Validation of a new State Test Anxiety Scale (STAS)

Alper Sahin [iD][1,*]

[1]Atılım University, School of Foreign Languages, Department of Basic English, Ankara, Turkey

Abstract: This study aimed to validate the English version of Sahin's (2019) State Test Anxiety Scale (STAS) which was originally developed in Turkish based on the latest bio-psychosocial model of test anxiety. For this purpose, the Turkish version of STAS was translated into English and administered to 360 (123 females, 237 males) students from 22 countries. The data were subjected to the confirmatory factor analysis to confirm the three-factor structure as in the Turkish version. Most of the fit indices examined (χ^2/df=1.94, CFI=.98, NFI=.96, NNFI=.98, IFI=.98, RMSEA=.05, SRMR=.05) indicated that the data of the STAS had a good fit to the three-factor structure as its Turkish version. The Alpha internal consistency coefficients were found to be .81, .77, .91, and .92, for Physiological effects, Psychosocial effects, Cognitive effects subscales, and the total scale respectively. Stratified Alpha was also calculated and was found to be .93. All in all, evidence collected in this study indicated that the English version of STAS was a valid and reliable scale as its Turkish version.

1. INTRODUCTION

Test anxiety is one of the common types of anxiety. It affects 25% to 40% of the individuals at a certain level (McDonald, 2001). Test anxious individuals can experience some symptoms that prevent their brain and body function properly. This can decrease their test performance dramatically. Especially in a high-stakes exam situation, it is nearly impossible for some students not to get anxious before or after the exam. The numbers reveal the truth that test anxiety is very common even since the early days of schooling. According to Gençdoğan (2006), 66% of the students experience exam anxiety. Moreover, it is found in recent studies that test anxiety starts in primary school (Tornio, 2019; Popa et al., 2019) and elevates as the number of failures increases (Karataş et al., 2013). It is also said that test anxiety experienced in early childhood can be transferred to adulthood if necessary precautions are not taken (Delvecchio et al., 2017). The reason for the test anxiety can be the test-oriented nature of educational systems in different countries. For example, children in the United States take tests even before being accepted to kindergarten. Individuals may experience test anxiety in variegated levels due to the pressure of the expectations to be successful in tests starting in the early years of their life, the uncertainty about the result of the tests, and the possibility of losing face among family members.

*CONTACT: Alper ŞAHİN ✉ alpersahin2@yahoo.com ▣ Atılım University, School of Foreign Languages, Department of Basic English, Ankara, Turkey

The first studies on test anxiety started after the mid-20th century with some pioneering studies from a couple of scholars (Mandler & Sarason, 1952; Sarason & Stoops, 1978; Suinn, 1969; Osterhouse, 1970). According to Spielberger (1972) anxiety has two types: Trait-level anxiety and state-level anxiety. The former is a serious mood disorder in which individuals take their experience as a threat. The latter is a version of trait-level anxiety that triggers some problematic physiological and cognitive responses and it emerges in some specific situations like tests. Test anxiety is a kind of state-level anxiety as it has observable symptoms in only testing situations (Sapp, Farrel, & Durand, 1995). It is commonly accepted that if an individual has high trait level anxiety, they will experience high test anxiety as well (Spielberger & Vagg, 1995).

Test anxiety was first accepted as a phenomenon manifesting itself with a couple of bodily symptoms such as nausea, sweating, faster heartbeats, trembling, dizziness, dry mouth, tense muscles, headache, stomachache (Hagtvet et al., 2001). Later, it was noticed that anxiety had some deeper symptoms which could not be attributed to bodily reactions only. They were thought of as the emotional and cognitive reactions that the body produces against an uncertain situation like a test. Therefore, test anxiety was, then, considered to have two commonly accepted main sub-dimensions (Liebert & Morris, 1967). They are "worry" and "emotionality". Worry constitutes the cognitive reactions that the body produces. They can be considered as the self-deprecative thoughts that may lead the test taker to failure (Bozkurt et al., 2017). Emotionality constitutes bodily reactions such as nausea, sweating, and faster heartbeats (Spielberger, 1972). Although this two-factor model was commonly accepted, there was another model of test anxiety taking the thoughts about social derogation of the individuals into account put forward by Friedman and Bendas-Jacob (1997) and it was then called the biopsychosocial model by Lowe et al. (2008). The Bio-psychosocial model takes social and environmental factors into account and considers the environment and the social interaction of the individuals with others as the sources of anxiety. These social factors may be observed as being teased by friends, getting negative reactions from parents, teachers and losing face with them.

There are plenty of scales developed to measure test anxiety. The test anxiety questionnaire (Mandler & Sarason, 1952) is one of the first instruments developed to measure test anxiety. It has 42 items accumulated under three subscales: "anxiety about group intelligence tests", "anxiety about individual intelligence tests", and "anxiety about course examinations". Responders reflect their anxiety by pinpointing their level of endorsement to the items on a 15-centimeter line. The state-trait anxiety scale (STAI) developed by Spielberger et al. (1970) is also one of the initial scales to measure test anxiety. It has 40 items with 20 items in state and 20 others in trait anxiety subscales. Driscoll (2007) has another test anxiety scale named Westside Test Anxiety Scale. It has ten items and it takes around five minutes to administer. Another test anxiety scale is the Revised Test Anxiety Scale (RTAS) developed by Benson and El-Zahhar (1994). It has subscales called Tension, Worry, Test-irrelevant thoughts, and Bodily symptoms. Test Anxiety Inventory (TAI) which was developed by Spielberger (1980) is another test anxiety scale and it is one of the most used test anxiety scales (Chapell et al., 2005). It has 20 items under two subscales of Worry, and Emotionality.

There are two trait-level test anxiety scales recently developed. One of them is the cognitive test anxiety scale revised (Cassady & Finch, 2014). It has 25 items in one common factor related to the cognitive dimension of test anxiety. The other one is the IDA test anxiety scale (Başol, 2017) which was developed in Turkish and has nine items in the cognitive and physiological reactions subscale and six items in a subscale targeting social and environmental triggers of test anxiety.

Test anxiety has the potential to cause individuals to experience undesired drawbacks before, during, and after test-taking situations. It has some drawbacks from the point of view of

educational measurement as well. The current reliability coefficients calculate the reliability from the point of view of the items on a scale or a test based on the responses by the test takers. However, when it is test anxiety in action during a test, it is not the test items threatening the reliability of the scores. It is the test anxiety the individuals experience threatening it. That is to say, the test items may be statistically sound items and 5% or 10% of the test takers experiencing test anxiety in severe levels may not be adequate to indicate a problem in the reliability coefficients via statistical means. As their total scores are affected negatively by test anxiety, highly test anxious students are seen as low or mid-ability students who are not being able to answer the items correctly (as all other students with similar scores) without taking the anxiety they experience into consideration. As a result, students who get 60 out of 100 points without experiencing test anxiety and students who get 60 out of 100 due to the anxiety they experience are considered as equal by statistical formulas and everyone in the society. However, the students experiencing high test anxiety get fewer points than they deserve due to the negative effects of test anxiety. According to Hembree (1988) highly anxious students get around 12 percentile points lower than their peers who experience lower anxiety levels. What is worse is that it is nearly impossible to measure the exact impact of test anxiety on their test scores as they experience test anxiety in varying degrees. Therefore, the score of the students experiencing test anxiety does not reflect the real level of their latent ability on the measured construct due to test anxiety. As this hidden test anxiety effect on test scores changes based on the severity level of the anxiety individuals experience during the test, it is currently not possible to devise a correction formula to eliminate the error in the test scores of these students caused by test anxiety. As a result, it can be said that until the effects of test anxiety are lifted from the test environment, the reliability of the test scores obtained via statistical means may be much less than calculated via legitimate formulas.

The accurate and early diagnosis of test anxiety is highly critical, and such an early and accurate diagnosis may yield more reliable results if scales developed considering the latest model of test anxiety (bio-psychosocial model) are used. As studies conducted on test anxiety in the early 1950s suggest that anxiety has a connection to different situations, it would be more valid to diagnose anxiety in specific situations which is called the "Specificity Theory of Anxiety" (Harper, 1971). There are some scales developed specifically measuring state test anxiety (e.g. Mandler & Sarason, 1952; Spielberger et al., 1970). It would not be wrong to state that the youngest of them is a half-century old. The recent scales developed to measure trait level test anxiety may be limited as they are administered in the relaxed classroom environment for which the test takers should consider their overall test anxiety experience in all tests cumulatively while responding. Such scales are useful as it is practical to collect data with them. Moreover, as it is commonly accepted that there is a high correlation between state and trait test anxiety, that is, individuals experiencing high trait level anxiety will probably experience high state-level test anxiety (Spielberger & Vagg, 1995), the scale scores obtained from these scales can be used to diagnose and support the individuals experiencing test anxiety. However, the author of this paper believes that there may be some specific tests (as mentioned in the Specificity Theory of Anxiety) that the individuals with trait level anxiety may experience state-level test anxiety before, during, and after some particular tests much more than some others. Moreover, there may be some tests that they do not experience test anxiety at all or experience it to a lower stance. The current trait level test anxiety scales administered in a classroom environment away from the threat or the perception of the threat influencing the emotional state of the individuals and getting data from individuals collectively for all tests may not always distinguish between test anxiety experienced before, during, or after specific test-taking scenarios. Moreover, it would not be wrong to claim that the existing state test anxiety scales developed half a century ago may not be accurate enough to be used for such a purpose. Therefore, there may be an immediate need for a state test anxiety scale that would include the latest bio-psychosocial

model of the test anxiety, and which would enable the researchers to collect data for each specific test situation individually. For this purpose, the Turkish version of the State Test Anxiety Scale (Sahin, 2019) was developed. However, the need for such a scale in the English language to close the aforementioned gap in the literature still exists. Its translation to English and validation study was conducted to provide the international literature with the English version of this up-to-date state test anxiety scale. Therefore, the purpose of this study was to validate the English version of the State Test Anxiety Scale (STAS) which was developed originally in Turkish by Sahin (2019).

2. METHOD

In this part of the paper, the data collection instruments, the participants, the data collection procedures, and the statistical analyses used to collect data for the validity, reliability, and linguistic equivalence of the scale with its Turkish version will be detailed.

2.1. Data Collection Instruments

This study was conducted at the Northern Cyprus Campus of a reputable Turkish university with the participation of the students taking an academic English course during their freshmen year. A couple of data collection tools were used due to the nature of this study. They are the Turkish version of the *State Test Anxiety Scale* (STAS-TR; Sahin, 2019), the English version of the *State Test Anxiety Scale (STAS-EN),* and the Turkish version of the *Revised Test Anxiety Scale* (RTAS; Akin et al., 2012).

2.1.1. *STAS-TR*

The Turkish version of STAS was developed by Sahin (2019) using 312 participants (129 Females, 183 males) who were students at the Preparatory School Program (PSP) of an English medium private university located in the Turkish Republic of Northern Cyprus. Students attend PSP to develop their English language skills so that they can easily follow their departmental courses which will be held in English. STAS-TR has 22 items accumulated under three subscales called Physiological Effects (PE), Psychosocial Effects (PSE), and Cognitive Effects (CE). Cronbach's Alpha internal consistency coefficient for each subscale was found to be .85, .84, .93 respectively, and .94 for the total scale (Sahin, 2019). The participants respond to STAS-TR through a 4-point Likert scale with the Turkish versions of "1-Not At all, 2-Slightly, 3-Moderately, 4-A lot". The PE subscale of STAS-TR has eight items covering physiological effects of test anxiety such as headache, tighter muscles, difficulty breathing, fever, nausea, and stomachache. Two of the items under this subscale are "I have difficulty breathing" and "I have a dry mouth".

The PSE subscale has five items covering the effects of test anxiety pushing the individuals to think over the reactions of the people in their social environment to their possible incompetence. This subscale has items about negative reactions from the family members, feeling of losing face with the teacher, fear of being teased by the classmates in case of failure. One of the items in this subscale is "I worry about my classmates making fun of me for getting a low grade on the exam".

The CE subscale has nine items like the fear of not being able to come up with the right answers, not being able to understand what is asked in the exam, getting a low grade, failure, not being able to finish the exam on time. One of the items under this subscale is "I am afraid of not being able to finish the exam on time".

As the purpose of this study was to validate the English version of STAS-TR, STAS-TR was used to compare the reliability, validity, and linguistic equivalence of STAS-EN in this study.

2.1.2. *STAS-EN*

In order to ensure the linguistic equivalence of STAS-EN and STAS-TR, STAS-TR, which was found to be a valid and reliable state test anxiety scale in its original study (Sahin, 2019), was given to two bilingual translators who were both highly competent in Turkish and English and it was translated to English by them. When the translations were finalized, the translations of each item were reviewed and the most appropriate translations for each item were selected by the author who is also bilingual and who has B.A. and M.A. degrees in English language teaching. After this stage, the initial form of STAS-EN was given to two separate competent bilingual translators and was translated back to Turkish. The back translations were reviewed by the author and amendments in the items of STAS-EN were done based on the back translations. Then, the draft version of the STAS-EN was obtained. Two separate bilingual language experts with backgrounds in linguistic studies reviewed the draft version of STAS-EN. Some statements were amended at this stage based on their feedback as well.

At the final stage, the draft version of STAS-EN was reviewed by a British, a Canadian, and two American native speakers one of which was a scholar engaged in anxiety research and two of which were psychologists and psychometricians at the same time. STAS-EN was edited one more time based on the native speakers' suggestions. Finally, STAS-EN was administered to a group of students with diverse nationality and English proficiency levels in order to get feedback about the clarity and comprehensibility of the items in the scale. The draft version was edited one last time based on the student feedback and questions.

2.1.3. *RTAS*

RTAS, developed by Benson and El-Zahhar (1994) and adapted to Turkish by Akin et al. (2012), was used to collect evidence on the concurrent validity of the STAS-EN as was done while STAS-TR was being developed. The first reason why RTAS was selected as the scale to be used for concurrent validity was that it had many subscales with a close relationship with the subscales of STAS-TR and STAS-EN. Moreover, it was a widely used scale with separate validity studies done by multiple samples (Egyptian, American, Spanish) and authors (Bados & Sanz, 2005; Benson & Al-Zahhar, 1994; Hagtvet & Benson,1995). More importantly, RTAS (Turkish version) was used in the development of STAS-TR and it was thought that using it while developing STAS-EN would yield comparable results with STAS-TR.

RTAS is a trait level test anxiety scale that has 20 items under four subscales entitled "Tension", "Bodily Symptoms (BS)", "Worry", and "Test-irrelevant Thoughts (TiT)" with a 4-point Likert scale. The Tension subscale has five items covering anxious thoughts of the individuals about the test like "I worry a great deal before taking an important exam", BS subscale has five items on the bodily reactions such as "I get a headache during an important test", Worry subscale has six items covering the anxiety about failure and self-deprecating thoughts like "During tests, I find myself thinking about the consequences of failing" and TiT subscale has 4 items about test-irrelevant thoughts like "While taking tests, I sometimes think about being somewhere else".

Benson and El-Zahhar (1994) reported alpha coefficients ranging between .76 and .91 with the American sample (as cited in Akin et al. (2012), p.107). Hagtvet and Benson (1995) found that the alpha reliability coefficients were .89, .81, .81, .85, and .91 for the Tension, Worry, BS, TiT, and the whole scale respectively. Akin et al. (2012) also reported .78, .71, .78, .80, and .88 for the Tension, Worry, BS, TiT, and the whole scale.

2.2. The Participants, Data Collection Procedures, and Statistical Analyses

STAS-EN was administered to students a couple of times just before different exams to collect adequate data as evidence of validity, reliability, and indirect evidence of linguistic equivalence of the scale. Apart from STAS-EN, STAS-TR, and RTAS (Turkish version) were also

administered to participants at different stages of the study. The data collection started with the administration of RTAS to 196 students (77 Females, 119 Males) in the classroom environment and an informed consent form detailing the purpose of the study, how the data collected will be used and how participants can leave the study anytime they wish was signed by the voluntary participants. STAS-EN was administered to these students right before their midterm exam for an academic English course and the Pearson product-moment correlation coefficient between the same 196 students' scale and subscale scores obtained from both RTAS and STAS-EN were calculated using SPSS 23 (IBM Corp, 2015) as evidence of concurrent validity of STAS-EN.

The main group of participants of the study (n=360) was selected from a population of students with which STAS-TR was developed. STAS-TR was developed with the participation of only students with Turkish origins due to the language of the scale. The students with Turkish origins constituted a large part of this study as well (282 including the 20 students who were citizens of the Turkish Republic of Northern Cyprus). Moreover, 78 international students from 21 different countries such as Pakistan, Bangladesh, Egypt, Jordan, Kenya, Syria, Kazakhstan, Azerbaijan, Ruanda, Nigeria, and Uganda also participated in the study reaching a total number of 360 students (123 females, 237 Males) from 22 countries. 257 (56 Females, 201 Males) of the 360 students were from Engineering programs and 103 (67 Females, 36 Males) of them were students registered to Social Sciences programs. The more detailed descriptive statistics about the gender and fields of the participants can be found in Table 1.

Data collected from the main participants of the study was also used to confirm the three-factor structure of STAS-EN that it inherited from STAS-TR. That is, the factorial structure of STAS-EN was already known and it was expected to maintain the three-factor structure of STAS-TR. Therefore, confirmatory factor analysis (CFA) was utilized to confirm this structure. Before the CFA, to detect multivariate outliers, Mahalanobis Distance (D^2) was calculated using SPSS 23 (IBM Corp, 2015) and their p values were examined. Moreover, the Q-Q plots of the Mahalanobis Distances were examined. In this examination, no multivariate outliers were detected in the data. Following this examination, data was taken into confirmatory factor analysis (CFA) based on the default options of LISREL 8.51 (Jöreskog & Sörbom, 2001) in order to confirm and analyze whether the three-factor model of STAS-TR fit the data collected using STAS-EN.

Table 1. *Descriptive statistics of the main participants of the study (n=360).*

	Department	Female	Male	Frequency	Percent
Social Sciences Programs	Economy	6	5	11	3.1
	Guidance and Psychological Counseling	14	1	15	4.2
	Business Administration	5	14	19	5.3
	Political Science and International Relations	6	11	17	4.7
	Psychology	36	5	41	11.4
	Social Sciences Total	67	36	103	28.7
Engineering Programs	Aerospace Engineering	16	27	43	11.9
	Electrical and Electronics Engineering	7	38	45	12.5
	Chemical Engineering	7	6	13	3.6
	Mechanical Engineering	4	47	51	14.2
	Petroleum and Natural Gas Engineering	6	8	14	3.9
	Computer Engineering	9	47	56	15.6
	Civil Engineering	7	28	35	9.7
	Engineering Total	56	201	257	71.3
	Grand Total	123	237	360	100.0

Evidence about the reliability and internal consistency of STAS-EN was assessed utilizing different analyses. For this purpose, Cronbach's Alpha, stratified Alpha, McDonald's Omega, the Spearman-Brown split-half reliability coefficients were calculated. Apart from these, test-retest reliability analysis was also conducted. While calculating the reliability coefficients, data collected from the main participants (n=360) of the study were used. For test-retest reliability analysis, a separate group of 98 students (32 Females, 66 Males) than the main participants of the study enrolled in an academic English course at the same institution where the study was held were asked to respond to STAS-EN before a quiz and a midterm exam which had around 4-week interval in between. The data for test-retest reliability analysis was collected from a different group of students because it was not possible to find an adequate number of exams within one single course to collect the necessary evidence to properly assess psychometric properties of the STAS-EN from a single group of students. The Pearson-product moment correlation between scale and subscale scores obtained from both administrations of the STAS-EN were used as evidence of test-retest reliability. Moreover, item-by-item correlation coefficients between the two administrations of STAS-EN were also calculated.

The test-retest reliability analysis was done one more time to evaluate whether there was a change in correlations when two language versions, STAS-EN and STAS-TR, were administered to a common group of bilingual Turkish students who were competent in both languages. For this purpose, first, STAS-EN was administered to a group of students (n=90), and then STAS-TR was administered to the same group of students before another exam three weeks later. Then, the correlations between the scale and subscale scores obtained from these administrations were calculated. This analysis was taken as an indirect proof of linguistic equivalence.

3. RESULTS

The results of the study will be presented here under three headings: Results as evidence of the validity of STAS-EN, Results as evidence of the reliability of STAS-EN, and Results as evidence of linguistic equivalence of STAS-TR and STAS-EN.

3.1. Results as Evidence of the Validity of STAS-EN

The first validity evidence collected for STAS-EN was out of the analysis conducted to evaluate its construct validity. STAS-TR had a three-factor structure. In order to confirm this three-factor structure in STAS-EN and the structural equivalence of STAS-EN with STAS-TR, the data collected was used to conduct a CFA. The path diagram obtained as an output of CFA can be found in Figure 1 below.

The suggested modifications were reviewed and added to the model if it was thought that the correlation between their residuals could be explained and if they improved the fit indices (Seçer, 2015). When reviewed, it was found that Q2, Q8, and Q19 were all about the stomach-related problems a test-taker experiences. Moreover, Q10 and Q21 were both items related to classmates seeing the individual's exam score, and Q16 and Q18 were both items related to anxiety experienced due to having low scores from the test. Moreover, it was observed that all these modifications contributed to the model fit indices. As a result, they were decided to be added to the model. Moreover, the model fit indices obtained out of CFA can be found in Table 2.

As can be seen in Table 2, out of nine fit indices calculated, STAS-TR had six of them indicating a good fit to the model; however, according to the data collected with STAS-EN, seven out of nine fit indices indicated a good fit of the data to the model. This was considered as the evidence of the construct validity of the STAS-EN. It is also important to note that STAS-EN got consistently better fit index values in many fit indices (e.g. CFI, AGFI, IFI, NFI, NNFI).

Figure 1. *The path diagram of the CFA conducted over the research data.*
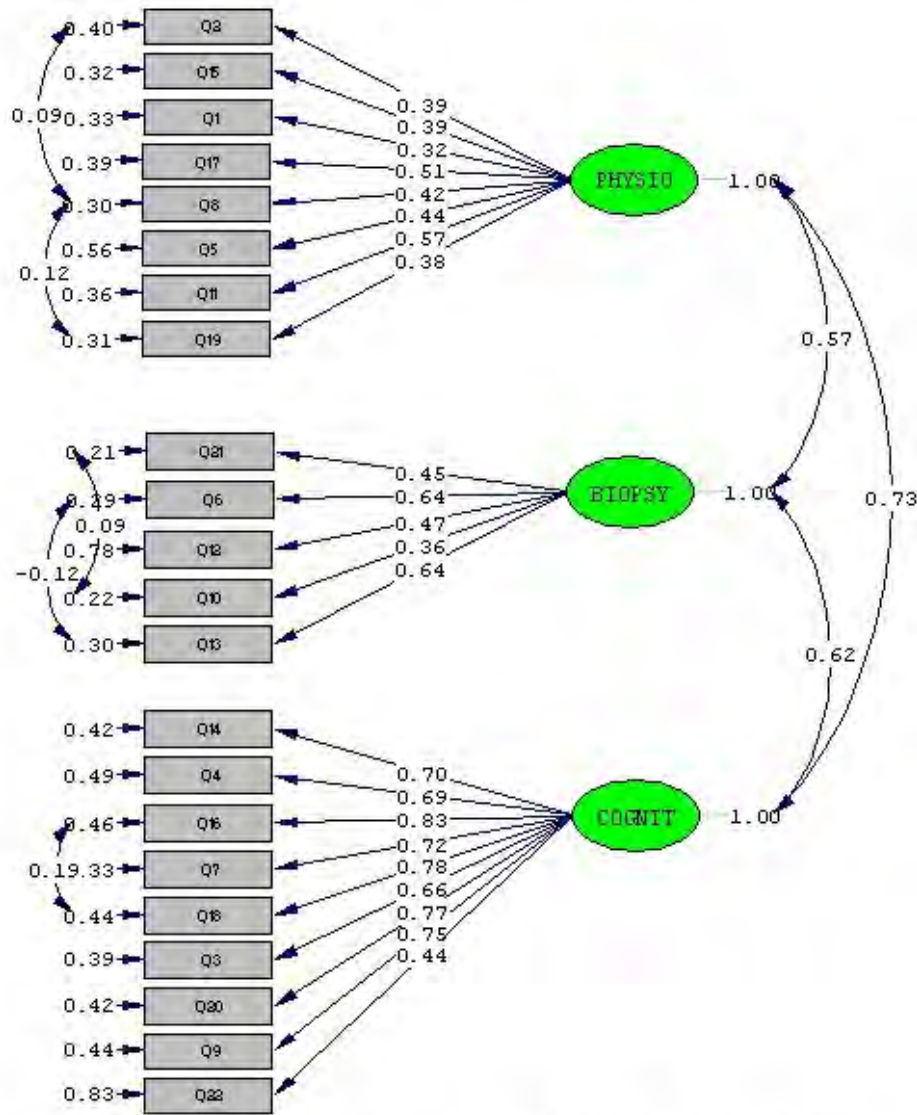
Chi-Square=390.81, df=201, P-value=0.00000, RMSEA=0.051

Table 2. *Fit values obtained from CFA.*

Fit Index	Good Fit Interval*	Adequate fit Interval*	STAS-TR / EN	Degree
χ^2/df	$0 \leq \chi 2/df \leq 2$	$2 \leq \chi^2/df \leq 3$	1.72 / 1.94	Good fit
CFI	$.95 \leq$ CFI ≤ 1.00	$.90 \leq$ CFI $\leq .95$	.96 / .98	Good fit
SRMR	$.00 \leq$ SRMR $\leq .05$	$.05 \leq$ SRMR $\leq .10$	.05 / .05	Good fit
GFI	$95 \leq$ GFI ≤ 1.00	$.90 \leq$ GFI ≤ 95	.91 / .91	Adequate fit
AGFI	$.90 \leq$ AGFI ≤ 1.00	$.85 \leq$ AGFI $\leq .90$	.88 / .89	Adequate fit
IFI	$.95 \leq$ IFI ≤ 1.00	$.90 \leq$ IFI $\leq .95$	.96 / .98	Good fit
NFI	$.95 \leq$ NFI ≤ 1.00	$.90 \leq$ NFI $\leq .95$	.92 / .96	Good fit
RMSEA	$.00 \leq$ RMSEA $\leq .05$	$.05 \leq$ RMSEA $\leq .08$	.05 / .05	Good fit
NNFI	$.95 \leq$ NNFI ≤ 1.00	$.90 \leq$ NNFI $\leq .95$	.96 / .98	Good fit

*Çokluk et al. (2010), Kahraman et al. (2018), Schermelleh-Engel et al. (2003).

In order to compare the performance of STAS-EN with STAS-TR, Pearson product-moment correlation coefficients between scale and subscale scores of STAS-EN were also calculated. You can find these coefficients in Table 3 with the same correlations obtained from STAS-TR in the original study (Sahin, 2019).

Table 3. *The correlations between the scale and subscale scores of STAS-TR and STAS-EN.*

Subscales	PSE TR / EN	PE TR / EN	Total TR / EN
CE-TR/EN	.65** / .66**	.65** / .59**	.94** / .92**
PSE-TR/EN		.49** / .57**	.78** / .81**
PE – TR / EN			.82** / .80**

**p < 0.001

As can be seen in Table 3, the CE subscale of STAS-EN had correlations of .66 with PSE subscale scores, .59 with PE subscale scores, and .92 with total scale scores of STAS-EN. These correlations were .65, .65, and .94 in the original study in which STAS-TR was developed. Moreover, the PSE subscale of STAS-EN had .57 correlation with PE subscale and .81 with the total scale score of STAS-EN. These correlations were .49 and .78 in the original study in which STAS-TR was developed. Lastly, the PE subscale of the STAS-EN had .80 correlation with the total scale score of STAS-EN which was .82 in the original study with STAS-TR. These can be taken as the validity evidence of the STAS-EN and its subscales.

As mentioned earlier, to evaluate the concurrent validity of STAS-EN, and RTAS (Turkish version), another valid and reliable scale on test anxiety which was validated with multinational samples was also administered to a group of students (n=196) in the classroom environment. Then, the scale and subscale scores of RTAS and STAS-EN were correlated. These correlations can be viewed in Table 4. Table 4 also indicates the correlations obtained between STAS-TR and RTAS in the original study (Sahin, 2019).

Table 4. *Correlations of STAS-TR and STAS-EN scale and subscale scores with RTAS.*

	Tension	BS	Worry	TiT	RTAS Total
PE–TR/EN	.61** / .56**	.69** / .63**	.52** / .45**	.25** / .18**	.66** / .61**
PSE–TR/EN	.43** / .52**	.39** / .43**	.49** / .53**	.24** / .30**	.50** / .61**
CE– TR/EN	.73** / .70**	.47** / .49**	.63** / .65**	.22** / .19**	.67** / .70**
STAS Total-TR/EN	.73** / .73**	.60** / .61**	.73** / .66**	.27** / .25**	.73** / .76**

**p < 0.001

When Table 4 is reviewed, it can be seen that STAS-TR and STAS-EN scale scores had very similar correlations with the subscales of RTAS. Another important finding to point out is that high correlation coefficients between STAS-EN scale and subscale scores and RTAS scale and subscale scores were obtained except the PE subscale giving hints of discriminant validity of STAS-EN. It is also important to emphasize that this was also the case in the original study of STAS-TR as can be seen in Table 4.

When the correlations between the subscales were analyzed, it can be seen that the highest correlation was between the PE subscale of STAS and the BS subscale of RTAS in both versions of STAS (.69 vs .63). This was expected as the corresponding scale which contains the physiological effects of test anxiety was the BS subscale of RTAS. It is also interesting to

state here is that there was around a .05-point decrease between the PE in STAS-TR and STAS-EN with some subscales of RTAS. For example, correlation with the Tension subscale decreased from .61 to .56 and the correlations with the BS subscale of RTAS decreased from .69 to .63. This may be due to the nature of different tests triggering different bodily reactions before each at this administration of the scale.

When the correlations of the PSE subscale of STAS with subscales of RTAS are analyzed, it can be noticed that in both versions of STAS, the highest correlations were obtained with the Worry and Tension subscales. It can be considered as normal as the source of the feelings regarding the pressure from family and friends are related to the items under Tension and Worry subscales of RTAS. Apart from this, it can also be seen that correlations with the Tension subscale of RTAS with STAS-TR and STAS-EN were quite similar.

The correlations with the CE subscale of STAS (both TR and EN) and the subscales of RTAS indicated that both versions of the scale had higher correlations with the Tension subscale of RTAS. This was also expected as they had similar items covering anxious thoughts regarding the test in both scales such as "I feel nervous" and "I am afraid of not being able to finish the exam on time" in STAS-EN's CE subscale and "I start feeling very uneasy just before getting a test paper back" and "I worry before the test because I do not know what to expect" from the Tension subscale of RTAS.

It is important to note that all subscales of STAS-TR and STAS-EN had very low correlations with TiT of RTAS. This was also an expected outcome as the TiT subscale of RTAS had items covering another dimension of test anxiety which was not covered in both STAS-TR and STAS-EN. This can be taken as evidence of the discriminant validity of STAS-TR and STAS-EN.

3.2. Results as Evidence of the Reliability of STAS-EN

Reliability evidence for STAS-EN was collected calculating Cronbach's alpha reliability coefficient, Stratified Alpha coefficient, McDonald's Omega, Spearman-Brown Split-half reliability coefficient, and test-retest reliability analysis (four-week interval in between). The results of these calculations can be seen in Table 5. Moreover, the reliability-related findings of STAS-TR in the original study (Sahin, 2019) were also reported in Table 5 to compare both versions of STAS.

Table 5. *Reliability evidence collected for STAS-EN.*

	Stratified Alpha	McDonald's Omega (ω)	C's Alpha (STAS-TR/EN)	Spearman-Brown Split Half (STAS-TR/EN)	Test-Retest STAS-TR (n=108) / STAS-EN (n=98)
PE		.81	.85 / .81	.89 / .80	.74[**] / .76[**]
PSE		.80	.84 / .77	.86 / .81	.80[**] / .76[**]
CE		.91	.93 / .91	.94 / .89	.78[**] / .76[**]
Total	.93	.92	.94 / .92	.96 /.91	.81[**] / .78[**]

[**] $p < 0.001$

As can be seen in Table 5, it can be said that Stratified Alpha and Mcdonald's Omega (ω) indicated high reliability. Moreover, it can also be said that the Cronbach's alpha reliability coefficients of STAS-EN were consistently lower compared to STAS-TR. However, they still indicated consistently high reliability. A similar decrease is also evident in the reliability coefficients obtained from Spearman-Brown's Split-half reliability. However, the Split-half reliability coefficients obtained still indicate high stability between the scores from both halves as well.

The negative shift observed in alpha reliability coefficients of STAS-EN compared to STAS-TR can be attributed to the multinational sample which was used to collect the data for STAS-EN. As mentioned earlier, the sample used in the development of STAS-TR was composed of only Turkish nationals; however, the sample used in this study was of a multinational one. Therefore, the highly diverse test-taker sample may have yielded some inconsistent response patterns in the data. However, STAS-EN's capability to accommodate high cultural diversity in return for a slight decrease in reliability coefficients should be praised and could even be taken as an indicator of its being a highly reliable and stable scale of test anxiety.

Another reliability-related evidence collected was the test-retest reliability analysis of STAS-EN. As mentioned earlier, for this purpose, STAS-EN was administered to a sample of 98 (32 Females, 66 males) students before a quiz and a midterm exam in a 4-week interval. As can be seen in Table 6, the correlations between these two administrations of the STAS-EN to a common group of test-takers before different tests indicated highly consistent results. Calculating test-retest reliability between these two tests was risky as one was a quiz that constituted 10% of the total score of the students for the course and the other one was a midterm exam that was more seriously taken by most students and constituted 20% of the total scores of the students. Although there was such a drawback of evaluating the stability of STAS-EN between a quiz and a midterm, it was thought that it still yielded highly stable results by producing correlations ranging between .76 and .78 compared to STAS-TR test-retest correlations ranging between .74 and .81 which were obtained from two administrations of STAS-TR before two midterm exams with a 4-week interval.

An item-by-item correlational analysis was done to see how consistent the test taker responses were in two consecutive administrations of STAS-EN to a common group of students. The findings of this analysis can be viewed in Table 6.

Table 6. *Item-by-Item Correlations Test-Retest Reliability for STAS-EN (n=98).*

Scale	Item #	r	Scale	Item #	r	Scale	Item #	r
	1	.43		3	.57		6	.51
	2	.30		4	.65		10	.43
	5	.47		7	.61		12	.64
	8	.70		9	.43		13	.56
PE	11	.66	CE	14	.48	PSE	21	.57
	15	.28		16	.55			
	17	.31		18	.55			
	19	.74		20	.57			
				22	.42			
	Ave. :	.49		Ave. :	.54		Ave. :	.54

When Table 6 is analyzed, it can be seen that there are some inconsistent items with correlations less than .40 especially in the PE subscale of STAS-EN. These items were items 2, 15, and 17. However, when these items were reviewed, it was noticed that they were about some bodily symptoms (2. I have nausea, 15. It feels like I have a fever, 17. I have a headache.) of test anxiety that may not manifest before or after each exam. Therefore, such fluctuations of correlations were expected because the symptoms experienced may alter in variety and magnitude before, or after each exam depending on the test type, the test environment, and some other factors. This item-by-item correlational analysis should be analyzed considering this variability in symptoms. However, the findings in Table 6 may be taken as an indication of high consistency for the items of STAS-EN despite this drawback.

3.3. Results as Evidence of Linguistic Equivalence of STAS-TR and STAS-EN

The findings as evidence of reliability and validity already had implications of the linguistic equivalence of both versions of STAS; however, apart from the direct evidence of forward and backward translations used during the development of STAS-EN, some indirect evidence regarding linguistic equivalence was also collected to evaluate whether the linguistic equivalence of STAS-EN and STAS-TR was maintained.

A correlational analysis was conducted between the scale and subscale scores of data obtained from the STAS-EN and STAS-TR. First, STAS-EN was administered to a new group of bilingual students (n=90) who were both competent in English and Turkish before a midterm exam, and then STAS-TR was administered to the same group of students before a quiz three weeks later. A comparison of correlation coefficients obtained from administration of STAS-EN to the same group of individuals (n=98) two times (r^{EE}) while conducting test-retest reliability analysis (last column in Table 5) and consecutive administration of STAS-EN and STAS-TR (r^{ET}) to a common group of students (n=90-only Turkish ones) who are competent Turkish and English speakers are presented in Table 7. Please note that these correlations are inter-scale correlations of the same subscales in these two administrations of the scales to a common group of individuals.

Table 7. *The inter-scale correlations of the subscales of STAS-EN and two administrations of STAS-EN and STAS-TR to common groups.*

	STAS PE (r^{EE} / r^{TE})	STAS CE (r^{EE} / r^{TE})	STAS PSE (r^{EE} / r^{TE})	STAS-EN Total (r^{EE} / r^{TE})
STAS PE	.76 /.71			
STAS CE		.76 / .85		
STAS PSE			.76 / .78	
STAS-EN Total				.78 / .85

As can be seen in Table 7, when STAS-EN and STAS-TR were administered to a common group of students, the correlations between the scale and subscale scores of the common students for these two administrations ranged between .71 and .85. More importantly, except for the one obtained from STAS-EN PE, all other correlations obtained from scale and subscale scores of STAS-EN and STAS-TR surpassed the ones obtained from the consecutive administration of STAS-EN two times. This can be taken as an indication of the high stability of the scores between the English and Turkish versions of STAS and taken as indirect evidence of linguistic equivalence of STAS-TR and STAS-EN.

The item-by-item correlational analysis was iterated to be able to evaluate whether the items in STAS-TR and STAS-EN functioned equally well or not after their administration of these two versions to a common group of students in a three-week interval. Another comparative table, Table 8, was prepared to present the findings of item-by-item correlations obtained from test-retest reliability analysis of STAS-EN (Table 6) and the same analysis done using student responses after the consecutive administration of STAS-EN and STAS-TR to a common group of students.

When Table 8 is analyzed, it can be seen that the item-by-item correlations between the two versions of the STAS range between .35 and .74. The lowest correlation obtained was from item 1 (I have difficulty breathing.) while the highest correlation was obtained from Item 21 (I worry about my classmates seeing my exam score). Moreover, it can also be seen that the correlations between STAS-EN and STAS-TR were quite like the ones obtained after the administration of STAS-EN two times during the test re-test reliability analysis. In most cases,

the correlations obtained from scale and subscale scores of STAS-TR and STAS-EN were higher than the ones obtained while conducting test-retest reliability analysis for STAS-EN. This was also taken as the indirect evidence of linguistic equivalence of STAS-EN and STAS-TR.

When all findings are put together, it would not be wrong to state that the STAS-TR and STAS-EN can be considered as linguistically equivalent forms of the same construct in different languages.

Table 8. *Comparison of Item-by-Item correlations.*

Scale	Item #	r^{EE}	r^{TE}	Scale	Item #	r^{EE}	r^{TE}	Scale	Item #	r^{EE}	r^{TE}
	1	.43	.35		3	.57	.65		6	.51	.69
	2	.30	.48		4	.65	.59		10	.43	.71
	5	.47	.49		7	.61	.65		12	.64	.66
	8	.70	.40		9	.43	.48		13	.56	.54
PE	11	.66	.67	CE	14	.48	.55	PSE	21	.57	.74
	15	.28	.55		16	.55	.68				
	17	.31	.36		18	.55	.58				
	19	.74	.48		20	.57	.61				
					22	.42	.49				
	Ave. :	.49	.47		Ave. :	.54	.59		Ave. :	.54	.67

3.4. Calculation of the Scale and Subscale Scores of STAS-EN

The scale and subscale scores of the STAS-EN can be calculated by simply adding the responses given to each item in each subscale. The items numbered 1, 2, 5, 8, 11, 15, 17, and 19 can be collected to obtain the subscale score for PE. The lowest score that can be obtained from this subscale is 8 and the highest score is 32. The score obtained from adding up the responses given to items numbered 3, 4, 7, 9, 14, 16, 18, 20, 22 will yield the CE subscale score. The lowest score that can be obtained from this subscale is 9 and the highest score is 36. The score obtained from items numbered 6, 10, 12, 13, 21 yields the subscale score for PSE. The lowest score for this subscale is 5 and the highest score is 20. Apart from all these subscale scores, all item scores can be collected to obtain the scale score. 22 is the lowest score and 88 is the highest score that can be obtained from the scale.

When the scale or subscale scores reach 1/2 of the highest score (e.g. 18/36 for CE) that can be obtained from that subscale, it may be considered that the subscale score is at its medium level. When it reaches 2/3 (e.g. 24/36 for CE) of the total scale score, the score may be considered as high. However, these figures should be approached with caution as there were no validation studies done to confirm them.

4. DISCUSSION & CONCLUSION

The need for a scale that can be used to measure and pinpoint the test anxiety levels of the individuals right before each exam triggered the development of STAS-TR (Sahin, 2019). The existing need for such an up-to-date scale in English to close the gap in the international literature constituted the motivation of this study. Therefore, the purpose of this study was to validate the English version of STAS to make it available to international researchers. For this reason, multiple pieces of evidence from a multinational group of individuals were collected. First, STAS-EN was administered to a multinational sample consisted of 360 freshman students who constituted the main participants of the study. Evidence of validity was collected at this stage. For the construct validity of the scale, data collected from the main sample of 360

students were subjected to a CFA in order to check whether STAS-EN had the same factorial structure with STAS-TR. Most of the fit indices for CFA indicated good fit (χ^2/sd=1.94, CFI=.98, NFI=.96, NNFI=.98, IFI=.98, RMSEA=.05, SRMR=.05) of the data to the three-factor model. The findings confirmed that STAS-EN had a three-factor structure as was the case in STAS-TR. Apart from construct validity, more data was collected to compare the performance of the STAS-EN to that of another valid scale, RTAS, as evidence of concurrent validity. The findings of this analysis also confirmed that STAS-EN was a valid scale as RTAS to measure test anxiety.

The evidence regarding the reliability of STAS-EN was also collected through Cronbach's Alpha, Stratified Alpha, McDonald's Omega, test-retest (n=98), and Spearman Brown's split-half formula. Data obtained from these analyses (α^{PE}= .81, α^{PSE}= .77, α^{CE}= .91, α^{Total}= .92) supported each other and confirmed that STAS-EN was a highly reliable and stable scale.

For the indirect evidence of linguistic equivalence, STAS-TR and STAS-EN were administered to a new common group of students (n=90). The correlations between the scale and subscale scores indicated that STAS-EN had similar linguistic properties. All in all, the evidence collected for reliability, validity, and linguistic equivalence of STAS-TR and STAS-EN indicated that STAS-EN was both valid and reliable equivalent of STAS-TR.

With the STAS-EN available in the literature and as it is a scale that can be administered before each exam separately, it will be possible for the international researchers to gain deeper insights into whether some factors such as the type of the course, the instructor, or even the proctor can leverage or soothe the test anxiety levels of individuals. Moreover, it will also be possible for the researchers to do research on some exam room arrangements like the use of soothing music, meditation, or different lighting before the exam and test their effects on the anxiety levels of test anxious individuals. Such research is missing in the literature and the researchers are kindly invited to research such topics. Moreover, international researchers are also invited to design studies to identify the cut scores of the STAS-EN scale and subscale scores and validate STAS-EN in different cultures.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s). The ethical committee approval dated 06.05.2019 numbered BAYEK 2019 / 04-06 was obtained from the Scientific Research and Publication Ethics Committee of the Middle East Technical University Northern Cyprus Campus.

ORCID

Alper Şahin ⓘ https://orcid.org/0000-0001-7750-4408

5. REFERENCES

Aydın, U., Bulgan, G. (2017). Adaptation of children's test anxiety scale to Turkish. *Elementary Education Online, 16*(2), 887-899. https://doi.org/10.17051/ilkonline.2017.304742

Akın, A., Demirci, İ., & Arslan, S. (2012). Revize edilmiş sınav kaygısı ölçeği: Geçerlik ve güvenirlik çalışması [Revised Test Anxiety Scale: Validity & Reliability Study]. *Educational Sciences and Practice, 11*(21), 103-118.

Bados, A. & Sanz, P. (2005). Validation of the revised test anxiety scale and the Friedben test anxiety scale in a Spanish sample. *Ansiedad y* Estrés, *11* (2/3), 163-174.

Başol, G. (2017). IDA test anxiety scale: Validity and reliability study. *The Journal International Education Science, 4*(13), 173-193. https://doi.org/10.16991/INESJOURNAL.1506

Benson, J., & El-Zahhar, N. (1994). Further refinement and validation of the revised test anxiety scale. *Structural Equation Modelling, 1*(3), 203-221. https://doi.org/10.1080/10705519409539975

Bozkurt, S., Ekitli, G. B., Thomas, C. L., & Cassady, J. C. (2017). Validation of the Turkish version of the cognitive test anxiety scale-Revised. *Sage Open, (7)*1, 1-9. https://doi.org/10.1177/2158244016669549

Cassady, J. C., & Finch, W. H. (2014). Confirming the factor structure of the Cognitive Test Anxiety Scale: Comparing the utility of three solutions. *Educational Assessment Journal, 19*(3), 229-242. https://doi.org/10.1080/10627197.2014.934604

Chapell, M. S., Blanding, 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*(2), 268-274. https://doi.org/10.1037/0022-0663.97.2.268

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2010). *Sosyal Bilimler için Çok değişkenli istatistik : SPSS ve LISREL uygulamaları [Multivariate Statistics for Social Sciences: SPPS and LISREL Applications]*. Pegem Academy Press.

Delvecchio, E., Cavallina, C., Di Riso, D., & Mazzeschi, C. (2017). Early evidence of Italian validation of the trait anxiety scale of the state trait anxiety inventory for children. *European Journal of Developmental Psychology*, *15*(2), 214-223. https://doi.org/10.1080/17405629.2017.1297227

Driscoll, R. (2007). *Westside test anxiety scale validation*. Education Research Information Center. https://files.eric.ed.gov/fulltext/ED495968.pdf

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

Gençdoğan, B. (2006). Lise öğrencilerinin sınav kaygısı ile boyuneğicilik düzeyleri ve sosyal destek algısı arasındaki ilişkiler [The relationship between the test anxiety, submissive behavior level and perception of social support of high school students]. *Atatürk University Journal of Social Sciences Institute, 7*(1), 153-164.

Hagtvet, K. A., & Benson, J. (1997). The motive to avoid failure and test anxiety responses: Empirical support for integration of two research traditions. *Anxiety, Stress and Coping*, *10*(1), 35-57. https://doi.org/10.1080/10615809708249294

Hagtvet, K.A., Man, F., & Sharma, S. (2001). Generalizability of self-related cognitions in test anxiety. *Personality and Individual Differences*, *31*(7), 1147-1171. https://dx.doi.org/10.1016/S0191-8869(00)00212-9

Harper, F. (1971). Specific Anxiety Theory and the Mandler-Sarason Test Anxiety Questionnaire. *Educational and Psychological Measurement, 31*(4), 1011-1014. https://doi.org/10.1177/001316447103100431

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. Review of Educational Research, 58 (1), 47-77. https://doi.org/10.3102/00346543058001047

IBM Corp. (2015). *IBM SPSS Statistics for Windows*, *Version 23.0*. IBM Corp.

Jöreskog, K. G., & Sörbom, D. (2001). *LISREL 8.5 for Windows* [Computer software]. Scientific Software International, Inc.

Kahraman, S., Özbaşı, D., & Özdemir, M. (2018). A Study on the Development of an Attitude Scale Towards the Use of PowerPoint in Classroom. *Kastamonu Education Journal*, *26*(4), 1237-1246. https://doi.org/10.24106/kefdergi.434168

Karataş, H., Alıcı, B., & Aydın, H. (2013). Correlation among high school senior students' test anxiety, academic performance and points of university entrance exam. *Educational Research and Reviews, 8*(13), 919-926. https://doi.org/10.5897/ERR2013.1462

Liebert, R. M., & Morris, L.W. (1967). Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological Reports*, *20*(3), 975-978. https://doi.org/10.2466/pr0.1967.20.3.975

Lowe, P. A., Lee, S. W., Witteborg, K. M., Prichard, K. W., Luhr, M. E., Cullinan, C. M. … Janik, M. (2008). The test anxiety inventory for children and adolescents (TAICA): Examination of the psychometric properties of a new multidimensional measure of test anxiety among elementary and secondary school students. *Journal of Psychoeducational Measurement, 26(3)*, 215-230.

Mandler, G., & Sarason, S. B. (1952). Some correlates of test anxiety. *Journal of Abnormal and Social Psychology, 47*(4), 810-817. https://doi.org/10.1037/h0060009

McDonald, A. S. (2001). The prevalence and effects of test anxiety in school children. *Educational Psychology, 21*(1), 89-101. https://doi.org/10.1080/01443410020019867

Osterhouse, R. A. (1970). Desensitization and study skills as treatment for two types of test-anxious students. *Journal of Counseling Psychology*, *19*(4), 301-307. https://doi.org/10.1037/h0034177

Popa, C., Bochis, L., & Clipa, O. (2019, July 3-5). *School assessment and test anxiety at primary school pupils* [Paper Presentation]. 4th International Conference on Lifelong Education and Leadership for All, Wroclaw, Poland.

Sahin, A. (2019). State Test Anxiety Scale (STAS): Validity and Reliability Study. *Trakya Journal of Education, 9*(1), *78-90.* https://doi.org/10.24315/tred.450423

Sapp, M., Farrel, W., & Durand, H. (1995). The effects of mathematics, reading and writing tests in producing worry and emotionality test anxiety with economically and educationally disadvantaged college students. *College Students Journal, 29*(1), 122- 125.

Sarason, I. G., & Stoops, R. (1978). Test anxiety and the passage of time. *Journal of Consulting and Clinical Psychology, 46*(1), 102- 109. https://doi.org/10.1037//0022-006x.46.1.102

Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*(2), 23-74.

Seçer, İ. (2015). Process of psychological test development and adaptation: SPSS and LISREL applications. Anı Publishing.

Spielberger, C. D. (1972). *Theory and research in anxiety.* Academic Press.

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

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). *Manual for state-trait anxiety inventory*. Consulting Psychologists 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). Taylor & Francis.

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). Routledge/Taylor & Francis Group.

Suinn, R. M. (1969). The STABS, a measure of test anxiety for behavior therapy: Normative data. *Behavior Research and Therapy, 7*(3), 335-339. https://doi.org/10.1016/0005-7967(69)90018-7

Tornio, S. (2019, March 14). *More kids than ever are dealing with test anxiety, and we need to help*. We Are Teachers. https://www.weareteachers.com/test-anxiety/