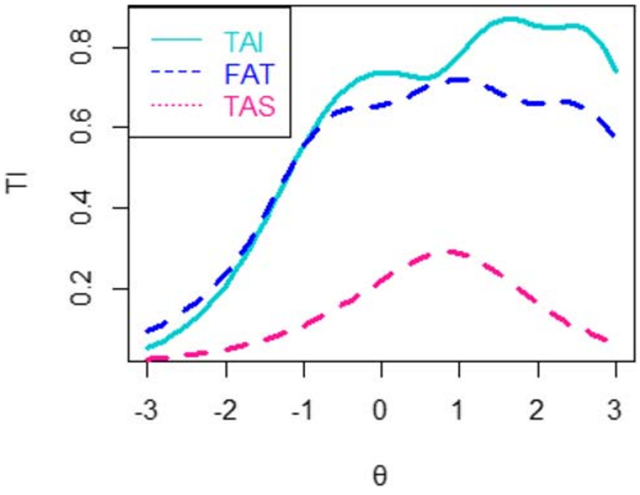

Fig. 2 Average test information curves of FAT, TAI, and TAS

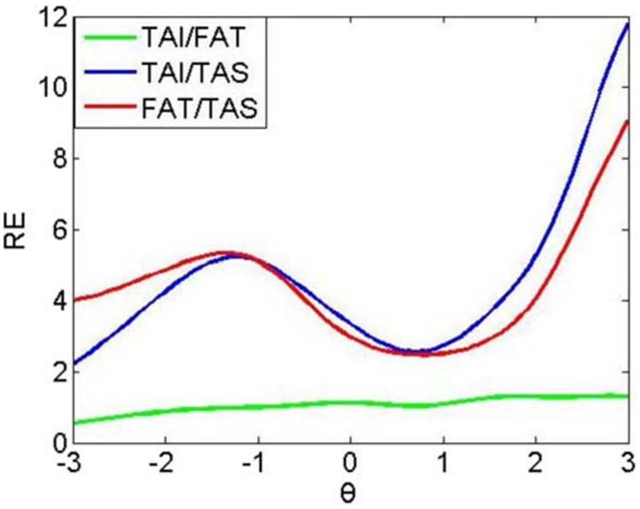

Fig. 3 Relative efficiency curves of FAT, TAI, and TAS

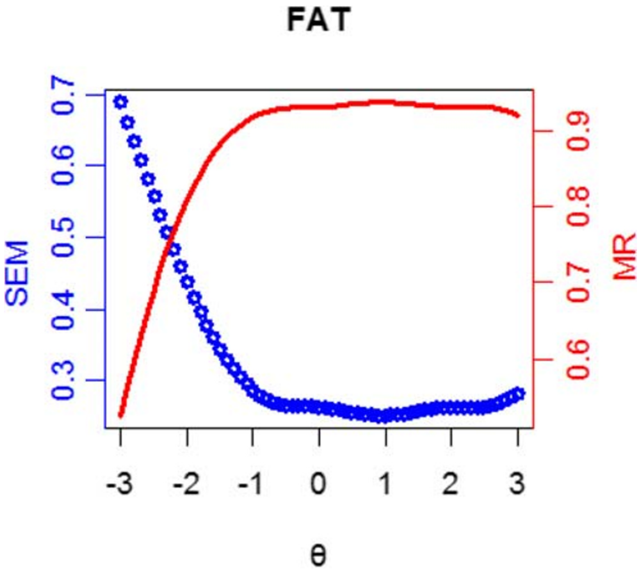

Fig. 4 Standard error of measurement and marginal reliability curves of FAT

investigated structures and simultaneously compared psychometric properties of three commonly used self-reporting TA instruments, including the TAI, the TAS, and the FAT.

The past researches were found the TAI score of female university students were always higher than that of male university students. In this study, it was also founded that TAI score of female university students (mean = 18.02) in the emotional subscale was significantly higher than that of male university students (mean = 16.39) with $t = -2.57$, $df = 788$, and $p < 0.05$, which was consistent with researches Benson and Tippets (1990) and Everson et al. (1991a, 1991b). As for TAS, 32% participants were at a low level of test anxiety with score of less than 12, 51% participants were moderate TA with score of between 12 and 20, and 17% participants were severely test anxiety with score of more than 20 (Newman 1996). Concerning FAT, the mean score (mean = 38.12) of female university students was significantly higher than that of male university students (mean = 33.47) with $t = -4.89$, $df = 788$, and $p < 0.01$. Descriptive results showed that both the Cronbach's alpha and the reliability of Spearman-Brown Half Coefficient for each scale were acceptable in Chinese university samples. The correlations of test scores among three scales ranged from 0.5 to 0.6 with significant moderate to high correlate ($p < 0.01$), which showed that they measured the similar latent trait. That is, there is comparability between the three scales. In addition, the result of the dimensionality and factor analysis showed that the TAI fitted two-Bifactor structure very well, and the FAT and TAS moderately fitted four-Bifactor structure. A correlated factors model did not include a general factor and attributes all explanatory variance to first-order factors (Morgan et al. 2015). A correlated factors model is conceptually ambiguous because it is not able to separate the specific or unique contributions of a factor from the effect of the overall construct shared by all interrelated factors (Chen et al. 2012), whereas a Bifactor model contains a general factor (G) and multiple specific factors (S). Because G and S are independent, a Bifactor model can disentangle how each factor contributes to the systematic variance in each item. The possibility of segmenting the variance in independent sources is one of the primary advantages of the Bifactor model (Reise 2012). In addition, the Bifactor structure has consistently

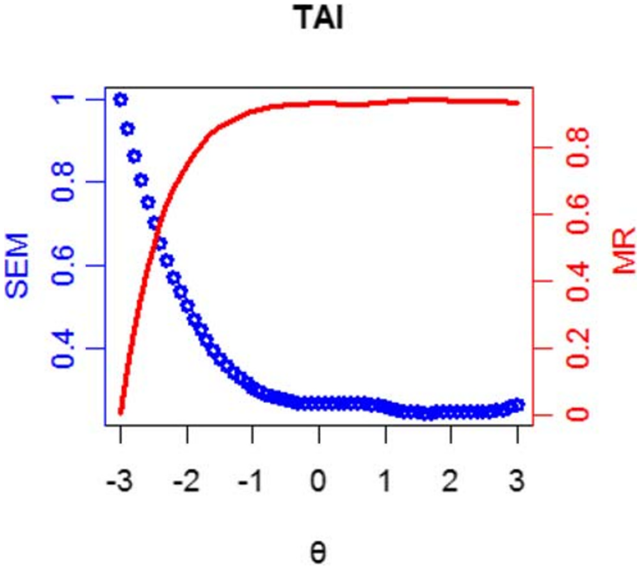

Fig. 5 Standard error of measurement and marginal reliability curves of TAI

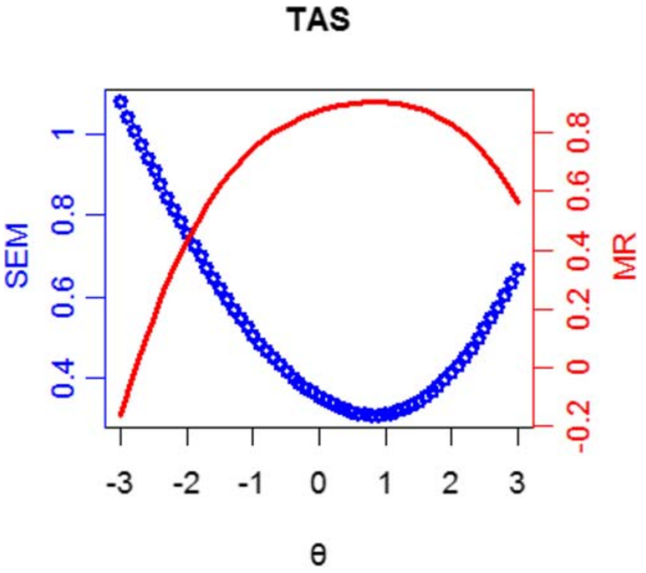

Fig. 6 Standard error of measurement and marginal reliability curves of TAS

proven to provide superior model fit for TA symptoms across measures in large samples, this finding lends further confidence to the phenomenon that this Bifactor solution offers a more optimal representation of the data than any of the previously suggested correlated-factors structures.

Additionally, psychometric properties of the three instruments by Bifactor IRT approach showed that the three scales had both high reliabilities and low SEMs at the broad range of TA severity, which indicated that the three scales performed well overall. The findings also provide suggestions for determining which scale to use in a given study design: the TAI evaluated TA along a wider range of severity with more precision than the other two scales. TAI can also be used to measure trait test anxiety and state test anxiety, depending on the time of the test. If using it outside the examination situations, the trait test anxiety is measured; if the scale is measured immediately at the post or last of an examination, the state test anxiety is measured (Dong et al. 2011). It may be pointed in this study TAI is a better instrument for the trait test anxiety. The FAT is performing a litter worse than TAI at the same levels of severity of TA. The TAS provided more information at the lower level of TA symptomatology. In conclusion, the TAI and the FAT evaluated information at greatly overlapping ranges; however, the TAI, performing a litter better at the same levels of severity of TA, may be a good choice when recruited those with various levels of TA severity to ensure a high precision. What's more, FAT may be a good choice for measuring those with moderate TA severity. Meanwhile, the TAS provided more information at the lower level of TA symptomatology, that is to say, TAS is suitable for epidemiological TA studies and for measuring those with lower TA severity. Of note, in fact, the study focused on the comparison of the general factor (i.e., TA) in the Bifactor Multi-IRT model while ignoring specific factors of the three scales in the current study. The FAT merely performed worse than the TAI on psychometric properties of the general TA factor; however, psychometric properties, including the reliability, the SEM, the TI, and the RE of specific factors for three scales were not investigated. Thus, the issue was confused as to whether the TAI is better or worse than the FAT on psychometric properties of specific factors.

Another contribution of this study was that a new approach of the Bifactor IRT model was used to fit the multidimensional structures of TA scales, while almost all of the prior studies used CTT approaches (which cannot offer specific information on the severity of TA symptomatology with respect to the differentiability levels) or unidimensionality IRT methods (the unidimensionality is difficult to be satisfied for TA scales). In a Bifactor IRT model, each item of the scale was able to not only load onto one specific factor but also a general factor (Osman et al. 2012), in which researcher could derive more information from the items and participants for both a general factor and specific factors. Therefore, compared with CTT and unidimensionality IRT approaches, the Bifactor Multi-IRT approach had natural advantages for analyzing psychological scales with multidimensional structures. There are some suggestions for conducting a Bifactor MIRT model. For example, the sample size needs to be large enough to accurately calibrate item parameters (Gignac 2016; Umegaki and Todo 2017). Meanwhile, the Bifactor MIRT model requires two or more specific factors in the structure (Cai et al. 2011; Li and Rupp 2011), and each specific factor needs to contain more than two items (Gomez and McLaren 2015; MacCallum et al. 1999; Velicer and Fava 1998; Zwick and Velicer 1986).

Although the IRT approach got the good result relatively, there also existed some limitations. First, the sample was not comprehensive and not representative, only selected from several universities, generating repeatedly better-fitting models across different samples of primary school students and adolescents. Second, considering the unidimensional IRT model applied will be robust to moderate degrees of multi-dimensionality (Drasgow and Parsons 1983; Olino et al. 2012). Therefore, trying to keep the unidimensional structure or the initial structure of the scales will get more information and provide more detailed and accurate suggestions for a given study. Third, inclusion of other commonly used self-report test anxiety scales may provide further suggestions for determining the usability of different self-report TA scales (e.g., State scale of state-trait anxiety inventory; STAI). At last, to use the TA instruments before an examination may ensure the reliability and validity of scales. The potential to use the TA scales in pre-post examination situations has been supported by previous research (Zeidner 1991). Development of a novel inventory that covers a wider range of TA severity and has the largest amount of test information at any point on the continuum or making a integration of the existed TA instruments are also a future direction.

Funding All of the authors agree with the submission to Current Psychology. This study was supported by the National Natural Science Foundation of China (31960186,31760288,31660278).

Compliance with Ethical Standards

Conflict of Interest All authors declare that they have no conflict of interest.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Informed consent was obtained from all individual participants included in the study.

Appendix 1

The index of test quality criteria

Index	Criteria	Meaning or Function
Reliability	Test-retest reliability	Test-retest reliability is the degree to which test scores remain unchanged when measuring a stable individual characteristic on different occasions (Vilagut 2014).
	Split-half coefficient	Refers to the correlation of the scores of all the subjects on the two halves after dividing one test into two equal parts.
	Cronbach's alpha	Alpha is an estimate of the correlation between two random samples of items from a universe of items like those in the test (Bland and Altman 1997).
Validity	Convergent validity	The degree to which the instrument correlates with other measures with which it should correlate (Ballatori et al. 2010).
	Divergent validity	Divergent validity is a term to describe evidence that measures of constructs that theoretically should not be highly related to each other are, in fact, not found to be highly correlated to each other. (Hubley 2014).
	Factorial validity	Factorial validity examines the extent to which the underlying putative structure of a scale is recoverable in a set of test scores. (Piedmont 2014).
Norm	–	Norm is more of a reference system for evaluating the position of the test score in the team, that is, the index used to evaluate the test score.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, AC-19*, 716–723. https://doi.org/10.1007/978-1-4612-1694-0_16.

Ballatori, E., Roila, F., Ruggeri, B., Bruno, A. A., Tiberti, S., & Orio, F. D. (2010). *Handbook of disease burdens and quality of life measures*. New York: Springer. https://doi.org/10.1007/978-0-387-78665-0_5907.

Benson, J., & Tippets, E. (1990). Confirmatory factor analysis of the test anxiety inventory. *Cross-culture Anxiety Research, 4*, 149–156. https://doi.org/10.4324/9781315825724-12.

Bi, Z. Z. (2002). Investigation and study on examination anxiety of high school students in Chongqing. *Journal of Southwest China Normal University (Natural Science Edition), 27*(4), 596–599.

Bland, J. M., & Altman, D. G. (1997). Cronbach's alpha. *Bmj, 314*(7), 572–572. https://doi.org/10.1136/bmj.314.7080.572.

Bodas, J., & Ollendick, T. H. (2005). Test anxiety: A cross-cultural perspective. *Clinical Child and Family Psychology Review, 8*, 65–88. https://doi.org/10.1007/s10567-005-2342-x.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item Bifactor analysis. *Psychological Methods, 16*, 221–248. https://doi.org/10.1037/a0023350.

Chapell, M. S., Blanding, Z. B., Silverstein, M. E., Takahashi, M., Newman, B., Gubi, A., & McCann, N. (2005). Test anxiety and academic performance in undergraduate and graduate students. *Journal of Educational Psychology, 97*, 268–274. https://doi.org/10.1037/0022-0663.97.2.268.

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of Bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41*, 189–225. https://doi.org/10.1207/s15327906mbr4102_5.

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J. P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the Bifactor model to other approaches. *Journal of Personality, 80*, 219–251. https://doi.org/10.1111/j.1467-6494.2011.00739.x.

Dave, P., Nathaniel, P. E., Natasha, S., & Stephen, P. K. (2013). Identification and validation of a brief test anxiety screening tool. *International Journal of School and Educational Psychology, 1*(4), 246–258. https://doi.org/10.1080/21683603.2013.826152.

Devellis R. F. (2005). Classical test theory. Applied Rasch Measurement*: A Book of Exemplars*. Springer Netherlands. https://doi.org/10.1007/1-4020-3076-2.

Dimitra, F., Despina, M., & Georgia, P. (2011). Psychometric properties of the Greek version of the test anxiety inventory. *Psychology, 2*(3), 241–247. https://doi.org/10.4236/psych.2011.23038.

Dong, Y. Y., Zhou, R. L., Gao, X., Jiao, F., & Guo, W. (2011). Reliability and validity of the Chinese version of test anxiety inventory (TAI) short form in college students. *Chinese Mental Health Journal, 25*(11), 872–876.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multi-dimensional data. *Applied Psychological Measurement, 7*, 189–199. https://doi.org/10.1177/014662168300700207.

Efklides, A., Papadaki, M., Papantoniou, G., & Kiosseoglou, G. (1997). Effects of cognitive ability and affect on school mathematics performance and feelings of difficulty. *The American Journal of Psychology, 110*, 225–258. https://doi.org/10.2307/1423716.

Efklides, A., Papadaki, M., Papantoniou, G., & Kiosseoglou, G. (1999). Individual differences in school mathematics performance and feelings of difficulty: The effects of cognitive ability, affect, age, and gender. *European Journal of Psychology of Education, 14*, 57–69. https://doi.org/10.1007/BF03173111.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Maheah: Lawrence Erlbaum Associates.

Embretson S.E., & Reise S. P. (2013). Item Response Theory: *Psychology Press*. https://doi.org/10.4135/9781412952651.n143.

Ergene, T. (2003). Effective interventions on test anxiety reduction: A meta-analysis. *School Psychology International, 24*, 313–328. https://doi.org/10.1177/01430343030243004.

Everson, H. T., Millsap, R. E., & Rodriguez, C. M. (1991a). Isolating gender differences in test anxiety: A confirmatory factor analysis of the test anxiety inventory. *Educational and Psychological Meas urement, 51*, 243–251. https://doi.org/10.1177/0013164491511024.

Everson, H. T., Millsap, R. E., & Rodriguez, C. M. (1991b). Isolating gender differences in test anxiety: A confirmatory factor analysis of the test anxiety inventory. *Educational and Psychological Measurement, 51*(1), 243–251. https://doi.org/10.1177/0013164491511024.

Fereshteh B., Alsadat S. M., Razieh E., & Shermin, R. (2012). Validation and standardization of persian version of friedben test anxiety scale (FAT). *Psychology Studies, 1*(29).

Friedman, I. A., & Bendas-Jacob, O. (1997). Measuring perceived test anxiety in adolescents: A self-report scale. *Educational and Psychological Measurement, 57*(6), 1035–1046. https://doi.org/10.1177/0013164497057006012.

Gignac, G. E. (2016). The higher-order model imposes a proportionality constraint: That is why the Bifactor model tends to fit better. *Intelligence, 55*, 57–68. https://doi.org/10.1016/j.intell.2016.01.006.

Gomez, R., & McLaren, S. (2015). The center for epidemiologic studies depression scale: Support for a Bifactor model with a dominant general factor and a specific factor for positive affect. *Assessment, 22*, 351–360. https://doi.org/10.1177/1073191114545357.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*, 47–77. https://doi.org/10.3102/00346543058001047.

Holzinger, K. J., & Swineford, F. (1937). The Bifactor method. *Psychometrika, 2*, 41–54. https://doi.org/10.1007/BF02287965.

Hong, E. (1998). Differential stability of individual differences in state and trait test anxiety. *Learning and Individual Differences, 10*, 51–69. https://doi.org/10.1016/S1041-6080(99)80142-3.

Hubley A. M. (2014). *Divergent Validity.* Springer Netherlands.

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology, 3*, 29–51. https://doi.org/10.1146/annurev.clinpsy.3.022806.091419.

Li, Y., & Rupp, A. A. (2011). Performance of the S-X statistic for fullinformation Bifactor models. *Educational and Psychological Measurement, 71*, 986–1005. https://doi.org/10.1177/0013164410392031.

Lori, T. R., & Richard A.R. (1998). Test anxiety and study behavior of community college students in relation to ethnicity, gender, and age. *Age Differences, 21*.

Lowe, P. A., Lee, S. W., Witteborg, K. M., Prichard, K. W., Luhr, M. E., Cullinan, C. M., Mildren, B. A., Raad, J. M., Cornelius, R. A., & Janik, M. (2008). The test anxiety inventory for children and adolescents (TAICA): Examination of the psychometric properties of a new multi-dimensional measure of test anxiety among elementary and secondary school students. *Journal of Psychoeducational Assessment, 26*, 215–230. https://doi.org/10.1177/0734282907303760.

Lowe, P. A., Ang, R. P., & Loke, S. W. (2011a). Psychometric analyses of the test anxiety scale for elementary students (TAS-E) scores among Singapore primary school students. *Journal of Psychopathology and Behavioral Assessment, 33*(4), 547–558. https://doi.org/10.1007/s10862-011-9250-9.

Lowe, P. A., Ang, R. P., & Loke, S. W. (2011b). Psychometric analyses of the test anxiety scale for elementary students (TAS-E) scores among Singapore primary school students. *Journal of Psychopathology and Behavioral Assessment, 33*(4), 547–558. https://doi.org/10.1007/s10862-011-9250-9.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*, 84–99. https://doi.org/10.1037/1082-989X.4.1.84.

Manavipour, D., Mohammadi, A., & Shahabi, P. (2013). Test anxiety inventory in Iranian culture. *Social Science Electronic Publishing.*

Mandler, G., & Sarason, S. B. (1952). A study of anxiety and learning. *JournaI of Abnormal and Social Psychology, 47*, 166–173. https://doi.org/10.1037/h0062855.

Marteau, T. M., & Bekker, H. (1992). The development of a six-item short form of the state scale of the Spielberger state—Trait anxiety inventory (STAI). *British Journal of Clinical Psychology, 31*, 301–305. https://doi.org/10.1111/j.2044-8260.1992.tb00997.x.

Metallidou, P., & Vlachou, A. (2007). Motivational beliefs, cognitive engagement, and achievement in language and mathematics in elementary school children. *International Journal of Psychology, 42*, 2–15. https://doi.org/10.1080/00207590500411179.

Morgan, G. B., Hodge, K. J., Wells, K. E., & Watkins, M. W. (2015). Are fit indices biased in favor of bi-factor models in cognitive ability research? : A comparison of fit in correlated factors, higher-order, and bi-factor models via Monte Carlo simulations. *Journal of Intelligence, 3*, 2–20. https://doi.org/10.3390/jintelligence3010002.

Morris, L. W., & Liebert, R. M. (1969). The effects of anxiety on timed and untimed intelligence tests: Another look. *Journal of Consulting and Clinical Psychology, 33*, 240–244. https://doi.org/10.1037/h0027164.

Mowbray, T., Jacobs, K., & Boyle, C. (2015). Validity of the German test anxiety inventory (TAI-G) in an Australian sample. *Australian Journal of Psychology, 67*(2). https://doi.org/10.1111/ajpy.12058.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x.

Newman, E. (1996). *No more test anxiety.* Los Angels: Learning Skills Publications.

Olino, T. M., Yu, L., Klein, D. N., Rohde, P., Seeley, J. R., Pilkonis, P. A., & Lewinsohn, P. M. (2012). Measuring depression using item response theory: An examination of three measures of depressive symptomatology. *International Journal of Methods in Psychiatric Research, 21*, 76–85. https://doi.org/10.1002/mpr.1348.

Osman, A., Wong, J. L., Bagge, C. L., Freedenthal, S., Gutierrez, P. M., & Lozano, G. (2012). The depression anxiety stress scales—21 (DASS-21): Further examination of dimensions, scale reliability, and correlates. *Journal of Clinical Psychology, 68*, 1322–1338. https://doi.org/10.1002/jclp.21908.

Piedmont, R. L. (2014). Factorial validity. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research*. Dordrecht: Springer. https://doi.org/10.1007/978-94-007-0753-5_3704.

Raju, P. M., Mesfin, M., & Alia, E. (2010). Test anxiety scale: reliability among Ethiopian students. *Psychological Reports, 107*(3), 939–948. https://doi.org/10.2466/03.11.17.PR0.107.6.939-948.

Reise, S. P. (2012). The rediscovery of Bifactor measurement models. *Multivariate Behavioral Research, 47*, 667–696. https://doi.org/10.1080/00273171.2012.715555.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 17*, 5–17. https://doi.org/10.1007/BF02290599.

Sarason, I. G. (1978). The test anxiety scale: Concept and research. *In Stress and Anxiety*, 193–216.

Sebastián, E. P., Daniel, E. H., & Luis, A. F. (2012). Psychometric properties of the revised Spanish version of the German test anxiety inventory (GTAI-AR) in Argentinean university students. *Universitas Psychologica, 11*(1), 177–186.

Sommer, M., & Arendasy, M. E. (2015). Further evidence for the deficit account of the test anxiety–test performance relationship from a high-stakes admission testing setting. *Intelligence, 53*, 72–80. https://doi.org/10.1016/j.intell.2015.08.007.

Song, W. Z., & Zhang, Y. (1987). The relationship between examination anxiety and personality traits of Chinese undergraduates. *Chinese Mental Health Journal, 1*(4), 165–168.

Song, F., & Zhang, X. J. (2008). The application of test anxiety scale (TAS) in middle school students in Beijing. *Chinese Journal of Clinical Psychology, 16*(6), 623–624.

Spielberger, C. D. (1980). *Test anxiety inventory: Preliminary professional manual*. Palo Alto: Consulting Psychology Press.

Spielberger, C. D., & Vagg, P. R. (1995). Test anxiety: A transactional process model. In C. D. Spielberger & P. R. Vagg (Eds.), *Test anxiety: Theory, assessment and treatment* (pp. 3–14). Washington, DC: Taylor & Francis.

Spielberger, C. D., Gonzalez, E. P., Taylor, C. J., Anton, W. D., Algaze, B., Ross, G. R., & Westberry, L. G. (1979). *Preliminary manual for the test anxiety inventory*. Palo Alto: Consulting Psychologists Press.

Spielberger, C. D., Gonzalez, H. P., Taylor, C. J., Anton, E. D., Algaze, B., Ross, G. R., & Westberry, L. G. (1980). *Manual for the test anxiety inventory ("test attitude inventory")*. Redwood City: Consulting Psychologists Press.

Sub, A., & Prabha, C. (2003). Academic performance in relation to perfectionism, test procrastination and test anxiety of high school children. *Psychological Studies, 48*, 7–81.

Szafranski, D. D., Barrera, T. L., & Norton, P. J. (2012). Test anxiety inventory: 30 years later. *Anxiety, Stress, and Coping, 25*(6), 667–677. https://doi.org/10.1080/10615806.2012.663490.

Taylor, J., & Deane, F. P. (2002). Development of a short form of the test anxiety inventory (TAI). *The Journal of General Psychology, 129*(2), 127–136. https://doi.org/10.1080/00221300209603133.

Umegaki, Y., & Todo, N. (2017). Psychometric properties of the Japanese CESD, SDS, and PHQ-9 depression scales in university students. *Psychological Assessment, 29*(3), 354–359. https://doi.org/10.1037/pas0000351.

Velicer, W. F., & Fava, J. L. (1998). Affects of variable and subject sampling on factor pattern recovery. *Psychological Methods, 3*, 231–251. https://doi.org/10.1037/1082-989X.3.2.231.

Vilagut, G. (2014). Test-retest reliability. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research*. Dordrecht: Springer. https://doi.org/10.1007/978-94-007-0753-5.

Wang, C. K. (2001). Test report of examination anxiety scale among university students. *Journal of Mental Health in China, 15*(2), 95–97. https://doi.org/10.3321/j.issn:1000-6729.2001.02.011.

Yazici, K. (2017). The relationship between learning style, test anxiety and academic achievement. *Universal Journal of Educational Research, 5*(1), 61–71. https://doi.org/10.13189/ujer.2017.050108.

Ye, B. J., & Wen, Z. L. (2012). Estimating homogeneity coefficient and its confidence interval. *Acta Psychologica Sinica, 44*, 1687–1694. https://doi.org/10.3724/SP.J.1041.2012.01687.

Zeidner, M. (1991). Test anxiety and aptitude test performance in an actual college admissions testing situation: Temporal considerations. *Personality & Individual Differences, 12*, 101–109. https://doi.org/10.1016/0191-8869(91)90092-P.

Zeidner, M. (1998). *Test anxiety: The state of art*. New York: Plenum Press. https://doi.org/10.1002/9780470479216.corpsy0984.

Zhu, Z. J., Wang, N., Zhou, H. B., Lv, S. B., & Zhao, Y. (2019). Mediating role of psychological elasticity in relationship between parent-child communication and examination anxiety of senior 3 students. *Journal of North China University of Technology(Medical Sciences), 21*(04), 317–320.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*, 432–442. https://doi.org/10.1037/0033-2909.99.3.432.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.