


May 2022

The Multidimensional Test Anxiety Scale: A Latent Profile Analysis and an Examination of Measurement Invariance

Gabrielle Francis
University of South Florida

Follow this and additional works at: <https://digitalcommons.usf.edu/etd>

 Part of the [Educational Assessment, Evaluation, and Research Commons](#), and the [Psychiatric and Mental Health Commons](#)

Scholar Commons Citation

Francis, Gabrielle, "The Multidimensional Test Anxiety Scale: A Latent Profile Analysis and an Examination of Measurement Invariance" (2022). *USF Tampa Graduate Theses and Dissertations*.
<https://digitalcommons.usf.edu/etd/9352>

This Ed. Specialist is brought to you for free and open access by the USF Graduate Theses and Dissertations at Digital Commons @ University of South Florida. It has been accepted for inclusion in USF Tampa Graduate Theses and Dissertations by an authorized administrator of Digital Commons @ University of South Florida. For more information, please contact scholarcommons@usf.edu.

The Multidimensional Test Anxiety Scale: A Latent Profile Analysis and an Examination of
Measurement Invariance

by

Gabrielle Francis

A thesis submitted in partial fulfillment
of the requirements for the degree of
Educational Specialist
in Curriculum and Instruction with a concentration in
School Psychology
Department of Educational and Psychological Studies
College of Education
University of South Florida

Major Professor: Nathaniel von der Embse, Ph.D.
Member: David Putwain, Ph.D.
Member: Eunsook Kim, Ph.D.

Date of Approval:
May 12, 2022

Keywords: classification, cut scores, usability, validation

Copyright © 2022, Gabrielle Francis

Table of Contents

List of Tables	iv
List of Figures	v
Abstract	vi
Chapter One: Introduction	1
Statement of the Problem	1
Purpose of Study	3
Research Questions/Purposes	4
Definition of Terms	4
Cognitive Interference	4
Worry	5
Tension	5
Physiological Indicators	5
Measurement Invariance	5
Classification Standards	5
Theoretical Framework	6
Significance of the Study	7
Chapter Two: Literature Review	8
Evolution of Test Anxiety Theory and Measurement	13
Operational Definitions	16
Measuring Test Anxiety	18
Critique of Test Anxiety Measures	19
Development and Initial Validation of the Multidimensional Test Anxiety Scale	22
Measurement Invariance	24
Examination of Response Profiles	25
Purpose of the Present Investigation	27
Chapter Three: Methods	28
Participants	28
Variables	29
Test Anxiety	29
Gender	29
Grades	29
Socio-economic Status	29
Measures	29
MTAS	29
Data Collection Procedure	30

Data Analysis	31
Reliability.....	31
Measurement Invariance.....	31
Latent profile Analysis.....	34
Chapter Four: Results	36
Missing Data Analysis and Treatment.....	36
Descriptive Statistics.....	36
Confirmatory Factor Analysis.....	39
Reliability Analysis.....	42
Measurement Invariance.....	43
Measurement Invariance Across Gender	44
Model 1	44
Model 2	44
Model 3	46
Model 4	46
Model 5	47
Model 6	47
Model 7	47
Measurement Invariance Across Socioeconomic Status	48
Model 1	48
Model 2	48
Model 3	48
Model 4	49
Model 5	49
Model 6	49
Model 7	51
Measurement Invariance Across Grades	51
Model 1	51
Model 2	51
Model 3	51
Model 4	52
Model 5	52
Model 6	52
Model 7	53
Latent Profile Analysis	53
Model Evaluation.....	53
Interpretation of LPA.....	57
Chapter Five: Discussion	59
Summary and Explanation of Findings.....	60
Limitations and Future Research	62
Implications for Practice	63
Conclusion	64
References	66

List of Tables

Table 1: Descriptive Statistics for MTAS Items	36
Table 2: Correlation Matrix for MTAS Scales	38
Table 3: Correlation Matrix for MTAS Items	38
Table 4: Confirmatory Factor Analysis	40
Table 5: Standardized Factor Loadings from the Combined Measurement Model	41
Table 6: Internal Consistency Reliability Statistics for MTAS	43
Table 7: Descriptive Fit Statistics for Measurement Invariance Across Gender	45
Table 8: Descriptive Fit Statistics for Measurement Invariance Across Socioeconomic Status	50
Table 9: Descriptive Fit Statistics for Measurement Invariance Across Grades	54
Table 10: Descriptive Fit Statistics for LPA Models	58
Table 11: Mean Subscale Scores and Sample Percentages for Latent Profiles	58

List of Figures

Figure 1: MTAS Path Diagram	39
Figure 2: Path Diagram	42
Figure 3: MPlus Profile Plot for Four Profile Model	56

Abstract

Standardized testing is an integral part of the English and American education systems. The objectives of these tests are to evaluate students, teachers, and schools. However, this evaluation has unintended consequences, one of which is test anxiety. Over the last 50 years, there has been an increase in studies on test anxiety because of the widespread use of standardized tests (Hembree, 1988; von der Embse et al., 2019). However, two areas that have seen little attention are the measurement invariance of test anxiety scales across demographic groups, and the creation of classification standards for these test anxiety scales to increase usability. Examining measurement invariance is needed to determine if assessments measure the same constructs across groups. The lack of research in this area makes it unclear whether groups truly differ in severity of test anxiety or if the measurement tools themselves are flawed. Additionally, many test anxiety instruments are created for research rather than practice and lack evidence for classification standards or cut scores. This study seeks to address these limitations by examining the MTAS for measurement invariance across gender, age, and socio-economic status and by examining the differences between cluster groups identified through a Latent Cluster Analysis. The data used in this study was collected in the 2019 – 2020 school year from 918 students attending secondary school in England. The analyses that will be completed are a Confirmatory Factor Analysis to determine measurement invariance and a Latent Profile Analysis to determine classifications.

Chapter One: Introduction

Statement of the Problem

Accountability and testing have become essential parts of education in England and the United States (US). In England, students undergo several rounds of testing that begin at age seven and continue throughout their education. These standardised tests are used to measure students' academic performance in areas such as reading, science, and math during important stages in their development. Testing is also prevalent in the US, where policies such as the Every Student Succeeds Act (ESSA, 2015) rely on standardised testing to determine the performance of students, teachers, and schools. There are several benefits to standardised testing, including the ability to compare schools against national benchmarks and each other and to identify students in need of academic support (Carter et al., 2016). However, there are unintended consequences. The pressure that students experience from teachers and parents may result in the development of test anxiety (Putwain et al., 2010). Students experience greater stress levels in high-stakes exams, like the National Curriculum Tests taken regularly by English students, than they do in regular class exams (Segool, et al., 2013). Up to 40% of students experience high test anxiety, especially for high-stakes examinations (Segool, et al., 2013).

Test anxiety has become an important area of research as standardised testing has become entrenched in education systems (von der Embse & Hasson, 2012). Some groups experience higher levels of test anxiety than others. Across middle and high school and college, girls have shown higher rates of test anxiety when compared to boys (Lowe, 2014). Similarly, there are differences in test anxiety levels between high and low socio-economic status students. Students

living in low-income families typically have high test anxiety (Putwain, 2007). The relationship between test anxiety and test performance may differ across grades or ages. The association between test anxiety and test performance increases from middle school to college, though the relationship remained negative with high levels of test anxiety related to poor test performance (von der Embse et al., 2018). However, measurement invariance must be examined before researchers can analyse differences in test anxiety across groups. Measurement invariance is a statistical analysis used to determine if an assessment is biased against racial/ethnic groups and gender groups. It examines scales for the underlying construct across groups to determine if these remain constant across these groups. A scale with measurement invariance assesses the same construct across groups. When scales that have not been examined for measurement invariance are used in research, there is potential for error and biased decision-making (Pendergast et al., 2017). The scales may be biased against certain groups making any outcomes erroneous (Pendergast et al., 2017). Though several studies analyse differences in test anxiety across gender, age, and socio-economic status (Everson et al., 1991; Aydin, 2019; Fayegh et al., 2010; Segool et al. 2013), these studies often rely on assessments that have not been examined for measurement invariance.

Another important aspect of assessment development is determining classifications and standards for inclusion in these categories. The creation of these cut scores increases an assessment's usability in schools. For example, cut scores help identify levels of test anxiety that are concerning and indicate a need for support compared to levels that are below an established threshold and thus not as concerning. Many test anxiety assessments do not provide these standards because they are primarily used for research instead of practice (von der Embse et al., 2021), e.g. the FRIEDBEN Test Anxiety Scale and the Test Anxiety Inventory (Friedman &

Bendas, 1997; Spielberger, 1980). The Multidimensional Test Anxiety Scale (MTAS) was developed from a modern theoretical framework of test anxiety and has had several psychometric studies to date in support of factor structure and reliability. The MTAS has initial evidence for cut scores to inform use (Putwain et al., 2020). These initial classification standards were created using a variable centred analysis approach. This study used a different method, a person-centred analysis approach, to create new classification standards.

Purpose of Study

The primary purpose of this study was to create profiles of responses for classification standards. This will increase the usability of the MTAS in schools as administrators will know which scores fall within 'at risk' versus 'not at risk.' Before these profiles could be created, however, the sample was examined for measurement. This study examined measurement invariance across gender (male and female), grades (Year 10 – 13), as well as socio-economic status (eligible for free school lunch or not). Then response profiles were created. Initial cut scores of 58 and 60 on the MTAS Total scale have been created for this assessment using a variable centred analysis approach, and these cut scores can be used to determine whether or not a student is experiencing high test anxiety (von der Embse, et al., 2021). This study used a person-centered approach to identify profiles and examine the significance of differences between the profiles, which is an important step in developing the MTAS tool. The creation and validation of classifications will increase the usability of the MTAS for both research and practice. Results from this study provide evidence to support use, allowing users to identify students with high test anxiety and thus facilitate intervention services. This is especially important for practice since many test anxiety scales (e.g. CTAS; Wren & Benson, 2004) do not

have classification standards. When these scales are used, the administrator must subjectively decide which range of scores is 'at risk' and thus deserving of intervention.

Research Questions/Purposes

1. Is there measurement invariance within different demographic groups (gender, grade, and socio-economic status)?
2. Are there significant differences in levels of test anxiety amongst classification clusters?

Definition of Terms

Test anxiety is the changes in behaviour, emotion, and physiology resulting from an individual's perception of evaluative situations as threatening (Spielberger & Vagg, 1995). This definition of test anxiety was used as a guiding principle for conceptualising the underlying domains of test anxiety. The focus of the MTAS tool is to measure responses to evaluative situations, not the responses and actions of individuals before or after such situations (Putwain et al., 2020). This led to the inclusion of two cognitive dimensions, worry and cognitive interference, and two affective-physiological dimensions, tension and physiological indicators. These domains are represented across the 16 items of the MTAS (Putwain et al., 2020). In addition, this study focused on two particular aspects of the MTAS, measurement invariance and classification standards.

Cognitive Interference. Cognitive interference refers to thoughts that are test irrelevant and are associated with avoidance coping, it is also often called test-irrelevant thinking (Schutz et al., 2004). It is measured with four items in the MTAS, an example of one of the items is: “During tests/exams, I find it hard to concentrate.”

Worry. Worry refers to thoughts focused on possible failure in a test and the consequences of this (Putwain, Connors, & Symes, 2010). It is measured with four items in the MTAS, an example of one of the items is: “Before a test/ exam, I am worried I will fail.”

Tension. Tension refers to feelings of nervousness and is often also referred to as emotionality (Sarason 1984). It is measured with four items in the MTAS, an example of one of the items is: “Even when I have prepared for a test/ exam I feel nervous about it.”

Physiological Indicators. Physiological Indicators refer to autonomic arousal and physical reactions that accompany test anxiety. It is measured with four items in the MTAS, an example of one of the items is: “My heart races when I take a test/exam.”

Measurement Invariance. Measurement invariance is a type of statistical analysis used to determine if the underlying construct being measured is the same across different groups or across time. This is the desired outcome in an assessment that is achieved when "the relationship between response to items and latent constructs are the same across groups" (Pendergast et al., 2017). This study focused on multi-group measurement invariance, which is concerned with the stability of the scale's underlying construct across multiple groups. Gender (male and female), age, and socio-economic status were the focus of measurement invariance analyses.

Classification Standards. This is a term used to specify the scores on an assessment that place students in different categories, typically 'at risk' vs. 'not at risk.' The methods used to determine cut scores typically fall into two groups, person centred approaches or variable centred approaches. They differentiate between these two groups and inform the test administrator of which students are at risk and in which specific areas. Receiver Operating Characteristics Area Under the Curve (ROC AUC) is a variable centred approach that is often used to determine cut scores; it is a norming approach where assessment developers collect a large and representative

sample to determine cuts. It does so by graphing the sensitivity and specificity of a scale at multiple thresholds, thus allowing the researcher to choose cut scores based on desired levels of specificity and sensitivity. The Latent Profile Analysis is a person centred approach that analyses the responses of participants and groups them in clusters determined by commonalities in item responses. This study used a Latent Profile Analysis to determine classification standards.

Theoretical Framework

The theoretical framework underpinning this study was the interpretation and use argument posited by Kane (IUA; 2013). This theory defines interpretation and use as "the sequence or network of inferences and assumptions involved in getting from a test taker's observed test performances to the conclusions and decisions based on these performances" (Kane, 2013). IUA refers to assessment scores and the specific decisions that can be made based on these scores. In this argument, interpretations, and uses are only valid when the assumptions underlying them are either feasible or supported by evidence. This study was most concerned with the second criterion, evidential support of underlying assumptions. More specifically, it focused on the score use inferences for the MTAS. According to Kane (2013), score-based decisions typically follow a sequence of steps. The performance in the assessment leads to associations with specific characteristics, which then leads to decisions based on the characteristics identified by the assessment. In the case of the MTAS, this sequence of steps would involve the assessment scores leading to a determination of high or low test anxiety, which would lead to the decision to provide services and testing accommodations.

Two major steps must be completed to validate an assessment's score use. These steps include examining the assessment for measurement invariance and validating the cut scores or classification standards of the assessment. Examining the tool for measurement invariance

supports the assumption that the same construct is measured across groups. This then provides evidence that the MTAS tool can be used with these populations and that the score is representative of the test anxiety construct. The identification of classification standards using a latent profile analysis will not only increase the usability of the MTAS tool in schools but will also provide evidence to support their use in the decision-making process. The combination of the cut scores of 58 and 60 produced by von der Embse et al. (2020) and the classifications created in this study provided the evidential support needed to validate the score use inferences of the MTAS.

Significance of the Study

This study's purpose was to create response profiles for the MTAS. Before those profiles were created the study examined the MTAS for measurement invariance. Measurement invariance was examined first to support the use of the MTAS across gender, age and socio-economic status. However, the focus was placed on the creation of response profiles because of its impact on usability. Many test anxiety scales do not have classification thresholds that indicate which scores are considered concerning versus not concerning. By creating these classifications or thresholds the study will increase the usability of the MTAS. Researchers and practitioners will be able to easily classify students as 'at risk' or 'not at risk' in test anxiety.

Chapter Two: Literature Review

This review of the literature will give further information on test anxiety and assessments that measure test anxiety. First, it will review the relationship between test anxiety and several outcomes for students. Then it will provide an overview of the evolution of test anxiety research over the last 50 years, followed by a comprehensive definition of the term test anxiety and its elements. This section of the study will also review the importance of measuring test anxiety and provide a critique of several currently available test anxiety scales. This will then be followed by a more in-depth description of the development and validation of the Multidimensional Test Anxiety Scale (MTAS). Then finally, this section will end with an overview of measurement invariance and the processes used to determine score classifications.

Context and Impact of Test Anxiety

Academic anxiety is a comprehensive term for the various types of anxieties that students may experience in the school environment (Cassady, 2010). These academic anxieties may be triggered by different contextual cues in the school environment (Cassady, 2010). Test anxiety falls under the umbrella of academic anxiety because students experience it in the school environment, and evaluative situations set it off. It is defined as the changes in behaviour, emotion, and physiology resulting from an individual's perception of the consequences of a test or exam (Zeidner, 1998). This is a topic that is especially important in today's educational environment, where assessments and tests are used throughout a student's academic life. Students who experience test anxiety may perform poorly in exams that they would otherwise succeed in due to the disruptive nature of test anxiety (Zeidner, 1998).

The primary sample used within this study was from the England and Wales, and thus a description of English and Welsh test-based accountability systems is briefly discussed and compared with the US system. Testing has become an integral part of the education system in England and Wales. In England and Wales, students take standardised assessment at several points throughout their education. These assessments are known as National Curriculum Tests and are given at ages seven, eleven, and sixteen ("The UK's Exam System Explained," 2018). In primary school students take the Early Years Foundation Stage (EYFS) at seven. At the end of primary school, at age eleven, students take reading, science, and math examinations. Then, at age sixteen, or at the end of Year 11, students take the General Certification of Secondary Education (GCSE; "The UK's Exam System Explained," 2018). Lastly, students take the GCE A Levels or BTEC at age 19 to determine entrance into tertiary education. These assessments are intended to gauge the students' academic performance at key stages in their educational career. The stakes associated with these assessments are high for both teachers and students. These assessments are used to evaluate the effectiveness of teacher and school performance (Segool et al. 2014). Additionally, these scores can determine if students attain a high school diploma and can access post-secondary education (Segool et al. 2014), and they are often used to support or make educational decisions, like retention.

Standardised tests are also an integral part of the American education system. These tests help states evaluate students, as well as teachers, schools and school districts. In the United States, due to policies like the Every Student Succeeds Act (ESSA), and its precursor, the No Child Left Behind Act, there standardized tests are an important component of state accountability plans. Most states in America have curriculum standards, which outline the specific skills and knowledge that students are expected to learn at the various grades. States

administer standardized assessments to determine student learning based on these standards. Florida, for instance, administers the Florida State Assessments (FSA) for English Language Arts, Mathematics, Algebra, and Geometry for students in grades three – ten (Florida Department of Education, n.d.). These assessments are considered high stakes in comparison to regular classroom instruction and assignments because the outcomes of these assessments are used to determine retention for the students as well as evaluate school effectiveness (ESSA).

However, there are unintentional consequences to standardised testing, like the negative effects on the teachers' working environment (Youn, 2018) as well as increases in teacher stress which can negatively impact their teaching performance (von der Embse et al., 2015). One major consequence of standardised testing is test anxiety. Children experience increased test anxiety with standardised tests, compared to classroom assessments (Segool et al., 2013). The relationship between test anxiety and negative academic and mental health outcomes is of increased importance due to the entrenchment of standardised testing in education, both in England and the US.

There is a relatively strong extant literature that has indicated a relationship between test anxiety and academic performance. Test anxiety has been associated with poor academic performance since research into the subject began in the 1950s (Sarason & Mandler, 1952). Hembree (1988) conducted a meta-analysis of 562 studies and examined the correlations between test anxiety and several different outcomes. When the authors looked at specific subjects (English, reading, math, and science), there was a negative relationship between academic performance and test anxiety, such that as test anxiety increased academic performance decreased. There was also a relationship reported between test anxiety and the need to achieve. In elementary school, the relationship between test anxiety and the need to achieve was negative;

as test anxiety increased the need to achieve decreased. However, in high school, this trend flipped so there was a positive relationship between the two constructs, as test anxiety increased the need to achieve also increased (Hembree, 1988). Several other associations were examined between test anxiety and self-esteem, well-being, self-acceptance, and self-control. Similar to the academic variables, as test anxiety increased these variables decreased, showing a negative relationship between test anxiety and these variables (Hembree, 1988).

A more recent meta-analysis was conducted in 2018 by von der Embse and colleagues. The authors examined 238 studies on test anxiety. Some of the relationships examined in this study were similar in the Hembree (1988) article and results were largely consistent. The von der Embse (2018) article examined the relationship between test anxiety and test performance in typical classroom exams. This relationship increased in strength between elementary school to middle school and decreased in high school (von der Embse et al., 2018). However, throughout these grades, there was still a negative relationship between test anxiety and exam performance, such that as test anxiety increased exam performance decreased (von der Embse et al., 2018). Predictably, this relationship was also noted between test anxiety and grade point average (GPA), showing that increases in test anxiety were associated with decreases in student GPA (von der Embse et al., 2018). The pattern was also seen with standardised tests, which were used to compare scores against a normative sample of scores, whereby high test anxiety levels were associated with poor performance in the tests (von der Embse et al., 2018). The correlation between test anxiety and standardised tests was also stronger than the others, even compared to the typical classroom examinations. There were also negative relationships between test anxiety and several personal variables, including motivation and coping skills (von der Embse et al., 2018). These negative associations provide evidence that test anxiety is related to lower

motivation and coping skills. Overall, the relationship between test anxiety and academic performance remains negative across the 50 years of research in this construct. The relationship between test anxiety and outcomes goes beyond academic performance. Steinmayr et al. (2016) examined the relationships between subjective well-being, test anxiety, and GPA. There was a negative relationship between the worry component of test anxiety (i.e., thoughts of failure or the consequences of a test, Putwain et al., 2010) and the life satisfaction component of subjective well-being, such that as test anxiety increased life satisfaction decreased (Steinmayr et al., 2016). This negative relationship was also noted between worry and GPA, whereby GPA decreased as worry increased. Beidel and Turner (1988) linked test anxiety with broader anxiety disorders. Their outcomes showed that children who experience test anxiety might exhibit the same behaviours (i.e., worry, cognitive interference, tension or physiological indicators) in other situations that may be evaluative (Beidel & Turner, 1988). Additionally, children with high test anxiety also indicated more concerns than their peers regarding other factors like health and safety, which may be related to a more general anxiety disorder (Beidel & Turner, 1988).

However, these studies have some limitations. The first study was done in Germany, and thus its findings may not be generalizable to an American population. Because the demographics and school experiences of German students may be very different from the demographics and school experiences of American and English students. These differences may make it difficult to assume that outcomes seen in one population are generalisable to another. However, the students in the German sample were on the highest academic track in Germany and were preparing for a major standardised assessment that impacted entrance into tertiary education. So, though there is a cultural difference between this population and American and English populations, the impact of test anxiety on well-being would still be similar. The more major weakness lies with the

second study, which was conducted almost forty years ago, and there have been substantial advancements in the understanding and measurement of the test anxiety construct. Another area of research that needs to be explored further is the relationship of test anxiety on mental health and its correlation with well-being and anxiety.

Evolution of Test Anxiety Theory and Measurement

Based on the theory behind the MTAS this study considers worry, cognitive interference, physiological indicators, and tension to be the domains that contribute to test anxiety. However, several domains or factors were theorised to make up test anxiety throughout the history of test anxiety research, and these influenced the measurements used to assess test anxiety.

Research into test anxiety began in the 1950s (Sarason and Mandler, 1952) and had been building on itself and progressing ever since. One of the first aspects of test anxiety researched was the difference between state and trait test anxiety. Trait test anxiety is a more general type of test anxiety and refers to the tendency to experience anxiety across different testing situations, while state test anxiety refers to the anxiety that is experienced in a specific test situation (Spielberger & Vagg, 1995; Zeidner, 1998). Trait test anxiety is considered stable, while state is variable and situation-specific (Hong, 1998). Trait and state test anxiety have also shown differences in outcomes; for instance, trait test anxiety has a weaker correlation to academic performance in elementary school children than state test anxiety (Hong, 1998). State test anxiety, however, is the focus of MTAS and thus the focus of this section of the literature review.

Overall, the identification of underlying domains is the aspect of state test anxiety that has gained the most attention from researchers. Early theorists hypothesised that test anxiety was caused by emotion and worry, which hindered students from utilising the information stored in

their memories and negatively impacted tests scores (Liebert & Morris, 1967). Thoughts that were irrelevant to the test, i.e. worry over one's performance, would act as distractors and lower academic performance (Wine, 1971). The Test Anxiety Inventory, which was created in 1980 (Spielberg), reflects this bidimensional theory of test anxiety as it measures the domains of worry and emotionality to assess test anxiety. However, theory and research continued to progress and the theory raised by Liebert and Morris (1967) was followed by another in the 1980s that test anxiety was instead related to deficits in knowledge rather than issues with retrieval of knowledge (Culler & Holahan, 1980). This new theory hypothesised that test anxiety and the associated poor academic performance were related to the students' behaviors, specifically their study habits.

The older theories cited above posited that hindrances caused this outcome in retrieving knowledge through worry and test irrelevant thoughts (Wine, 1971) or by deficits in knowledge (Culler & Holahan, 1980). Current theories on test anxiety have built upon this early conceptualisation of test anxiety and its relationship with knowledge and cognition. Theorists have looked further into the reasoning behind test anxiety's relationship with test performance to examine deficiencies in processing (Cassady, 2004a). Students with high test anxiety have issues with processes like attention, working memory, metacognitive skills as well as retrieval of information from long-term memory (Cassady, 2004a). Recent research has gone further and looked into the relationship between test anxiety and behaviours before and after the test or assessment.

This movement towards examining student behaviour before, during and after the test has suggested that students with high test anxiety engage in behaviours that negatively impact their test performance at all three stages (Cassady, 2004b). In the study by Cassady (2004b), students

with high test anxiety also reported low levels of self-efficacy. This low self-efficacy could result in behaviours like avoiding test preparation and setting low goals for themselves, which can have detrimental effects on test performance (Cassady, 2004b). This research highlights the importance and possible benefits of test anxiety assessment. It shows that the benefits of test anxiety assessment go beyond the obvious function of identifying students who are at a disadvantage in evaluative situations. It also identifies students struggling with test preparation, self-esteem, self-efficacy, and cognitive processes. Thus, using test anxiety assessments will identify students in need of support so schools can address this need.

Consequently, test anxiety is a complex concept influenced by multiple factors (Hong, 1998). Current theories and studies have built on and progressed past the bidimensional theories of the past. These theories are aware of this multidimensionality and have begun to incorporate multiple factors into their measurement models. Segool et al. (2014) proposed a cognitive behavioural model of test anxiety that examined several different factors across dimensions as predictors of test anxiety. This included self-efficacy, which was conceptualised as a cognitive perception that acted as a predictor of test anxiety and was influenced by environmental factors like the student's learning experiences and demographics. Gender and academic achievement were also direct predictors, with socio-economic status identified as an indirect predictor of self-efficacy and academic achievement. This expansion to include factors like self-efficacy shows both a shift in thinking and an area within the test anxiety research that needs to be further studied.

Shifts in theory and the development of new models do not happen in a vacuum. As research progresses, these shifts inform the development of measurement tools. This shift in theory towards a multidimensional understanding of test anxiety has also impacted the produced

measurement tools. Lowe et al. (2008) developed a multidimensional model of test anxiety that included biological factors, like intelligence and academic ability, psychological factors, like social-emotional functioning, and social systems, like school communities. This test anxiety scale examined four domains of test anxiety, including cognitive obstruction, physiological hyperarousal, social humiliation, and worry. The shifts reviewed in this section show the progress that test anxiety research has made since its conception in the 1950s, going from bi-dimensional to multidimensional.

Operational Definitions

There have been several shifts in test anxiety research over the last 50 years as the focus shifts from an unidimensional model to a multidimensional one. As these shifts occur, there has been an accompanying development of test anxiety scales that incorporate these theories into their design. The MTAS is one such scale. It was developed in 2020 (Putwain et al., 2020) and reflects the current theories on test anxiety. The factors underlying the items of this scale are worry, cognitive interference, physiological indicators, and tension.

Worry. Worry refers to thoughts that are focused on the consequences of the test or assessment (Putwain, Connors, & Symes, 2010). When students engage in this aspect of test anxiety, their attention is on the consequences of the test instead of taking the test (Sarason, 1986). There are some theories on what causes this worry to occur. Sarason and Sarason (1990) posited that worry occurs in evaluative situations due to the student determining that their readiness for the assessment is inadequate and likely to do poorly. Worry in itself can be a helpful emotion as it can help prepare individuals for possible outcomes or act as a motivator to work hard towards a goal (Zeidner, 1998). However, in the case of test anxiety, worry instead serves to increase students' stress past their coping abilities (Zeidner, 1998).

Cognitive Interference. Cognitive interference is a concept that is similar to worry because it also refers to thoughts that draw attention away from the test-taking task. However, unlike worry, these thoughts can be about anything (Zeidner, 1998). According to Deffenbacher (1978), students who experience test anxiety would spend 60% of their time attending to test-relevant tasks and 40% of their time attending to intrusive thoughts that were irrelevant to the test they were taking. Students may find it difficult to dismiss these thoughts and waste valuable time and effort trying to refocus themselves (Zeidner, 1998).

Physiological Indicators. Physiological indicators are what they sound like, the physical reactions to test anxiety. The most common and obvious physical response associated with test anxiety is changes in autonomic arousal (Zeidner, 1998). Students who experience test anxiety may have symptoms like increased heart and breathing rates, trembling in hands, dry mouth, or increased sweating (Galassi et al., 1981). These responses to being in an evaluative situation are derived from the flight or fight response. Students who experience fight or flight during exams are at a disadvantage because students have to sit for hours while experiencing these disrupted bodily functions (Zeidner, 1998).

Tension. Tension is sometimes referred to as emotionality in the test anxiety literature. It is seen as the connection between physiological indicators and cognitive processes (Zeidner, 1998). Students with test anxiety who experience the physical changes are more likely to perceive these responses as negative compared to students who do not experience test anxiety, who may perceive them as motivation for increased effort (Zeidner, 1998).

An important step in both researching test anxiety and supporting this student is accurately identifying students experiencing test anxiety. Thus, the development of test anxiety scales is an essential aspect of test anxiety research. The MTAS, with its current theoretical

framework, is one tool that can be utilised in research and practice. However, before it can be used in either of these areas, its psychometric properties must be examined. This study's main goal is to examine the measurement invariance of the MTAS as well as develop classifications to improve its usability, thus adding to the research on test anxiety scales.

Measuring Test Anxiety

It is important to identify students who are experiencing test anxiety because of the negative relationship between test anxiety and academic performance (von der Embse et al., 2018). Teachers are often the people held responsible for referring students that they suspect may have social, emotional or behavioural issues (Green, et al., 2017). Unfortunately, these same teachers feel under-equipped to identify these students (Splett et al., 2019) and are not confident in their ability to do so (Askill-Williams & Lawson, 2013). For example, teachers often under-identify students with internalising behaviours (Dowdy, Doane, Eklund, & Dever, 2013). Since test anxiety is an internalising issue and is usually experienced by students before, during, or after a test, this reduces the likelihood of educators accurately identifying students who experience high levels of test anxiety.

Assessments of test anxiety address this issue by providing a tool that can be utilised to identify these students rather than relying solely on teacher referrals. Additionally, while test anxiety does fall within the umbrella of academic anxieties it is specifically triggered by evaluative situations (Cassady, 2010) and thus needs measurement tools that take this into consideration. The literature on test anxiety shows that scales for general anxiety and test anxiety are assessing two different constructs even if the two seem intuitively similar (Alpert & Haber, 1960). The specificity of the items in a test anxiety scale allows for a higher sensitivity for identifying students with test anxiety, while the lack of focus on evaluative situations in general

anxiety scales makes it a poor measure of test anxiety (Alpert & Haber, 1960). Thus, general anxiety assessments or tools that help identify students with general anxiety disorders cannot be used to also identify students with test anxiety as they are separate constructs. Test anxiety has several negative relationships with academic performance, a major reason to support the students experiencing it. However, the first step to providing that support is identifying students with test anxiety.

There are several assessments that measure test anxiety, including the most widely used assessment, the Test Anxiety Index (TAI; Spielberg, 1980). Other assessments have been created since the TAI, including the Friedben Test Anxiety Scale (FTAS; Friedman & Bendas-Jacob, 1997), the Children's Test Anxiety Scale (Wren & Benson, 2004), and the Test Anxiety Scale for Elementary Students (Lowe et al., 2011). These assessments were developed alongside the evolving theories on test anxiety and are varied in their focus (von der Embse et al., 2018). Some of the assessments measure test anxiety through a two-factor model, cognition and emotionality, like the TAI (Spielberg, 1980). In contrast, others have included social factors to incorporate the multidimensionality of test anxiety, like the FTAS (Friedman and Bendas-Jacob, 1997).

Critique of Test Anxiety Measures

There are several test anxiety scales that are commonly used in research to study this phenomenon. In this section of the literature review, four widely used assessments will be critically examined, including the Test Anxiety Inventory (Spielberger, 1980), the Children's Test Anxiety Scale (Wren & Benson, 2004), the Test Anxiety Scale for Elementary Students (Lowe et al., 2011) and the Friedben Test Anxiety Scale (FTAS; Friedman & Bendas-Jacob, 1997). These four test anxiety scales were chosen because they display limitations common among the test anxiety scales currently available for use.

The Test Anxiety Inventory (Spielberg, 1980) is the most frequently used test anxiety scale. This scale consists of two domains, worry, and emotionality, as well as a total score which looks at a combination of these two domains. It has reportedly good reliability and validity, but there are certain limitations to this scale that warrant further examination, mainly the fact that the scale was created about 40 years ago. The age of this scale is related to two main issues, the theories that the scale is based on are outdated, and the population that it was normed on is not representative of current populations. As was previously mentioned, test anxiety theory has developed and progressed over the years. The current theories and models view it as multidimensional, and this can be seen in the more current test anxiety scales like the MTAS, which includes multiple domains. This is a different theory from earlier years which proposed that test anxiety consists of worry and emotion only as is reflected in the domains of the TAI. The TAI is a product of its time and reflects this outdated mindset. Given the progress in test anxiety research, the Test Anxiety Inventory is now being questioned. Does it truly capture all the aspects of test anxiety? Researchers have called for assessments based on evidence from the 21st century and address current needs (Rajagopalan & Gordon, 2016).

The TAI is not the only test anxiety scale that is not consistent with recent theoretical advancements (e.g. Children's Test Anxiety Scale (CTAS; Wren & Benson, 2004). Current research has indicated that test anxiety is impacted by the student's perception of the test or assessment they are about to take (Segool et al., 2013). Students experience higher levels of test anxiety for high stakes tests, e.g. the A-Levels exam taken in the sixth form, compared to classroom assessments (Segool et al., 2013). However, the CTAS does not take the importance of the students' perceptions of test importance into consideration with the development of the scale. This calls the validity of this assessment into question as well since this is an important

aspect of test anxiety. The Test Anxiety Scale for Elementary Students (TAS-E; Lowe et al., 2011) is a more current test anxiety scale that, unlike the TAI, is multidimensional. The TAS-E reflects the current research on test anxiety in its creation through its inclusion of the multiple dimensions of test anxiety. This test anxiety scale measures four domains "physiological hyperarousal, social concerns, task-irrelevant behavior, and worry" (Lowe et al., 2011). The TAS-E thus incorporates the multidimensionality of test anxiety in its development. However, similar to the CTAS, it does not take student perceptions into consideration in its development. In addition to this weakness, there is also little research or support for measurement invariance across groups. The TAS-E has been validated for use in multiple cultures (Lowe et al., 2011), however, there is no known evidence for measurement invariance across various demographic groups.

Lastly, the Friedben Test Anxiety Scale (FTAS; Friedman & Bendas-Jacob, 1997) has evidence that supports its use for measuring test anxiety but does not provide cut scores or classifications. The FTAS is another multidimensional test scale, it measures social derogation which is associated with negative social feedback, cognitive obstruction and tenseness (Friedman & Bendas-Jacob, 1997). However, the authors of this test anxiety scale did not provide cut scores to differentiate between students with low, moderate or high test anxiety. This impacts usability in schools as the scores derived from the scale cannot be objectively differentiated between a student in need of support versus one who's test anxiety is below an established threshold.

Though there are many different test anxiety assessments there are two main limitations of these assessments. First, most of the assessments were created for research purposes. In the case of the FTAS, the authors of the scale did not provide cut scores or classifications for applied use. This is also an issue with the TAS-E, which has not been tested for the various measurement

invariances that would make it acceptable for use across multiple populations. Second, some test anxiety scales are not reflective of modern advancements in theoretical conceptualizations of test anxiety. This was apparent with the TAI, which is 40 years old and thus does not incorporate current test anxiety research, e.g. the multidimensionality of test anxiety, into its development. This was also seen in the CTAS, which does not take student perceptions of tests into consideration, this is another recent finding that makes the CTAS outdated compared to other more recent test anxiety assessments.

Development and Initial Validation of the Multidimensional Test Anxiety Scale

The Multidimensional Test Anxiety Scale (MTAS) was created to address the aforementioned limitations of several of the most frequently used test anxiety scales. To do this they followed the principles of content validation posited by Haynes et al. (1995). The first step was choosing the components to be included in the test anxiety scale. The guiding principle in the selection of these components was the focus on factors representative of evaluative threat and thus could not be related to elements that precede or follow the evaluative situation. This was in accordance with Spielberger and Vagg's (1995) definition of test anxiety. This guiding principle led to the exclusion of several theorised components of test anxiety, including the social components of test anxiety, like worry over judgement by family and peers (Putwain, 2009), and low self-esteem on academic performance (Lowe et al., 2008). The components that were included following this principle were the cognitive aspects of worrying, thoughts around failure, and cognitive interference (Lowe et al., 2008; Segool et al., 2014; Spielberger & Vagg, 1995; Zeidner & Matthews, 2005) and the autonomic aspects of physiological indicators and tension. The second step to validate the content of the MTAS tool was collecting expert feedback. To do this the authors began by consulting a panel of test anxiety experts to provide feedback on the

item pool for their assessment. This item pool consisted of items that represented the four chosen domains of test anxiety- cognitive interference, worry, physiological indicators and tension. This resulted in a sixteen item assessment, four items for each of the domains, with responses across a five-point Likert scale (strongly disagree to strongly agree, See Appendix). The authors examined the factor structure, measurement invariance and created cut scores in a recent study using 918 secondary school students (von der Embse, et al., 2021). The MTAS had a good model fit ($\chi^2 = 450.77$ (100); RMSEA= .062; SRMR= .050; CFI= .958) and had measurement invariance across gender (Scalar Invariance: $\chi^2=585.06$ (227); RMSEA=.059; SRMR=.071; CFI=.943; TLI=.940).

The interpretation and use argument given by Kane (2013b) states that there are several components that must be included when validating measurement tools. This includes an examination of factor structures across groups to ensure that it remains consistent and that the measurement is assessing the same construct. Additionally, it includes the development and validation of cut scores across different groups to aid in the interpretation of results. Analysed through this measurement framework, this scale also has its own weaknesses that need to be addressed. Though measurement invariance was examined, this was only done across gender (male and female). However, if this scale is used in schools and research, which is the goal of its creation, then we must ensure that the same construct is being measured across multiple groups beyond gender. To address this limitation, the current study will examine measurement invariance across gender, age and socio-economic status. The second limitation is the need for more evidence in support of the cut scores developed by von der Embse et al. (2021). The current study utilised a Latent Profile Analysis to examine clusters and determine if these clusters support the cut scores previously produced.

Measurement Invariance

One important consideration that must be examined when a self-report scale is developed is whether or not that scale has measurement invariance. There are two types of measurement invariance, multi-group measurement invariance which examines the stability of the underlying construct across different groups and longitudinal measurement invariance, which looks at the stability of the construct across time. This study focused on multi-group measurement invariance mainly due to the data available for analysis, because the test anxiety data was only collected at one time longitudinal measurement invariance cannot be examined in this study. Measurement invariance is a statistical property that indicates whether or not the underlying construct being measured is the same across various groups. In the case of the MTAS, the underlying constructs are the four domains of test anxiety, worry, cognitive interference, physiological indicators, and tension.

This statistical property is important because the absence of multi-group measurement invariance can lead to errors. Comparisons may be inaccurate if the scale exhibits variance across groups. For instance, if a test anxiety scale did not have measurement invariance but was used to examine the difference in test anxiety levels between males and females, the results of that comparison would be inaccurate and meaningless. The underlying construct being measured across those two groups would be different, meaning that conceptually the two groups were measured on two different scales. Thus, the results of the scale could not be compared across groups. Decisions and comparisons should not be made with scales that do not have measurement invariance. For these reasons, this study examined the measurement invariance of the MTAS across gender (male and female) as well as socio-economic status across the four domains. This will inform the usability of the MTAS for research as well as practice and add to

the pool of research on the MTAS tool in particular and test anxiety in general. Lastly, this study highlighted the need for scales to be thoroughly vetted for measurement invariance before they are used for decision-making or research.

Examination of Response Profiles

One of the factors that affects the usability of an assessment is the inclusion of guidelines for the interpretation of results. This is where cut scores or classification standards come into play. They outline the ranges of scores that fall within different classifications. With regards to test anxiety, these classifications are typically ‘at risk’ versus ‘not at risk.’ Students who fall within a specific score range can then be identified as one of these classifications for each of the four domains, cognitive interference, worry, physiological indicators, and tension, as well as the overall scale, test anxiety.

There are various methods used to determine cut scores or standards for assessments. The two main types that these methods fall under are person centred analyses or variable centred analyses. The variable approach to analyses focuses on the variables themselves, as the name suggests. In this analysis approach, the focus is on associations among variables (Laursen & Hoff, 2006). The person centred approach instead identifies groups of respondents who responded in similar ways (Laursen & Hoff, 2006). This analysis identifies groups of individuals or clusters who share common attributes (Laursen & Hoff, 2006). One of the most commonly used variable centred analyses, and the one used to determine the initial cut scores for the MTAS, is the Receiver Operating Characteristics Curve (ROC) Area Under the Curve (AUC).

The ROC is a probability curve that plots the sensitivity versus specificity at various classification thresholds (Narkhede, 2018). Sensitivity is a statistical term that refers to the scale's ability to correctly identify all at-risk students (Parikh et al., 2008). For example, if 100

children with social issues are assessed by a screener for social-emotional risk and this screener only identifies 70 students, then the sensitivity of the screener is 70%. In this example, the screener identified 70 true positives but 30 false negatives. Specificity is a statistical term that refers to the scale's ability to correctly identify students who are not at risk. For example, if 100 children without social problems are assessed by the screener and 85 are found to be not at risk then the screener has 85% specificity. In this example, the screener identified 85 true negatives but 15 false positives. As sensitivity increases specificity decreases and vice versa (Parikh et al., 2008). The AUC indicates how well the model is at distinguishing between binary classifications, in the case of the MTAS 'at risk' vs. 'not at risk,' at different classification thresholds (Narkhede, 2018).

When using ROC curve analysis, researchers typically compare the newly developed assessment to a "gold standard" assessment examining the AUC at different classification thresholds. In a previous study, cut scores of 58 and 60 in MTAS Total score were created for the MTAS compared to a panic and an anxiety scale using ROC (von der Embse, et al., 2021). There are several benefits to using the ROC AUC analysis to create cut scores. This measurement system provides graphical data on the diagnostic accuracy of the scale, it also allows the researcher to easily compare scales against each other. But the main advantage of ROC is the fact that it reports the changes in specificity and sensitivity across different cut of scores. This is useful when creating cut scores because different thresholds can be chosen based on the level of sensitivity and specificity desired by the researcher. However, that freedom to determine thresholds and cut scores is also a weakness. The level of specificity and sensitivity is dependent on the researcher and is thus subjective. Additionally, as a variable centred analysis it assumes that the sample is homogenous and focuses on the variables.

This study examined classification thresholds through a different method, Latent Profile Analysis. Latent Profile Analysis (LPA) is a person centred analysis based on the principle that responses to items in the assessment form distinct and mutually exclusive subgroups called latent profiles. This method can be used to determine both the number of classification groups formed through the assessment as well as the standards for inclusion in each of these classifications. Thus it will expand on the process already started in von der Embse et al. (2021) to increase the usability of the MTAS by providing classification standards. By using a person centred analysis instead of the variable centred analysis this study compared the outcomes of this analysis to the outcomes of the variable centred ROC AUC approach. While both analyses are accepted and valid using and comparing the two will increase the accuracy of the classifications.

Purpose of the Present Investigation

There was one main limitation in test anxiety research that this study addressed. This limitation was the absence of classification standards or cut of scores for currently available test anxiety scales. The MTAS already had initial cut scores based on an earlier study (von der Embse et al., 2021), but this study used a different approach to create classification thresholds which were then be compared to the initial cut scores. This built evidence for classifications in the MTAS as well as added to the body of research on classification development. With the addition of this study to the research on test anxiety scales, hopefully future scales for test anxiety and other constructs will engage in this type of analysis as well to increase the usability of measurement tools and decrease the research-practice gap.

Chapter Three: Methods

This study analysed pre-existing data from secondary school students in England. This data was collected by Dr. David Putwain, who is a professor at Liverpool John Moores University in England, in partnership with Dr. Nathaniel von der Embse, a professor in the school psychology program at the University of South Florida. The original data was collected by Dr. Putwain to pilot and assess the psychometric properties of the new test anxiety measure during the 2019 – 2020 academic years. This study examined the measurement invariance of the MTAS looking within different groups (gender, socio-economic status and grades). Additionally, this study used a Latent Profile Analysis to identify unique profiles of respondents.

Participants

This study used a pre-existing dataset. This dataset consisted of 918 participants, 217 self-identified as male and 694 self-identified as female while 7 declined to disclose their gender. All participants were secondary school students from eight schools in England and Wales, two of which were girls' schools which explains the higher number of females in the sample. The mean age of this sample was 15.76 years old. The grades in this sample were between Year 10 – Year 13. There were 100 participants in Year 10, 481 participants in Year 11, 158 participants in Year 12 and 179 participants in Year 13. The racial/ethnic breakdown of this sample was 3% Asian, 5% Black, 87% White and 2% multiracial. Additionally, 15% of the sample were eligible for free school meals.

Variables

Test Anxiety. In this study test anxiety referred to the cognitive and physical aspects of testing anxiety, specifically worry, cognitive interference, tension and physiological indicators. This variable was measured using the MTAS assessment which provides a total test anxiety score as well as subscale scores for each of the four domains.

Gender. In this study gender referred to the gender that the participants choose on the questionnaire between female and male. The participants in the sample were 24% male and 76% female.

Grades. In this study, age referred to the different grades that the participants chose on the questionnaire. The breakdown of grades in this sample were Year 10 = 100 participants, Year 11 = 481 participants, Year 12 = 158 participants and Year 13 = 179 participants.

Socio-economic Status. In this study socio-economic status referred to whether or not the student was eligible for free school meals. Those who were eligible for this service were categorised in the low socio-economic status group, and those who were not eligible were categorised in the high socio-economic status group. There were 137 participants in this sample who were eligible for free and reduced lunch making up 15% of the sample.

Measures

MTAS. The Multidimensional Test Anxiety Scale (MTAS: Putwain & von der Embse, 2020) was developed to assess test anxiety. The assessment was developed through several steps. The developers first reviewed the literature around test anxiety to determine which constructs should be included in their assessments and which should be excluded. They then surveyed a panel of experts in test anxiety on the relevance of the different items within the four constructs of cognitive interference, worry, physiological indicators and tension. The item pool was piloted

with secondary school students in England. Initial data were used to complete exploratory and confirmatory factor analyses as well as examined for measurement invariance. The end result was the 16-item assessment used in this study, which uses a Likert scale from 1 - “Strongly Disagree” to 5 “Strongly Agree” and has 4 items for each of the four constructs listed above. The authors have examined the MTAS for internal consistency, factorial validity, predictive validity, measurement invariance and for the creation of cut scores in previous studies (Putwain et al., 2020; von der Embse, et al., 2020). The MTAS had good internal consistency ($\omega_s = .85 - .91$; Putwain et al., 2020). It also had good factorial validity with items loading onto predicted factors ($\lambda_s = .46$ to $.92$; Putwain et al., 2020). It also showed predictive validity through its positive relationship with mental health ($r_s = .13$ to $.46$) and negative relationships with academic performance ($r_s = .01$ to $.41$) and wellbeing ($r_s = .01$ to $.41$; Putwain et al., 2020). It also had a good model fit ($\chi^2 = 450.77$ (100); RMSEA= .062; SRMR= .050; CFI= .958) and had measurement invariance across gender (Scalar Invariance: $\chi^2=585.06$ (227); RMSEA=.059; SRMR=.071; CFI=.943; TLI=.940).

The current study examined measurement invariance as one of its preliminary steps before conducting a latent profile analysis. The main purpose of this study was to create classifications based on a latent profile analysis which also provided support for the use of the MTAS tool.

Data Collection Procedure

The data was collected from eight secondary schools. Once IRB approval was obtained , the eight schools were enrolled into the study and consent was solicited from the principal, students and parents of students who were under 18 and considered minors. The data was then collected during a ‘free’ period in the students’ timetables when instruction was not being

provided. The teachers who administered the MTAS followed a script whereby they informed the students that they were not being tested, but to fill out the questionnaire honestly.

Data Analysis

Reliability. The internal consistency of the four factors of the scale was examined to show the validity of the scale. Cronbach's alpha and McDonald's omega were calculated for the different factors as well as the overall scale. With Cronbach's alpha and McDonald's omega values between .7 and .9 indicate acceptable to exceptional internal consistency (Tavakol & Dennick, 2011).

Measurement Invariance. For this data analysis a Confirmatory Factor Analysis (CFA) was run. A CFA was chosen due to the breadth of information gained from this analysis as it allows one to examine several components including factor loadings. However, there are several criteria that need to be met before an assessment can be examined for measurement invariance through a CFA (Pendergast et al., 2017). The first is the sample size which should be at least 200 participants because CFAs require samples of 200 or more to give accurate results (MacCallum, et al., 1999). This study meets this criteria with its sample size of 851 students. Additionally, a CFA evaluates existing factors and the items that are related to these different factors. The MTAS has four domains with four items related to each, which was determined by the Exploratory Factor Analysis run in a previous study by the scale authors (Putwain, et al., 2020) and reported above in the description of the MTAS.

The model for the MTAS is a second-order factor model. Both the first order and second order must be examined to determine measurement invariance. The examination of measurement invariance of a second-order factor model followed the steps recommended by Chen et al. (2005). Configural invariance examined which items load on to which factors and invariance is

accepted if the same items load on to the same factors across groups. The next step was to examine the factor loadings of the items to the factors across the groups. The model fit indices of this model (Model 2) was compared to the previous model (Model 1), if the two models were not statistically different, then the scale was determined to have metric invariance. This was followed by examining the second-order factor loadings which assessed the factor loadings of the first-order factors (worry, cognitive interference, tension, and physiological indicators) on to the second-order factor (total test anxiety) to see if they were the same across groups. The model fit indices of this model (Model 3) was also compared to the previous model (Model 2) to determine if there were significant differences. If the scale has metric invariance for first and second-order factors then the analysis moved on to the next step.

The item intercepts of the first-order factors were examined to see if they were the same across the groups. The fit indices of this model (Model 4) was then compared to the fit indices of the previous model (Model 3), if the two models were not statistically different, then the analysis moved on to the next step. The intercepts of the first-order factors were examined to see if they are the same across the groups, and this model (Model 5) was then compared to the previous model (Model 4). If there was no significant difference in model fit indices between the two models, then the analysis moved on to the next step. In this step, the residual variance of the items was examined to see if they are the same across groups. As with all of the other models, this model (Model 6) was then compared to the previous model (Model 5), and if there were no significant differences, then the next step was completed. The last step was an examination of the disturbances of the first-order factors to determine if they are the same across groups. This model (Model 7) was also compared to the previous model (Model 6), if there were no significant differences then strict invariance was accepted.

The general CFA was evaluated with a set of model fit indices which include: Comparative Fit Index (CFI), which compared the observed model to a null model. High CFIs mean that the observed model was a better fit than the null model. Root Mean Square Error of Approximation (RMSEA), which examined “the difference between the hypothesised model and the population covariance model” (Hu & Bentler, 2009). Low RMSEAs mean the hypothesised model fit well (Hu & Bentler, 2009). And, lastly, the Standardized Root Mean Square Residual (SRMR), which also examined the difference between the hypothesised model and the data in the sample. Like RMSEA a low SRMR is needed to support a good model fit. When comparing models using a CFA, two methods that are typically used are the Likelihood Ratio Test (LRT) and differences in CFI and RSMEA (Kim et al., 2017). The LRT examined two different models and compared the fit between these two models by comparing the log-likelihoods of the models (Woolf, 1957). If there is a statistically significant difference between the two models, then the model with more variables or parameters was considered the better fit for the data than the other model (Woolf, 1957). However, this test has a major limitation. It is influenced by sample size, whereby large samples are likely to get a significant Chi-squared even if the scale has measurement invariance (Cheung & Rensvold, 2002). This study also examined changes in CFI and RSMEA across groups. Using this test the scale had measurement invariance if the change in CFI across the groups is less than .01 (Cheung & Rensvold, 2002) and a change of RMSEA of less than .015 (Chen, 2007).

The cut-off scores advised by Hu and Bentler (2009) were used to determine model fit. The cut-off score for CFI is .95 and since high scores in this index indicate better model fit, scores above this also indicated good model fit. For the RMSEA the cut-off score is .06 and since lower scores in this criterion indicate good model fit, scores lower than this indicated good

model fit. Lastly, for the SRMR, the cut-off score is .08 and, like RMSEA, lower scores indicate better fit, so scores lower than this indicated good model fit. Additionally, the study examined differences in these fit indices across models. The chi-square difference test determined if the differences between the constrained model and the more relaxed model were statistically significant. A *p*-value greater than .05 indicated that the difference between the models was not statistically significant and that the more constrained model was a good fit. However, the chi-square difference test is not always accurate and may reject models with minor violations, especially with large sample sizes >300 (Chen, 2007). Because of this, the chi-square difference test should be accompanied by other criteria like the differences in fit indices (Chen, 2007). The criteria for the differences in fit indices were a .01 change in CFI, .015 change in RMSEA, .030 change in SRMR for metric invariance, and .015 change in SRMR for scalar and residual invariance (Chen, 2007).

Through these various steps, the MTAS scale was examined for measurement invariance across gender, specifically female vs. male, socio-economic status, specifically eligible for free school meal vs. ineligible, and grades, specifically Years 10 through 13.

Latent Profile Analysis. LPA was used to identify patterns of risk among the four factors, cognitive interference (CI), worry (W), physiological indicators (PI) and tension (T), and overall test anxiety (OTA). The interpretation and use of the MTAS is directly impacted by this analysis. According to the IUA theory (Kane, 2013), which is the theoretical framework for this study, assessments are more useful in practical settings when observed test scores are associated with some descriptive classification. These classifications help administrators to make informed decisions with the test scores regardless of their personal knowledge on the area being tested.

The purpose of the LPA is to support the creation of classifications in the MTAS to improve its usability in practice.

According to previous research into test anxiety classification and cut scores there would be three profile model of low, moderate, and high as the classifications for test anxiety scores (Thomas et al., 2017; Segool et al., 2013; von der Embse et al., 2014). However, to ensure that the proper number of profiles were selected for the model a two profile model was run first, followed by a three profile, a four profile and a five profile model. The analysis stopped at a five profile model because the main purpose of the LPA was to increase the usability of the MTAS by creating classifications. Five profiles would be usable in a school setting, however profiles greater than five may become too complicated and nuanced for a lay person to interpret.

There were several different criteria that were utilised to compare these models and choose the most accurate one. These included criteria like Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC) and sample size adjusted BIC, Lo-Mendall-Rubin likelihood ratio test (LMR-LRT), adjusted LMR-LRT, bootstrap LRT, entropy, profile proportions, and qualitative distinctiveness of profiles (Kim et al., 2016; Nylund et al., 2007; Tein et al. 2013; von der Embse et al., 2021). For the information criteria (IC) the model with the smallest IC is typically chosen. With regards to the LMR-LRT, adjusted LMR-LRT and bootstrap LRT models were compared to the next model with one less profile. If there was a statistical difference between the compared models then the model with the additional profile was considered necessary. Entropy examined the separation between the profile profiles, scores closer to 1 are desired as this indicated that the profiles were distinct and the assignment of individuals to these profiles was accurate.

Chapter Four: Results

Missing Data Analysis and Treatment

There was no missing data in the dataset, and no adjustments or modifications were necessary.

Descriptive Statistics

Tables 1, 2, and 3 display the descriptive statistics and correlation matrices for the items and scales of the MTAS. Based on the skewness and kurtosis of the items and subscales of the MTAS the data was normally distributed. According to (Chou & Bentler, 1995), absolute values of skew greater than three are considered extreme. For kurtosis, absolute values greater than ten are considered problematic and absolute values over 20 are deemed extreme (Kline, 2005). There were no variables with skewness greater than three or any with kurtosis greater than 10, which means that the data is approximately normally distributed.

Table 1
Descriptive Statistics for MTAS Items

Variable	Mean	Variance	Skewness	Kurtosis
MTAS1	3.919	0.192	-1.005	-0.194
MTAS2	3.581	0.956	-0.655	0.695
MTAS3	3.961	0.962	-1.157	-0.225
MTAS4	2.650	0.935	0.291	1.086
MTAS5	3.870	1.578	-1.037	-1.068
MTAS6	3.597	0.936	-0.716	0.942
MTAS7	3.695	0.994	-0.805	-0.095
MTAS8	3.180	1.175	-0.168	-0.107
MTAS9	3.797	1.476	-0.929	-0.991
MTAS10	3.574	1.098	-0.595	0.381
MTAS11	3.582	1.027	-0.682	-0.324
MTAS12	2.803	1.350	0.179	-0.439
MTAS13	3.552	1.640	-0.606	-1.096
MTAS14	3.317	1.382	-0.231	-0.634
MTAS15	3.905	1.667	-1.152	-1.172
MTAS16	2.463	0.992	0.478	1.070
Worry	15.139	11.712	-0.845	0.195

Table 1 (Continued)

Variable	Mean	Variance	Skewness	Kurtosis
Tension	15.143	11.722	-0.956	-0.025
Cog Inter	14.069	11.7	-0.535	0.586
Phys Ind	11.096	13.068	0.114	0.636
TA Total	55.447	1.336	-0.534	-0.581

Table 3 displays the correlations between items. Correlation coefficients measure the relationship between two variables (Ratner, 2009). Coefficients between 0 and $\pm .3$ indicate a weak relationship, coefficients between $\pm .3$ and $\pm .7$ indicate a moderate relationship, and coefficients between $\pm .7$ and ± 1 indicate a strong relationship (Ratner, 2009). The correlations between the items ranged from 0.3 to 0.7, which indicates that the relationships between the items are weak to moderate. Moderate correlations were found between items measuring the same subscale, for instance the relationship between items two and six which measures cognitive interference was 0.612. In the MTAS, there are four items to measure each subscale (worry, cognitive interference, and physiological indicators), and these items all had correlation coefficients between $+ .7$ and $+ .9$ with their subscale, which indicates a strong relationship between the items and the subscales.

Table 2 displays the correlations between the subscales. The correlations between the subscales ranged from 0.491 – 0.763 which indicates that the relationship between the subscales are moderate. The strongest correlations were seen between worry and tension ($r=0.763$) and tension and physiological indicators ($r=0.732$). The strength of these correlations may indicate that tension and worry, and tension and physiological indicators are measuring similar constructs and may need to be re-examined.

Table 2

Correlation Matrix for MTAS Scales

	W	CI	T	PI
W	1.000			
CI	0.571	1.000		
T	0.763	0.436	1.000	
PI	0.648	0.491	0.732	1.000

Table 3

Correlation Matrix for MTAS Items

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1.000															
2	0.387	1.000														
3	0.577	0.277	1.000													
4	0.434	0.325	0.453	1.000												
5	0.559	0.359	0.515	0.448	1.000											
6	0.413	0.612	0.380	0.361	0.436	1.000										
7	0.516	0.225	0.579	0.500	0.498	0.292	1.000									
8	0.501	0.294	0.551	0.620	0.493	0.347	0.618	1.000								
9	0.672	0.354	0.574	0.446	0.618	0.428	0.512	0.512	1.000							
10	0.404	0.621	0.343	0.340	0.410	0.752	0.286	0.330	0.423	1.000						
11	0.574	0.311	0.624	0.557	0.532	0.372	0.668	0.675	0.608	0.370	1.000					
12	0.398	0.281	0.420	0.546	0.427	0.331	0.488	0.550	0.459	0.295	0.555	1.000				
13	0.491	0.306	0.469	0.382	0.502	0.341	0.484	0.475	0.555	0.343	0.544	0.428	1.000			
14	0.335	0.384	0.263	0.366	0.355	0.412	0.271	0.334	0.386	0.415	0.327	0.312	0.358	1.000		
15	0.606	0.229	0.643	0.469	0.542	0.322	0.685	0.593	0.587	0.286	0.691	0.447	0.477	0.242	1.000	
16	0.379	0.249	0.392	0.547	0.437	0.326	0.512	0.582	0.413	0.319	0.546	0.591	0.391	0.342	0.458	1.000
W	0.818	0.426	0.648	0.519	0.805	0.490	0.612	0.603	0.864	0.479	0.688	0.522	0.796	0.437	0.671	0.493
CI	0.477	0.793	0.391	0.437	0.485	0.845	0.337	0.409	0.497	0.849	0.430	0.382	0.432	0.730	0.335	0.390
T	0.661	0.304	0.819	0.579	0.608	0.398	0.858	0.713	0.664	0.375	0.879	0.560	0.577	0.323	0.875	0.560
PI	0.519	0.349	0.551	0.826	0.547	0.414	0.642	0.833	0.555	0.389	0.708	0.820	0.509	0.410	0.596	0.817
TA	0.731	0.548	0.717	0.714	0.723	0.630	0.733	0.772	0.763	0.613	0.809	0.692	0.682	0.559	0.739	0.685

Confirmatory Factor Analysis

A CFA was run before testing the sample for measurement invariance, to test the second-order model proposed by the MTAS authors. This examined the model's fit proposed by the MTAS authors to the sample data. Confirming the model fit was an important first step before examining the model for measurement invariance because if the model did not fit the data, this would negatively impact the measurement invariance, and would need to be addressed before analysing the model for measurement invariance. The proposed model for the MTAS is as follows:

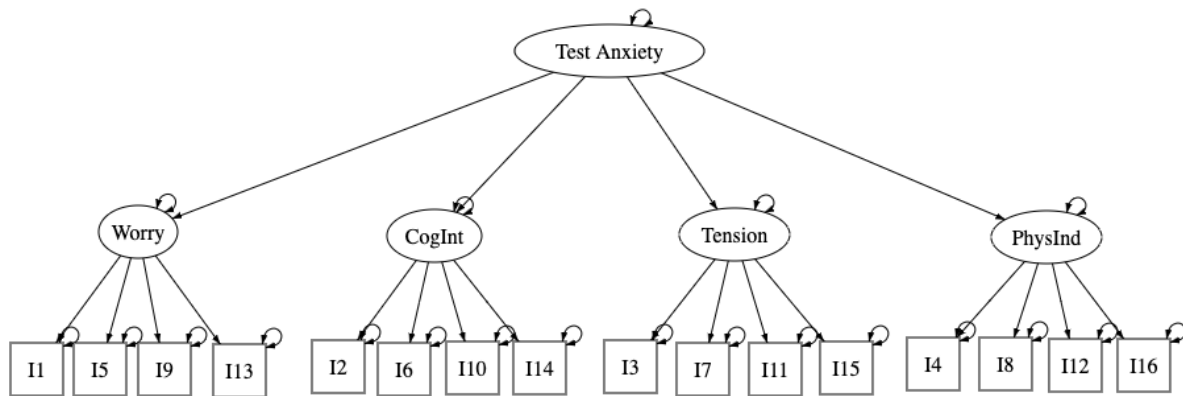


Figure 1. MTAS Path Diagram

Examinations of assumptions for analytical approaches. Higher-order factor models must meet several rules. The three-indicator rule states that each higher-order factor must have at least three factors (Kline , 2016). This rule is satisfied in this test anxiety model. The higher-order factor of test anxiety has four indicators, worry, cognitive interference, tension, and physiological indicators. There is also a two-indicator rule for first-order factors, which states that these factors must have two or more indicators (Kline , 2016). This is also satisfied by this test anxiety model as each first-order factor (worry, cognitive interference, tension, and physiological indicators) has four indicators. The counting rule specifies that the model's degrees of freedom must be equal to or more than zero (Kline , 2016). In Confirmatory Factor Analyses,

a unit must be identified for each latent variable in the model. In MPlus this is done automatically through unit loading, which fixes one of the factor loadings of the observed variables that load on a target factor at one.

The chi-square test of model fit had a p -value of 0.0000 ($\chi^2=450.32$, $df=100$) which is less than .05 and thus indicates poor model fit. However, most CFA research do not use chi-square as a test for model fit because it is overly sensitive to large samples (Cheung & Rensvold, 2002). The CFI was .959, which is more than .95, which indicates a good model fit. The RMSEA was .062, over .05, but under .08, indicating acceptable model fit. The SRMR was .051 and slightly greater than the recommended .05. Overall, model fit indices indicated an acceptable model fit to the data. The chi-square test of model fit is sensitive to sample size. With large sample sizes, minor differences between the observed and model-implied matrix may lead to a rejection of the null hypothesis and result in an indication of poor model fit (Cheung & Rensvold, 2002).

Table 4
Confirmatory Factor Analysis

	χ^2 (df)	RMSEA	SRMR	CFI
Test Anxiety	450.31 (100) p-value 0.000	.062	.051	.959

High factor loadings indicate alignment to the proposed latent variable. Thus standardized factor loadings that are closer to one are good indicators that the assessment items are suitable indicators (Kline, 2015). Table 5 shows the factor loadings of each item in the MTAS with the associated factor for the dataset. All of the loadings are above .7 except items 13 and 14. Item 13 is loaded onto the latent variable of worry and has a factor loading of .675, which, while not as large as the other factor loadings, is still acceptable. The same is true for

item 14 which is loaded onto the latent variable of cognitive interference and has a factor loading of .503. Additionally, for the latent variable of test anxiety which is the total anxiety scale, their latent variable of cognitive interference has the lowest factor loading of .577. Though these loadings are lower than the other factor loadings, which are all over .7, and thus have a stronger relationship with the latent factor of test anxiety, a factor loading of .577 is still high and indicates alignment to the proposed latent variable (Kline, 2015).

Table 5
Standardized Factor Loadings from the Combined Measurement Model

Items	TA	W	CI	T	PI
1.Before a test/ exam, I am worried I will fail.		.781			
2.I forget previously known material before taking a test/exam.			.717		
3.Even when I have prepared for a test/ exam I feel nervous about it.				.756	
4.Before I take a test/ exam my hand trembles.					.745
5.During tests/ exams, I worry about the consequences of failing.		.742			
6.I forget facts I have learnt during tests/exams.			.863		
7.I feel tense before taking a test/exam.				.789	
8.My heart races when I take a test/exam.					.825
9. After a test/exam, I am worried I have failed.		.825			
10.During tests/exams, I forget things that I have learnt.			.862		
11.Just before I take a test/exam, I feel panicky.				.852	
12.During a test/ exam I experience stomach discomfort.					.717
13.During a test/ exam, I worry that I gave the wrong answers.		.675			
14.During tests/exams, I find it hard to concentrate.			.503		
15.Before a test/exam, I feel nervous.				.823	
16.During a test/ exam, my muscles are tight					.726

Table 5 (continued)

Items	TA	W	CI	T	PI
Worry (W)	.914				
Cognitive Interference (CI)	.577				
Tension (T)	.959				
Physiological Indicators (PI)	.880				

Path diagram

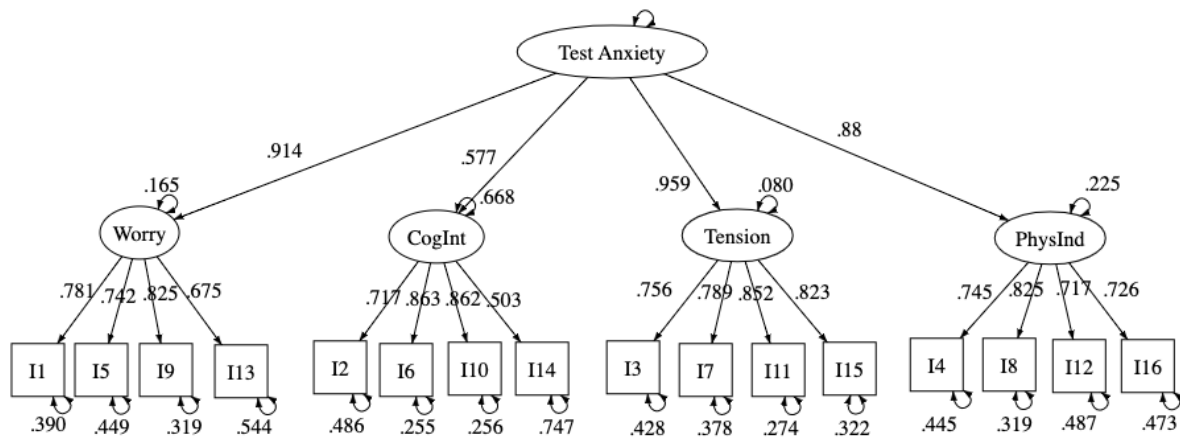


Figure 2. Path Diagram

Reliability Analysis

The MTAS was measured for internal consistency for all four subscales and for overall test anxiety to ensure the of the scale. Internal consistency was measured using Cronbach's (1951) alpha with the criteria that .9 indicates great internal consistency, .8 indicates good internal consistency, and .7 indicates acceptable internal consistency. The Cronbach's (1951) alpha for the MTAS scales were worry $\alpha=.835$, cognitive interference $\alpha=.804$, tension = .879, physiological indicators $\alpha=.842$ and total test anxiety $\alpha=.927$. The Cronbach's Alpha statistics

for the scales were between .8 and .9, which indicates that there is good to great internal consistency in the MTAS tool.

Table 6
Internal Consistency Reliability Statistics for MTAS

Scale	Cronbach's Alpha	McDonald's Omega
Worry	.835	.837
Cognitive Interference	.804	.807
Tension	.879	.882
Physiological Indicators	.842	.842
Total Test Anxiety	.927	.868

However, Cronbach's alpha has been criticised because it assumes constant variance for true scores which unrealistic. This can lead to several issues including alpha underestimating score reliability (Graham, 2006) or reliability being inflated (Cortina, 1993; Yuan & Bentler, 2002). One alternative to Cronbach's alpha is McDonald's omega which allows the means and variances of the true scores and the error variances to vary (Joreskog, 1971). The McDonald's omega for the subscales were worry $\omega=.837$, cognitive interference $\omega=.807$, tension $\omega = .882$, physiological indicators $\omega=.842$ and total test anxiety $\omega=.868$. The McDonald's omega statistics for the scales were between .8 and .9, which indicates that there is good to great internal consistency in the MTAS tool.

Measurement Invariance

Measurement invariance testing “assesses the psychometric equivalence of a construct across groups or across time” (Putnick & Bornstein, 2016). This statistical property examines the underlying structure of an assessment and indicates if this underlying structure is the same across groups. The MTAS was examined for measurement invariance across gender (male and female), grade (Year10-13) and socio-economic status (eligible or not eligible for free lunch) before the

LPA. This analysis was completed first to determine if there was invariance across these groups before the creation of profiles using an LPA.

Measurement Invariance Across Gender

Model 1. Configural invariance examines the loading of items onto the factors. A model is determined to have configural invariance if the number of factors is the same across the groups (male and female) and if the items that load onto those factors are the same across the groups (male and female). To test this form of invariance the model constrained each group to have the same structure. The results from Table 7 show that the $\chi^2=581.066$ ($df=200$), $p = .000$, RMSEA=0.065, SRMR=0.054 and CFI=0.951. These model fit indices indicate that there is good model fit.

Model 2. Metric invariance examines the factor loadings between the items and the factors. To test this form of invariance in the first-order model, the factor loadings between the items and the factors were constrained to be equal across groups (male and female). The results from Table 7 show that $\chi^2=599.138$ ($df=212$), $p=.000$, RMSEA=0.063, SRMR=0.058 and CFI=0.950. These indices indicate a good model fit with this more constrained model. The comparisons between the two models show that, $\Delta\chi^2 =18.072$ ($df=12$), $p= .114$. According to the chi-square difference test model the differences between the two models are not statistically significant. The criteria for the differences in fit indices are .01 change in CFI, .015 change in RMSEA, .030 change in SRMR for metric invariance, and -.015 change in SRMR for scalar and strict invariance (Chen, 2007). In model two Δ RMSEA=-0.002, Δ SRMR=0.004 and Δ CFI=-0.001. Thus the differences between the two models are not significant and there is not a considerable decrease in model fit from model one to model two which means that constraining the factor loadings across groups does not have a significant impact on the model fit.

Table 7
Descriptive Fit Statistics for Measurement Invariance Across Gender

Model	χ^2	<i>df</i>	RMSEA	SRMR	CFI	Model comparison	$\Delta\chi^2$	Δdf	<i>p</i>	$\Delta RMSEA$	$\Delta SRMR$	ΔCFI
Model 1 configural invariance	581.066	200	0.065	0.054	0.951	1	-	-	-	-	-	-
Model 2 First-order factor loadings invariant	599.138	212	0.063	0.058	0.950	1 vs. 2	18.07	12	.114	-0.002	0.004	-0.001
Model 3 First- and second-order factor loadings invariant	632.419	215	0.065	0.058	0.946	2 vs. 3	33.28	3	.000	0.002	0	-0.004
Model 4 First- and second-order factor loadings and intercepts of measured variables invariant	666.236	227	0.065	0.071	0.943	3 vs. 4	33.81	12	.001	0	0.013	-0.001
Model 5 First- and second-order factor loadings and intercepts of measured variables and first-order factors invariant	678.938	230	0.065	0.074	0.942	4 vs. 5	12.70	3	.005	0	0.003	-0.001
Model 6 First- and second-order factor loadings, intercepts, and disturbances of first-order factors invariant	737.776	246	0.066	0.082	0.936	5 vs. 6	58.83	16	.000	0.001	0.008	-0.006
Model 7 First- and second-order factor loadings, intercepts, disturbances of first-order factors, and residual variances of measured variables invariant	752.071	250	0.066	0.089	0.935	6 vs. 7	14.29	4	.006	0	0.007	-0.001

Model 3. The first-order factors (worry, cognitive interference, physiological indicators, and tension) and the second-order factor (total test anxiety) were examined for metric invariance. To test this form of invariance in the second-order model, the factor loadings between the first-order factors and the second-order factor were constrained to be equal across groups (male and female). The results from Table 7 show that $\chi^2=632.419$ ($df=215$), $p=.000$, RMSEA=0.065, SRMR=0.058 and CFI=0.946. These indices indicate a good model fit with this more constrained model. The comparisons between the two models show that $\Delta\chi^2 =33.281$ ($df=3$), $p=.000$. According to the chi-square difference test the differences between the two models are statistically significant. However, the chi-square difference test is known to reject models with minor violations, especially when the sample size is >300 , and should be accompanied with other criteria like the differences in fit indices (Chen, 2007). In model three Δ RMSEA=0.002, Δ SRMR=0 and Δ CFI=-0.004. According to the criteria for differences in fit indices there is not a considerable decrease in model fit from model two to model three and thus model three is accepted.

Model 4. Strict invariance examines the item intercepts. To test this form of invariance, the intercepts of the measured variables were constrained to be the same across the groups (male and female). The results from Table 7 show that $\chi^2=666.236$ ($df=227$), $p=.000$, RMSEA=0.065, SRMR=0.071 and CFI=0.943. These indices indicate a good model fit with this more constrained model. The comparisons between the two models show that $\Delta\chi^2=33.817$ ($\Delta df=12$), $p=.001$. According to the chi-square difference test the differences between the two models are statistically significant. However, the Δ RMSEA=0, Δ SRMR=0.013, and Δ CFI=-0.001. This indicates that there was not a considerable decrease in model fit from model three to model four (Chen, 2007) and thus model four is accepted.

Model 5. This model examined the intercepts of the measured variables and the intercepts of the first-order factors. The results from Table 7 show that $\chi^2=678.938$ ($df=230$), $p=.000$, RMSEA=0.065, SRMR=0.074 and CFI=0.942. These indices indicate a good model fit with this more constrained model. The comparisons between the two models show that $\Delta\chi^2=12.702$ ($\Delta df=3$), $p=.005$. According to the chi-square difference test the differences between the two models are statistically significant. However, the $\Delta RMSEA=0$, $\Delta SRMR=0.003$, and $\Delta CFI=-0.001$. According to the criteria for differences in fit indices there is not a considerable decrease in model fit from model four to model five and thus model five.

Model 6. Strict invariance examines the residual variance and constrains the residual variance to be the same across the groups (male and female). In model 6 the residual variance of the first order factors were constrained to be the same across groups (male and female). The results from Table 7 show that $\chi^2=737.776$ ($df=246$), $p=.000$, RMSEA=0.066, SRMR=0.082 and CFI=0.936. These indices indicate that there is adequate model fit with this more constrained model. The comparisons between the two models show that $\Delta\chi^2=58.838$ ($\Delta df=16$), $p=.000$. According to the chi-square difference test the differences between the two models are statistically significant. However, the $\Delta RMSEA=0.001$, $\Delta SRMR=0.008$ and $\Delta CFI=-0.006$. According to the criteria for differences in fit indices there is not a considerable decrease in model fit from model five to model six and thus model six is accepted.

Model 7. This model examined the residual variance of the first order factors and the measured variables, which was done by constraining these to be equal across the groups (male and female). The results from Table 7 show that $\chi^2=752.071$ ($df=250$), $p=.000$, RMSEA=0.066, SRMR=0.089 and CFI=0.935. These indices indicate a good model fit with this more constrained model. The comparisons between the two models show that $\Delta\chi^2=14.295$ ($\Delta df=4$),

$p=.006$. According to the chi-square difference test the differences between the two models are statistically significant. However, the $\Delta RMSEA=0$, $\Delta SRMR=0.007$, and $\Delta CFI=-0.001$.

According to the criteria for differences in fit indices there is not a considerable decrease in model fit from model six to model seven and thus model seven is accepted.

Measurement Invariance Across Socioeconomic Status

Model 1. The results from Table 8 show that the $\chi^2=569.356$ ($df=200$), $p = .000$, $RMSEA=0.063$, $SRMR=0.053$ and $CFI=0.956$. These model fit indices indicate that there is good model fit.

Model 2. The results from Table 8 show that $\chi^2=585.896$ ($df=212$), $p = .000$ $RMSEA=0.062$, $SRMR=0.057$ and $CFI=0.956$. These indices indicate a good model fit with this more constrained model. The comparisons between the two models show that $\Delta\chi^2= 16.54$ ($\Delta df=12$, $p = .168$). According to the chi-square difference test the differences between the two models are not statistically significant. Additionally, the $\Delta RMSEA=-0.001$, $\Delta SRMR=0.004$, and $\Delta CFI=0$. This indicates that there was not a considerable decrease in model fit from model one to model two (Chen, 2007) and thus model two is accepted.

Model 3. The results from Table 8 show that $\chi^2=587.922$ ($df=215$), $p = .000$, $RMSEA=0.061$, $SRMR=0.057$ and $CFI=0.956$. These indices indicate a good model fit with this more constrained model. The comparisons between the two models show that $\Delta\chi^2=2.026$ ($\Delta df=3$, $p = .567$). According to the chi-square difference test the differences between the two models are not statistically significant. Additionally, the $\Delta RMSEA=-0.001$, $\Delta SRMR=0$, and $\Delta CFI=0$. This indicates that there was not a considerable decrease in model fit from model two to model three (Chen, 2007) and thus model three is accepted.

Model 4. The results from Table 8 show that $\chi^2=599.028$ ($df=227$), $p = .000$, RMSEA=0.060, SRMR=0.057 and CFI=0.956. These indices indicate a good model fit with this more constrained model. The comparisons between the two models show that $\Delta\chi^2=11.106$ ($\Delta df=12$, $p = .520$). According to the chi-square difference test the differences between the two models are not statistically significant. Additionally, the $\Delta RMSEA=-0.001$, $\Delta SRMR=0$, and $\Delta CFI=0$. This indicates that there was not a considerable decrease in model fit from model three to model four (Chen, 2007) and thus model four is accepted.

Model 5. The results from Table 8 show that $\chi^2=605.984$ ($df=230$), $p = .000$, RMSEA=0.060, SRMR=0.057 and CFI=0.955. These indices indicate a good model fit with this more constrained model. The comparisons between the two models show that $\Delta\chi^2=6.956$, ($\Delta df=3$, $p = .073$). According to the chi-square difference test the differences between the two models are not statistically significant. Additionally, the $\Delta RMSEA=0$, $\Delta SRMR=0$, and $\Delta CFI=-0.001$, which are not significant and indicate that the item intercepts are invariant across socio-economic status. This indicates that there was not a considerable decrease in model fit from model four to model five (Chen, 2007) and thus model five is accepted.

Model 6. The results from Table 8 show that $\chi^2=630.092$ ($df=246$), $p = .000$, RMSEA=0.058, SRMR=0.060 and CFI=0.954. These indices indicate a good model fit with this more constrained model. The comparisons between the two models show that $\Delta\chi^2=24.108$ ($\Delta df=16$, $p = .087$). According to the chi-square difference test the differences between the two models are not statistically significant. Additionally, the $\Delta RMSEA=-0.002$, $\Delta SRMR=0.003$, and $\Delta CFI=-0.001$. This indicates that there was not a considerable decrease in model fit from model five to model six (Chen, 2007) and thus model six is accepted.

Table 8

Descriptive Fit Statistics for Measurement Invariance Across Socioeconomic Status

Model	χ^2	<i>df</i>	RMSEA	SRMR	CFI	Model comparison	$\Delta\chi^2$	Δdf	<i>p</i>	Δ RMSEA	Δ SRMR	Δ CFI
Model 1 configural invariance	569.356	200	0.063	0.053	0.956	-	-	-	-	-	-	-
Model 2 First-order factor loadings invariant	585.896	212	0.062	0.057	0.956	1 vs. 2	16.54	12	.168	-0.001	0.004	0
Model 3 First- and second-order factor loadings invariant	587.922	215	0.061	0.057	0.956	2 vs. 3	2.026	3	.567	-0.001	0	0
Model 4 First- and second-order factor loadings and intercepts of measured variables invariant	599.028	227	0.060	0.057	0.956	3 vs. 4	11.106	12	.520	-0.001	0	0
Model 5 First- and second-order factor loadings and intercepts of measured variables and first-order factors invariant	605.984	230	0.060	0.057	0.955	4 vs. 5	6.956	3	.073	0	0	-0.001
Model 6 First- and second-order factor loadings, intercepts, and disturbances of first-order factors invariant	630.092	246	0.058	0.060	0.954	5 vs. 6	24.108	16	.087	-0.002	0.003	-0.001
Model 7 First- and second-order factor loadings, intercepts, disturbances of first-order factors, and residual variances of measured variables invariant	632.881	250	0.058	0.063	0.955	6 vs. 7	2.789	4	.594	0	0.003	0.001

Model 7. The results from Table 8 show that $\chi^2=632.881$ ($df=250$), $p = .000$, RMSEA=0.058, SRMR=0.063 and CFI=0.955. These indices indicate a good model fit with this more constrained model. These indices indicate a good model fit with this more constrained model. The comparisons between the two models show that $\Delta\chi^2=2.789$ ($\Delta df=4$), $p = .594$. According to the chi-square difference test the differences between the two models are not statistically significant. Additionally, the $\Delta RMSEA=0$, $\Delta SRMR=0.003$, and $\Delta CFI=0.001$. This indicates that there was not a considerable decrease in model fit from model six to model seven (Chen, 2007) and thus model seven is accepted.

Measurement Invariance Across Grades

Model 1. The results from Table 9 show that the $\chi^2=758.099$ ($df=400$) $p=0.000$, RMSEA=0.065, SRMR=0.057 and CFI=0.955. These model fit indices indicate that there is good model fit.

Model 2. The results from Table 9 show that $\chi^2=849.793$ ($df=436$) $p=0.000$, RMSEA=0.064, SRMR=0.074 and CFI=0.951. These indices indicate a good model fit with this more constrained model. The comparisons between the two models show that $\Delta\chi^2=91.694$ ($\Delta df=36$), $p=0.000$. According to the chi-square difference test the differences between the two models are statistically significant. However, the $\Delta RMSEA=-0.001$, $\Delta SRMR=0.017$, and $\Delta CFI=-0.004$, which are not significant, which means that the factor loadings are invariant across grades. According to the criteria for differences in fit indices there is not a considerable decrease in model fit from model one to model two and thus model two is accepted.

Model 3. The results from Table 9 show that $\chi^2=874.389$ ($df=445$), $p=0.000$, RMSEA=0.065, SRMR=0.082 and CFI=0.950. These indices indicate a good model fit with this more constrained model. The comparisons between the two models show that $\Delta\chi^2=24.596$

(Δdf)=9, $p=0.003$. According to the chi-square difference test the differences between the two models are statistically significant. However, the $\Delta RMSEA=0.001$, $\Delta SRMR=0.008$, and $\Delta CFI=-0.001$. According to the criteria for differences in fit indices there is not a considerable decrease in model fit from model two to model three and thus model three is accepted.

Model 4. The results from Table 9 show that $\chi^2=939.590$ ($df=481$), $p=0.000$, $RMSEA=0.064$, $SRMR=0.084$ and $CFI=0.946$. These indices indicate a good model fit with this more constrained model. The comparisons between the two models show that $\Delta\chi^2=65.201$ (Δdf)=36, $p=0.002$. According to the chi-square difference test the differences between the two models are statistically significant. However, the $\Delta RMSEA=-0.001$, $\Delta SRMR=0.002$, and $\Delta CFI=-0.004$. According to the criteria for differences in fit indices there is not a considerable decrease in model fit from model three to model four and thus model four is accepted.

Model 5. The results from Table 9 show that $\chi^2=970.614$ ($df=490$), $p=0.000$, $RMSEA=0.065$, $SRMR=0.087$ and $CFI=0.944$. These indices indicate a good model fit with this more constrained model. The comparisons between the two models show that $\Delta\chi^2=31.024$ (Δdf)=9, $p=0.000$. According to the chi-square difference test the differences between the two models are statistically significant. However, the $\Delta RMSEA=0.001$, $\Delta SRMR=0.003$, and $\Delta CFI=-0.002$. According to the criteria for differences in fit indices there is not a considerable decrease in model fit from model four to model five and thus model five is accepted.

Model 6. The results from Table 9 show that $\chi^2=1032.834$ ($df=538$), $p=0.000$, $RMSEA=0.063$, $SRMR=0.091$ and $CFI=0.942$. These indices indicate a good model fit with this more constrained model. The comparisons between the two models show that $\Delta\chi^2=62.22$ (Δdf)=48, $p=0.081$. According to the chi-square difference test the differences between the two models are not statistically significant. Additionally, the $\Delta RMSEA=-0.002$, $\Delta SRMR=0.004$, and

$\Delta\text{CFI}=-0.002$. According to the criteria for differences in fit indices there is not a considerable decrease in model fit from model five to model six and thus model six is accepted.

Model 7. The results from Table 9 show that $\chi^2=1040.443$ ($df=550$), $p=0.000$, $\text{RMSEA}=0.062$, $\text{SRMR}=0.091$ and $\text{CFI}=0.943$. These indices indicate a good model fit with this more constrained model. The comparisons between the two models show that $\Delta\chi^2=7.609$ ($\Delta df=12$, $p=0.815$). According to the chi-square difference test the differences between the two models are not statistically significant. Additionally, $\Delta\text{RMSEA}=-0.001$, $\Delta\text{SRMR}=0$, and $\Delta\text{CFI}=0.001$. According to the criteria for differences in fit indices there is not a considerable decrease in model fit from model six to model seven and thus model seven is accepted.

Latent Profile Analysis

An LPA was utilized with the first dataset to identify profiles of risk among the four factors of the MTAS, worry (W), cognitive interference (CI), tension (T), and physical indicators (PI). This study hypothesised there would be three profiles of low test anxiety, moderate test anxiety, and high test anxiety. This hypothesis was based on existing research in test anxiety classification and cut scores (Thomas et al., 2017; Segool et al., 2013; von der Embse et al., 2014).

Model evaluation. Table 10 shows the fit statistics for a one profile to four profile model. The AIC, BIC, Sample-adjusted BIC, LMR-LRT, adjusted LMR-LRT, Bootstrap LRT, and entropy were all considered to decide which model fit best with the data. For AIC, BIC, and Sample-adjusted BIC, the model with the smallest number is regarded as the best fit. According to these criteria, the five profile model was the best fit for the data (AIC = 17784.367, BIC = 17919.389, Sample-adjusted BIC = 17830.464). However, when the five profile model was examined profile one only accounted for three percent of the sample. This percentage is

Table 9
Descriptive Fit Statistics for Measurement Invariance Across Grades

Model	χ^2	<i>df</i>	RMSEA	SRMR	CFI	Model comparison	$\Delta\chi^2$	Δdf	<i>p</i>	Δ RMSEA	Δ SRMR	Δ CFI
Model 1 configural invariance	758.099	400	0.065	0.057	0.955	-	-	-	-	-	-	-
Model 2 First-order factor loadings invariant	849.793	436	0.064	0.074	0.951	1 vs. 2	91.694	36	.000	-0.001	0.017	-0.004
Model 3 First- and second-order factor loadings invariant	874.389	445	0.065	0.082	0.950	2 vs. 3	24.596	9	.003	0.001	0.008	-0.001
Model 4 First- and second-order factor loadings and intercepts of measured variables invariant	939.590	481	0.064	0.084	0.946	3 vs. 4	65.201	36	.002	-0.001	0.002	-0.004
Model 5 First- and second-order factor loadings and intercepts of measured variables and first-order factors invariant	970.614	490	0.065	0.087	0.944	4 vs. 5	31.024	9	.000	0.001	0.003	-0.002
Model 6 First- and second-order factor loadings, intercepts, and disturbances of first-order factors invariant	1032.834	538	0.063	0.091	0.942	5 vs. 6	62.22	48	.081	-0.002	0.004	-0.002
Model 7 First- and second-order factor loadings, intercepts, disturbances of first-order factors, and residual variances of measured variables invariant	1040.443	550	0.062	0.091	0.943	6 vs. 7	7.609	12	.815	-0.001	0	0.001

less than ten percent and indicates that this class was not meaningful or practically useful for classification purposes. This means that one of the profiles in the five profile model is not meaningful and the four profile model would be more appropriate. The four profile model also had the next best model fit according to the AIC, BIC and Sample-adjusted BIC criteria (AIC = 17925.013, BIC = 18035.924, Sample-adjusted BIC = 17962.879).

For LMR-LRT, adjusted LMR-LRT, and bootstrap LRT each model is compared to the next model with one less profile. If there is a statistical difference between the compared models, the model with the additional profile is considered the best fit. For these methods of profile evaluation, p -values that are $<.05$ indicate that the model has a better fit than the next model with one less profile. According to these criteria, the five profile model was the best fit for the data (LMR-LRT = 150.646, $p=.0023$, Adjusted LMR-LRT = 146.355, $p=.0026$, Bootstrap LRT = 150.646, $p=.000$). However, as was mentioned previously, profile one in the five profile model was not meaningful.

Entropy is the last class of methods for evaluating latent profile models, and for this criteria scores closer to one indicate that the profiles are distinct and the assignment of individuals to these profiles are accurate. The two profile model had the most distinct profiles with .867 entropy. However, entropy is influenced by sample size, such that as sample sizes increase, entropy decreases (Collins & Lanza, 2010). So, while these fit statistics are useful in determining which models are the best fit for the interpretation of this LPA data will also account for the theory behind test anxiety and the ultimate goal of these profiles, which is classification standards.

The goal of this study was to increase the usability of the MTAS, so this assessment can be used in schools to make decisions based on student performance in the evaluation. One way to

do this is to create classifications of the scores so school personnel can quickly identify scores that indicate a high level of test anxiety and provide the resources needed. However, the creation of these classifications should also be guided by test anxiety theory. According to recent test anxiety theory, four factors influence test anxiety, cognitive interference, worry, tension, and physiological tension. High levels of any of these factors may be indicative that the student is experiencing stress past their ability to cope (Zeidner, 1998).

The study originally hypothesised that the three model profile would be the best fit model for the data. This hypothesis was based on the existing research on test anxiety assessments and their development of classifications which used a three profile model of low, moderate, and high (Thomas et al., 2017; Segool et al., 2013; von der Embse et al., 2014). However, the four profile model has better fit statistics (AIC = 17925.013, BIC = 18035.924, Sample-adjusted BIC = 17962.879, LMR-LRT = 186.953, $p=.0429$, Adjusted LMR-LRT = 181.628, $p=.0458$, Bootstrap LRT = 186.953, $p=.000$) and each profile accounted for a meaningful percentage of the sample. Because of this the four profile model was chosen over the two, three and five profile models.

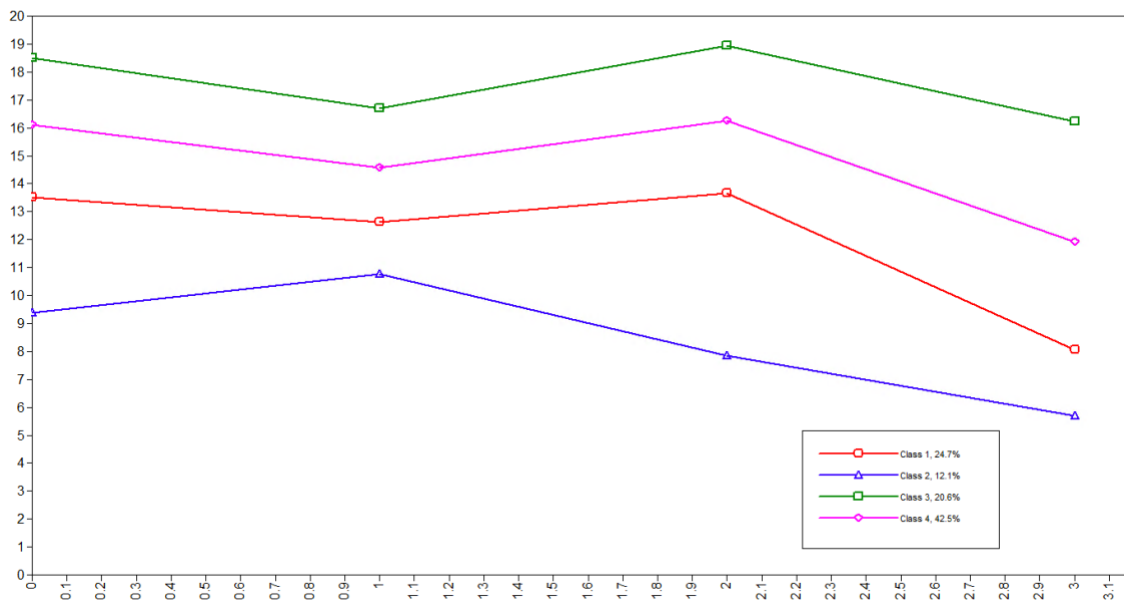


Figure 3. MPlus Profile Plot for Four Profile Model

Interpretation of LPA. The four profile model was chosen given the good model fit criteria (AIC = 17925.013, BIC = 18035.924, Sample-adjusted BIC = 17962.879, LMR-LRT = 186.953, $p=.0429$, Adjusted LMR-LRT = 181.628, $p=.0458$, Bootstrap LRT = 186.953, $p=.000$). Additionally, the four profile model would be usable in the school setting . Though the previous research indicates that three profile models are the norm in test anxiety scales, four profiles are still easily interpretable to school personnel. Table 11 shows the mean scores and sample percentages of the profiles in each model. Profile two was labelled low anxiety. It had low average scores in all four of the subscales with the scores ranging from six to eleven out of twenty. Additionally, the average total test anxiety of profile two was 34 out of 80. The general low scores of this profile is why it was labelled low test anxiety. Profile one was labelled average test anxiety. It had moderate average scores in the worry, cognitive interference and tension subscales, which all fell between 13-14 out of 20, while it had a low score in the physiological indicator subscale. The average total test anxiety score for profile one was 49. Profile four was labelled above average in test anxiety. This profile had high average scores across the subscales with scores ranging from 12-16 out of 20 and the average total test anxiety was 59 out of 80. Profile three was labelled high test anxiety because the average scores across the subscales were very high, ranging from 16-19 out of 20 and the average total score was also very high at 71 out of 80.

Table 10
Descriptive Fit Statistics for LPA Models

# of Profiles	Loglikelihood	# of Free Parameters	AIC	BIC	Sample-Size Adjusted BIC	LMR-LRT	Adjusted LMR-LRT	Bootstrap LRT	Entropy
1	-9931.710	8	19879.421	19917.998	19892.591	-	-	-	-
2	-9324.464	13	18674.927	18737.616	18696.329	1214.493 <i>p</i> =.000	1179.903 <i>p</i> =.000	1214.493 <i>p</i> =.000	0.867
3	-9032.983	18	18101.966	18188.766	18131.600	582.961 <i>p</i> =.000	566.358 <i>p</i> =.000	582.961 <i>p</i> =.000	0.824
4	-8939.507	23	17925.013	18035.924	17962.879	186.953 <i>p</i> =.0429	181.628 <i>p</i> =.0458	186.953 <i>p</i> =.0000	0.800
5	-8864.184	28	17784.367	17919.389	17830.464	150.646 <i>p</i> =.0023	146.355 <i>p</i> =.0026	150.646 <i>p</i> =.0000	0.821

Table 11
Mean Subscale Scores and Sample Percentages for Latent Profiles

	2 Profile Model						3 Profile Model						4 Profile Model						5 Profile Model					
	W	CI	T	PI	Total	% of sample	W	CI	T	PI	Total	% of sample	W	CI	T	PI	Total	% of sample	W	CI	T	PI	Total	% of sample
1	11	11	10	7	39	24	15	13	15	10	53	50	14	13	14	8	49	25	6	8	6	5	25	3
2	17	15	17	13	62	76	10	11	9	6	36	15	9	11	8	6	34	12	11	12	9	6	38	11
3							18	16	18	15	67	35	19	17	19	16	71	21	14	13	14	8	49	25
4													16	15	16	12	59	43	16	15	16	12	59	41
5													19	17	19	16	71	20	19	17	19	16	71	20

*Each of the subscales scores was out of a total score of 20 and each of the total scores were out of a total score of 80.

Chapter Five: Discussion

Test anxiety refers to the changes in behaviour, emotion, and physiology that occur due to a student's perceptions of the consequences of a test or exam (Zeidner, 1998). This phenomenon has been linked with negative outcomes, including low student performance in standardised tests (Zeidner, 1998), with significant downstream consequences such as grade retention (Shwerdt et al., 2017) or denial of admission to university (Phelps, 2017). There are a number of standardised tests that students take throughout their education, like the GCSE, and these test scores are used to evaluate students' academic performance as well as the effectiveness of teachers and schools (Segool et al. 2014). Test anxiety impacts how well students do on these exams, so identifying and addressing test anxiety should be a high priority in schools.

Test anxiety must first be reliably identified to facilitate early intervention services. Test anxiety scales are one of the most systematic ways to identify students experiencing test anxiety. There are several test anxiety scales that are being used, however, these scales have two primary limitations, 1) the scales were created decades ago and are not consistent with recent advancements in test anxiety theory and conceptualization, and 2) the scales were made for research purposes and are difficult to apply in practical settings, like a school. The MTAS was developed in 2020 to address these limitations and aligns with current test anxiety research. This assessment includes four factors of test anxiety, worry, cognitive interference, tension, and physiological indicators. While the MTAS has undergone initial validation studies that examined the scale's factor structure, measurement invariance, and created initial cut scores, the MTAS needs further evaluation before it can be used in the schools.

Summary and Explanation of Findings

Interpretation and use include all the steps between administering an assessment and using its results to make decisions (Kane, 2013). The main goal of the present study is to further evaluate the MTAS, increase its usability in schools and other practical settings and provide guidelines for interpretation of scores. These three research questions were put forth: 1) Is there measurement invariance within different demographic groups (gender and socio-economic status) across the four domains of the Multidimensional Test Anxiety Scale (cognitive interference, worry, tension, and physiological indicators)? 2) Is there measurement invariance in the Multidimensional Test Anxiety Scale across grades? 3) Is there significant differences in levels of test anxiety amongst classification profiles?

The first analysis that was run on the MTAS model was a CFA was run on the sample to test the second-order model proposed by the creators of the MTAS. According to the model fit indices ($\chi^2=450.31(df=100)$ p -value=0.000, CFI=.959, RMSEA=.062, and SRMR=.051), the proposed model had a good model fit. This means that the model proposed by the developers fits the data collected in the sample. This also meant that further analyses could be done since it was confirmed that the model fits the data. The study also used Cronbach's Alpha to examine the internal consistency of the subscales. The Cronbach's Alpha for the MTAS scales were, worry $\alpha=.835$, cognitive interference $\alpha=.804$, tension = .879, physiological indicators $\alpha=.842$, and total test anxiety $\alpha=.927$. These statistics indicate that there is good to great internal consistency in the subscales of the MTAS. McDonald's omega was also used to examine the scale for internal consistency. The McDonald's omega for the subscales were, worry $\omega=.837$, cognitive interference $\omega=.807$, tension $\omega = .882$, physiological indicators $\omega=.842$ and total test anxiety $\omega=.868$. These statistics also indicated good internal consistency.

The first research question focused on the measurement invariance of the MTAS tool. The MTAS had measurement invariance across gender (strict invariance of first and second-order factors: $\chi^2=678.938$ ($df=230$), $p=0.000$, RMSEA=0.065, SRMR=0.074 and CFI=0.942). It also had measurement invariance across socio-economic status (strict invariance of first and second-order factors: $\chi^2=605.984$ ($df=230$), $p=0.000$, RMSEA=0.060, SRMR=0.057 and CFI=0.955). And it had measurement invariance across grades (Years 10 – 13) (strict Invariance of first and second-order factors: $\chi^2=970.614$ ($df=490$), $p=0.000$, RMSEA=0.065, SRMR=0.087 and CFI=0.944).

The second research question focused on the usability of the MTAS. Several factors affect the usability of an assessment (e.g., cost, length, target population, etcetera), and one major factor is the inclusion of guidelines for the interpretation of results. Cut scores or classifications outline the range of scores that fall within different categories like 'at risk' versus 'not at risk.' When an assessment includes cut scores or classifications, it allows stakeholders (e.g., school psychologists, teachers, administration) to easily identify students for additional support.

The MTAS was examined for cut scores using a variable centred approach called the ROC AUC. This method allows researchers to plot the sensitivity versus specificity at various classification thresholds and distinguishes between binary classifications, like 'at risk' vs. 'not at risk,' at different classification thresholds (Narkhede, 2018). There are initial cut scores for the MTAS Total scale (58 and 60) using a variable centred analysis approach. These cut scores can be used to determine whether or not a student is experiencing high test anxiety (von der Embse, et al., 2021). This study took a different approach to examine the MTAS for classification clusters. An LPA was used, which is a person-centred approach. An LPA identifies clusters of

responses that have common attributes (Laursen & Hoff, 2006) based on the theory that responses form distinct and exclusive latent profiles.

The original hypothesis indicated that the LPA would identify a three profile model: low test anxiety, moderate test anxiety and high test anxiety. However, examining the fit indices and the profiles themselves showed that a four profile model was more appropriate. The four profile model had good model fit indices (AIC = 17925.013, BIC = 18035.924, Sample-adjusted BIC = 17962.879, LMR-LRT = 186.953, $p=.0429$, Adjusted LMR-LRT =181.628, $p=.0458$, Bootstrap LRT =186.953, $p=.000$) and it also made sense based on the practical use of the classification in schools. Upon examination of the four profile model it was apparent that the average scores in each of the latent profiles fell into four major categories. These categories were low test anxiety (average total score=34/80), average test anxiety (average total score=49/80), above average test anxiety (average total score=59/80) and high test anxiety (average total score=71/80). These four categories would be useful for schools and other stakeholders using the MTAS. They would provide relevant information to determine service provision in the schools. Additionally, these response profiles are comparable to the initial cut scores of 58 and 60 created using a variable centred analysis approach (von der Embse, et al., 2021). Though the profiles created by this study include low, average, above average, and high test anxiety profiles, it also identifies scores of 58 and 60 as an indication of concern since these scores fall within the above average test anxiety classification that was outlined in this study.

Limitations and Future Research

While there are strengths to the current study, there were also several limitations. The first limitation was the use of pre-existing data. This meant that the study could not modify the

data collection procedures. However, the dataset had a large sample and rich information that were utilised in this study.

The other limitation of this study is the sample. The participants were all secondary school students from the United Kingdom. While it is likely that the results from this study are generalisable to other cultures (like American high school students), this will have to be tested. Further research is needed with a more diverse population of students from across different cultures and countries to examine the measurement invariance of the MTAS. Such a study would increase the population of student that can use the MTAS. Further research could also examine the latent profiles that emerge with different samples of students. The latent profiles will likely remain the same, but a study that uses samples from different cultures and countries would support the classification system of the MTAS.

Lastly, further research needs to be done on the identified latent profiles. This study determined the number of latent profiles of MTAS responders and labelled these profiles but the characteristics of these profiles need to be examined further. Future research could look at the outcomes associated with membership in these different latent profiles. For instance, does one group score higher on general anxiety scales? These further analyses will add support to the identified profiles and give additional information to users of the MTAS on the needs of students identified as ‘at risk’ by this assessment tool.

Implications for Practice

The most important implication for practice is the increased usability of the MTAS. One major critique of the current test anxiety scales is the lack of classifications or cut scores. Classifications are important because they increase the usability and relevance of an assessment tool. It aids in interpreting a score, letting the administrator know whether a student's score is

concerning and indicates the need for additional services or whether the score is average and doesn't need further action. Assessments equipped with classifications or cut scores are easier to use and facilitate better decision-making. This study also helped to further the research into test anxiety. While the focus of the study was validating and creating classifications for the MTAS, this study also adds to the general knowledge of test anxiety. It can be built on by future researchers. Test anxiety is an important topic in education that is increasingly depending on standardised tests for decision-making because of the impact it has on test performance. Research in this area must continue, and the development and validation of assessment tools for schools is an essential part of this research. Once these students are identified, they can receive the tools or services they need to be successful and increase the accuracy of standardised tests.

Conclusion

The main goals of this study were to examine the MTAS for measurement invariance across gender, grade, and socio-economic status (based on eligibility for free school meals) and to determine the number of latent profiles found within the data to create classifications. This study found that the MTAS did have measurement invariance across gender, grade, and socio-economic status. This supports the use of the MTAS with these populations of secondary school students and increases its validity. This study also found that there are four latent profiles of respondents. These profiles were labelled low test anxiety, average test anxiety, above average test anxiety and high test anxiety, respectively. The creation of classifications like these increases the usability of the MTAS. Administrators of the assessment will be able to determine if a student needs services for test anxiety based on their scores. Overall, this study added to the validation and usability of the MTAS.

References

- Alpert, R., & Haber, R. N. (1960). Anxiety in academic achievement situations. *The Journal of Abnormal and Social Psychology, 61*(2), 207–215. <https://doi.org/10.1037/h0045464>.
- Askell-Williams, H., & Lawson, M. (2013). Teachers' knowledge and confidence for promoting positive mental health in primary school communities. *Asia-Pacific Journal of Teacher Education, 41*(2), 126–143. doi:10.1080/1359866X.2013.777023.
- Aydin, U. (2019). Test Anxiety: Gender Differences in Elementary School Students. *European Journal of Educational Research, 8*(1), 21-30.
- Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of school psychology, 48*(1), 5-37. <https://doi.org/10.1016/j.jsp.2009.10.001>.
- Beidel, D. C., & Turner, S. M. (1988). Comorbidity of test anxiety and other anxiety disorders in children. *Journal of Abnormal Child Psychology, 16*(3), 275-287. <https://doi.org/10.1007/BF00913800>.
- Carter, M. G., Klenowski, V., & Chalmers, C. (2016). Who pays for standardised testing? A cost-benefit study of mandated testing in three Queensland secondary schools. *Journal of Education Policy, 31*(3), 330-342.
- Cassady, J. C. (2004a). The impact of cognitive test anxiety on text comprehension and recall in the absence of external evaluative pressure. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 18*(3), 311-325. <https://doi.org/10.1002/acp.968>

- Cassady, J. C. (2004b). The influence of cognitive test anxiety across the learning–testing cycle. *Learning and instruction, 14*(6), 569-592.
<https://doi.org/10.1016/j.learninstruc.2004.09.002>
- Cassady, J.C. (2010). *Anxiety in Schools: The Causes, Consequences, and Solutions for Academic Anxieties* (2nd ed.) Peter Lang.
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural equation modeling: A multidisciplinary journal, 14*(3), 464-504.
<https://doi.org/10.1080/10705510701301834>
- Chen, F. F., Sousa, K. H. & West, S. G. (2005) .Teacher's corner: Testing measurement invariance of second-order factor models. *Structural Equation Modeling, 12*(3), 471-492,
https://doi.org/10.1207/s15328007sem1203_7
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling, 9*(2), 233-255.
- Chou, C. P., & Bentler, P. M. (1995). Estimates and tests in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 37–55). Sage Publications, Inc.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104. doi:10.1037/0021-9010.78.1.98
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.
- Culler, R.E., Holahan, C.J., (1980). Test anxiety and academic performance: The effects of study-related behaviors. *Journal of Educational Psychology, 72*(1), 16–20.
<http://dx.doi.org/10.1037/0022-0663.72.1.16>.

- Donnelly, A., Fitzgerald, A., Shevlin, M., & Dooley, B. (2018). Investigating the psychometric properties of the Revised Child Anxiety and Depression Scale (RCADS) in a non-clinical sample of Irish adolescents. *Journal of Mental Health, 15*, 1–12.
<https://doi.org/10.1080/09638237.2018.1437604>
- Dowdy, E., Doane, K., Eklund, K., & Dever, B. V. (2013). A comparison of teacher nomination and screening to identify behavioral and emotional risk within a sample of underrepresented students. *Journal of Emotional and Behavioral Disorders, 21*(2), 127–137. doi:10.1177/1063426611417627.
- Everson, H. T., Millsap, R. E., & Rodriguez, C. M. (1991). Isolating gender differences in test anxiety: A confirmatory factor analysis of the Test Anxiety Inventory. *Educational and Psychological Measurement, 51*(1), 243-251. <https://doi.org/10.1177/0013164491511024>.
- Every Student Succeeds Act (ESSA) of 2015, Pub. L. No. 114-95, § 129 Stat. 1802. (2015). Retrieved from <https://www.gpo.gov/fdsys/pkg/BILLS-114s1177enr/pdf/BILLS-114s1177enr.pdf>
- Fayegh, Y., Rumaya, J., & Talib, M. A. (2010). The effects of family income on test anxiety and academic achievement among Iranian high school students. *Asian Social Science, 6*(6), 89-93.
- Florida Department of Education. (n.d.). *K-12 student assessment*.
<http://www.fldoe.org/accountability/assessments/k-12-student-assessment/#:~:text=Florida%20Standards%20Assessments%3A%20The%20Florida,for%20Algebra%201%20and%20Geometry.>

- Friedman, I.A., & Bendas-Jacob, O. (1997). Measuring perceived test anxiety in adolescents: A self-report scale. *Educational Psychology Measurement* 57(6), 1035–1046.
<http://dx.doi.org/10.1177/0013164497057006012>.
- Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability – What they are and how to use them. *Educational and Psychological Measurement*, 66, 930–944. doi:10.1177/0013164406288165
- Green, J. G., Keenan, J. K., Guzmán, J., Vinnes, S., Holt, M., & Comer, J. S. (2017). Teacher perspectives on indicators of adolescent social and emotional problems. *Evidence-Based Practice in Child and Adolescent Mental Health*, 26(4), 96-101.
doi:10.1080/23794925.2017.1313099.
- Hair, J. F., Hult, G. T. M., Ringle, C. M., and Sarstedt, M. (2021). *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. 3rd Ed. Thousand Oaks, CA: Sage
- Hembree, R. (1988). Correlates, Causes, Effects, and Treatment of Test Anxiety. *Review of Educational Research*, 58(1), 47–77.
- Hong, E. (1998). Differential stability of individual differences in state and trait test anxiety. *Learning and Individual Differences*, 10(1), 51–69. [https://doi.org/10.1016/S1041-6080\(99\)80142-3](https://doi.org/10.1016/S1041-6080(99)80142-3).
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, 6(1), 1-55. <https://doi.org/10.1080/10705519909540118>
- Joreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, 36, 109–133. doi:10.1007/BF02291393

- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. <https://doi.org/10.1111/jedm.12000>
- Kelley, T. L. (1927). *Interpretation of educational measurements*. Oxford, England: World Book Co.
- Kline, R. B. (2016). *Principles and practice of structural equation modeling (4th ed.)*. New York: Guilford Press
- Laursen, B., & Hoff, E. (2006). Person-centered and variable-centered approaches to longitudinal data. *Merrill-Palmer Quarterly 52*(3), 377-389.
- Liebert, R. M. & Morris, L. W. (1967). Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological Reports, 20*(3), 975-978.
[doi:10.2466/pr0.1967.20.3.975](https://doi.org/10.2466/pr0.1967.20.3.975).
- 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*(6), 503–514.
- Lowe, P. A. (2014). Should test anxiety be measured differently for males and females? Examination of measurement bias across gender on measures of test anxiety for middle and high school, and college students. *Journal of Psychoeducational Assessment, 33*(3), 238-246. <https://doi.org/10.1177/0734282914549428>.
- 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. <http://dx.doi.org/10.1177/0734282907303760>.

- Narkedede, S. (2018, June 26). *Understanding AUC - ROC Curve*. Towards Data Science.
<https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>
- Oetting, E. R., & Deffenbacher, J. L. (1980). *Test Anxiety Profile Manual*. Fort Collins, CO: RMBSI, Inc.
- The UK's Exam System Explained*. (2018, January 2). Point to Point Education. Retrieved 2020, August 26 from <https://www.pointtopointeducation.com/blog/uks-exam-system-explained/>
- Parikh, R., Mathai, A., Parikh, S., Sekhar, G. C., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology*, 56(1), 45.
- Pendergast, L. L., von der Embse, N., Kilgus, S. P., & Eklund, K. R. (2017). Measurement equivalence: A non-technical primer on categorical multi-group confirmatory factor analysis in school psychology. *Journal of School Psychology*, 60, 65-82.
<https://doi.org/10.1016/j.jsp.2016.11.002>
- Phelps, R. P. (2017). *Kill the messenger: The war on standardized testing*. Routledge.
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41, 71-90. doi: 10.1016/j.dr.2016.06.004
- Putwain, D. W. (2007). Test anxiety in UK schoolchildren: Prevalence and demographic patterns. *British Journal of Educational Psychology*, 77(3), 579-593.
<https://doi.org/10.1348/000709906X161704>.
- Putwain, D. W., Connors, L., & Symes, W. (2010). Do cognitive distortions mediate the test anxiety-examination performance relationship? *Educational Psychology*, 30(1), 11-26.
<https://doi.org/10.1080/01443410903328866>.

- Putwain, D., & Daly, A. L. (2014). Test anxiety prevalence and gender differences in a sample of English secondary school students. *Educational Studies, 40*(5), 554-570.
<https://doi.org/10.1080/03055698.2014.953914>.
- Putwain, D. W., Nathaniel, P., Rainbird, E. C., & West, G. (2020). The development and validation of a new Multidimensional Test Anxiety Scale (MTAS). *European Journal of Psychological Assessment* . <https://doi.org/10.1027/1015-5759/a000604>
- Putwain, D. W., Woods, K. A., & Symes, W. (2010). Personal and situational predictors of test anxiety of students in post-compulsory education. *British Journal of Educational Psychology, 80*(1), 137-160. <https://doi.org/10.1348/000709909X466082>
- Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592.
- Rajagopalan, K., & Gordon, E. W. (2016). *The testing and learning revolution: The future of assessment in education*. Springer.
- Ratner, B. (2009). The correlation coefficient: Its values range between +1/-1, or do they?. *Journal of targeting, measurement and analysis for marketing, 17*(2), 139-142.
- Sarason, I. G. (1984). Stress, Anxiety and Cognitive Interference: Reactions to Tests. *Journal of Personality and Social Psychology, 46*(4): 929–938. doi:10.1037/0022-3514.46.4.929.
- Sarason, S. B., & Mandler, G. (1952). Some correlates of test anxiety. *The Journal of Abnormal and Social Psychology, 47*(4), 810–817. <https://doi.org/10.1037/h0060009>.
- Sarason, S. B., Davidson, K. S., Lighthall, F. F., Waite, R. R., & Ruebush, B. K. (1960). *Anxiety in elementary school children: A report of research*. John Wiley & Sons Inc.
<https://doi.org/10.1037/14349-000>.

- Schutz, P. A., Di Stefano, Ch., Benson, J., & Davis, H. A. (2004). The emotional regulation during test-taking scale. *Anxiety, Stress, and Coping: An International Journal*, *17*, 253–269. <https://doi.org/10.1080/10615800410001710861>.
- Schwerdt, G., West, M. R., & Winters, M. A. (2017). The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida. *Journal of Public Economics*, *152*, 154-169. <https://doi.org/10.1016/j.jpubeco.2017.06.004>
- 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*, 489-499.
- Segool, N., von der Embse, N.P., Mata, A., Gallant, J., (2014). Cognitive-behavioral model of test anxiety in a high-stakes context: An exploratory study. *School Mental Health* *6*(1), 50–61. <http://dx.doi.org/10.1007/s12310-013-9111-7>.
- Sommer, M., & Arendasy, M. E. (2014). Comparing different explanations of the effect of test anxiety on respondents' test scores. *Intelligence*, *42*, 115–127. doi: 10.1016/j.intell.2013.11.003.
- Spielberger, C. D. (1980). The test anxiety inventory. *Psychology*, *4*(6A).
- Spielberger, C. D., & Vagg, R. P. (1995). *Test anxiety: Theory, assessment and treatment*. Bristol, UK: Taylor & Francis.
- Splett, J. W., Garzona, M., Gibson, N., Wojtalewicz, D., Raborn, A., & Reinke, W. M. (2019). Teacher recognition, concern, and referral of children's internalizing and externalizing behavior problems. *School Mental Health*, *11*(2), 228-239

- Steinmayr, R., Crede, J., McElvany, N., & Wirthwein, L. (2016). Subjective well-being, test anxiety, academic achievement: Testing for reciprocal effects. *Frontiers in psychology*, 6, 1994. <https://doi.org/10.3389/fpsyg.2015.01994>
- Tein, J., Coxe, S., & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural Equation Modeling*, 20(4), 640–657. <https://doi.org/10.1080/10705511.2013.824781>.
- von der Embse, N., & Hasson, R. (2012). Test anxiety and high-stakes tests: Implications for educators. *Preventing School Failure*, 56(3), 180–187.
- von der Embse, N., Kim, E., Jenkins, A., Sanchez, A., Kilgus, S. P., & Eklund, K. (2021). Profiles of rater dis/agreement within universal screening in predicting distal outcomes. *Journal of Psychopathology and Behavioral Assessment*, 1-14. <https://doi.org/10.1007/s10862-021-09869-0>
- von der Embse, N. P., Kilgus, S. P., Solomon, H. J., Bowler, M., & Curtiss, C. (2015). Initial development and factor structure of the Educator Test Stress Inventory. *Journal of Psychoeducational Assessment*, 33(3), 223-237. <http://dx.doi.org/10.1177/0734282914548329>.
- von der Embse, N., Putwain, D. & Francis, G. (2020). *Interpretation and Use of the Multidimensional Test Anxiety Scale (MTAS)*. Manuscript submitted for publication.
- 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. <https://doi.org/10.1016/j.jad.2017.11.048>
- Wine, J. (1971). Test anxiety and direction of attention. *Psychology Bulletin*, 76(2), 92–104. <http://dx.doi.org/10.1037/h0031332>.

- Woolf, B. (1957). The log-likelihood ratio test (the G-test). *Annals of Human Genetics*, 21(4), 397-409. <https://doi.org/10.1111/j.1469-1809.1972.tb00293.x>
- Young, M. (2018). The influence of standardized testing pressure on teachers' working environment. *KEDI Journal of Educational Policy*, 15(2), 3-22.
- Yuan, K., & Bentler, P. M. (2002). On robustness of the normal-theory based asymptotic distributions of three reliability coefficient estimates. *Psychometrika*, 67, 251-259. doi:10.1007/BF02294845
- Zeidner, M. (1998). *Test anxiety: The state of the art*. New York, NY: Plenum Press.
- Zeidner, M., & Matthews, G. (2005). Evaluation anxiety: Current theory and research. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 141-163). New York, NY: Guilford Press.