Anxiety, Stress, and Coping,
September 2004, Vol. 17, No. 3, pp. 227–240

BrunnerRoutledge Taylor & Francis healthsciences

MEASURING TEST ANXIETY IN CHILDREN: SCALE DEVELOPMENT AND INTERNAL CONSTRUCT VALIDATION

DOUGLAS G. WREN[a,]* and JERI BENSON[b]

[a]*DeKalb County School System, Kittredge Magnet School 2383 N, Druid Hills Road, Atlanta, GA 30329, USA;* [b]*University of Georgia, Athens, GA, USA*

Given the increased testing of school-aged children in the United States there is a need for a current and valid scale to measure the effects of test anxiety in children. The domain of children's test anxiety was theorized to be comprised of three dimensions: thoughts, autonomic reactions, and off-task behaviors. Four stages are described in the evolution of the *Children's Test Anxiety Scale* (CTAS): planning, construction, quantitative evaluation, and validation. A 50-item scale was administered to a development sample ($N = 230$) of children in grades 3–6 to obtain item analysis and reliability estimates which resulted in a refined 30-item scale. The reduced scale was administered to a validation sample ($N = 261$) to obtain construct validity evidence. A three-factor structure fit the data reasonably well. Recommendations for future research with the scale are described.

Keywords: Children's test anxiety; Test anxiety; Scale development; Construct validity

The testing of school-aged children in the United States has continued to increase over the past 25 years. This trend is primarily due to an emphasis on the accountability of schools to increase their students' achievement scores. The American public schools are often perceived to be doing an inferior job of educating children compared with the schools of most other industrialized nations (Valverde and Schmidt, 1997). American politicians and education leaders have consequently mandated the use of standardized tests to hold schools accountable for student achievement. This accountability movement has led to the federal legislation, *No Child Left Behind* (NCLB; No Child Left Behind Act, 2001) which mandates: (a) a minimum of 95% of all children in each school district in each state will be tested at grades 3–8 annually and at least once in grades 10–12, and (b) test scores for all subgroups of students must improve each year so that by 2013 no child is achieving below grade level. The NCLB legislation has resulted in a tremendous increase in standardized achievement testing in American schools.

To provide a context for the amount of testing, one local school district will be profiled. The district is a large suburban system with approximately 99,500 students in grades K-12. During the 2003–2004 school year, the system was comprised of 76.9%

*Corresponding author.
E-mail: dgwren@aol.com

ISSN 1061-5806 print: ISSN 1477-2205 online © 2004 Taylor & Francis Ltd
DOI: 10.1080/10615800412331292606

African-American students, 10.4% White (non-Hispanic) students, 6.9% Hispanic students, and 3.4% Asian students. The remaining 2.4% fell under the classification of multiracial or other. During the school year, all students in grades 1–8 were required to take both a standardized criterion-referenced test and a norm-referenced test. Students in grades 3, 5, and 8 were administered a writing assessment, and students in grades 1, 3, 5, and 7 took a standardized aptitude test. With this testing agenda, students who enter first grade in this school system can expect to take 23 standardized tests by the time they finish eighth grade. The testing will involve about 90 school days during this 8-year period.

The increase in the amount of standardized testing brought on by the requirements of NCLB has led to pressure on school administrators and teachers to raise students' test scores (Herman and Golan, 1993; Paris *et al*., 1991) and the pressure is passed on to the students (Hembree, 1988; Hill and Wigfield, 1984). The increase in standardized testing will likely lead to an increase in test anxiety in elementary school children. When achievement test scores are influenced by test anxiety, especially for subgroups of students, the scores will be biased. In this situation, issues about the validity of the score's interpretation will be raised. Given the importance of the test scores in assessing the effectiveness of the schooling experience under NCLB, any question of bias in the scores due to test anxiety will be equally important to evaluate. Therefore having a reliable, valid, and updated measure of children's test anxiety is necessary to study this form of potential bias in children's test scores.

Unfortunately, the majority of instruments for measuring test anxiety have been developed for use with an adult population (Anderson and Sauser, 1995). Some researchers have employed these instruments or have used adapted versions in studies of test anxiety involving children (Crocker *et al*., 1988; Swanson and Howell, 1996). When an adult test anxiety scale is administered to children, all of the assumptions concerning the instrument's reliability and validity become speculative since the earlier evidence was obtained with samples of different aged subjects. Due to developmental differences, it cannot be assumed that the construct of test anxiety is the same in adult and child populations. An instrument for measuring children's test anxiety must be grounded in theory specific to the ages for which the instrument is intended.

The *Test Anxiety Scale for Children* (TASC; Sarason *et al*., 1960) was developed specifically for school-aged children and has been the most widely used self-report instrument for measuring children's test anxiety. Sarason *et al.* (1960) hypothesized that test anxiety resulted from interactions early in life between the child and the parents. The TASC is verbally administered and was conceptualized as a single dimension measured by 30 items using a yes/no response format. The scale was considered appropriate for children in grades 1–6.

Although the TASC has continued to appear in studies through the 1990s, its continued use has been questioned over the past four decades due to three major factors: outdated and/or overly complex wording of some items, outdated domain definition, and dimensionality issues. Ludlow and Guida (1991) suggested modifying certain words and phrases in TASC items as changes in teaching styles since the 1960s have rendered some of the original items obsolete. One example of an outdated item is "When the teacher asks you to write on the blackboard in front of the class, does the hand you write with sometimes shake a little?" Today it is uncommon for teachers to ask children to go to the board to work problems for the class. In addition, blackboards are being replaced in American classrooms by dry erase boards. Wigfield and Eccles (1989) questioned the validity of the TASC scores because several items appeared to be too complicated for young children. An example is, “If you are sick and miss school, do you worry that you will do more poorly in your schoolwork than other children when you return to school?” The wordiness of the TASC items is evident by the fact that nearly two-thirds of the items contain 20 or more words, including two items comprised of 34 and 36 words.

The continued use of the TASC has also been questioned with respect to the currency of the domain’s definition and dimensionality issues. Wigfield and Eccles (1989) pointed out that “the measurement of anxiety in children has not kept pace with theoretical advances in conceptualizing anxiety” (p. 164); specifically, the content domain of the TASC does not take into account more recent components such as worry-emotionality or cognitive interference. Also, Nicholls (1976) noted some of the TASC items confound test anxiety with children’s self-concept of ability. For example, “Do you sometime dream at night that other boys and girls in your class can do things that you cannot do?” The multidimensionality of the TASC has been studied by Dunn (1964) and Feld and Lewis (1967) who found four factors in their respective analyses, rather than a single dimension as theorized by Sarason *et al.* (1960). Dunn labeled the factors Test Anxiety, Generalized School Anxiety, Recitation Anxiety, and Physiological Arousal in Anticipated Recitation Situations. The four factors yielded by the Feld and Lewis (1967) study were Test Anxiety, Remote School Concern, Poor Self-Evaluation, and Somatic Signs of Anxiety. Rhine and Spaner (1973) replicated the results from Feld and Lewis (1967), using a sample of children from low-income White and African-American families. Lastly, a four-factor structure was also found in a sample of grade 7 and 8 students by Ludlow and Guida (1991), which were Test Anxiety, Somatic Signs of Anxiety, Recitation Anxiety, and Manifest Dream Anxiety. While each study seemed to identify a test anxiety factor along with several other common themes given the factor labels reported (e.g., Recitation Anxiety, Somatic Signs) there appeared to be content differences across these four studies as well (e.g., Generalized School Anxiety, Poor Self-Evaluation, and Manifest Dream Anxiety factors).

Not mentioned in any of these studies is an important change that has taken place in many American classrooms since the 1960s, namely the number of students from culturally diverse groups has grown enormously, in particular Latino children (National Center for Education Statistics, 1996). Thus, children with a wide range of language skills, some non-English speaking, are entering American schools. As such the TASC may not be relevant for children whose cultural experiences are radically different from those of the mainstream group at the time the TASC was developed. Furthermore, the TASC was designed to be administered verbally, which was probably not the best choice, especially for the primary grades 1–3. Given there are various levels of proficiency regarding listening skills, and the fact that many children as well as adults are not accomplished listeners, the likely result is in an increase in measurement error and reduced validity in TASC scores due to the introduction of construct-irrelevant variance (Messick, 1989).

The present study addressed the need for a current and reliable self-report instrument with validity evidence for measuring the construct of test anxiety in children from various ethnic groups. In this paper we describe the development and factorial validation of a new measure, the *Children's Test Anxiety Scale* (CTAS), which included four phases (Benson and Clark, 1982). The first three phases related to scale development, and in the fourth phase, we present the initial construct validity evidence.

DEVELOPMENT OF THE CTAS

Planning Phase

The planning phase involved specifying the target group for which the instrument was intended and defining the theoretical and empirical domains of the construct. The target group selected was an ethnically diverse sample of students in grades 3 through 6 (equivalent to ages 8–12). These grades were chosen for three reasons. First, a meta-analysis of test anxiety studies indicated the prevalence of test anxiety increased considerably in grades 3–5 (Hembree, 1988). Second, given Stone and Lemanek (1990) found most self-report measures for children are written at the third grade reading level, the new instrument was targeted to a child with at least a third grade reading level so that students could complete the self-report instrument with minimal assistance. Third, the bulk of standardized testing across the United States begins in grade 3.

The next step of the planning phase involved defining the theoretical and empirical domains of children's test anxiety (Benson, 1998). The theoretical domain was formulated from both an observational perspective and the research literature. As an elementary school teacher with 13 years experience, the first author has had the opportunity to observe children in various testing situations, conduct studies with children, and interview fifth and sixth graders on the topic of test anxiety for the present study. Based upon the test anxiety literature, we viewed test anxiety in children to be a situation-specific trait, which is manifest during formal evaluative situations and is experienced as an unpleasant emotional state. The manifestations of test anxiety in children are thought to include cognitions, somatic symptoms, and test-irrelevant behaviors (Dusek, 1980; Sarason, 1984; Spielberger *et al*., 1978; Wine, 1982).

Theoretical domain of test anxiety The cognitive dimension of test anxiety emerges in the literature as the most consequential component with regard to test performance. Early test anxiety researchers such as Sarason *et al.* (1960) perceived the construct as being comprised of a single dimension. Liebert and Morris (1967) advanced the theory that test anxiety was comprised of two dimensions – worry and emotionality. Worry was defined as cognitive concern about test performance, while emotionality referred to autonomic reactions that occur during test taking. Their formulation of two dimensions of test anxiety has been operationalized by Spielberger *et al.* (1978) in the development of the Test Anxiety Inventory. However, based on a comprehensive review of test anxiety studies, Wine (1982) concluded "worry scores are negatively related to performance expectancies and to actual performance, while emotionality scores bear no consistent relationship to expectancies or test performance" (p. 212).

Drawing upon the research literature in test anxiety for adults and children as well as recent observations and interviews, we considered the cognitive component of children's test anxiety to be similar to – but more encompassing than – the worry component of adult test anxiety. The *thoughts* component we propose includes the various worry cognitions that occur during testing, such as self-critical thoughts, test-related concerns, and test-irrelevant thoughts. We chose to combine the worry cognitions and test-irrelevant thoughts because we hypothesized that children would see these two areas as less differentiated than what has previously been found in adult reactions to test anxiety (Benson and El-Zahaar, 1994; Sarason, 1984).

Wigfield and Eccles (1989) have suggested the Physiological Arousal and Somatic Signs factors extracted in the factor analyses of the responses to the TASC (Dunn, 1964; Feld and Lewis, 1967; Rhine and Spaner, 1973) were similar to emotionality in the adult test anxiety literature. Because children interviewed in a pilot for the present study referred more often to somatic responses to test-related stress (e.g., perspiring, stomach problems) than to their emotions, we theorized a second component of children's test anxiety would be *autonomic reactions*.

Numerous authors have suggested a behavioral component as part of children's test anxiety (Dusek, 1980; Sieber, 1980). Early on Nottelmann and Hill (1977) referred to several *off-task behaviors* as being symptomatic of the high test-anxious child's dependency needs. Later, Fleege, Charlesworth, Burts, and Hart (1992) helped to operationally define several components we refer to as *off-task behaviors*: (a) automanipulation (rocking, playing with clothes or hair), (b) object manipulation (playing with or biting pencils), and (c) inattentive or distracted behaviors (looking around room, not focused on the test). The attentional aspect of the behavioral component has long been recognized by Wine (1982) and Dusek (1980) who noted the high test-anxious child attends to task-irrelevant stimuli more than the low test-anxious child. Thus, the third component we propose to comprise the construct of children's test anxiety is *off-task behaviors*, which includes nervous habits and other distracting behaviors that can be observed in test-anxious children. Very little has been devoted to test-anxious behaviors in the literature, perhaps due to the fact that most of the research is conducted with various adult populations and these groups do not outwardly show nervous behaviors as children do. To conclude, we propose a theoretical definition of children's test anxiety as being comprised of three correlated components: *thoughts*, *autonomic reactions*, and *off-task behaviors*.

Operationally defining children's test anxiety Corresponding to the theoretical domain of children's test anxiety was the need to write items that would operationally define the theoretical domain. In order to come up with the wording of the items for a children's scale, an open-ended questionnaire was administered to elementary school students as an optional writing assignment in order to elicit words and language children use with respect to test anxiety.

Questionnaire responses were collected from 218 students (56 third graders, 47 fourth graders, 60 fifth graders, and 55 sixth graders). Although additional demographic data were not available for the sample, the school's student body consisted of approximately 50% White students, 30% African-American students, and 20% Asian students. A content analysis of the questionnaire data was used to categorize the student responses in accordance with the three dimensions of the construct as conceptualized earlier. The content analysis was used to develop items during the construction phase of the study.

Construction Phase

The construction phase involved the creation of the initial item pool, review of the items, preliminary item tryouts, and final editing of the items. The Likert method of summated ratings was chosen as the item response format, with four response options (i.e., *almost never* = 1, *some of the time* = 2, *most of the time* = 3, *almost always* = 4).

A pool of items was constructed based upon the content analysis of the questionnaire responses obtained during the planning phase. From this data, 107 items were written to reflect the three dimensions of children's test anxiety (i.e., 42 *thoughts*, 34 *autonomic reactions*, and 31 *off-task behavior* items). All of the items were written in first person (e.g., "I worry about failing," "My heart beats fast").

A panel of eight public school teachers judged the items: two each from grades 3, 4, 5, and 6. The teachers had at least 5 years experience teaching in the grade level for which they were judging the item pool. The purpose of having the teachers judge the items was to gather evidence as to the content-representativeness of the items by a group who were familiar with teaching and observing the testing behaviors of children in these grade levels. In addition to teacher judgement of the items, an analysis of the reading level of the items was completed by a reading specialist who worked individually with students. Two students from each of grades 3, 4, 5, and 6 were selected to read all of the items. The reading level analysis revealed a number of words and phrases the students were unable to read or comprehend (e.g., *sweat*, *fidget*, *butterflies in my stomach*).

Using the data from the teacher ratings and the reading level analysis, decisions were made as to whether items should be retained, revised, or discarded. Earlier, several third grade teachers had recommended that no more than 50 items could be administered to a group of 8- and 9-year-olds in a single session without occurrences of inattention or weariness. Thus, the item pool was reduced to 50 items (23 *thoughts*, 14 *autonomic reactions*, and 13 *off-task behavior* items) after discarding items containing difficult words or phrases, and items judged by the teachers as not relevant for the content areas thought to represent the theoretical domain.

Quantitative Evaluation Phase

The quantitative evaluation phase of the study was designed to (a) obtain data for estimating the internal consistency of the new scale and subscales, (b) obtain an initial indication of the plausibility of the three-factor structure proposed for children's test anxiety and how well the items map on to the theoretical domain, and (c) assess the relationship among the factors.

Sample and procedures The 50-item scale was administered to a sample of 230 children. The sample included 61 third graders, 58 fourth graders, 63 fifth graders, and 48 sixth graders. The sample distribution was 50% female and in terms of ethnicity was composed of 40% African Americans, 6% Asians, and 54% White students. This sample will be referred to as the development sample and will be used to evaluate the three objectives of the quantitative evaluation phase mentioned above.

Each student received a copy of the 50-item instrument, which included written instructions, questions pertaining to demographics, and a sample item for practice. The instructions asked the students to respond in terms of how they think, feel, or act during a test. Each question was responded to with the stem, "While I am taking tests . . .." The term *test anxiety* was intentionally omitted from the instrument; instead the title "Test Attitude Survey" was used. All administrations were completed in the students' classrooms during regular school hours with no assistance from the staff at the schools.

Data analysis Response frequencies and distributions for each item were considered for the entire sample as well as for the subsamples of gender, race, and grade level. Some items showed non-normal distributions for the entire sample as well as for specific subsamples. All of these items were noted for further review. Descriptive statistics and reliability of the 50-item instrument was estimated using alpha coefficients for the overall scale and for each of the three subscales. Item-total correlations were also calculated to determine how well each item contributed to the measurement of its respective dimension, as operationally defined.

All 50 items had correlations ranging from 0.22 to 0.71 within their subscales. The item-total correlations were examined for each subsample of gender, race, and grade level. Ten items that correlated less than 0.20 with their subscale were flagged. A total of 20 items were subsequently discarded based upon low item–total correlations, non-normal distributions, or wording problems. Thus, a reduced 30-item CTAS was retained which included 9 items on the Autonomic Reactions subscale, 8 items on the Off-Task Behaviors subscale, and 13 items on the Thoughts subscale. The reliability estimates for the 30-item CTAS was 0.92 and its subscales were 0.85 for Autonomic Reactions, 0.78 for Off-Task Behaviors, and 0.89 for the Thoughts subscale.

Confirmatory factor analyses Maximum likelihood estimation using LISREL 8.30 was used to test the theoretical model of interest. This model was specified as a simple structure three-factor model using all 30 items (Model 1: 30-items 3 factors). Model 1 did not permit any complex loadings. As noted in Table I, the overall model-data fit of Model 1 showed a significant chi-square ($\chi^2 = 853$, $df = 402$) indicating the model-data fit could be improved. However, the other indicators of fit showed the model to be a reasonable approximation to the data. The standardized item loadings and factor intercorrelations for Model 1 are reported in Table II. All loadings and correlations were statistically significant. The relationships among the factors revealed the Thoughts factor was highly related to the Autonomic Reactions factor (0.64) and less so with the Off-Task Behaviors factor (0.53). The relationship between the Autonomic Reactions factor and Off-Task Behaviors was higher (0.60).

While the fit of Model 1 seemed reasonable, we sought to simplify the 30-item scale by seeing if additional items could be removed and still maintain the internal consistency and theoretical three-factor simple structure. In confirmatory analyses there are indications as to what parts of the model fit might be improved. However, after the first test, the subsequent analyses are no longer confirmatory even though the program will still produce indices of statistical fit. Thus, the indices obtained in

TABLE I Confirmatory factor analyses for measurement models

Models	χ^2	*df*	χ^2/df	*TLI*	*RMSEA*	*CI*
Development sample						
Model 1 (30-item 3 factors)	853	402	2.12	0.806	0.074	0.068–0.080
Model 2 (25-item 3 factors)	529	272	1.94	0.851	0.067	0.058–0.075
Validation sample						
Model 3 (cross-validation of Model 1)	880	402	2.19	0.812	0.069	0.063–0.075
Model 4 (cross-validation of Model 2)	589	272	2.17	0.837	0.066	0.059–0.074

Note. TLI, Tucker–Lewis Index; RMSEA, Root Mean Square Error of approximation; CI, 90% confidence interval for RMSEA.

TABLE II Standardized factor loadings of CTAS

Items	*Thoughts (13)*	*Off-Task Behaviors (8)*	*Autonomic Reactions (9)*
5. I think I am going to get a bad grade.	0.76 (0.71)		
24. I think about what will happen if I fail.	0.75 (0.72)		
29. I worry about what my parents will say.	0.72 (0.64)		
9. I worry about failing.	0.70 (0.72)		
11. I worry about doing something wrong.	0.68 (0.67)		
1. *I wonder if I will pass.	0.65 (0.48)		
13. I think about what my grade will be.	0.63 (0.60)		
19. I think most of my answers are wrong.	0.60 (0.75)		
27. I think about how poorly I am doing.	0.58 (0.62)		
21. I worry about how hard the test is.	0.55 (0.59)		
16. I think that I should have studied more.	0.48 (0.52)		
15. *I wonder if my answers are right.	0.49 (0.59)		
6. It is hard for me to remember the answers.	0.37 (0.45)		
3. I look around the room.		0.74 (0.68)	
18. I look at other people.		0.71 (0.61)	
30. I stare.		0.60 (0.63)	
12. I check the time.		0.55 (0.38)	
14. *I find it hard to sit still.		0.54 (0.57)	
26. I tap my feet.		0.53 (0.41)	
7. I play with my pencil.		0.42 (0.59)	
22. I try to finish up fast.		0.40 (0.39)	
4. I feel nervous.			0.69 (0.71)
2. My heart beats fast.			0.69 (0.67)
28. I feel scared.			0.67 (0.75)
17. My head hurts.			0.63 (0.55)
8. *My face feels hot.			0.62 (0.44)
20. I feel warm.			0.61 (0.51)
23. My hand shakes.			0.60 (0.53)
10. My belly feels funny.			0.57 (0.66)
25. *I have to go to the bathroom.			0.41 (0.38)
Factor correlations			
Thoughts	1.00		
Off-Task Behaviors	0.53 (0.56)	1.00	
Autonomic Reactions	0.64 (0.81)	0.60 (0.47)	1.00

Note. Results from the validation sample are in parentheses; *; deleted items.

subsequent reanalyses will be interpreted descriptively only and will require cross-validation to ensure their stability. With that perspective in mind, the CFA results from Model 1 indicated the model-data fit could be improved if several items were deleted from the 30-item model.

Through a sequence of removing one item at a time using information from the standardized residuals and modification indices we systematically deleted five items (two from the Thoughts subscale, two from the Autonomic Reactions subscale, and one from the Off-Task Behaviors subscale), leaving a 25-item scale. This model, reported as Model 2 in Table I, showed a statistically significant chi-square ($\chi^2 = 529$, $df = 272$); however additional indices of fit revealed a modest improvement over Model 1. All of the loadings for the 25 items were statistically significant as were the correlations among the factors. The same relative relationships among the factors remained for the 25-item model. The relationship between the Thoughts factor and the Autonomic Reactions factor was 0.61 and with the Off-Task Behaviors factor was 0.49, whereas the relationship between the Autonomic Reactions factor and Off-Task Behaviors was 0.55.

Internal construct validation of the CTAS The fourth phase of the study was the validation phase, which involved administering the reduced 30-item scale to a different sample of students in grades 3–6. It was important to cross-validate the findings from the quantitative evaluation phase since 20 items were removed from the 30-item scale in an effort to clarify the three-factor structure. The three purposes of the internal validation phase were to cross-validate with a new sample (a) the reliability estimates of the 30-item CTAS and subscales and (2) the findings from the development sample of the instrument's structure, and (3) to evaluate whether the 25- or 30-item model best fit the data. With a stable internal structure, an external domain study should be conducted where the focal construct, *children's test anxiety*, is evaluated against other similar and dissimilar constructs in a nomological network (Benson, 1998). Thus, an external domain study to evaluate the convergent and/or discriminate validity of the focal construct would be premature at this stage of scale development.

Sample and procedures Data were collected from 261 students at four different schools from those in used the development sample. This validation sample included 46 third graders, 55 fourth graders, 94 fifth graders, and 66 sixth graders. The sample was made up of 53% females and 44% African-Americans, 51% White students, and 5% Asian students. The 30-item CTAS was administered in the same manner as in the previous phase. The number of students who required assistance reading or understanding items was notably fewer, and all of the students finished the CTAS in 5–12 minutes. Thus the readability of the instrument seemed to be adequate. A copy of the CTAS is available from the first author.

Data analysis Item data from the 261 respondents were used to ascertain the instrument's reliability and to cross-validate the factor structure of the responses. The internal consistency estimates for the CTAS were satisfactory and highly similar to the results from the development sample: 0.92 for the overall scale, 0.76 for the 8-item Off-Task Behaviors subscale, 0.82 for the 9-item Autonomic Reactions subscale, and 0.89 for the 13-item Thoughts subscale. The alphas for the subsamples of gender, race, and grade level ranged from 0.61 to 0.93. All but two of the subsample alphas were greater than 0.70 (grade 3 Off-Task Behaviors = 0.61; Asians Off-Task Behaviors = 0.68), and over two-thirds were greater than 0.80.

The means and standard deviations for the entire sample and each subsample are presented in Table III. At the bottom of the table are the overall descriptives for the 30-item scale from the development sample for comparative purposes. A three-way ANOVA on the total score revealed there were significant main effects for gender and race only ($p < 0.05$). None of the interactions or grade level were significant. A review of the means indicated that girls reported higher levels of test anxiety than boys and a post hoc analysis on race revealed the differences were only between African-American and White students where the African-Americans reported greater levels of test anxiety. These findings replicate those reported by Silverman *et al*. (1995) for children's worries across race, gender, and grade level for children ages 7–12.

Looking across the three subscales means, it was found that only the Autonomic Reactions and Thoughts subscales showed significant differences, again where the African-American students endorsed to a greater extent the Thoughts items and

TABLE III Descriptive statistics for Children's Test Anxiety Scale (CTAS) based on validation sample

Subsample				*Subscales (items)*					
				AR (9)		*OTB (8)*		*T (13)*	
	n	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Gender									
Girls	138	65.78	16.93	17.11	5.90	17.33	5.11	31.34	8.95
Boys	123	57.71	14.94	14.67	5.03	16.40	5.16	26.63	7.92
Race									
Asian	14	57.36	12.49	13.79	3.81	15.86	4.09	27.71	7.63
African-American	115	66.38	16.57	17.02	6.19	17.67	5.44	31.70	8.30
White	132	58.62	15.97	15.27	5.12	16.32	4.90	27.03	8.78
Grade									
3	46	62.76	14.81	17.15	4.53	16.07	4.36	29.54	8.51
4	55	60.98	16.61	16.25	5.25	16.65	5.06	28.07	8.65
5	94	62.95	17.98	15.94	6.37	17.27	5.37	29.74	9.62
6	66	60.86	15.52	14.92	5.42	17.12	5.42	28.82	7.91
Total *N*	261	61.97	16.49	15.96	5.63	16.89	5.14	29.12	8.79
Devel. sample	230	55.77	15.54	15.12	5.66	14.15	4.71	26.50	8.42

Note. AR, Autonomic Reactions subscale; OTB, Off-Task Behaviors subscale; T, Thoughts subscale; Number items per scale in parentheses. Devel. sample, results from the development sample for comparative purposes.

Autonomic Reactions items than the White students. Similarly for the gender comparisons, the girls reported higher levels of Autonomic Reactions and Thoughts than boys did. None of the means for the Off-Task Behavior subscale were significantly different across gender, grade level, or race. Given that there were only 14 Asian students in the sample, the above analyses were rerun with this group removed. The above findings remained the same with the exception that the means for the Off-Task Behaviors subscale differed significantly ($p < 0.05$). With the 14 Asian students removed, the African-American students reported greater levels of Off-Task Behaviors than did the White students.

Considering the stability of the three components of children's test anxiety, two models were evaluated: Model 3, the theoretical simple structure three-factor 30-item model and Model 4, the reduced 25-item model obtained from the development sample. Models 3 and 4 represented a cross-validation of the prior results to determine how generalizable and stable each model was for this population.

The overall fit of Model 3 showed a significant chi-square ($\chi^2 = 880$, $df = 402$), which was slightly higher than Model 1 and indicated the model-data fit could be improved (see the lower half of Table I). However, the other indicators of fit (Tucker–Lewis Index and Root Mean Square Error of Approximation) again revealed Model 3 was a reasonable approximation to the data. The standardized item loadings and factor intercorrelations for the 30-item three-factor models (Models 1 and 3) are reported in Table II, the validation sample results are in parentheses. All loadings and correlations were statistically significant. The relationships among the factors revealed the Thoughts factor was highly related to the Autonomic Reactions factor (0.81) and less so with the Off-Task Behaviors factor (0.56), whereas the relationship among the Autonomic Reactions factor and Off-Task Behaviors was moderate (0.47).

To determine if the model that resulted from the deletion of five items using the development sample would cross-validate, the 25-item three-factor Model 4 was tested using the validation sample. The results are given in Table I and reveal a pattern similar to Model 2, although the results are somewhat higher (meaning a somewhat worse fit) than Model 2. This type of result is expected in cross-validation studies and is similar to shrinkage in multiple R, a regression phenomenon. It was noted that the expected cross-validation index from the development sample was 2.737 (95% confidence interval = 2.472–3.036); this index in the validation sample was 2.766, which was well within the confidence interval. Thus, the 25-item CTAS seems to have cross-validated just as well as the 30-item CTAS. Given the internal consistency estimates did not vary appreciably, the decision to use a 25-item scale could be made on the basis of parsimony and efficiency. It was noted the correlation between Thoughts and Autonomic Reactions factors decreased slightly (0.77) for the 25-item scale; the remaining factor correlations were highly similar to the 30-item scale (ranging from 0.45 to 0.59).

DISCUSSION AND FUTURE RESEARCH

Reliability of the Instrument

Initial reliability estimates were obtained from administration of the 50-item instrument to the development sample ($N = 230$). Based upon an item analysis, coefficient alpha for the reduced 30-item instrument was 0.92, and the subscale alphas ranged from 0.78 to 0.89. The reliability of the 30-item CTAS was replicated using the validation sample ($N = 261$) where the alpha estimate for the overall scale was again 0.92 and the alphas for the subscales ranged from 0.76 to 0.89. Similar values were noted for the 25-item version ($\text{CTAS}_{25} = 0.89$; subscales ranged from 0.73 to 0.86). Thus, the measurement of children's test anxiety with the CTAS seemed to be quite consistent across these two samples.

While we experimented with trying to reduce the number of items on the CTAS, we feel that the fit of the 25-item model did not improve the instrument appreciably. Therefore, the 30-item scale should continue to be used until additional data are collected which would allow more refined subgroup comparisons such as those mentioned previously.

Internal Construct Validity of the Instrument

Ascertaining validity evidence is a complicated process, and the answer can only be approximated, since "validation is a continual process, one in which an end point is rarely achieved" (Benson and Clark, 1982, p. 800). The CFAs in both the development and validation samples showed all item-factor loadings were statistically significant and ranged from 0.37 to 0.76 (over half were ≥ 0.60). However, in comparing specific loadings across the two samples, it was noted that 9 of 30 items had loadings that differed by ≥ 0.10 (6 of the 9 were lower in the validation sample). Thus, while the three-factor structure appears stable over the two samples, individual item loadings varied somewhat. Furthermore, the correlation among the Thoughts and Autonomic Reactions factors is perhaps too highly correlated in the validation sample (0.81) to say the factors are distinct, while the other factor correlations were moderately correlated as theory would predict, ranging from 0.47 to 0.64.

The main objectives of this study were to define, develop, and provide initial construct validation for an updated and more refined measure of test anxiety in children. We believe we have developed a measure that overcomes some of the noted shortcomings of the TASC (e.g., obsolete and overly complex item wording, outdated domain definition, and dimensionality issues). The mean words per item in the TASC is 20.7 whereas the mean words per item in the CTAS is 10.7, which includes the 5 words in the stem. Removing the words in the stem results in an average of 5.7 words per item in the CTAS. While the two scales are similar in terms of the number of items, the wording of the CTAS items was developed using the language of children in grades 3–6, and the format of the CTAS uses four Likert anchor points compared to a yes/no format in the TASC. Overall, the CTAS is a theoretically more updated measure of children's test anxiety that appears to hold promise as a reliable and potentially valid measure for use with a multi-ethnic elementary school population. With these initial positive findings in mind, several suggestions for future research for the CTAS are presented.

Future Research Directions

First, while factor analysis is one of the most widely used methods for assessing construct validity, factor analysis (either exploratory or confirmatory) itself is a circular process (Benson, 1998). The confirmatory factor analyses of the CTAS data indicated three factors were present, and the three factors appear to correspond well with the dimensions specified in the theoretical definition of children's test anxiety that guided the development of the CTAS. Although factor analytic studies can provide important information on the internal structure of the CTAS, external domain studies of how children's test anxiety functions with other latent variables in a nomological network are needed to show that children's test anxiety is a separate construct (Benson, 1998). External domain studies might consider how predictive the CTAS and its subscales are of student achievement and how antecedents such as classroom climate or ability level interact with test anxiety to influence student achievement in predicted ways (Zatz and Chassin, 1985). Finally, the issue of social desirability affecting children's responses needs to be investigated to ensure the validity of the scores are not influenced by this source of "construct-irrelevant" test variance (Messick, 1989). Several types of external domain studies described above are important next steps in the continued construct validation of the CTAS scores.

Second, the three-dimensional theory of test anxiety in children is an initial attempt to define currently the theoretical content domain of the construct for this population. It is possible the content domain of children's test anxiety may have been defined too narrowly in the present study and other components may have been overlooked. Messick (1989) has referred to "construct underrepresentation, [when] the test is too narrow and fails to include important dimensions or facets of the construct" (p. 34). Given the evidence that children from the ages of 7 to 11 years are able to accurately recognize negative emotions and use mental cues to understand their own emotions (Stone and Lemanek, 1990), the existence of an emotional component within the construct of children's test anxiety is possible. Future studies along the lines of that conducted by Hodapp and Benson (1997) are needed where alternative content domains for children's test anxiety are posited and systematically evaluated.

Third, it is strongly recommended that larger samples be employed to enable a more refined analysis by different student subgroupings. With larger samples confirmatory factor analysis could be used to detect any item bias in the CTAS over different subgroups of students (Benson, 1987). That is, the factor structure of CTAS for boys and girls and various racial groups could be statistically compared to determine if the construct is invariant over these student groupings or if the differences noted in the item loadings across the two samples in this study could be associated with different student subgroups. If the construct was found to be invariant, then more meaningful interpretations of differences in mean scores could be made among girls and boys or African-American and White children's responses to the items. Until we know if the construct is manifested similarly across various student subgroups, it is difficult to fully explain the meaning of the higher mean scores of the African-American and female students in the present study. Finally, having larger samples would permit one to conduct analyses by grade level to determine if the three-dimensional model of children's test anxiety would be stable developmentally.

In sum, the CTAS appears to be a reliable measure of the three components of children's test anxiety, as theoretically and operationally defined in this study. The scale was purposefully developed and initial validity evidence using a racially diverse sample of elementary students is promising. However, additional validity evidence for the scores is clearly needed. With the continued increase in mandated state and district standardized testing, an associated increase in children's test anxiety is expected. Therefore, an updated instrument with evidence of reliability and construct validity for measuring children's test anxiety is needed.

References

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

Benson, J. and Clark, F. (1982). A guide for instrument development and validation. *American Journal of Occupational Therapy*, **36**, 789–800.

Benson, J. (1987). Detecting item bias in affective scales. *Educational and Psychological Measurement*, **47**, 55–67.

Benson, J. and El-Zahaar, N. (1994). Further refinement and validation of the revised Test Anxiety Scale. *Structural Equation Modeling*, **1**, 203–221.

Benson, J. (1998). Developing a strong program of construct validation: a test anxiety example. *Educational Measurement: Issues and Practice*, **17**, 10–17, 22.

Crocker, L., Schmitt, A. and Tang, L. (1988). Test anxiety and standardized achievement test performance in middle school years. *Measurement and Evaluation in Counseling and Development*, **20**, 149–157.

Dunn, J.A. (1964). Factor structure of the Test Anxiety Scale for Children. *Journal of Consulting Psychology,* **28**, 92.

Dusek, J.B. (1980). The development of test anxiety in children. In: Sarason, I.G. (Ed.), *Test anxiety: Theory, research, and applications*, pp. 87–110. Lawrence Erlbaum Associates, Hillsdale, NJ.

Feld, S. and Lewis, J. (1967). Further evidence on the stability of the factor structure of the Test Anxiety Scale for Children. *Journal of Consulting Psychology*, **31**, 434.

Fleege, P.O., Charlesworth, R., Burts, D.C. and Hart, C.H. (1992). Stress begins in kindergarten: A look at behavior during standardized testing. *Journal of Research in Childhood Education*, **7**, 20–26.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research*, **58**, 47–77.

Herman, J.L. and Golan, S. (1993). The effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice*, **12**, 20–25, 41–42.

Hill, K.T. and Wigfield, A. (1984). Test anxiety: A major educational problem and what can be done about it. *The Elementary School Journal*, **85**, 105–126.

Hodapp, V. and Benson, J. (1997). The multidimensionality of test anxiety: A test of different models. *Anxiety, Stress and Coping*, **10**, 219–244.

Liebert, R.M. and Morris, L.W. (1967). Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological Reports*, **20**, 975–978.

Ludlow, L.H. and Guida, F.V. (1991). The Test Anxiety Scale for Children as a generalized measure of academic anxiety. *Educational and Psychological Measurement*, **51**, 1013–1021.

Messick, S. (1989). Validity. In: Linn, R.L. (Ed.), *Educational Measurement*, 3rd ed., pp. 13–103. American Council on Education, Washington, DC.

National Center for Education Statistics. (1996). *Urban schools: The challenge of location and poverty* (NCES Publication No. 96–184). US Department of Education, Washington, DC.

Nicholls, J.G. (1976). When a scale measures more than its name denotes: The case of the Test Anxiety Scale for Children. *Journal of Consulting and Clinical Psychology*, **44**, 976–985.

No Child Left Behind Act of 2001. (2001) Pub. L. No 107–110. (Web address: www.ed.gov/nclb/landing.jhtml).

Nottelmann, E.D. and Hill, K.T. (1977). Test anxiety and Off-Task Behavior in evaluative situations. *Child Development*, **48**, 225–231.

Paris, S.G., Lawton, T.A., Turner, J.C. and Roth, J.L. (1991). A developmental perspective on standardized achievement testing. *Educational Researcher*, **20**, 12–20.

Rhine, W.R. and Spaner, S.D. (1973). A comparison of the factor structure of the Test Anxiety Scale for Children among lower- and middle-class children. *Developmental Psychology*, **9**, 421–423.

Sarason, I.G. (1984). Stress, anxiety, and cognitive interference: Reactions to tests. *Journal of Personality and Social Psychology*, **46**, 929–938.

Sarason, S.B., Davidson, K.S., Lighthall, F.F., Waite, R.R. and Ruebush, B.K. (1960). *Anxiety in elementary school children*. Wiley, New York.

Sieber, J.E. (1980). Defining test anxiety: Problems and approaches. In: Sarason, I.G. (Ed.), *Test anxiety: Theory, research, and applications*, pp. 15–40. Lawrence Erlbaum Associates, Hillsdale, NJ.

Silverman, W.K., LaGreca, A. and Wasserstein, S. (1995). What do children worry about? Worries and their relation to anxiety. *Child Development*, **66**, 671–686.

Spielberger, C.D., Gonzalez, H.P., Taylor, C.J., Algaze, B. and Anton, W.D. (1978). Examination stress and test anxiety. In: Spielberger, C.D. and Sarason, I.G. (Eds.), *Stress and anxiety*, Vol. 5, pp. 167–191. Hemisphere, Washington, DC.

Stone, W.L. and Lemanek, K.L. (1990). Developmental issues in children's self-reports. In: La Greca, A.M. (Ed.), *Through the eyes of a child: Obtaining self-reports from children and adolescents*, pp. 3–17. Allyn & Bacon, Boston.

Swanson, S. and Howell, C. (1996). Test anxiety in adolescents with learning disabilities and behavior disorders. *Exceptional Children*, **62**, 389–397.

Valverde, G. and Schmidt, W. (1997). Refocusing US math and science education. *Issues in Science and Technology*, **14**, 60–66.

Wigfield, A. and Eccles, J.S. (1989). Test anxiety in elementary and secondary school students. *Educational Psychologist*, **24**, 159–183.

Wine, J. (1982). Evaluation anxiety: A cognitive-attentional construct. In: Krohne, H.W. and Laux, L. (Eds.). *Achievement, stress, and anxiety*, pp. 207–219. Hemisphere, Washington, DC.

Zatz, S. and Chassin, L. (1985). Cognitions of test-anxious children. *Journal of Consulting and Clinical Psychology*, **51**, 526–534.