*Psicothema (2022) 34*(4) 562-570

Psicothema

https://www.psicothema.com/es • ISSN 0214–7823

Colegio Oficial de Psicólogos del Principado de Asturias

Methodology

# Development of a Revised Version of the Statistical Anxiety Scale

Urbano Lorenzo-Seva[1], Andreu Vigil-Colet[1] and Pere Joan Ferrando[1]

1 Universitat Rovira i Virgili.

ARTICLE INFO

Received: May 13, 2022
Accepted: July 11, 2022

***Keywords:***
Statistics anxiety
Response biases
Academic performance

## ABSTRACT

**Background:** Statistics anxiety is a common problem in students taking statistics courses in the social sciences. It is most widely measured by the statistical anxiety scale. The various adaptations of this instrument have shown certain problems in the replication of its factorial structure and do not have a system to control possible response bias effects. The objective of our study was to propose a short test to measure statistical anxiety that also includes a scale to control social desirability bias. **Method:** We developed a revised version of the statistical anxiety scale using procedures for controlling response biases and examined its factorial structure using exploratory and confirmatory analysis in a sample of 531 students. **Results:** The revised version showed a clear four-factor structure in exploratory and confirmatory factor analyses with the expected three content factors plus one social desirability factor. The scales showed no acquiescence effects and moderate social desirability effects, and had a clear relationship with academic success. **Conclusions:** The revised version of the statistical anxiety scale improves on the psychometric properties of the original version and may overcome the problems detected in some adaptations of the previous version.

## Desarrollo de una Versión Revisada de la Escala de Ansiedad Estadística

## RESUMEN

***Palabras clave:***
Ansiedad Estadística
Sesgos de Respuesta
Rendimiento Académico

**Antecedentes:** La ansiedad estadística es un problema habitual en los estudiantes que cursan materias relacionadas con la estadística en las ciencias sociales. Una de las escalas más utilizadas en su evaluación es la Escala de Ansiedad Estadística. En algunas adaptaciones se han detectado problemas en la replicación de su estructura factorial y no controlan los sesgos de respuesta. El objetivo de nuestra investigación fue proponer un test para la evaluación de la ansiedad estadística incluyendo una escala para el control de la deseabilidad social. **Método:** Se desarrolló una versión revisada de la escala utilizando procedimientos para el control de la deseabilidad social analizándose su estructura factorial en una muestra de 531 estudiantes. **Resultados:** La versión revisada mostró un ajuste adecuado tanto a nivel exploratorio como confirmatorio a una estructura de cuatro factores; los tres de contenido esperados y un factor de deseabilidad social. Las escalas no mostraron efectos de la aquiescencia y un moderado efecto de la deseabilidad social, además las escalas de contenido mostraron una clara relación con el rendimiento académico. **Conclusiones:** La versión revisada de la escala mejora las propiedades de la versión precedente y puede solventar los problemas detectados en algunas adaptaciones de la misma.

Cite as: Lorenzo-Seva, U., Vigil-Colet, A. & Ferrando, P. J. (2022). Development of a Revised Version of the Statistical Anxiety Scale. *Psicothema, 34*(4), 562-570. https://doi.org/10.7334/psicothema2022.231

Corresponding author: Andreu Vigil-Colet, andreu.vigil@urv.cat

Statistics is extremely important in society for understanding news, government policies and research outcomes (Chew & Dillon, 2014b). For this reason, students in most social science programs must take courses in statistics so that they can conduct research and interpret outcomes (O'Bryant, et al., 2021; Steinberger, 2020). Nevertheless, these students are often studying non-mathematical disciplines and do not have an extensive background in mathematics. Therefore, these courses frequently produce high levels of anxiety in students of the social sciences, and are viewed as a negative experience (Onwuegbuzie & Wilson, 2003; Vigil-Colet, et al., 2008). The specific characteristics of this situation led to the development of statistics anxiety as a concept different from others such as test anxiety or mathematical anxiety (Zeidner, 1991). Although it is related to mathematics anxiety, it requires an understanding not only of mathematical symbols but also of data processing (Chew & Dillon, 2014b; Cui, et al., 2019).

There are various definitions of statistics anxiety, but most of them agree that it may be defined as a feeling of anxiety when encountering statistics in any form and at any level that generates worry, tension and mental disorganization (Onwuegbuzie, et al., 1997; Zeidner, 1991). The most important consequence of it is low performance in statistics-related courses. Many studies using a range of measures have found a negative relationship between statistics anxiety and achievement in different cultures (Cantinotti et al., 2017; Oliver et al., 2014; Onwuegbuzie & Wilson, 2003; Vigil-Colet et al., 2008).

Because of the difficulties students have with this kind of subject, various measures have been developed to assess statistics anxiety. One of the best known is the Statistical Anxiety Rating Scales (STARS) developed by Cruise et al. (1985), consisting of 6 subscales. Nevertheless, only three subscales specifically address the measurement of statistics anxiety: Test and Class anxiety, Fear of Asking for Help Anxiety and Interpretation Anxiety. The others are related to student attitudes towards statistics (Chew & Dillon, 2014b; Cui et al., 2019). Taking all this into account, Vigil-Colet et al. (2008) developed the statistical anxiety scale (SAS) with three main aims: to focus specifically on statistics anxiety, to be shorter than STARS and to be better adapted to statistics in the social sciences.

The SAS consisted of 24 items measuring examination anxiety, asking for help anxiety and interpretation anxiety. The study by Vigil-Colet et al. (2008) showed that SAS had a good fit to the proposed three-dimensional structure, good reliability (ranging from α=.819 to α=.924) and the expected negative relationship with academic performance in statistics.

Despite the good properties of SAS, two issues seem to indicate that it needs to be revised and a new version developed that improves the original.

Firstly, SAS has been adapted to other languages such as Bangladeshi (Paul, et al., 2018), English (O'Bryant et al., 2021), French (Cantinotti et al., 2017), Italian (Chiesi et al., 2011), Portuguese (Hernandez et al., 2015) and Turkish (Durak & Karagöz, 2021). Although some studies have clearly confirmed the factorial structure proposed for SAS (Cantinotti et al., 2017; Oliver et al., 2014), many others have shown a good fit to the three-factor structure only after allowing correlations between some items' error terms (Durak & Karagoz, 2021; Chew & Dillon, 2014a; Chiesi et al., 2011; Frey-Clark et al., 2019; Hernandez et al., 2015; Paul et al., 2018), and other authors such as O'Bryant (2017) and O'Bryant et al. (2021) have proposed a modified bifactorial structure with correlated errors, which was also found by Frey-Clark et al. (2019). Therefore, a revision of SAS may provide a factorial structure that is even clearer than that of the original version.

Secondly, SAS has no procedure to control response biases, defined as a systematic tendency to answer items on some other basis than specific item content (Paulhus, 1991). Of the different response biases, the most important are social desirability (SD), defined as the tendency for people to present themselves in a generally favourable fashion (Holden, 2010), and acquiescence (AC), defined as the tendency of respondents to agree with statements without regard to their content (Paulhus & Vazire, 2005).

Both response biases play an important role in measuring statistics anxiety. SD has been shown to affect the scores of personality traits related to anxiety and to measures of trait anxiety and negative affect (Soubelet & Salthouse, 2011; Vigil-Colet et al., 2013). So it is important to determine whether SD has any effect on the assessment of statistics anxiety and whether it can distort SAS' scores. AC can generate distortions in the factorial structure of self-reports and decrease the predictive validity of measures (Hernández-Dorado et al., 2021; Navarro-Gonzalez et al., 2016; Rammstedt & Farmer, 2013; Vigil-Colet et al., 2020). Therefore, a revised version of SAS may determine whether AC is present or not and, if it is, control its effects.

Taking everything into account, we believe that a revised version of SAS needs to be developed to override these limitations by providing a clearer factorial structure based on the same three scales as in SAS and controlling the two most important response biases (SD and AC).

In recent decades, various approaches have been proposed for controlling response biases but most of them make assumptions that are almost never true and may remove meaningful variance from the trait they intend to measure (Leite & Cooper, 2010; Li & Bagger, 2006).

To solve these problems, Ferrando et al. (2009) developed a restricted FA model, which simultaneously assesses the effects of AC and SD, models them as additional factors that can be distinguished from content factors. Therefore, the procedure removes the effect of both response biases from the factor structure, and makes it possible for the item structure to be analysed once the distortion generated by SD and AC has been removed so the participants' estimated factor scores are free of response biases effects.

The procedure involves using a few items as SD markers, so the SD loadings of the content items can be computed, and direct and reversed items to assess AC and remove its effects.

For this reason, SAS-R will incorporate items designed to measure SD and adapted to the context of a teaching situation in statistics courses, and it will also have direct and reversed items.

To develop the SAS-R, we produced 64 items, which were either new or adapted from the SAS. In this initial pool, each dimension of statistics anxiety had 16 items and there were a further 16 items for assessing SD. Each dimension also had eight directly worded items and eight reversed items. In a second step, six judges with experience in developing typical response measures rated each item. We chose the five items from each scale that had in the highest judges' ratings, with the restriction that each dimension had to have three items in one direction and two in the other in order to control the effects of AC. An additional, dummy item was included as the first item on the scale and was used as a training item when the test was administered online (Ferrando & Lorenzo-Seva, 2005). Thus, the SAS-R consisted of 21 items, five for each dimension and for SD and one dummy item (See table 2).

In the original SAS (Vigil-Colet et al., 2008) each item was a positive sentence describing typical situations that students enrolled on a statistics course might find. The participants had to indicate the level of anxiety that they would feel in these situations using a five-point scale. These items were incompatible with items measuring SD and difficult to reverse. For this reason, the items in the SAS-R were statements and the participants had to indicate their level of agreement on a five-point scale ranging from completely disagree to completely agree. The objective of our research was to propose a short test to measure statistical anxiety that also includes a scale to control for social desirability bias.

## Method

### Participants

Data was collected from students studying a degree in Psychology at six universities in Spain: Universitat Rovira i Virgili (70%), Universitat de les Illes Balears (6%), Universitat de Barcelona (12%), Universitat de València (6%), and Universidad de Oviedo (6%). The sample consisted of 531 first-year undergraduates, whose ages ranged from 18 to 60 ($M$ = 20.4, $SD$ = 4.2). Most of the participants were women (82%), reflecting the gender distribution of the population of psychology students in Spain (Chiesi et al., 2011). In order to compute exploratory and confirmatory factor analyses, the SOLOMON method (Lorenzo-Seva, 2021b) was used to split the sample into two halves. The first subsample (266 participants) was used to compute the EFA, and the second subsample was used to compute the CFA.

For a subsample of 299 participants, the criterion measure described below was available from one of the universities. This subsample was used for collecting the external validity evidence.

### Instruments

Data analysis was computed using FACTOR (Ferrando & Lorenzo-Seva, 2017), Psychological Test Toolbox (Navarro et al., 2019), and Mplus version 5.1 (Muthén, & Muthén, 2007).

As for measures, the SAS-R is described in detail above. The criterion of statistical performance for assessing external validity was the numerical grade in the final examination of the subject Statistics from one of the universities. This grade was always given by the same teacher.

### Procedure

The study was approved by the University's Ethics Committee (Ref: CEIPSA-2021-PR-0028). Respondents stated that they agreed to fill out the instrument online on a website belonging to the Department of Psychology of one of the universities. All students participated on a voluntary basis after they had been informed about the general aim of the research. All the questions on the form were obligatory to ensure that there was no missing data.

The SAS-R was answered by students in the classroom under the supervision of their teacher. The same teacher who supervised the administration of the SAS-R collected the information in order to match the test responses to the academic grades in a course of introductory statistics.

Once the test responses and the academic grades had been matched, alltudents' personal information was discarded from the dataset.

As (a) the item response format was a five-point Likert scale, and (b) most item scores showed asymmetrical distributions with excess kurtosis (normalized estimates above 1 in absolute value for both indices), it was decided to assess the dimensionality and structure of the SAS using the non-linear FA model based on the underlying-variables approach (UVA; Ferrando & Lorenzo-Seva, 2014). At the structural level, this choice essentially entails fitting the common FA model to the polychoric inter-item correlation matrix (Lorenzo-Seva & Ferrando, 2021a).

The procedure used to assess the presence of acquiescent variance in unbalanced scales (Lorenzo-Seva & Ferrando, 2009) found that participants' responses were not contaminated by acquiescent response style in our sample so we made the following analyses without considering acquiescence.

Response variance due to Social Desirability was assessed using the five marker items in the SAS-R. The specific method reported in the literature to control this kind of variance (see Ferrando, Lorenzo-Seva & Chico, 2009) assumes that SD items are factorially pure measures and are not influenced at all by bstantive content items (i.e., anxiety to statistics). However, as the SD items in the SAS-R are contextualized in statistics teaching situations, they may be influenced by some of the anxiety dimensions. For this reason, we decided to allow the SD items to be an extra dimension in the overall factor space that defines SAS-R. In other words, we considered the statistical anxiety dimensions to be substantially correlated with one another, but the SD dimension to be independent from the content dimensions. However, we do not consider the SD items to be pure measures of the SD factor. Rather, some of them are complex, with the most salient loading on SD but also some non-negligible cross-loadings on the content factors.

To make a preliminary exploration of the advisable dimensionality, we used HULL to analyse the correlation matrix (Lorenzo-Seva et al., 2011). HULL is advised when the sample is large. Next, the polychoric correlation matrix was factor analysed using Robust Unweighted Least Squares (RULS), and the four extracted factors were obliquely rotated using Robust Promin (Lorenzo-Seva & Ferrando, 2019). Robust Promin aims to find simple and stable solutions: to do so, it gives greater importance to the loadings related to the most consistent correlations.

Three facets of goodness of model data fit were ascertained (see Ferrando, et al., 2022): (a) absolute fit (GFI and RMSR indices), (b) relative fit with respect to the degrees of freedom (RMSEA index), and (c) comparative fit with respect to the null model of independence (CFI index).

As the correlations between anxiety factors were substantial, we decided to further explore if the SAS-R scores could be regarded as essentially unidimensional (Ferrando & Lorenzo-Seva, 2018).

We computed the indices Unidimensional Congruence (UniCo), Explained Common Variance (ECV) and Mean of Item Residual Absolute Loadings (MIREAL) only for anxiety items. Threshold values of UniCo above .95, ECV above .85, and MIREAL below .30 suggest that the hypothesis of essential unidimensionality is tenable.

In order to assess whether the model obtained in the first subsample (a) was generalizable to the target population, and (b) could be specified in a more restricted way, we fitted a CFA solution in the second subsample with the following specifications. First, the loading values that were observed to be salient in the EFA were set as free parameters in the factor model and the rest were constrained to zero. Second, the inter-factor correlations between the SD factor and the content factors were set to zero in the population. The summarized CFA solution was fitted in the second subsample using ULSMV estimation as implemented in Mplus.

As the same clear and interpretable structure was obtained in both sub-samples, we decided that the best final structural assessment was to fit a semi-confirmatory solution based on a target matrix to the total sample of 265 respondents. The Polychoric correlation matrix based on the total sample was analysed using RULS, and the direct factor solution was rotated against the target matrix specified according to the cross-validation results obtained from oblique Procrustean rotation.

To assess how strong and replicable the estimated factor solution was, the generalized H (G-H) indices were computed. A G-H value above .80 suggests a strong, well-defined factor that is expected to be stable and replicable across studies.

In order to estimate the levels of individuals on the underlying factors, factor score estimates derived from the UVA-FA solution were computed. These scores have two main properties (e.g. Ferrando & Lorenzo-Seva, 2016). First, they provide different amounts of accuracy at different trait levels (conditional reliability). Second, they are nonlinearly related to the usual unit-weight sum scores. So, the following points need to be assessed: (a) the amount of general accuracy of the factor score estimates (marginal reliability); (b) the range of trait levels at which the SAS-R scores provide accurate measurement in the target population (information profile), and (c) the extent to which the simple sum scores are appropriate proxies for the trait levels they attempt to measure. As for points (a) and (b) above, factor score estimates were ORION scores (Ferrando & Lorenzo-Seva, 2016), which are based on fully-informative prior Bayes expected a posteriori estimation, and are recommended when the true factor scores are correlated with each other. In order to inspect conditional accuracy, information curves for the content primary factor scores were computed.

To estimate the appropriateness of unit-weight sum scores as proxies for the true trait levels, we computed the DIANA procedure (Ferrando & Lorenzo-Seva, 2021).

Finally, the external validity of the SAS-R scores was assessed by computing the correlation between the factor score estimates and the numerical marks in the statistics exam.

Overall, in order to further study the methodological approaches used in the different analyses, we invite the reader to consult the Decalogue for the factor analysis of test items proposed by Ferrando et al. (2022). Our analyses are actually based on this proposal.

## Results

The first analyses assessed whether the SAS-R items were appropriate for the test. In order to explore location properties, it can be observed that raw item means ranged from 1.73 to 4.38. Proportional means or relative difficulty indices (RDI) are shown in Table 1, most of which are in the range [.40 - .60]. It must be noted that this is precisely the range usually advised if maximized individual differentiation with medium inter-item correlation values is to be achieved (e.g. Lord, 1952). The second preliminary measure of item appropriateness was the Normed Measure of Sampling Adequacy (N-MSA). N-MSA values in the order of .50 suggest that the item behaves almost at random and is disconnected from the other items in the pool. Items with values above the .50 threshold suggest that they share a substantial proportion of common variance with the other items (Lorenzo-Seva & Ferrando, 2021b). Table 1 shows that all the N-MSA estimates were significantly above this threshold. As an overall measure of sampling adequacy, the Kaiser-Meyer-Olkin (KMO) test estimate was .841 (95th confidence interval values of .807 and .866). To conclude this preliminary stage, all the items can be regarded as suitable for inclusion in the test, and the polychoric correlation matrix as suitable for undergoing factor analysis.

**Table 1.**
Preliminary Item Statistics.

| Item | Dimension | RDI | N-MSA | | |
|---|---|---|---|---|---|
| | | | Point estimate | 95th Confidence interval | |
| 6 | EA- | .430 | .838 | .778 | .865 |
| 14 | EA- | .488 | .833 | .753 | .863 |
| 10 | EA+ | .660 | .851 | .786 | .871 |
| 2 | EA+ | .721 | .861 | .812 | .880 |
| 18 | EA+ | .844 | .861 | .732 | .880 |
| 15 | AH+ | .387 | .872 | .834 | .886 |
| 7 | AH+ | .420 | .907 | .868 | .915 |
| 3 | AH+ | .427 | .881 | .845 | .894 |
| 19 | AH- | .573 | .889 | .842 | .908 |
| 11 | AH- | .636 | .888 | .844 | .904 |
| 20 | IA- | .313 | .653 | .505 | .734 |
| 16 | IA+ | .469 | .844 | .727 | .871 |
| 4 | IA- | .493 | .827 | .724 | .861 |
| 8 | IA+ | .521 | .837 | .757 | .871 |
| 12 | IA- | .545 | .800 | .697 | .846 |
| 5 | SD+ | .183 | .708 | .526 | .774 |
| 21 | SD+ | .233 | .698 | .627 | .735 |
| 13 | SD+ | .370 | .709 | .631 | .744 |
| 17 | SD- | .417 | .707 | .596 | .743 |
| 9 | SD- | .486 | .773 | .636 | .814 |

Note. EA: examination anxiety; AH: asking for help anxiety; IA: interpretation anxiety; SD: social desirability; RDI: Relative Difficulty Index; N-MSA: Normed Measure of Sampling Adequacy.

As already pointed out, the first subsample was used to explore the dimensionality and structure of the SAS-R based on the UVA exploratory factor analysis model, HULL recommended four dimensions.

Goodness-of-fit index values were: (a) *GFI* = .989, *RMSR* = .0380 (Kelly's threshold = .0614); (b) *RMSEA* = 0.018, and (c) *CFI* = .997. These values suggest an excellent fit in all the facets considered. Inspection of the factor loading estimates showed that the four extracted factors consisted of salient loadings, which were congruent with the theoretically expected SAS structure. Therefore, we can safely label the factors Examination anxiety (EA), Asking for help (AH), Interpretation anxiety (IA), and Social Desirability (SD).

Only one content item showed a substantial cross loading (i.e., the item had more than one salient loading value on at least one factor): it was item 10 (I feel like I have a knot in my stomach on the morning of a statistics exam) and it loaded on the EA and SD factors. The items that showed most cross loadings were SD markers. For example, item five (I use a statistics book to expand on the material covered in class) had a salient loading on the SD factor, but also on the IA and EA factors.

The estimated inter-factor correlation matrix showed that the correlations between anxiety factors were between .23 and .42. However, in agreement with the expectations above, the correlations between the SD factors and the anxiety factors were only between .10 and .01.

As already explained, indices to assess the essential unidimensionality were computed. The observed values were *UniCo* = .879 (95th confidence interval values of .815 and .943), *ECV* = .704 (95th confidence interval values of .653 and .758), and *MIREAL* = .366 (95th confidence interval values of .326 and .398). So the EFA-bases conclusion was that regarding the SAS-R as essentially unidimensional would entail a considerable loss of information, and that the multidimensional factor model considered has a coherent substantive interpretation.

The CFA solution in the second subsample produced goodness of model-data fit results that were *GFI* = .984, *RMSR* = .0448, *RMSEA* = 0.046, and *CFI* = .943. They suggest that the proposed SAS structure is tenable in the population, and clear enough to be specified in a restricted way. However, as usually occurs when restricted solutions are fitted to personality data, the additional simplicity in the factorial pattern was achieved at the cost of higher inter-factor correlation estimates among the content factors.

As already explained, the total sample was finally factor-analysed. Table 2 shows the loading estimates. As can be seen, the items related to statistical anxiety showed a simple pattern with a single salient loading on the expected factor.

**Table 2.**
Factor Loading Values Estimated for the Total Sample and Proposed Items in English.

| Item | Four-factor solution | | | | One-factor solution |
|---|---|---|---|---|---|
| | EA | AH | IA | SD | SA |
| 1.Preparing for a statistics exam is a stimulating challenge that I enjoy. | | | | | |
| 2.I get nervous just thinking about taking the final exam for a statistics course. | **.743** | .077 | .138 | .076 | **.685** |
| 10.I feel like I have a knot in my stomach on the morning of a statistics exam. | **.711** | .023 | .024 | .173 | **.551** |
| 18.The day right before a statistics exam, I get very nervous if I realise that I can't do some exercises that I thought were going to be easy. | **.506** | -.001 | .029 | .097 | **.397** |
| 14.The day before a statistics exam I feel calm and focused on studying as much as I need. | **-.753** | .071 | .042 | .129 | **-.460** |
| 6.When going into a statistics exam I feel calm and confident that I will do well. | **-.800** | -.043 | .013 | .103 | **-.606** |
| 15.I get very nervous if I have to ask the statistics professor for help interpreting a results table. | .048 | **.888** | .001 | .036 | **.767** |
| 3.I feel very anxious if I have to ask the statistics professor about how to use a probability table. | ,063 | **.887** | -.051 | .062 | **.744** |
| 7.I find it nerve wracking to go to the statistics professor's office to ask questions. | .005 | **.869** | -.013 | .040 | **.708** |
| 19.If I have not understood an explanation given by the statistics professor, I simply ask for a repetition without getting uptight about it. | -.005 | **-.779** | .006 | .188 | **-.641** |
| 11.I feel at ease when I have to ask the statistics professor a question. | .064 | **-.838** | -.116 | .014 | **-.714** |
| 16.I find trying to understand lottery odds an impossible challenge and I get nervous just thinking about it. | .066 | .021 | **.496** | .104 | **.367** |
| 8.I find having to interpret the meaning of a table in a journal article very distressing. | .106 | .121 | **.492** | .139 | **.479** |
| 12.If I don't understand the statistical analyses described in a journal article, I keep calm and study them until I understand. | -.067 | .038 | **-.493** | .149 | **-.309** |
| 4.Trying to understand a mathematical proof is a challenge I like. | -.115 | .025 | **-.498** | .105 | **-.359** |
| 20.I like car ads that include graphs on consumption (litres/km), compliance with pollution standards, etc. | .194 | .002 | **-.514** | .007 | -.164 |
| 21.I study statistics every day of the week, even if I have not had a statistics class. | -.033 | .036 | .042 | **.798** | |
| 13.Before entering a statistics class, I always review the contents from the previous class. | -.020 | -.011 | -.017 | **.702** | |
| 5.I use a statistics book to expand on the material covered in class. | **.214** | .003 | **-.407** | **.268** | |
| 9.If presented with an opportunity where I was certain I wouldn't get caught, I would cheat on a statistics exam without hesitating. | .068 | -.099 | **.311** | **-.252** | |
| 17.I only study statistics in the days leading up to an exam. | .004 | .133 | -.076 | **-.618** | |

Note. EA: examination anxiety; AH: asking for help anxiety; IA: interpretation anxiety; SD: social desirability; SA: statistical anxiety.

Two SD items (5 and 9) showed a complex pattern. The loading value of item 5 (I use a statistics book to expand on the material covered in class) on IA means that students who do not complement their study with a statistical handbook are the ones with the highest interpretation anxiety scores. In addition, the salient loading of this item on EA means that the students who do complement their study with a statistical handbook are the ones who feel anxious about the examination. The loading value of item 9 (If presented with an opportunity where I was certain I wouldn't get caught, I would cheat on a statistics exam without hesitating) on IA means that students with the highest scores on interpretation anxiety would prefer not to cheat during the examination.

We also computed the unidimensional solution using only the anxiety items. It is in the column Statistical Anxiety (SA). As mentioned above, this solution entails a loss of information, but it is simpler, and could be useful for preliminary usages such as quick screenings. It is apparent that the IA items do not contribute to the general factor as strongly as the EA and AH items do, which is why the full set was not considered to be essentially unidimensional above. However, the column of loadings shows positive manifold and its estimated values are all substantial, which supports the use of general scores in situations in which they could be appropriate.

The estimated inter-factor correlation matrix is in Table 3. As can be seen, the correlations between the statistical anxiety factors are substantial. On the other hand, the correlations of SD factors with the anxiety factors are not, and only one of them was significant.

Generalized H (G-H) indices were computed next. The outcome is displayed in Table 4 where it is noted that IA still does not reach the above threshold. However, all the other factors, including the general factor, can be regarded as very strong, well determined and replicable.

**Table 3.**
Interfactor Correlation Estimates for the Total Sample.

| Factors | Correlation | 95th Confidence interval | |
|---|---|---|---|
| EA -- AH | **.321** | .227 | .417 |
| EA -- IA | **.521** | .585 | .480 |
| AH -- IA | **.260** | .370 | .164 |
| EA -- SD | .059 | -.050 | .163 |
| AH -- SD | .019 | -.107 | .118 |
| IA -- SD | **-.126** | -.034 | -.288 |

Note. EA: examination anxiety; AH: asking for help anxiety; IA: interpretation anxiety; SD: social desirability.

**Table 4.**
Construct Replicability.

| Factor | G-H | 95th Confidence Intervals | |
|---|---|---|---|
| Statistical anxiety | **.895** | .870 | .906 |
| Examination anxiety | **.880** | .856 | .898 |
| Asking for help anxiety | **.941** | .921 | .950 |
| Interpretation anxiety | .730 | .683 | .752 |
| Social desirability | **.796** | .733 | .829 |

Table 5 shows the general accuracy of ORION score estimates. Overall, all the score estimates in table 5 are well determined and accurate. However, if scores are to be used in individual assessment to make fine differentiations among individual levels with minimum error, FDI values need to be above .90, and marginal reliabilities above .80 (Ferrando & Lorenzo-Seva, 2018). So, IA and SD scores must be interpreted with great care because they have most associated error.

We turn now to conditional accuracy. Information curves for the content primary factor scores (i.e., EA, AH, & IA) are shown in Figure 1. To relate the profiles in Figure 1 to the marginal reliabilities in table 5, we regard the marginal reliability as being proportional to the area under the corresponding curve. Thus, the IA information curve is the lowest, which agrees with the smaller marginal reliability estimate (and also with the smallest G-H estimate). However, the profile in Figure 1 also shows that IA scores provide almost constant information with values around .8 across the trait range that contains virtually all the population. EA and AH scores are more reliable, but the curves are more peaked and provide maximal accuracy at different levels. Thus, EA scores are more accurate at low levels while AH scores are more accurate at high levels.

**Table 5.**
Quality and Effectiveness of Factor Score Estimates.

| Index | SA | EA | AH | IA | SD |
|---|---|---|---|---|---|
| Factor Determinacy Index (FDI) | .970 | .954 | .971 | .888 | .894 |
| ORION EAP marginal reliability | .941 | .910 | .943 | .789 | .799 |

Note. EA: examination anxiety; AH: asking for help anxiety; IA: interpretation anxiety; SD: social desirability; SA: statistical anxiety.

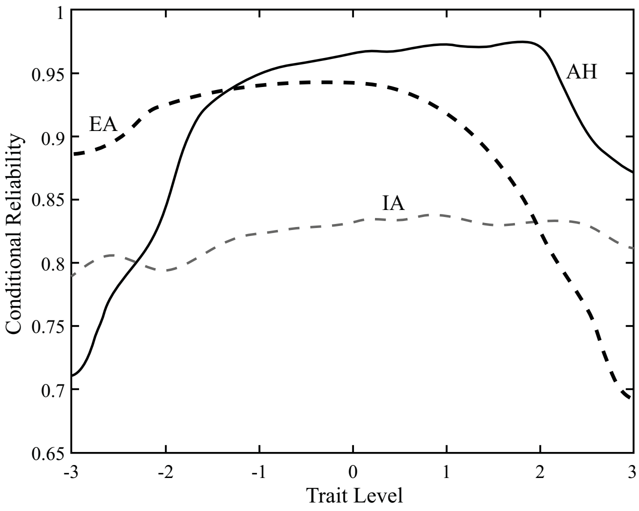

**Figure 1.**
Information Curves. SAS-R Content Primary Factor Scores.

Figure 2 shows the information curve for the general factor score estimates (i.e., SA). It is a well-filled curve and the conditional reliability is above .87 at all trait levels. However, it provides the most accurate measurement at high trait levels (about one and a half standard deviations above the mean).

The results of the DIANA procedure are in Table 6 and contain two main pieces of information. First, the coefficient of fidelity (O-COF): the estimated correlation between the scores and the 'true' trait levels they measure. Second, the amount of stability of the score estimates. As expected, factor score estimates are more precise (larger O-COF values), but somewhat less stable. However, as there is little difference in stability and the precision is much greater for factor scores estimates, we conclude that the optimal scoring schema for the SAS-R is to use factor score estimates. Even so, sum scores have reasonable fidelity values and can be justifiably used as simpler proxies at the cost of a considerable loss in accuracy.

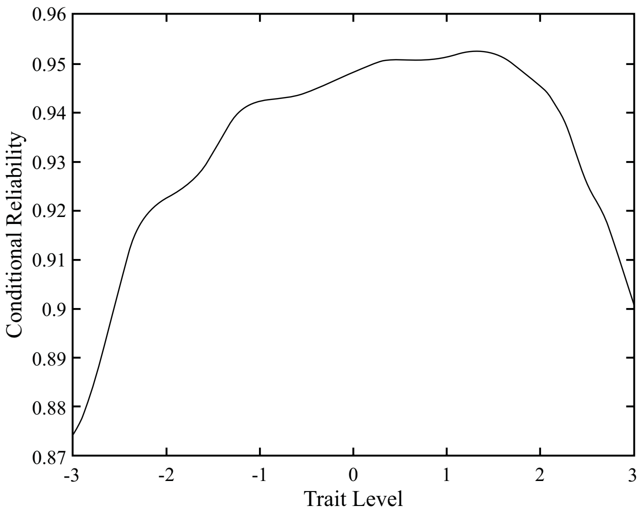

**Figure 2.**
Information Curves. SAS-R General Factor Scores.

**Table 6.**
Quality and Effectiveness of Factor Score Estimates.

| Factors | DIANA sum scores | | Factor score estimates | |
|---|---|---|---|---|
| | O-COF | Stability | O-COF | Stability |
| Statistical anxiety | .858 | 1 | .970 | .997 |
| Examination anxiety | .897 | .989 | .957 | .989 |
| Asking for help anxiety | .944 | 1 | .972 | .994 |
| Interpretation anxiety | .788 | .899 | .897 | .908 |
| Social desirability | .820 | 1 | .865 | .988 |

Note. O-COF: Ordinal Coefficient of Fidelity

Finally, we assessed the external validity of the SAS-R scores. We expected SA, EA, AH and IA, but not SD, to correlate substantially and negatively with the numerical grade. Table 7 shows the disattenuated zero-order correlations, corrected using the ORION marginal reliability estimates. As expected, all the anxiety scores correlate negatively with academic performance, while SD does not. The scores that best predict student grades on the exam are EA and SA. Although IA was the measure that showed less quality and effectiveness, it would still be an acceptable predictor of academic performance. Nevertheless, Fisher's z test did not show any significant difference between the magnitude of the correlations.

**Table 7.**
External Validity Study.

| Factors | *r* | 95th Confidence interval | | *dis-r* | 95th Confidence interval | |
|---|---|---|---|---|---|---|
| Statistical anxiety | -.319 | -.430 | -.178 | -.328 | -.444 | -.183 |
| Examination anxiety | -.328 | -.437 | -.211 | -.344 | -.458 | -.221 |
| Asking for help anxiety | -.220 | -.336 | -.091 | -.227 | -.346 | -.094 |
| Interpretation anxiety | -.279 | -.390 | -.150 | -.314 | -.440 | -.168 |
| Social desirability | .024 | -.118 | .158 | .027 | -.132 | .176 |

Note. r: Zero-order correlation; dis-r: Disattenuated zero-order correlation.

## Discussion

The aim of this study was to develop and validate a revised version of the SAS scale, to improve its properties and override its limitations. The analyses discussed above showed that, as expected, the EFA yielded a four-dimensional structure in which all the items had their salient loading on the expected dimension: three dimensions related to statistical anxiety, and one related to SD. This four-dimensional structure was then confirmed in a second sample by means of a CFA, which showed a good fit without any need to introduce correlations between error terms, which was a problem found in different adaptations of SAS (Durak & Karagoz, 2021; Chiesi et al., 2011; Frey-Clark et al., 2019; Hernandez et al, 2015; Paul et al, 2018).

The final factorial solution found in the whole sample showed high factor simplicity, a reflection of the fact that the items tended to show high loadings on their factor and close-to-zero loadings on the other factors. The factors were also well defined, stable and replicable. As far as the factor scores are concerned, the results showed that they were much more precise than raw scores (marginal reliability estimates ranging from .79 to .94) with no great loss in stability. Therefore, we advise that factor scores be used instead of scores based on the sum of items, although this second option is also acceptable. We have developed an Excel correction system that computes factor score estimates for the SAS-R which can be found as supplementary materials.

In addition, as the three dimensions related to statistical anxiety showed noticeable correlations with each other, they can be viewed as related subscales that stem from an overall scale. In this regard, the three SAS scales showed moderate correlations between r = .26 and r = .52, a range that is similar to the ones reported for the original version (Vigil-Colet et al., 2008), and its different adaptations (Cantinotti et al., 2017; Chew & Dillon, 2014; Chiesi et al., 2011).

It is also worth mentioning that the overall SAS-R score gives information in a wide range of trait levels. The specific scales also give information in a wide range but EA is more informative at low levels and AH at high levels.

As far as external validity is concerned, SAS scores showed good prediction of academic success with correlations in the range *r* = -.23 to *r* = -.34. These values were similar to or higher than those reported in previous studies (Cantinotti et al., 2017; Oliver et al., 2014; Onwuegbuzie & Wilson, 2003; Vigil et al., 2008). It should be noted that the best predictor of academic performance was EA (*r* = -.34) and the worst AH (*r* = -.23). Nevertheless, the difference in magnitude between those correlations is not significant so differential predictive power between the three subscales and the measure of academic performance cannot be assumed. This is similar to the results reported in studies using the original version of SAS (Cantinotti et al., 2017; Vigil-Colet et al., 2008), but other authors such as Oliver et al. (2014) only found significant correlations for AH.

As we stated above, one of the reasons we developed SAS-R was to introduce response bias control into the test. Our results seem to show that SAS-R is free of AC in this version of the test. However, in future adaptations of the scale in other cultures AC effects may be non-negligible so we recommend that these effects be tested in future versions. As far as SD is concerned, the SAS items showed low-to-moderate loadings on the SD factor. The average loading of SD on each scale was $\lambda_m = .116$ for EA, $\lambda_m = .07$ for AH and $\lambda_m = .11$ for IA. Therefore, although SAS-R is not deeply impacted by SD, it should be controlled because its effects are non-negligible and of a similar magnitude to those observed in anxiety-related measures, such as emotional stability (Vigil-Colet et al., 2013).

The study reported here has some limitations that future research will have to explore. First, regarding validity evidence, we studied how effective SAS-R scores are at predicting academic performance in some introductory courses on the psychology degree, but further research will have to assess if the results are generalizable to other fields and to more advanced statistical courses. As well as external validity, it would be of great importance to explore other sources of evidence such as: (a) convergent validity with related measures, (b) multiple-group invariance, (c) incremental external validity, and possible mediating effects, when more broad-bandwidth anxiety measures (e.g. neuroticism) or cognitive measures (e.g. intelligence tests) are used in conjunction with the SAS-R.

As far as administration is concerned, in the present study the administration of SAS-R was computerized so further research will have to analyse the equivalence of the computerized and paper-and-pencil administration of SAS-R. This equivalence was established for the original SAS (Frey-Clark et al., 2019), but has not yet been established for SAS-R.

As far as response bias control is concerned, SAS-R did not show any AC effects and only moderate SD effects. Nevertheless, these effects have been found in a specific culture and there may be cross-cultural differences. So future research should analyse whether SAS-R adaptations to other cultures show the same or different degrees of response bias effects (Johnson et al., 2005).

Courses on statistics are commonly taught at universites. However, they are also commonly taught to students preparing for university degrees. In the future, it would be interesting to determine the properties of SAS-R in this population of students, so that the test can be properly validated for this population.

## Acknowledgements

The authors are in debt to colleagues who have helped us obtain data from various universities: Montserrat Domènech, Albert Sesé, Joan Guàrdia, Inés Tomás, Salvador Algarabel, and Marcelino Cuesta. This paper has been made possible by the support of the Ministerio de Ciencia e Innovación, the Agencia Estatal de Investigación (AEI) and the European Regional Development Fund (ERDF) (PID2020-112894GB-I00).

## References

Cantinotti, M., Lalande, D., Ferlatte, M. A., & Cousineau, D. (2017). Validation de la version francophone du Questionnaire d'anxiété statistique (SAS-F-24) [Validation of the French version of the statistical anxiety scales]. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement, 49*(2), 133–142. https://dx.doi.org/10.1037/cbs0000074

Chew, P. K., & Dillon, D. B. (2014a). Reliability and validity of the Statistical Anxiety Scale among students in Singapore and Australia. *Journal of Tropical Psychology, 4*(e7), 1-14. https://doi.org/10.1017/jtp.2014.7

Chew, P. K., & Dillon, D. B. (2014b). Statistics anxiety update: Refining the construct and recommendations for a new research agenda. *Perspectives on Psychological Science, 9*(2), 196-208. https://doi.org/10.1177/1745691613518077

Chiesi, F., Primi, C., & Carmona, J. (2011). Measuring statistics anxiety: Cross-country validity of the Statistical Anxiety Scale (SAS). *Journal of psychoeducational assessment, 29*(6), 559-569. https://doi.org/10.1177/0734282911404985

Cui, S., Zhang, J., Guan, D., Zhao, X., & Si, J. (2019). Antecedents of statistics anxiety: An integrated account. *Personality and Individual Differences, 144*, 79-87. https://doi.org/10.1016/j.paid.2019.02.036

Cruise, R. J., Cash, R. W., & Bolton, D. L. (1985). Development and validation of an instrument to measure statistical anxiety. *American Statistical Association Proceedings of the Section on Statistical Education, 4*(3), 92-97.

Durak, I., & Karagöz, Y. (2021). Adaptation of Statistics Anxiety Scale to Turkish: Validity and Reliability Study. *International Journal of Assessment Tools in Education, 8*(3), 667-683. https://doi.org/10.21449/ijate.863225

Ferrando, P.J., & Lorenzo-Seva, U. (2005). IRT-related factor analytic procedures for testing the equivalence of paper-and-pencil and Internet administered questionnaires. *Psychological Methods, 10*(2), 193-205. https://doi.org/10.1037/1082-989X.10.2.193

Ferrando, P. J., & Lorenzo-Seva, U. (2014). Exploratory item factor analysis: Additional considerations. *Anales de psicología, 30*(3), 1170-1175. https://doi.org/10.6018/analesps.30.3.199991

Ferrando, P. J., & Lorenzo-Seva U. (2016). A note on improving EAP trait estimation in oblique factor-analytic and item response theory models. *Psicologica, 37*(2), 235-247.

Ferrando, P.J., & Lorenzo-Seva, U. (2017). Program FACTOR at 10: origins, development and future directions. *Psicothema, 29*(2), 236-241. https://doi.org/10.7334/psicothema2016.304

Ferrando, P. J., & Lorenzo-Seva U. (2018). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educational and Psychological Measurement, 78*(5), 762-780. https://doi.org/10.1177/0013164417719308

Ferrando, P. J., & Lorenzo-Seva, U. (2021). The Appropriateness of Sum Scores as Estimates of Factor Scores in the Multiple Factor Analysis of Ordered-Categorical Responses. *Educational and Psychological Measurement, 81*(2), 205-228. https://doi.org/10.1177/0013164420938108

Ferrando, P. J., Lorenzo-Seva, U., & Chico, E. (2009). A general factor-analytic procedure for assessing response bias in questionnaire measures. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(2), 364-381. https://doi.org/10.1080/10705510902751374

Ferrando, P. J., Lorenzo-Seva, U., Hernández-Dorado, A., & Muñiz, J. (2022). Decalogue for the Factor Analysis of Test Items. *Psicothema, 34*(1), 7-17. https://doi.org/10.7334/psicothema2021.456

Frey-Clark, M., Natesan, P., & O'Bryant, M. (2019). Assessing statistical anxiety among online and traditional students. *Frontiers in Psychology, 10*, Article 1440. https://doi.org/10.3389/fpsyg.2019.01440

Hernandez, J. A. E., Santos, G. R. D., Silva, J. D. O. D., Mendes, S. L. L., & Ramos, V. D. C. B. (2015). Validity of the Statistics Anxiety Scale in Psychology Students. *Psicologia: Ciência e profissão, 35*(3), 659-675. https://doi.org/10.1590/1982-3703000362014

Hernández-Dorado, A., Vigil-Colet, A., Lorenzo-Seva, U., & Ferrando, P. J. (2021). Is correcting for acquiescence increasing the external validity of personality test scores? *Psicothema, 33*(4), 639-646. https://doi.org/10.7334/psicothema2021.131

Holden, R. (2010). Social desirability. *Corsini Encyclopedia of Psychology.* John Wiley & Sons, Inc.

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-cultural psychology, 36*(2), 264-277. https://doi.org/10.1177/0022022104272905

Leite, W.L., & Cooper, L.A. (2010). Detecting social desirability bias using factor mixture models. *Multivariate Behavioral Research, 45*(2), 271-293. https://doi.org/10.1080/00273171003680245

Li, A., & Bagger, J. (2006). Using the BIDR to distinguish the effects of impression management and self-deception on the criterion validity of personality measures: A meta-analysis. *International Journal of Selection & Assessment, 14*(2), 131-141. https://doi.org/10.1111/j.1468-2389.2006.00339.x

Lord, F. (1952). A theory of test scores. *Psychometric Monographs 7*.

Lorenzo-Seva, U. (2021). SOLOMON: a method for splitting a sample into equivalent subsamples in factor analysis. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01750-y

Lorenzo-Seva, U., & Ferrando, P. J. (2009). Acquiescent responding in partially balanced multidimensional scales. *British Journal of Mathematical and Statistical Psychology, 62*(2), 319-326. https://doi.org/10.1348/000711007X265164

Lorenzo-Seva, U., & Ferrando, P.J. (2019). Robust Promin: a method for diagonally weighted factor rotation. LIBERABIT, *Revista Peruana de Psicología, 25*, 99-106. https://doi.org/10.24265/liberabit.2019.v25n1.08

Lorenzo-Seva, U., & Ferrando, P. J. (2021a). Not positive definite correlation matrices in exploratory item factor analysis: causes, consequences and a proposed solution. *Structural Equation Modeling: A Multidisciplinary Journal, 28*(1), 138-147. https://doi.org/10.1080/10705511.2020.1735393

Lorenzo-Seva, U., & Ferrando, P. J. (2021b). MSA: The Forgotten Index for Identifying Inappropriate Items Before Computing Exploratory Item Factor Analysis. *Methodology, 17*(4), 296-306. https://doi.org/10.5964/meth.7185

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H.A.L. (2011). The Hull method for selecting the number of common factors. *Multivariate Behavioral Research, 46*(2), 340-364. https://doi.org/10.1080/00273171.2011.564527

Muthén L.K. & Muthén, B.O. (2007). *Mplus user's guide* (5th Ed.). Muthén & Muthén.

Navarro-González, D., Lorenzo-Seva, U., & Vigil-Colet, A. (2016). How response bias affects the factorial structure of personality self-reports. *Psicothema, 28*(4), 465-470. https://doi.org/10.7334/psicothema2016.113

Navarro-González, D., Vigil-Colet, A., Ferrando, P. J., & Lorenzo-Seva, U. (2019). Psychological Test Toolbox: A New Tool to Compute Factor Analysis Controlling Response Bias. *Journal of Statistical Software, 91*(6), 1-21. https://doi.org/10.18637/jss.v091.i06

O'Bryant, M. J. (2017). *How attitudes towards statistics courses and the field of statistics predicts statistics anxiety among undergraduate social science majors: a validation of the Statistical Anxiety Scale.* [Doctoral dissertation, University of North Texas]. ProQuest Dissertations & Theses Global. https://search.proquest.com/docview/2009455494

O'Bryant, M., Natesan Batley, P., & Onwuegbuzie, A. J. (2021). Validation of an Adapted Version of the Statistical Anxiety Scale in English and Its Relationship to Attitudes Toward Statistics. *SAGE Open, 11*(1), 1-15. https://doi.org/10.1177/21582440211001378

Oliver, A., Sancho, P., Galiana, L., & Cebrià i Iranzo, M. A. (2014). Nueva evidencia sobre la Statistical Anxiety Scale (SAS) [New evidence on the Statistical Anxiety Scale (SAS)]. *Anales de psicología, 30*(1), 150-156. https://doi.org/10.6018/analesps.30.1.151341

Onwuegbuzie, A. J., Da Ros, D., & Ryan, J. (1997). The components of statistics anxiety: A phenomenological study. *Focus on Learning Problems in Mathematics, 19*(4), 11-35.

Onwuegbuzie, A. J., & Wilson, V. (2003). Statistics anxiety: Nature, etiology antecedents, effects, and treatments—A comprehensive review of the literature. *Teaching in Higher Education, 8*(2), 195-209. https://doi.org/10.1080/1356251032000052447

Paul, L., Parveen, T., Ahmed, O., & Aktar, R. (2018). Adaptation study of the statistical anxiety scale on a Bangladeshi sample. *Bulgarian Journal of Science & Education Policy, 12*(2), 380-401.

Paulhus D.L. (1991). Measurement and Control of Response Bias. In J. P. Robinson, P. R. Shaver, & L. S. Wright (Eds.), *Measures of personality and social psychological attitudes* (pp. 17-59). Academic Press.

Paulhus, D. L., & Vazire, S. (2005). The Self-Report Method. In R. W. Robins, R. Fraley & R. F. Krueger (Eds.). *Handbook of research methods in personality psychology* (pp. 224-239). Guilford Press.

Rammstedt, B., & Farmer, R. F. (2013). The impact of acquiescence on the evaluation of personality structure. *Psychological Assessment, 25*(4),1137-1145. https://doi.org/10.1037/a0033323

Soubelet, A., & Salthouse, T.A. (2011). Influence of social desirability on age differences in self-reports of mood and personality. *Journal of Personality, 79*(4), 741-762. https://doi.org/10.1111/j.1467-6494.2011.00700.x

Steinberger, P. (2020). Assessing the statistical anxiety rating scale as applied to prospective teachers in an Israeli teacher-training college. *Studies in Educational Evaluation, 64*, Article 100829. https://doi.org/10.1016/j.stueduc.2019.100829

Vigil-Colet, A., Lorenzo-Seva, U., & Condon, L. (2008). Development and validation of the statistical anxiety scale. *Psicothema, 20*(1), 174–180.

Vigil-Colet, A., Morales-Vives, F., Camps, E., Tous, J., & Lorenzo-Seva, U. (2013). Development and validation of the Overall Personality Assessment Scale (OPERAS). *Psicothema, 25*(1), 100-106. https://doi.org/10.7334/psicothema2011.411

Vigil-Colet, A., Navarro-González, D., & Morales-Vives, F. (2020). To reverse or to not reverse Likert-type items: That is the question. *Psicothema, 32*(1), 108-114. https://doi.org/10.7334/psicothema2019.286

Zeidner, M. (1991). Statistics and mathematics anxiety in social science students: Some interesting parallels. *British journal of educational psychology, 61*(3), 319-328. https://doi.org/10.1111/j.2044-8279.1991.tb00989.x