Abstract

The increasing use of tests to assessing academic competences has been associated with higher levels of test anxiety (TA) in children, underlining the importance of identifying children who suffer from moderate-to-high levels of TA in order to help them achieve their goals at school. This study aimed to contribute to the extant literature on the assessment of TA by examining the psychometric properties of the Test Anxiety Questionnaire for Children (TAQ-C), in primary and middle school children. In Study 1 (*N*=123), we selected 24 items from a wider initial pool, dividing them into scales measuring Thoughts, Autonomic Reactions, Off-Task Behaviors, and Social Derogation, to develop the TAQ-C. In Study 2 (*N*=899), the psychometric properties of this set of scales were assessed in students attending primary and middle school. Analyses supported the bifactor latent structure of the TAQ-C, invariance across educational levels and gender, concurrent and convergent validity and test-retest reliability. Overall, the TAQ-C seems to be a promising tool for assessing TA in primary and middle school students. Implications and directions for future research are discussed.

Keywords: Test Anxiety Questionnaire for Children; factorial structure; bifactor model; measurement invariance; psychometric properties.

THE TEST ANXIETY QUESTIONNAIRE FOR CHILDREN

Test anxiety (TA) has been defined as the tendency to experience evaluative situations, such as examinations, as threatening for a person, who may worry about failing exams, and develop negative physiological responses and avoidant behaviors (Zeidner, 1998). Much research suggests that children who report higher levels of TA are more likely to have lower grades (McDonald, 2001; Segool, Carlson, Goforth, von der Embse, & Barterian, 2013), and a worse test performance (Putwain, 2008; von der Embse & Witmer, 2014). It is consequently very important to identify children who suffer from moderate or severe TA by means of reliable and up-to-date tools. Then such children can be given emotional support so that their academic performance can reflect their real academic ability (Erford, Mckechnie, & Moore-Thomas, 2004; von der Embse & Hasson, 2012). There are some self-report tools that aim to assess TA in children, but several have been adapted from those designed for the adult population (Anderson & Sauser, 1995; see also Wren & Benson, 2004), and others fail to capture some dimensions, such as TA relating to social concerns (see Lowe, Grumbein, & Raad, 2011). Hence the need to improve the assessment of TA in school-age children (Weems et al., 2010). The present study aims to contribute to the extant research by examining the psychometric properties of the Test Anxiety Questionnaire for Children (TAQ-C), a self-report tool for assessing TA in primary and middle school students.

Test anxiety in children

Theories about TA have changed since it was first conceptualized in the 1950s (Sarason & Mandler, 1952). An initial contribution in this field initially suggested the presence of a single factor (Sarason, Davidson, Lighthall, Waite, & Ruebush, 1960), while subsequent theories considered TA as a multidimensional construct (Lowe et al., 2008; Zeidner, 1998). Concerning the study of TA in children, a prominent contribution came from Wren and Benson (2004) who, in the light of previous research on adults, proposed a reconceptualization of the available evidence on TA in children. They suggested that worry and task-irrelevant thoughts were indistinguishable in children, and that both these components could reflect the cognitive dimension of TA (what the authors called *thoughts*). They consequently proposed to operationalize TA in children in terms of different dimensions related to: thoughts (i.e., worries about themselves, task and test-irrelevant thoughts); off-task behaviors (i.e., nervous habits, such as playing with pencils, and distracting avoidant behaviors); and autonomic reactions (i.e., increased heart rate or stomach problems). Although this work made a huge contribution to the research field, the authors failed to include the social component of TA, defined as the presence of concerns about the social consequences of failing a test (Friedman & Bendas-Jacob, 1997; see also von der Embse, Jester, Roy, & Post, 2018). A recent theoretical model has added an important contribution to the topic by providing some comprehensive theoretical background on the role of individual and environmental factors related to the development of TA (Lowe et al., 2008). This model includes and discusses the social and educational factors relating to the influence of schools, parents and families that may be involved in TA, underscoring the importance of considering these aspects.

Some studies considered the interplay between factors that could promote or prevent the onset of TA. Researchers found TA positively related to other forms of anxiety, i.e., general anxiety and mathematics anxiety (Hembree, 1988, 1990; see also Carey, Devine, Hill, & Szűcs, 2017; Devine, Fawcett, Szűcs, & Dowker, 2012). Recent studies have also suggested that children who suffer from higher general anxiety are more at risk of developing specific forms of anxiety, including TA (Carey et al., 2017; Mammarella, Donolato, Caviola, & Giofrè, 2018).

Other research has shown that additional factors that may be negatively related to TA include self-concept and resilience. Previous evidence pointed to children with higher academic self-concept being more likely to report lower levels of TA (Arens, Becker, & Möller, 2017; Bandalos, Yates, & Thorndike-Christ, 1995; Hembree, 1988; Raufelder & Ringeisen, 2016). As concerns resilience, little research is available as yet on the relationship between TA and this protective factor (Martin & Marsh, 2006; Putwain, Nicholson, Connors, & Woods, 2013). One of the few studies on the topic found a negative association between resilience and TA, suggesting that resilient students would experience less TA (Putwain et al., 2013). Overall, self-concept and resilience need to be further explored to better clarify the relationship between these factors and their role in supporting children's success at school.

Test anxiety scales for children

The existing literature includes descriptions of some self-report tools developed to assess TA in school-aged children. One of the first was the Test Anxiety Scale for Children (TASC; Sarason et al., 1960), a self-report tool for assessing TA in children from grades 1 to 6 based on 30 items with a two-choice (yes/no) format that provide a single score. Although this measure has been widely used for many years, some authors have raised concerns about the presence in the items of outdated and complicated words (Lowe, Grumbein, & Raad, 2011). It is also important to consider that, as well as limiting respondents' chance to qualify their choices, the use of a yes/no format has been accused of being more susceptible to acquiescence response bias (see Krosnick, 1999). Other tools developed over time consider TA as a multidimensional construct. For instance, the Children's Test Anxiety Scale (CTAS) is a self-report tool for children from grades 3 to 6, based on 30 items rated on a 4-point Likert scale, which provides scores on scales for Thought, Off-Task Behaviors and Autonomic Reactions (Wren & Benson, 2004). The CTAS reportedly has a good internal consistency (Wren & Benson, 2004; see also Spence, 2018). Other more recently developed self-report tools for measuring TA include the Test Anxiety Scale for Elementary Students (TAS-E; Lowe et al., 2011), and the Test Anxiety Measure for Adolescents (TAMA; Lowe, 2014). Both the TAS-E and the TAMA have been validated in the US, and shown adequate internal consistency and good convergent validity (Lowe, 2014; Lowe et al., 2011). On the whole, the research on self-report questionnaires for assessing children's TA seems to be relatively limited, especially in Europe, suggesting that further research is needed to examine these tools in the light of theoretical advances in TA research (Weems et al., 2010; see also von der Embse et al., 2018).

The present study

The present study aimed to contribute to what is known about the assessment of TA in school-age children by examining the psychometric properties of another self-report tool, the TAQ-C. This questionnaire measures different dimensions of TA in primary and middle school students, focusing on the cognitive, behavioral, physiological and social components of this construct (called 'Thought', 'Off-Task Behaviors', 'Autonomic Reaction', and 'Social Derogation' from here on). This investigation was developed in two interrelated phases: i) a pilot phase in which we selected and adapted a first set of items from existing self-reports (Study 1); and ii) a testing phase in which we tested the psychometric properties of the resulting questionnaire (Study 2). In Study 1, an initial pool of items was drawn, and several confirmatory factor analyses (CFAs) were conducted to obtain a shorter self-report tool. In Study 2, the psychometric properties of the items selected were tested in a large sample of school-age children. In the light of the extant literature, the following models were tested: a) a single factor (Sarason et al., 1960); b) a four-factor model based on the assumption that TA is a multidimensional construct with cognitive, physiological, behavioral and social dimensions (Lowe et al., 2008); and c) a bifactor model, considering a single general factor (Sarason et al., 1960), and the specific dimensions of TA (Lowe et al., 2008). Once the best model had been identified, measurement invariance was tested simultaneously across educational levels (primary vs. middle school students) and gender. The concurrent and convergent validity, and the test-retest reliability were assessed. For the convergent validity, we examined to what extent TA was associated with general anxiety, mathematics anxiety, academic self-concept, and resilience judging from the available literature (Arens et al., 2017; Hembree, 1988; Martin & Marsh, 2006; Putwain et al., 2013; Raufelder & Ringeisen, 2016). Our aim was to test different hypotheses. Concerning the factor structure, despite the handful of studies assessing the bifactor model in TA, we expected TA to be best explained by a single general factor reflecting the variance shared by the four specific factors considered (i.e., Thoughts, Autonomic Reactions, Off-Task Behaviors, and Social Derogation). Specifically, the bifactor model might combine different theoretical conceptualizations, considering TA as a single general factor (Sarason et al., 1960) or multidimensional construct (Lowe et al., 2008; Zeidner, 1998). As for the factor invariance, we expected our questionnaire to assess TA similarly across educational levels (i.e., primary and middle school) and gender. As regards convergent validity, we assumed that TA would correlate positively with general anxiety (Hembree, 1988; see also Carey et al., 2017), and negatively with academic self-concept and resilience (Arens et al., 2017; Hembree, 1988; Putwain et al., 2013; Raufelder & Ringeisen, 2016). Finally, we expected the test-retest reliability to be good because that TA seems to be relatively stable over periods of one or two months (Lowe et al., 2011).

Study 1

We conducted a pilot study to reduce an initial pool of items derived from previous self-reports measuring TA in school-age students.

Method

Participants

The study involved 128 children in grade 5 at several primary schools in the north-east of Italy. Children with intellectual disabilities or other developmental disorders (*n* = 5) were not included in the data analyses. No data were missing in the questionnaires. The final sample thus consisted of 123 children (boys = 46.3%; mean age = 10.89 years, *SD* = 3.94 months).

Procedure

The study was approved by the Ethics Committee on Psychology Research at the University of [blind]. Signed consent was obtained from both parents, and verbal assent was obtained from each child before starting any data collection. Participants were tested in their classrooms during a single collective session lasting approximately 20 minutes. All children completed a socio-demographic form (date of birth, gender), and the draft (37-item) version of the TAQ-C.

Measures

Item selection and design. An initial pool of 37 items was considered, drawing from the Children's Test Anxiety Scale (CTAS; Wren & Benson, 2004), and the FRIEDBEN Test

Anxiety Scale for adolescents (FTA; Friedman & Bendas-Jacob, 1997) on the basis of the items' content and formulation. We considered: a) 13 items assessing worry about giving wrong answers or failing a school test (e.g., "*I think I'm going to get a bad mark*"); b) 8 items assessing somatic responses, such as heart rate or stomach ache (e.g., "*My heart beats fast*"); c) 9 items on nervous habits and distracting behaviors as indexes of avoidance behaviors that might occur during a school test (e.g., "*I play with my pencil*"); and d) 7 items on social concerns about failing in a school test (e.g., "*I'm worried all my friends will get high scores in the test and only I will get low ones*"). The 37 items were identified, translated and adapted from English to Italian using a forward and back translation method. A 4-point Likert scale (from 1= "never" to 4= "always") was used to rate each item, as this format is considered particularly suitable for children (e.g., Borgers, Hox, & Sikkel, 2003). Participants were also prompted to judge how often they experienced each of the situations described during school tests.

Analysis

Analyses were run using R (R Development Core team, 2017).

Descriptive analysis. The item response distributions, the mean, standard deviation, range and skewness of the initial 37 items were first explored at descriptive level.

Item selection. A semi-confirmatory approach based on a series of CFAs was used to reduce the initial pool of items. First, a unidimensional 37-item model was run, in which all items were loaded on a single factor. This model was used as a baseline for comparison with a multidimensional model in which each item was loaded on the corresponding latent factor (i.e., the Thoughts, Autonomic Reactions, Off-Task Behaviors and Social Derogation subscales).

All models were estimated using the diagonal weighted least squares estimator (DWLS), which is designed specifically for ordinal data (Flora & Curran, 2004). For each model, we computed and evaluated several fit indices including the degrees of freedom ratio (χ^2/df), the Root-Mean Square Error of Approximation (RMSEA), the Comparative Fit Index (CFI), the Tucker Lewis Index (TLI), and the Standardized Root Mean Square Residual (SRMR). A χ^2/df lower than 3, a CFI and a TLI higher than .95, a RMSEA lower than .08, and a SRMR lower than .10 were considered acceptable (Schermelleh-Engel, Moosbrugger, & Müller, 2003).

Results

Descriptive analysis. Descriptive statistics (mean, standard deviation, range and skewness) for the 37 items are provided in the supplementary material section (see Table S1).

Factor structure and item selection. First, a baseline one-factor model was tested and compared with the theoretical 4-factor model (see Table 1). Results showed a poor fit for the one-factor solution, and a good fit for the four-factor model (see M1 and M2, Table 1). So item selection based on the four-factor model was conducted to reduce the number of items from 37 to 24 (i.e., 6 items for each theoretical dimension). The item removal process was guided by the following criteria: i) each factor would have 6 items; ii) items with standardized factor loadings of less than .40 were considered weak; iii) each factor had to show face validity. Following this procedure, items were removed in several steps. First, three items were dropped from the Thoughts subscale (see M3, Table 1): items 1 and 18 were found unrelated to the latent structure (i.e., not statistically significant at the 5% level), while item 16 reached a weak factor loading (< .40). Then (see M4, Table 1) three items were removed from the Thoughts, Off-Task Behaviors, and Autonomic Reactions subscales, one for each subscale (M4: 7-17-32). This procedure was repeated (see M5, Table 1) and another three items were removed (M5: 31-37-9). As a result, M5 included 28 items (i.e., 7 items for each subscale). In the final step (see M6, Table 1), 4 more items (M6: 21-28-26-25) were removed, one from each subscale, on the grounds of their relevance to the construct. This procedure yielded a final model comprising 24 items with 6 items for each of the Thoughts, Off-Task Behaviors, Autonomic Reactions, and Social Derogation subscales.

Study 2

In Study 2, we aimed to test the 24 items selected in Study 1 in terms of: a) factor structure and internal consistency; b) invariance across gender (boys vs. girls) and educational levels (primary vs. middle school students); c) concurrent and convergent validity; and d) test-retest reliability.

Method

Participants

The study involved 936 children attending primary (*n* = 594) or middle school (*n* = 342) recruited at State schools in north-eastern Italy. The children came from urban areas and middle-class families. Self-report measures completed by children with intellectual disabilities or developmental disorders, or revealing a critical score in the Cattell Culture Fair Intelligence Test (Cattell & Cattell, 1981) (*n* = 31), or with more than two missing responses for each subscale of the target instrument (*n* = 6) were excluded from analyses. The remaining missing values (< .1%, see Table 2) were handled with the pairwise maximum likelihood (PML) estimation method available in the R *lavaan* package (Rosseel, 2012), and developed for factor analyses with ordinal data (Katsikatsou, Moustakib, Yang-Wallentina, & Jöreskog, 2012). The final sample consisted of 899 children attending primary school from grades 3 to 5 (*N*= 571, 50.4% girls; mean age = 9 years and 2 months, *SD* = 10 months, range = 7.75 – 11.33 years), and middle school students in grades 6 to 8 (*N*=328, 46.48% girls; mean age = 12 years and 4 months, *SD* = 11 months, range = 10.17 – 16.67 years). The majority of the children in the sample were Italian (88%), and the whole sample was fluent in Italian. A subsample of 347 students was reassessed after two months to assess test-retest reliability: 238 were attending primary school (52.9% girls; mean age = 9.11 years, *SD* = 9.88 months, range = 7.75 – 11.33 years), and 109 were at middle school (45.9% boys; mean age = 12.45 years, *SD* = 5.63 months, range = 11.6 –14.8 years).

Procedure

Ethical approval for the study was obtained from the Ethical Committee of the University of [blind]. After obtaining the school's approval, written informed parental consent and the children's verbal assent were obtained prior to the test. Children were tested in their classrooms during two separate collective sessions lasting approximately 45 minutes each. In all sessions, the tests were administered by a trained assistant researcher using a standardized procedure, and in the presence of a teacher. Children were allowed more time to complete the self-report if they needed it, and also given a few minutes to rest between the tasks as necessary. During the first collective session participants completed the 24-item TAQ-C derived from Study 1, a socio-demographic form (date of birth, gender), the CFIT (Cattell & Cattell, 1981), the SC-Academic scale (Bracken, 2003), and the RCMAS-2 (Reynolds & Richmond, 2012). During the second collective session they completed the RASP (Hurtes & Allen, 2001), the TASC (Sarason et al., 1960), and the AMAS (Hopko, Mahadevan, Bare, & Hunt, 2003; see also Caviola, Primi, Chiesi,

& Mammarella, 2017). Students in the subsample considered for the purpose of assessing test-retest reliability were asked to complete the TAQ-C again two months later.

Measures

Test Anxiety Questionnaire for Children (TAQ-C). The 24-item version of the TAQ-C derived from Study 1 was used.

The Test Anxiety Scale for Children (TASC). The TASC (Sarason et al., 1960) is a 30-item measure designed to assess TA in children. Statements (e.g., "*During tests I find myself thinking of the consequences of failing*", "*I freeze up on things like intelligence tests and final exams*") are rated as true or false and scored as 0 or 1. The sum of the scores for all the items provides a total score, with higher scores indicating greater TA. In our sample, the CFA on the items showed an acceptable fit to the data ($\chi 2(629) = 2072.242$, CFI=.943, TLI=.940, RMSEA=.051[.048– .053]).

The Abbreviated Math Anxiety Scale (AMAS). The AMAS (Hopko et al., 2003) is a 9-item questionnaire for judging mathematics anxiety in children from statements such as "*Thinking about an upcoming math test on the day before*". The Italian version of the AMAS was used in the present study (Caviola et al., 2017). Participants were asked to judge how anxious they would be in each math-related situation using a 5-point scale ranging from 1= "strongly agree" to 5 = "strongly disagree". Scores across items were added up to obtain a total score, with higher scores indicating greater mathematical anxiety. The AMAS has shown a good internal consistency and external validity across several countries, including Italy (Caviola et al., 2017). In the present sample, the CFA on the 9 items showed acceptable fit indices ($\chi 2(18)=$ 118.303, CFI=.988, TLI=.981, RMSEA=.079[.065– .093]).

The Revised Children's Manifest Anxiety Scale: Second Edition (RCMAS-2). The RCMAS-2 (Reynolds & Richmond, 2012) is a 37-item self-report designed to assess general anxiety in children and adolescents. It consists of three subscales including worries (e.g., "*I feel nervous when things don't go as I want*"), physiological anxiety ("*I often have stomach-ache*"), and social anxiety ("*I am worried that my classmates could make fun of me*"). Each item is rated as "yes/no" and scored as "yes = 1" and "no = 0", such that higher total scores indicate greater levels of general anxiety. The authors reported that the scale showed a good factor structure, external validity and test-retest reliability (Reynolds & Richmond, 2012). Overall, the CFA in our sample showed an acceptable fit to the data ($\chi 2$(737)= 3299.733, CFI=.904, TLI=.898, RMSEA=.062 [.060 –.064]).

The Multidimensional Self-Concept Scale (MSCS). The MSCS (Bracken, 2003) is a questionnaire devised to assess self-concept in children and adolescents. For the purpose of the present study, children completed the Self-Concept – Academic scale (i.e., *"Studying is difficult for me"*) comprising 25 items rated on a 4-point Likert scale from 1 = "absolutely false" to 4 = "absolutely true". Scores across items were added up to provide a total score, where higher scores corresponded to higher academic self-concept. The Academic scale has shown a good internal consistency, external validity and test-retest reliability (Bracken, 2003). In the present sample, the CFA on the 25 items showed acceptable fit indices (χ2(275)= 1217.072, CFI=.949, TLI=.945, RMSEA=.061[.058–.065]).

The Resiliency Attitudes and Skills Profile (RASP). The RASP (Hurtes & Allen, 2001) is a 34-item questionnaire designed to assess different aspects that enable individuals to rise above adversity (i.e., "*I can change my behavior to match the situation*"). Each item is rated on a 6-point Likert scale ranging from 1= "strongly disagree" to 6 = "strongly agree". Scores on each item are added up to provide a total score, with higher values corresponding to greater resilience. The RASP has shown a good internal and convergent validity (Hurtes & Allen, 2001). In the present sample, the CFA on the 34 items showed a good fit to the data ($\chi 2$(506)= 1559.867, CFI=.941, TLI=.935, RMSEA=.048[.045–.051]).

The Cattell Culture Fair Intelligence Test (CFIT). The CFIT (Cattell & Cattell, 1981) is a pen-and-pencil standardized measure of non-verbal fluid intelligence. It includes 46 multiple-choice items divided into four subtests covering judgments and reasoning, each with a specific time constraint (from 2 to 4 minutes). For each correct answer, respondents score one point. The CFA on the CFIT showed good fit indices ($\chi 2$(2)=4.725, CFI=.996, TLI=.988, RMSEA=.039[.000–.086]).

Analysis

Analyses were run using R (R Development Core team, 2017). All models (CFAs, MG-CFAs and SEMs) were developed using the *lavaan* package (Rosseel, 2012).

Descriptive analysis. Distributions, mean, standard deviation, range and skewness were calculated for each item.

Factor structure. Several CFAs were run to test: a) a single-factor model (Sarason et al., 1960); b) a four-factor model (Lowe et al., 2008) (i.e., the Thoughts, Autonomic Reactions, Off-Task Behaviors, and Social Derogation subscales); and c) a bifactor model which assumed a single general factor (Sarason et al., 1960) while recognizing the multidimensionality of the construct (Lowe et al., 2008). In particular, the bifactor model was based on the assumption that a single general factor reflected the variance shared by all the items and the four orthogonal factors (i.e., the Thoughts, Autonomic Reactions, Off-Task Behaviors, and Social Derogation).

All models were estimated using the diagonal weighted least squares estimator (DWLS) recommended for ordinal data (Flora & Curran, 2004). Several goodness of fit indices were computed (i.e., the χ^2/df, the RMSEA, the CFI, the TLI, and the SRMR) and assessed according to the guidelines established in Study 1. If the bifactor model was supported, some complementary analyses were performed. In particular, following the recommendations of Reise, Scheines, Widaman, and Haviland (2013) and Bonifay, Reise, Scheines, and Meijer (2015), we computed: a) the explained common variance (ECV) which represents the percent of common variance attributable to the general factor in the bifactor model (Reise, Moore, & Haviland, 2010); and b) the percentage of uncontaminated correlations (PUC), which can be defined as the percentage of correlations that reflect general factor variance (Bonifay et al., 2015; Reise et al., 2013), and it "represents the degree to which multidimensional data can be modeled in a unidimensional structure without being affected by large parameter bias" (Ebesutani, Kim, & Park, 2016, p.119). When data follow a bifactor structure, correlations among the items within the construct reflect both general and specific factor variance, whereas correlations among the items within group factors reflect variance from the general factor and are uncontaminated by multidimensionality (Rodriguez, Reise, & Haviland, 2015). This means that, when the multidimensional data are largely unidimensional, the percentage of correlations affected by the general factor increases, and higher value of PUC reflects little parameter bias (Reise, Bonifay, & Haviland, 2018). Then, the hierarchical omega (ωH) was computed following the procedure explained by Reise et al (2013). A PUC greater than .80, ECV values greater than .60 and ωH higher than .70 suggest that "the presence of some multidimensionality is not severe enough to disqualify the interpretation of the instrument as primarily unidimensional" (p. 22, Reise et al., 2013). Finally, structural equation modeling (SEM) was used to examine to what extent the

TAQ-C subscales (i.e., Thoughts, Autonomic Reactions, Off-Task Behaviors, and Social Derogation) significantly predicted an external criterion of TA, while controlling for the general TAQ-C factor in the bifactor model (see Ebesutani et al., 2016 for the same procedure). In particular, we considered mathematics anxiety as a dependent variable (i.e., an endogenous variable), and the Thoughts, Autonomic Reactions, Off-Task Behaviors, and Social Derogation subscales, and the general TAQ-C factor as predictors (i.e., exogenous variables) (see Figure 1). Given that previous research found that girls scored higher for TA than boys (Hembree, 1988; von der Embse et al., 2018; Zeidner, 1998), and that TA levels increased from primary to middle school (Hembree, 1988; von der Embse et al., 2018) the model was implemented controlling for these variables.

Measurement invariance across educational levels and gender. Once the best model had been selected, Multigroup Confirmatory Factor Analyses (MG-CFAs) were run to examine the measurement invariance of the bifactor structure across educational levels and gender. Consistently with this procedure, the model was tested separately on girls attending primary school (G1), boys attending primary school (G2), girls attending middle school (G3), and boys attending middle school (G4). Configural invariance was first tested by allowing the parameters to remain free across the groups considered. Then the metric and scalar invariance were tested by constraining the factor loadings and thresholds to be equal across the groups (Muthén & Muthén, 2010). Several fit indices (i.e., CFI, TLI, RMSEA) were examined, and the difference in CFI (Δ CFI) was computed between the two proximal models (i.e., configural vs metric and scalar invariance). Acceptable model fit indices and a change in the CFI (Δ CFI) of less than .01 between models were considered evidence of model invariance (Chen, 2007; Cheung & Rensvold, 2002).

Concurrent and convergent validity. A SEM framework was used to test concurrent and convergent validity. For the former, we examined the latent correlation between TA measured with the TAQ-C (general factor of the bifactor model) and with the TASC; for the latter, we performed a series of SEMs (one for each external measure) to investigate the association between the general TAQ-C factor and a number of external measures, including general anxiety (RCMAS-2), resilience (RASP), and academic self-concept (SC-Academic scale), after controlling for educational levels and gender. In particular, we considered each external measure as a dependent (endogenous) variable, and the TA factor as a predictor (i.e., an exogenous variable). A graphical representation is provided in the supplementary material section.

Test–retest reliability. Pearson's correlations were performed to assess the test-retest reliability of the total TAQ-C score over a 2-month period.

Results

Descriptive analysis. Distributions and descriptive statistics for each of the 24 items in the TAQ-C used in Study 2 are provided in Table 2.

Factor structure and measurement invariance across educational levels and gender. Table 3 shows the results of all the CFA models tested. The initial CFA (M1) performed on the one-factor model obtained acceptable CFI and TLI values, but the other indices considered exceeded the cut-off values (Schermelleh-Engel et al., 2003). The four-factor model (M2) was also tested. It showed a good fit in all the fit indexes considered, and the factor loadings were all higher than .50. Then the bifactor model was tested (M3), yielding a good fit in all the indexes considered. All factor loadings were associated with the general factor and, beyond being statistically significant, most of them had a large effect size (see Figure 2). The vast majority of the items showed statistically significant loadings on both the general and the specific factors (see Figure 2 and Table S2), but the items for Off-Task Behaviors had higher factor loadings in this specific factor than in the general factor. To further investigate the appropriateness of the bifactor model, the PUC, the ECV and the ωH were calculated for both the general and the specific factors. The results indicated a PUC of 78%. The ECV for the general TAQ-C factor indicated that the common variance was 67%, while the ECV attributed to the specific factors ranged from 2% to 14% (Thoughts: 2%; Off-Task Behaviors: 14%; Autonomic Reactions 10%; Social Derogation 7%). The ωH was .86 for the general factor, and .01, .57, .33 and .26, respectively, for the Thoughts, Off-Task Behaviors, Autonomic Reactions, and Social Derogation factors. This goes to show that the item response variance accounted for by the general factor was highly reliable, and relatively small for the single subscales overall, once the effect of the general factor had been taken into account. This would point to the presence of a general TA factor. To further examine the appropriateness of the bifactor model, we ran a SEM to test to what extent each factor (i.e., the Thoughts, Autonomic Reactions, Off-Task Behaviors, and Social Derogation subscales) significantly predicted mathematics anxiety while controlling for the general TAQ-C factor. The SEM model reached a good fit (CFI=.982, TLI=.980, RMSEA=.047 [.045–.050], SRMR=.051). Among the specific TAQ-C factors (i.e., the Thoughts, Autonomic Reactions, Off-Task Behaviors, and Social Derogation subscales), only the Off-Task Behaviors subscale explained further variance over and above the general factor, as suggested by the significant association with our criterion measure after controlling for the general TAQ-C factor (β = -.082, z = - 2.049, p =.036).

Measurement invariance across educational levels and gender. A MG-CFAs analysis across educational levels and gender was performed to test the measurement invariance on the

TAQ-C bifactor model. First, the model was tested separately on the four groups, and all models yielded an excellent fit (see Table 3). In the next step, the configural invariance revealed a good fit (see Table 3), suggesting that the pattern of factor-indicator relationships was the same across boys and girls in primary and middle school. Then we kept the loadings and thresholds invariant across groups. As shown in Table 3, the model fit was also good with the Δ CFI between the two models lower than .01, supporting metric and scalar invariance. Descriptive statistics of TAQ-C by educational levels and gender are provided in the supplementary materials (see Table S3 and Table S4).

Concurrent and convergent validity. For concurrent validity, a highly positive and significant association was found between the TA obtained with the general TAQ-C score and the TA assessed with the TAS (r = .864, p<.001). As for convergent validity, the SEMs performed to assess the association between the general TAQ-C factor and the external measures considered (i.e., RCMAS-2, RASP, and SC-Academic scale) showed a good fit to the data after controlling for educational levels and gender. The results showed that higher levels of the general TAQ-C factor were significantly associated with stronger features of general anxiety (social: β = .637, p <.001; worry: β = .612, p <.001; physiological: β = .556, p <.001). As expected, negative associations were found between the general TAQ-C factor and resilience (β = -.257, p <.001), and the same results emerged for the academic self-concept scale (β = -.491, p <.001). Fit indices for each model are provided in the supplementary material section (see Table S5).

Test-retest reliability. Pearson's correlations performed on the total TAQ-C score over a 2-month period showed a strong correlation (r [345] = .74), supporting a good test-retest reliability of the scale over a 2-month period.

Discussion

The present study aimed to contribute to the assessment of TA in school-age children by examining the psychometric properties of the TAQ-C, a self-report tool for primary and middle school students. For this purpose, the present research included a pilot phase, in which an initial pool of items was selected and adapted to generate the TAQ-C (Study 1), and a test phase to assess the psychometric properties of this questionnaire (Study 2).

In Study 1, we used CFAs to examine an initial set of 37 items, identifying 24 items related to four dimensions of TA. The psychometric properties of the 24 items included in the TAQ-C were then tested in Study 2 on a large sample of children attending primary and middle school. In this stage, our findings concerning the factorial structure of the tool indicated the bifactor model as the best solution. This means that the TAQ-C comprises a general factor (i.e., test anxiety) as well as four specific orthogonal factors (i.e., the cognitive, behavioral, physiological and social components of TA). The SEMs performed to examine the association between each TAQ-C subscale and an external criterion measure (mathematics anxiety), while controlling for the general TAQ-C factor provide further support for the presence of a general TA factor. On the other hand, the Off-Task Behavior dimension seemed to provide a specific contribution to the prediction of mathematics anxiety, given its relation with our external measure. Further research should better elucidate the relevance of the Off-Task Behavior dimension in assessments of the behavioral component of children’s TA. Overall, our findings regarding the TAQ-C’s factorial structure are consistent with the results of a recent study supporting the presence of a bifactor structure in a self-report tool for assessing students’ anxiety (Lohbeck & Petermann, 2018). Our results expand on the previous literature (Lowe et al., 2008;

Zeidner, 1998), suggesting that the cognitive, behavioral, physiological and social components of TA are distinct aspects that could reflect a unidimensional factor.

The TAQ-C showed good measurement invariance across educational levels and gender, suggesting that this self-report is a reliable tool for assessing TA in boys and girls across primary and middle school (see van de Schoot, Lugtig, & Hox, 2012). Our findings also demonstrated the questionnaire's good concurrent and convergent validity, and test-retest reliability. As concerns convergent validity, a positive association was found between the general TAQ-C factor and general anxiety, consistently with previous research (Hembree, 1988; see also Carey et al., 2017). A negative correlation emerged instead for the general TAQ-C factor with academic self-concept and resilience (Arens et al., 2017; Hembree, 1988; Putwain et al., 2013; Raufelder & Ringeisen, 2016). This was true after controlling for educational levels and gender, but also when TA was considered in terms of its cognitive, behavioral, physical and social dimensions (Hembree, 1988; von der Embse et al., 2018; Zeidner, 1998).

Our study provides broad evidence of the TAQ-C questionnaire's good psychometric properties, but there are some limitations to consider in this respect. First, our findings concerning the tool's factorial structure are based on Italian children. Therefore, our findings would need to be replicated in other countries before we can be confident about the generalizability of our results. This would be very important, not only to test the TAQ-C's psychometric properties, but also to shed light on the factors related to the development of TA, as experienced and reported in different cultures (Bodas & Ollendick, 2005; Nyroos et al., 2015; Zeidner, 1998). Second, our study relied entirely on the use of self-report measures, which may suffer from respondents tending to present a positive image of themselves in their answers for reasons of social desirability (Furnham, 1986; Holtgraves, 2004; van de Mortel, 2008). Although children have consistently proved valid and reliable reporters of their own internalizing distress (Seligman, Ollendick, Langley, & Baldacci, 2004; Weems, Zakem, Costa, Cannon, & Watts, 2005), future studies should include other informants (i.e., parents and teachers), and multiple measures (e.g., interviews and behavior observations) in assessments of children's different forms of anxiety (Mohr & Schneider, 2013).

Despite these limitations, our results have important clinical and educational implications. From an applied perspective, total TAQ-C scores seem to be useful for assessing TA, given the tool's high reliability and predictive validity. Our findings suggest that the Thoughts, Autonomic Reactions, and Social Derogation scores relate more to the general TAQ-C factor, while the Off-Task Behavior score seems to be informative even after taking the general TAQ-C factor into account. In other words, it would seem more appropriate to use the total TAQ-C score rather than calculating the Thoughts, Autonomic Reactions, and Social Derogation scores separately, whereas the Off-Task Behavior score seems to be more informative when it comes to assessing the behavioral component of a child's TA. The TAQ-C could also be used in prevention programs as a quick and easy screening tool for detecting children who experience moderate or higher levels of TA. This could be particularly important at primary school where children could benefit the most from early intervention programs to reduce TA (Ergene, 2003; von der Embse et al., 2013). Previous research showed that few TA interventions are available for children, but those based on behavioral or cognitive, or cognitive-behavioral theory (CBT) have proved effective in containing TA (Lang & Lang, 2010; Larson, Ramahi, Conn, Este, & Ghibellini, 2010; von der Embse et al., 2013; Weems et al., 2010). Such programs should be proposed to whole classrooms or groups of students showing high levels of TA, and could be particularly useful to children with learning difficulties, who are more likely to exhibit emotional difficulties (see Nelson & Harwood, 2011; Mammarella et al. 2016; Mugnaini, Lassi, La Malfa, & Albertini, 2009).

In conclusion, the TAQ-C appears to be psychometrically sound, with a good validity, so it could prove a useful tool for researchers and psychologists assessing TA in children.

References

Anderson, S. B., & Sauser, W. J. (1995). Measurement of test anxiety: An overview. In C. D. Spielberger & P. R. Vaag (Eds.), *Test anxiety: Theory, Assessment, and Treatment* (pp. 15–33). Washington, DC: Taylor & Francis.

Arens, A. K., Becker, M., & Möller, J. (2017). Social and dimensional comparisons in math and verbal test anxiety: Within-and cross-domain relations with achievement and the mediating role of academic self-concept. *Contemporary Educational Psychology*, *51*, 240–252. doi:10.1016/j.cedpsych.2017.08.005

Bandalos, D. L., Yates, K., & Thorndike-Christ, T. (1995). Effects of math self-concept, perceived self-efficacy, and attributions for failure and success on test anxiety. *Journal of Educational Psychology*, *87*(4), 611–623.

Bodas, J., & Ollendick, T. H. (2005). Test anxiety: A cross-cultural perspective. *Clinical Child and Family Psychology Review*, *8*(1), 65-88. doi:10.1007/s10567-005-2342-x

Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(4), 504–516. doi:10.1080/10705511.2014.938596

Borgers, N., Hox, J., & Sikkel, D. (2003). Response quality in survey research with children and adolescents: the effect of labeled response options and vague quantifiers. *International Journal of Public Opinion Research*, *15*(1), 83–94. doi:10.1093/ijpor/15.1.83

Bracken, B. A. (2003). *TMA - Test di valutazione multidimensionale dell'autostima*. Trento, Italy: Erickson.

Carey, E., Devine, A., Hill, F., & Szűcs, D. (2017). Differentiating anxiety forms and their role in academic performance from primary to secondary school. *Plos One, 12*(3), 1–20. e0174418. doi:10.1371/journal.pone.0174418

Cattell, R. B., & Cattell, A. K. S. (1981*). Misurare l'intelligenza con i test "Culture Fair".* Florence, Italy: Giunti O.S.

Caviola, S., Primi, C., Chiesi, F., & Mammarella, I. C. (2017). Psychometric properties of the Abbreviated Math Anxiety Scale (AMAS) in Italian primary-school children. *Learning and Individual Differences*, *55*, 174-182. doi:10.1016/j.lindif.2017.03.006

Chen, FF. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*, 464–504. doi:10.1080/10705510701301834

Cheung, GW., & Rensvold, RB. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 233–255. doi:10.1207/S15328007SEM0902_5

Devine, A., Fawcett, K., Szűcs, D., & Dowker, A. (2012). Gender differences in mathematics anxiety and the relation to mathematics performance while controlling for test anxiety. *Behavioral and Brain Functions, 8*(33), 1–9. doi:10.1186/1744-9081-8-33

Ebesutani, C., Kim, M., & Park, H. H. (2016). The utility of the bifactor model in understanding unique components of anxiety sensitivity in a South Korean sample. *Asian Journal of Psychiatry*, *22*, 116–123. doi:10.1016/j.ajp.2016.06.005

Erford, B. T., Mckechnie, J. A., & Moore-Thomas, C. (2004). Program assessment and evaluation. In B. T. Erford (Ed.), *Professional school counseling: A handbook of theories, programs & practices* (pp. 303–309). Austin, TX: CAPS Press.

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

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466–491. doi:10.1037/1082-989X.9.4.466

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

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and individual differences, 7*(3), 385-400. doi:10.1016/0191-8869(86)90014-0

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research*, *58*(1), 47–77. doi:10.2307/1170348

Hembree, R. (1990). The nature, effects, and relief of mathematics anxiety. *Journal for Research in Mathematics Education, 21*, 33–46. doi:10.2307/749455

Hopko, D. R., Mahadevan, R., Bare, R. L., & Hunt, M. K. (2003). The Abbreviated Math Anxiety Scale (AMAS) construction, validity, and reliability. *Assessment*, *10*, 178–182. doi:10.1177/1073191103252351

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin, 30*(2), 161-172. doi:10.1177/0146167203259930

Hurtes, K. P., & Allen, L. R. (2001). Measuring resiliency in youth: The resiliency attitudes and skills profile. *Therapeutic Recreation Journal*, *35*(4), 333–347.

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., & Jöreskog, K. G. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics & Data Analysis, 56*(12), 4243–4258. doi:10.1016/j.csda.2012.04.010

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*(1), 537–567. doi:10.1146/annurev.psych.50.1.537

Lang, J., & Lang, J. (2010). Priming competence diminishes the link between cognitive test anxiety and test performance: Implications for the interpretation of test scores. *Psychological Science, 21*, 811–819. doi:10.1177/0956797610369492

Larson, H., Ramahi, M., Conn, S., Este, L., & Ghibellini, A. (2010). Reducing test anxiety among third-grade students through the implementation of relaxation techniques. *Journal of School Counseling, 8*(19), 1–19.

Lohbeck, A., & Petermann, F. (2018). Factorial Validity of the Anxiety Questionnaire for Students (AFS): Bifactor Modeling and Measurement Invariance. *Journal of Psychoeducational Assessment*, 1–12. doi:10.1177/0734282918794834

Lowe, P. A. (2014). The Test Anxiety Measure for Adolescents (TAMA) examination of the reliability and validity of the scores of a new multidimensional measure of test anxiety for middle and high school students. *Journal of Psychoeducational Assessment*, *32*(5), 404–416. doi:10.1177/0734282913520595

Lowe, P. A., Grumbein, M. J., & Raad, J. M. (2011). Examination of the psychometric properties of the Test Anxiety Scale for Elementary Students (TAS-E) scores. *Journal of Psychoeducational Assessment*, *29*, 503–514. doi:10.1177/0734282910395894

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 Assessment, 26(*3), 215–230. doi:10.1177/0734282907303760

Mammarella, I. C., Donolato, E., Caviola, S., & Giofrè, D. (2018). Anxiety profiles and protective factors: A latent profile analysis in children. *Personality and Individual Differences, 124*, 201–208. doi:10.1016/j.paid.2017.12.017.

Mammarella, I. C., Ghisi, M., Bomba, M., Bottesi, G., Caviola, S., Broggi, F., …& Nacinovich, R. (2016). Anxiety and depression in children with nonverbal learning disabilities, reading disabilities, or typical development. *Journal of Learning Disabilities, 49*(2), 130–139. doi:10.1177/0022219414529336

Martin, A. J., & Marsh, H. W. (2006). Academic resilience and its psychological and educational correlates: A construct validity approach. *Psychology in the Schools*, *43*(3), 267-281. doi:10.1002/pits.20149

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

Mohr, C., & Schneider, S. (2013). Anxiety disorders. *European Child & Adolescent Psychiatry*, *22*, 17-22. doi:10.1007/s00787-012-0356-8

Mugnaini, D., Lassi, S., La Malfa, G., & Albertini, G. (2009). Internalizing correlates of dyslexia. *World Journal of Pediatrics, 5*(4), 255-264. doi:10.1007/s12519-009-0049-7

Muthén, L.K., and Muthén, B.O. (2010). *Mplus User's Guide*, 4th Edn. LosAngeles, CA: Muthén & Muthén.

Nelson, J. M., & Harwood, H. (2011). Learning disabilities and anxiety: A meta-analysis. *Journal of Learning Disabilities, 44*(1), 3–17. doi:10.1177/0022219409359939

Nyroos, M., Korhonen, J., Peng, A., Linnanmäki, K., Svens-Liavåg, C., Bagger, A., & Sjöberg, G. (2015). Cultural and gender differences in experiences and expression of test anxiety among Chinese, Finnish, and Swedish grade 3 pupils. *International Journal of School & Educational Psychology*, *3*(1), 37-48. doi:10.1080/21683603.2014.915773

Putwain, D. W. (2008). Deconstructing test anxiety. *Emotional and Behavioural Difficulties, 13*(2), 141-155. doi:10.1080/13632750802027713

Putwain, D. W., Nicholson, L. J., Connors, L., & Woods, K. (2013). Resilient Children Are Less Test Anxious and Perform Better in Tests at the End of Primary Schooling. *Learning and Individual Differences*, *28*, 41–46. doi:10.1016/j.lindif.2013.09.010

R Development Core Team. (2017). R: *A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.Rproject.org/

Raufelder, D., & Ringeisen, T. (2016). Self-perceived competence and test anxiety. *Journal of Individual Differences, 37*(3), 159–167. doi:10.1027/1614-0001/a000202

Reise, S. P., Bonifay, W., & Haviland, M. G. (2018). Bifactor modelling and the evaluation of scale scores. In P. Irwing, T. Booth & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development* (pp. 675-707). Hoboken, NJ: Wiley-Blackwell.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*, 544–559. doi:10.1080/00223891.2010.496477

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement*, *73*, 5-26. doi:10.1177/0013164412449831

Reynolds, C. R., & Richmond, B. O. (2012). *RCMAS-2 Revised Children's Manifest Anxiety Scale — Second Edition.* Florence, Italy: Giunti O.S.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*(3), 223-237. doi:10.1080/00223891.2015.1089249

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*, 1–36. doi:10.1080/10705519909540118

Sarason, S. B., Davidson, K. S., Lighthall, F. F., Waite, R. R., & Ruebush, B. K. (1960). *Anxiety in Elementary School Children.* New York, NY: Wiley.

Sarason, S. B., & Mandler, G. (1952). Some correlates of test anxiety. *The Journal of Abnormal and Social Psychology*, *47*(4), 810-817.

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.

Segool, N., Carlson, J., Goforth, A., von der Embse, N., & Barterian, J. (2013). Heightened test anxiety among young children: Elementary school students' anxious responses to high-stakes testing. *Psychology in the Schools, 50*, 57–71. doi:10.1002/pits.21689

Seligman, L. D., Ollendick, T. H., Langley, A. K., & Baldacci, H. B. (2004). The utility of measures of child and adolescent anxiety: a meta-analytic review of the Revised Children's Manifest Anxiety Scale, the State–Trait Anxiety Inventory for Children, and the Child

Behavior Checklist. *Journal of Clinical Child and Adolescent Psychology*, *33*(3), 557-565. doi:10.1207/s15374424jccp3303_13

Spence, S. H. (2018). Assessing anxiety disorders in children and adolescents. *Child and Adolescent Mental Health*, *23*(3), 266–282. doi:10.1111/camh.12251

Van de Mortel, T. F. (2008). Faking it: social desirability response bias in self-report research. *Australian Journal of Advanced Nursing, 25*(4), 40-48.

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492. doi:10.1080/17405629.2012.686740

von der Embse, N., & Hasson, R. (2012). Test anxiety and high-stakes test performance between school settings: Implications for educators. *Preventing School Failure: Alternative Education for Children and Youth*, *56*(3), 180–187. doi:10.1080/1045988X.2011.633285

von der Embse, N., & Witmer, S. E. (2014). High-stakes accountability: Student anxiety and large-scale testing. *Journal of Applied School Psychology*, *30*(2), 132-156. doi:10.1080/15377903.2014.888529.

von der Embse, N., Barterian, J., & Segool, N. (2013). Test anxiety interventions for children and adolescents: A systematic review of treatment studies from 2000–2010. *Psychology in the Schools, 50*(1), 57–71. doi:10.1002/pits.21660

von der Embse, N., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of Affective Disorders*, *227,* 483–493. doi:10.1016/j.jad.2017.11.048

Weems, C. F., Scott, B. G., Taylor, L. K., Cannon, M. F., Romano, D. M., Perry, A. M., … & Triplett, V. (2010). Test anxiety prevention and intervention programs in schools: Program development and rationale. *School Mental Health, 2*, 62–71. doi:10.1007/s12310-010-9032-7

Weems, C. F., Zakem, A. H., Costa, N. M., Cannon, M. F., & Watts, S. E. (2005). Physiological response and childhood anxiety: Association with symptoms of anxiety disorders and cognitive bias. *Journal of Clinical Child and Adolescent Psychology, 34*(4), 712–723. doi:10.1207/s15374424jccp3404_13

Wren, D. G., & Benson, J. (2004). Measuring test anxiety in children: Scale development and internal construct validation. *Anxiety, Stress & Coping*, *17*, 227–240. doi:10.1080/10615800412331292606

Zeidner, M. (1998). *Test Anxiety: The State of the Art*. New York, NY: Plenum Press.

Table 1

Fit indices for the confirmatory factor models tested for the TAQ-C (Study 1).

Model	*Item deleted*	*χ2/ df*	*p*	*CFI*	*TLI*	*SRMR*	*RMSEA [90% CI]*
One-factor model							
M1	-	2.118	<.001	.907	.901		.096 [.089–.103]
Four-factor models							
M2	-	1.226	<.001	.981	.980	.102	.043 [.031–.053]
M3	1-16-18	1.245	<.001	.983	.981	.103	.045 [.032–.045]
M4	7-17-32	1.177	.007	.988	.987	.099	.038 [.021–.051]
M5	31-37-9	1.070	.174	.996	.995	.095	.024 [.000–.042]
M6	21-28-26-25	1.128	.081	.993	.992	.095	.032 [.000–.051]

Note. *N* = 123. Four-factor models (Thoughts, Autonomic Reactions, Off-Task Behaviors, and Social Derogation subscales). χ2/df = chi-square/degree of freedom; CFI = Comparative Fit Index; TLI =Tucker-Lewis Index; SRMR=Standardized Root Mean Square Residual; RMSEA = Root Mean Square of Approximation.

Table 2

Item response distributions and descriptive statistics for the final TAQ-C (Study 2).

		Frequency (%)				M	SD	Range	Skewness
Item	**N**	**1**	**2**	**3**	**4**				
Thoughts									
Item 1	896	23.21	42.75	16.52	17.52	2.28	1.01	1–4	0.43
Item 2	898	15.37	37.31	22.49	24.83	2.57	1.02	1–4	0.08
Item 3	894	14.65	45.75	23.83	15.77	2.41	0.92	1–4	0.32
Item 4	896	27.57	47.66	15.07	9.71	2.07	0.90	1–4	0.66
Item 5	887	21.87	42.50	20.63	14.99	2.29	0.97	1–4	0.38
Item 6	892	56.61	28.36	8.18	6.83	1.65	0.90	1–4	1.31
Off-Task Behaviors									
Item 7	897	22.41	42.14	22.41	13.04	2.26	0.95	1–4	0.37
Item 8	887	59.42	24.22	7.74	8.63	2.50	0.74	1–4	0.20
Item 9	892	38.97	25.20	15.23	20.60	1.66	0.95	1–4	1.34
Item 10	894	51.12	29.08	8.72	11.07	1.80	1.00	1–4	1.08
Item 11	889	28.79	46.90	15.86	8.43	2.04	0.89	1–4	0.65
Item 12	895	35.42	32.29	14.64	17.65	2.15	1.09	1–4	0.53
Autonomic Reactions									
Item 13	898	22.49	45.32	18.93	13.25	2.23	0.95	1–4	0.47
Item 14	887	25.03	39.80	19.84	15.33	2.25	1.00	1–4	0.40
Item 15	889	56.13	28.23	9.11	6.75	1.66	0.90	1–4	1.26
Item 16	895	54.86	35.08	5.59	4.47	1.60	0.79	1–4	1.39
Item 17	888	60.14	21.40	7.83	10.57	1.69	1.00	1–4	1.28
Item 18	892	48.99	33.41	9.19	8.41	1.77	0.93	1–4	1.10
Social Derogation									
Item 19	896	32.59	34.49	17.75	15.18	2.16	1.04	1–4	0.49
Item 20	896	46.88	27.73	13.95	11.94	1.91	1.04	1–4	0.82
Item 21	896	61.72	18.97	9.60	9.71	1.67	1.00	1–4	1.27
Item 22	898	63.14	20.27	7.61	9.31	1.63	0.97	1–4	1.41
Item 23	898	66.15	2060	6.35	6.90	1.54	0.89	1–4	1.62
Item 24	897	20.29	34.89	18.17	26.64	2.51	1.09	1–4	0.12

Table 3

Fit indices for the confirmatory factor models tested for the TAQ-C, and invariance by educational levels and gender (Study 2).

	χ2/ df	*p*	*CFI*	*TLI*	*SRMR*	*RMSEA [90% CI]*	*Δ CFI*
M1: One-factor	10.348	<.001	.956	.952	.093	.102 [.099–.106]	
M2: Four-factor	3.993	<.001	.986	.985	.057	.058 [.054–.062]	
M3: Bifactor	2.811	<.001	.992	.991	.050	.045 [.041–.049]	
MG-CFAs							
G1	1.477	<.001	.994	.993	.065	.041 [.031–.050]	
G2	1.596	<.001	.992	.991	.068	.046 [.037–.055]	
G3	1.025	.382	.999	.999	.074	.013 [.000–.037]	
G4	1.066	.234	.998	.998	.072	.019 [.000–.038]	
Configural invariance	1.291	<.001	.995	.994	.069	.036 [.030–.042]	
Metric and scalar invariance	1.604	<.001	.987	.980	.077	.052 [.048–.056]	-.008

Note. *N*=899. G1: primary-school girls (*N*=288); G2: primary-school boys (*N*=283); G3: middle-school girls (*N*=152); G4: middle-school boys (*N*=176). χ2/df = chi-square/degree of freedom; CFI = Comparative Fit Index; TLI =Tucker-Lewis Index; SRMR=Standardized Root Mean Square Residual; RMSEA = Root Mean Square of Approximation; *Δ CFI* = difference between CFIs. In bold, the model selected and considered in the MG-CFAs.

Figure captions

Figure 1. Structural equation model predicting mathematics anxiety.

Figure 2. Standardized factor loadings for the bifactor model (*N*=899).