1-1-2017

Factor Invariance of the Standard Achievement Admission Test by Gender and School Types

Mohammed Alqabbaa
University of Denver

Factor Invariance of the Standard Achievement Admission Test by Gender and School Types

---

A Thesis

Presented to

the Faculty of the Morgridge College of Education

University of Denver

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

---

by

Mohammed Alqabbaa

August 2017

Advisor: Dr. Duan Zhang

Author: Mohammed Alqabbaa
Title: Factor Invariance of the Standard Achievement Admission Test by Gender and School Types
Advisor: Dr. Duan Zhang
Degree Date: August 2017

Abstract

This study examined the factorial structure of the SAAT with exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Specifically, a CFA of a four-factor model (the hypothesized model) was tested to see if the model had good fit as well as a model generated from an EFA. The EFA showed that the SAAT test only measured two factors (biology and chemistry), not the four posited by the test developers. The CFA provided good fit to a two-factor model; however, a CFA showed that the hypothesized four-factor model also fit the data well and so the hypothesized model was selected as the most appropriate. Based on the CFA four-factor model, measurement invariance (configural, metric and scalar) were examined across school type (public versus private) and gender (male and female) on the test. The results revealed that the metric invariance model fit the data best compared to the other models. Finally, latent means differences were tested by using two-way ANOVA on all the four factor subjects (biology, chemistry, physics, and mathematics). The results revealed that female students in high schools did better than males on all four sections of the SAAT test. On the other hand, male students in public schools did not achieve well on the test compared to males in private high schools. Also, male students in public school did not achieve well compared to females in both schools.

Acknowledgements

It has been a great experience as a student in the Research Methods and Statistics program at the University of Denver. I learned from wonderful and very well educated faculty members. I am very grateful that I got all the help I needed from chair members Dr. Duan Zhang and Dr. Kathy Green. Without them I do not think I would be where I am right now. Their ideas, thoughts, and suggestions helped me succeed on my thesis paper. Also, I want to thank Lilian Chimuma and Abdullah Al-Sayar for being very patient with me and giving me support every time I ask for help.

In addition, I want to thank my amazing friend, Jeffery Chandler, for being always there for me no matter what, and providing me with unfailing support, love, and continuous encouragement throughout my years of study. This accomplishment would not have been possible without him. Also, I must express my very profound gratitude to my lovely father who provided me with everything he could to make sure that I got all the assistance I needed through my years of studies. Finally, I want to thank Ahmad Al-Khalifa for being my mentor through my education. His ideas, opinions, and thoughts always guided me to achieve my academic goals. Thank you.

Table of Content

List of Tables

Table **Page**

List of Figures

Figure **Page**

Chapter One: Introduction

Over time, the Saudi National Education Organization ceased using their own local tests for measuring proficiency at the high school level in Saudi Arabia, and replaced them with a test called the "Standard Achievement Admission Test "(SAAT). The test measures students' achievement levels in math, physics, chemistry, and biology (National Center for Assessment in Higher Education, 2016). This test is required for all students in high school in Saudi Arabia in order to apply to universities. When the test began to be used in the country in 2001, there were complaints from private high school students. Many of them disliked the test because they thought it was too difficult to get an acceptable score on it. Students in public high schools did better than students in private high schools on the SAAT (Abrqawie, 2015).

Only one study investigated the strengths and weaknesses of private and public high schools. This study was conducted in 2013 by Al-Garieb in Saudi Arabia. Another study was conducted in 2015 on the structure of the SAAT test and gender differences (Tsaousis, 2015). Neither of the two studies referenced above was published in a journal. In addition, neither of the studies reviewed the SAAT test structure in relationship to the education system itself. No studies have examined whether the test structure differs by school type (public versus private). This study addressed that gap by investigating more deeply the SAAT test structure and examining if the test structure differs by gender and school type. Further, latent mean differences based on school type and gender were investigated to compare the differences in performance based on these groups.

The lack of research on the SAAT makes it important to examine the SAAT test structure. The intent of this study was to provide answers to the problem stated above. Exploratory factor analysis (EFA) and Confirmatory factor analysis (CFA) were used to examine the test structure. The EFA determined how many factors were present and which variables they explained, without making any prior assumptions about measure structure. In addition, the CFA is a measurement model that describes the relationship between indicators and constructs (or factors). The purpose of conducting a CFA in the current study was to evaluate whether the proposed relationships between a set of measured variables was supported by the data (Vehkalahti, 2011). Further, multi-group CFAs were conducted to see if the test structure of the SAAT test differed by school type and gender. Finally, two-way ANOVA was used to measure the differences between gender and school type based on their latent means (Vehkalahti, 2011).

Results of this study are of interest to researchers who are invested in the education system of Saudi Arabia and whether there are gender differences on achievement tests, useful to parents who are concerned about schooling options for their children, and helpful for high school students, who can benefit from this study by helping them to decide which school system is most suitable for them. Also, results are of significant value to policymakers, and program administrators in the National Education Organization in Saudi Arabia, as they plan any necessary changes for better education for public and private high school students based on the findings. Finally, this study may be of value to people who are interested in broader aspects of the Saudi Arabian socio-cultural world.

The research questions for this study were as follows:

1- Does the four-factor model of the SAAT adequately fit the data? And if not, what alternative model has adequate fit?
2- Does the test factor structure differ for public and private schools?
3- Does the SAAT test structure differ by gender?
4- Are there statistically significant differences by school type and gender on SAAT test latent means?

Literature Review

Changes in the education system in Saudi Arabia

In the past decade, the Saudi National Education Organization (2016) has instigated a series of large scale changes in the high school system in Saudi Arabia. For example, in both private and public, males' and females' schools, a requirement was created that asked students to choose from three areas of study. These areas are the literary, the scientific, and the administrative areas of study (Al-Misharim, 2012). Students who choose to do a concentration in the literary area of study, focus more on religion, history, Arabic language, poetry, and geography. For the scientific concentration, students focus on sciences classes, such as biology, chemistry, physics, and mathematics. Finally, students who choose an administrative focus for their studies concentrate more on finance and the banking system in Saudi Arabia (NCA, 2016). With this in mind, students can only apply to specific majors in schools based on their area of choice. Specifically, students who are in the literary area cannot major in science or administration because they do not have enough knowledge in the area. But, students who are in the scientific concentration can apply to science, literary, and administrative majors (2016).

Another significant change occurred in the high school system in Saudi Arabia in 2001, when the Saudi National Education Organization (NCA, 2016) stopped using their own tests for measuring proficiency at the high school level, and replaced them with various tests for different purposes. They used to send their tests to high schools; they sent the final exam for each course that the students in the high school were taking. When the students finished the test, their schools sent the students' answers to the National Education Organization for grading and sent the scores back to the schools. The reason why the Saudi National Education Organization stopped using their test is because they wanted more involvement of the teachers themselves. In other words, they wanted to let teachers write the final exam questions and score the students themselves. By doing that, the teachers might have more confidence in themselves and the students would see their teachers as having a more important role in school (Al-Algpary & Al-Shapani, 2008).

With this in mind, the Saudi National Education Organization replaced their own test with tests called the 'Standard Achievement Admission Test' (SAAT) in the science and literary areas, and the General Aptitude Test (GAT) and Standardized Test of English Proficiency (STEP). These tests were created and administered by the Saudi Arabian National Center for Assessment in Higher Education (NCA, 2017). Each test listed above measures a different area. For example, the SAAT (science) measures student levels in math, physics, chemistry, and biology. Each of the four sections of the test is equally weighted at 25%. The GAT measures a student's analytic and deductive skills. It emphasizes testing the high school student's capacity for learning in general. The test measures the students' abilities with regard to: reading comprehension, recognizing logical relations, solving problems in basic math, inference skills, and measuring capacity. Finally, the STEP test measures students' knowledge of English. It measures students' skills in reading comprehension, sentence structure, listening comprehension, and composition analysis. Each test listed above has a different goal. For example, students who choose to be in a scientific area of study, and want to apply for scientific majors in universities, must take the SAAT (science) (NCA, 2017). Students who want to participate in an English concentration are required by the university to take the STEP test. In addition, the GAT test is required for all students in order to apply to universities in Saudi Arabia (NCA, 2017).

Studies related to the NCA achievement tests.

Some studies about the achievement tests listed above have been conducted and supported by the NCA. The first study investigated the SAAT test structure factor and gender differences on the test (Tsaousis, 2015). Three confirmatory factor models were tested (one factor, four factor, and bifactor). It was found that a bifactor model had the best fit compared to the other models [$\chi2$ =2591.87 (df=188), $p < .001$; CFI = .992; TLI = .990; RMSEA = .014 (C.I. = .014 - .015); SRMR = .010]. Also, the paper stated that the bifactor model showed good fit for both genders. Finally, the latent means gender difference was tested and it showed that women had higher means compared to men on the biology, chemistry, and global achievement latent constructs. On the other hand, men had higher means than women on the physics and mathematics latent constructs (Tsaousis, 2015).

In addition, Tsaousis used the CFA to test the structure of the GAT test. It was found that a two-factor model (i.e., verbal vs. numerical) had good fit to the data [$\chi 2$ =307.39 (df=13), $p < .001$; CFI = .984; TLI = .975; RMSEA = .052 (C.I. = .047 - .057); SRMR = .0214] (Tsaousis, 2014). The correlation between latent factors (i.e., verbal vs. numerical) was $r = .86$, $p = .001$. Next, configural invariance showed that there was good model fit for both genders [$\chi 2$ =302.20 (df=26), p<.000; CFI = .985; TLI = .976; RMSEA = .036 (C.I. = .032 - .040); SRMR = .0349]. Finally, in a test of latent mean differences it was found that males had higher means than females on the numerical domain. On the other hand, females scored higher than males on the verbal domain. At the sub-scale level, it was found that females scored significantly higher than males on the word meaning, sentence completion, and analogy latent domains, while males scored significantly higher than females on the arithmetic and geometry latent domains. No statistically significant difference between males and females was found for reading comprehension and analysis latent domains (Tsaousis, 2014).

Another study investigated whether private or public high school students performed better on the GAT test (Tsaousis, 2014). It was found that a second-order model fit the data better than a bifactor model [$\chi 2$ =1575.92 (df=428), $p < .001$; CFI = .972; TLI = .969; RMSEA = .018 (C.I. = .017 - .019); SRMR = .017]. Also, it was found that students from private schools scored higher than students from public schools on all

GAT latent domains except the arithmetic and geometry latent domains for which there was no statistically significant difference between the two types of schools.

Tsaousis (2014) examined the reliability and validity of the GAT scores for postgraduate students. It was found that the GAT-Post scales (i.e., verbal and quantitative) had acceptable internal consistency indices. Also, there was some evidence to support the validity of the GAT test. Finally, Tsaousis and Sideridis (2014) investigated the STEP test's validity, which showed that overall the STEP had good psychometric properties and had a good blueprint for measuring achievement in language. While some analyses have been conducted with the GAT, no such work has been done with the SAAT.

Private high schools' problems

Currently, more than a half million students in Saudi Arabia attend private schools, which represents almost a fifth of the students in Saudi Arabia (Al-Hagbani, 2013). Families choose to send their children to private schools, and despite this large a population attending this type of institution, we do not know if the private schools are well prepared to educate their students (Al-Hagbani). In fact, in Saudi Arabia, there are issues of concern for parents who send their children to these schools. In a mixed-method study that explored why parents are not satisfied with the private school system in Saudi Arabia, it was reported that more than 71% of parents think that private schools should be less flexible with student's grades, and they should focus more on the student's math and science knowledge (Al-Garieb, 2013). In addition, more than the half of the parents stated that teachers in private schools are less experienced and have less education. Also, 15%

of the parents decided to take their children out of private schools because their children were not educated in basic level English or beyond. Parents interviewed stated that teachers in private schools are not very well qualified to teach their children (Al-Garieb). Another study, using a comparative method, measured the difference between young women students' achievement in English language classes in both public and private universities in Saudi Arabia. It showed that students who graduated from public secondary schools had difficulty learning English at the beginning of their academic career, but once these students learned the English language while attending public schools, their achievement was better than people who graduated from private secondary schools (Deraney & Abdelsalam, 2015).

The efficacy of private secondary schools vs. public secondary schools in the preparation of students for success on tests for university admissions has been illustrated by a statistical table from the NCA (2016). This table shows the top ten highest all-male schools' averages for students' SAAT achievement in Saudi Arabia for the last three years. The information on this table illustrates that a private men's high school in Al-Khobar city had the highest average student SAAT achievement among all of the private and public schools in Saudi Arabia. Also, it was found that another private high school in Riyadh city ranked number three in terms of achievement on the SAAT, among the top 10 schools. Looking closely at the table, one finds that five cities in Saudi Arabia had high average scores on the SAAT for students who attended private secondary schools. On the other hand, public high schools in more than forty cities had high average student scores on the SAAT achievement test. So, while students' score averages on the SAAT test were among the highest in the nation for students who attended private schools, this achievement level was not consistent for the majority of students who attended private schools, while higher test score averages on the SAAT were more consistently produced by students who attended public schools.

Even though there are some studies in the Saudi Arabian context, such as Al-Garieb's (2013) study of "Why Parents Enroll Their Children in Private and International Schools," the research does not offer a robust discussion about students' grades and their level of achievement in private schools, and the reasons why 15% of the parents in the study eventually chose to take their children out of private schools. Furthermore, while the Saudi Arabian National Center for Assessment in Higher Education's table cited above provides information on success rates on standardized tests for both public and private schools, there still is no explanation about why men's private schools have lower test score averages in most of the cities in Saudi Arabia. Though there has been research done on standardized testing for high school students in Saudi Arabia, and this research has included the three tests mentioned above, the SAAT, STEP and GAT, there has not been specific research done on the influence of school type, namely private vs. public, on achievement on the SAAT test.

When the high school system in Saudi Arabia changed in 2001, and the Saudi National Education Organization stopped using their own tests for measuring proficiency at the high school level and replaced it with the SAAT, this policy change was enacted in both the private and public schools' systems. The test has become an issue for high school students in Saudi Arabia because many students in private schools did poorly on the test compared to the students in public schools, who achieved better scores on the test. As mentioned above, averages which illustrate this can be found on the statistical table from the Saudi Arabian National Center for Assessment in Higher Education (2016). Also, the SAAT test was a difficulty for many students who attended private schools, because achieving well on the test was dependent on their knowledge from science courses in school (Al-Ghamdi, 2012). This difficulty was explained by the notion that private schools are known for only considering students' grades and not their knowledge in classes (Al-Garieb, 2013). Ultimately, there have been questions asked about the reasons why students received high GPA's in school, while they did poorly on the SAAT tests.

In addition to this problem, some students have difficulty gaining admission to universities in Saudi Arabia because admissions depend on their scores on the SAAT test (Al-Falih, 2010). Among the changes enacted in 2001 by the Saudi National Education Organization was the policy that universities in Saudi Arabia begin granting admissions to students based on two types of tests, plus evidence of having earned a certain GPA (Al-Falih). These two tests are the SAAT and General Aptitude Test (GAT). Both of the tests are created and administered by a private institution, the NCA (2016). Universities' admission requirements included a percentage for the way that each test and the GPA is weighted in admissions decisions. For example, King Saud University in Saudi Arabia (2015), states that the weighted percentage for materials required for admission would be 40% for the SAAT test, 30% for the GAT test, and the remaining 30% for students' GPA. Students should have a score of 60to 70 out of 100 on the SAAT test in order to be included in the average category on test score. The weighted percentages for materials for entrance to university differ from institution to institution. Some universities require a higher score on the SAAT test in order to meet university admission standards. With this in mind, one might infer that the effectiveness of the school system plays an important role in students' futures because their scores on the SAAT test, which made up to 20 - 50% of their total materials submitted for admissions, depends on the knowledge gained in school (Al-Maliki, 2015). While the studies above are important, the relative lack of research on the SAAT in Saudi Arabia makes it important to examine the test structure itself. This study addressed this gap by investigating more deeply the SAAT test structure and examining if the test structure differs by gender and school type.

Chapter Two: Method

Participants

The SAAT test is administered to about 60,000 high school students per year in Saudi Arabia. The current study utilized data from 14,003 male and female students from both public and private high schools from 2016. Note that the NCA usually have 4 data versions of the test and each version has around 15000 students' responses. This dataset came from one of the 4 versions of the test itself. Specifically, this sample included 7,248 male and 6,755 female students. There were 5,431 male and female students from private high schools and 8,572 from public schools. All students took the SAAT test on paper. These data were delivered by and approved from the NCA organization on February 25, 2017 (Table 1).

Table 1

Description of the High School Student Sample

High school	Male	Female	Total
High school students	7,248	6,755	14,003
Students in private high schools	3,400	2,031	5,431
Students in public high schools	3,848	4,724	8,572

Using SPSS, the data were randomly split into two halves (50%). EFA analysis was used on the first half, and CFA analysis was used on the second half. A total of 1,074 cases had missing data and were excluded, resulting in a final sample size in this study of 12,389. Specifically, the EFA sample included 3,217 male and 3,063 female students. There were 2,429 male and female students from private high schools and 3,851 from public schools (Table 2).

Table 2

Descriptions of Students in the EFA Sample

High school	Male	Female	Total
High school students	3,217	3,063	6280
Students in private high schools	1508	921	2,429
Students in public high schools	1709	2142	3,851

As for the CFA sample, it contained 3,144 male and 2,965 female students. There were 2,416 male and female students from private high schools and 3,693 from public schools (Table 3).

Table 3

Descriptions of Students in the CFA Sample

High school	Male	Female	Total
High school students	3,144	2,965	6,109
Students in private high schools	1,518	898	2,416

Students in public high schools	1,626	2,067	3,693

Instrument

The Standard Achievement Admission Test (SAAT) was created by the NCA, which is the agency in charge of administering the test. It is an 88-item (multiple choice questions) admission test that involves four sections: Biology, Chemistry, Physics, and Mathematics (Table 4). Specifically, the SAAT test emphasizes the material of the official three-year (scientific) curriculum of the Saudi High Schools. Each of the four sections is worth 25% of the test score. The items on the SAAT are issued as follows: 20% of each school subject for the first year of the high school syllabus and 40% of each subject for the second and third years of the high school syllabus. The test is approximately two hours long, which means that students have only 25 minutes to finish a section of the SAAT test, and the students are not allowed to use an electronic calculator because the test questions do not require many complex calculations. There are two options for taking the test: students can take the paper version of the SAAT test or they can take the computerized version. (Examples of the test questions are provided in Appendix A.) Note that this current study only included students who took the paper version test.

The NCA sends the test to several schools in each city in Saudi Arabia, where high school students can take it. Each school has a team of people to deliver it to the students. The team collects the SAAT test answers when the students are done and sends them back to the NCA. After receiving the students' answers on the SAAT test, the NCA organization enters answers for each question into the SPSS program. Scores are analyzed by experts to make sure that the test has support for reliability and validity, and then the final test scores are made available for students to see online. It takes around three weeks for students to receive scores after the test is taken.

Table 4

SAAT Items by Section

Items	Section	Items Per Section
i 1 to i 24	Biology	24 Items
i 25 to i 44	Chemistry	20 Items
i 45 to i 64	Physic	20 Items
i 65 to i 88	Math	24 Items
Total	4 sections	88 Items

Procedure

An application was submitted to the University of Denver Institutional Review Board (IRB). The IRB committee decided that there was no need for a committee review on February 22, 2017.

Analysis

Exploratory Factor Analysis (EFA)

Principal axis factoring with varimax rotation was used to determine the likely measure structure, which was then tested later with a CFA (Tabachnick & Fidell, 2013).

Specifically, varimax rotation was used to simplify the factors through the differences between large and small loadings within each factor (Tabachnick & Fidell, 2013). In addition, a parallel analysis was used to determine the maximum number of factors that should be retained in the EFA. Also, a cut-off value of .30 for item loadings in the EFA was used.

The reason why the EFA was used initially was to determine the factor structure from the EFA to confirm with the CFA sample. However, the four-factor theoretical model proposed by the NCA was also confirmed with the CFA sample. Based on whether the empirical structure derived from the EFA or the theoretical structure had the better fit, measurement invariance and ANOVA were conducted. With this in mind, if the EFA empirical structure and the four-factor theoretical model showed similarity, and were confirmed with the CFA, it means that the SAAT test is measuring what it is supposed to measure and the theoretical structure is clearly supported in empirical test response data. However, if EFA and theoretical solutions were different, confirmation with the CFA indicated which structural model was more appropriate. Failure of data fit in the CFA to either model might mean that the test itself has a validation problem.

Confirmatory Factor Analysis (CFA)

Confirmatory factor analysis (CFA) was used to confirm measure structure. This measurement model describes the relationship between indicators and constructs (or factors). The purpose of conducting a CFA is to evaluate whether the proposed relationships among a set of measured variables are supported by the data. Specifically, a CFA four-factor model was tested (Biology, Chemistry, Physics, and Mathematics) to see if the model had a good fit in the data. Specifically, I examined the root mean square error of approximation (RMSEA), where I want the RMSEA to be less than .05 in order to have a good fit, or at least a reasonable fit from .05 to .08 (Vehkalahti, 2011). Also, the Comparative Fit Index (CFI) and Tuker-lewis Index (TLI) should be equal to or higher than .90 to show good fit (Bentler, 1990). Finally, the Standardized root mean square residual (SRMR) of less than .05 shows good model fit. If the SRMR is less or equal than 0.08 it means that there is reasonably good fit (Vehkalahti, 2011).

In addition, multi-group CFA analysis was examined to see if the factor structure of SAAT test differed by school type and gender (research questions 2 and 3). A test of measurement invariance was investigated in the following hierarchical order for nested models: configural invariance, metric invariance, and scalar invariance (Brown, 2006). Deciding on which test had a better fit, the *p*-value significance level of chi-square difference test between the three measurement invariance models were considered (Brown, 2006).

ANOVA

Two-way ANOVAs were used to determine the significance of mean differences associated with gender and school types and their interaction in terms of the four latent construct factors (biology, chemistry, physics, and math).

Programs

IBM 24 SPSS was used for to run the EFA analysis (Vehkalahti, 2011), and Rstudio was used to run the both CFA models and ANOVA analyses (Albert & Rizzo, 2012).

Chapter Three: Result

This section includes results in order as follow: first, the EFA results; second, the CFA results including the measurement invariance tests; third, latent means differences by gender and school type using ANOVA.

Exploratory Factor Analysis (EFA)

Results of a EFA are reviewed in four ways—factorability, number of factors (scree plot, parallel analysis, eigenvalues and variance), factor loadings after factor rotation, and factor retention. These are described below.

Factorability

Principal Axis Factoring (PAF) was used to determine how many factors were present and which variables they explained, without making any prior assumptions about the data (Bentler, 1990). The results showed that the determinant for factor analysis was -0.00002 which is a nonzero value, indicating that the dataset was factorable. Moreover, according to the Kaiser-Meyer-Olkin measure (.972) and Bartlet's Test of Sphericity (χ^2 (67308.68) = 3828, $p < .001$), the dataset with these items was factorable.

Scree Plot

A scree plot (Figure 1) displays the eigenvalues associated with each component in descending order versus the number of the components or factors (Tabachnick & Fidell, 2013). The scree plot indicates that there might be three factors underlying the

SAAT test because the leveling off of the eigenvalues curve starts with factor three (Figure 1). While the scree plot provides insight about the number of factors to anticipate, the researcher needs to consider other approaches to understanding the factor structure (Tabachnick & Fidell, 2013). Kaiser's rule which is based on the rule of thumb of interpreting factors with eigenvalues greater than 1.00 showed that the SAAT test consisted of possibly 20 factors. A parallel analysis was conducted to determine the maximum number of likely interpretable factors before making a final decision.

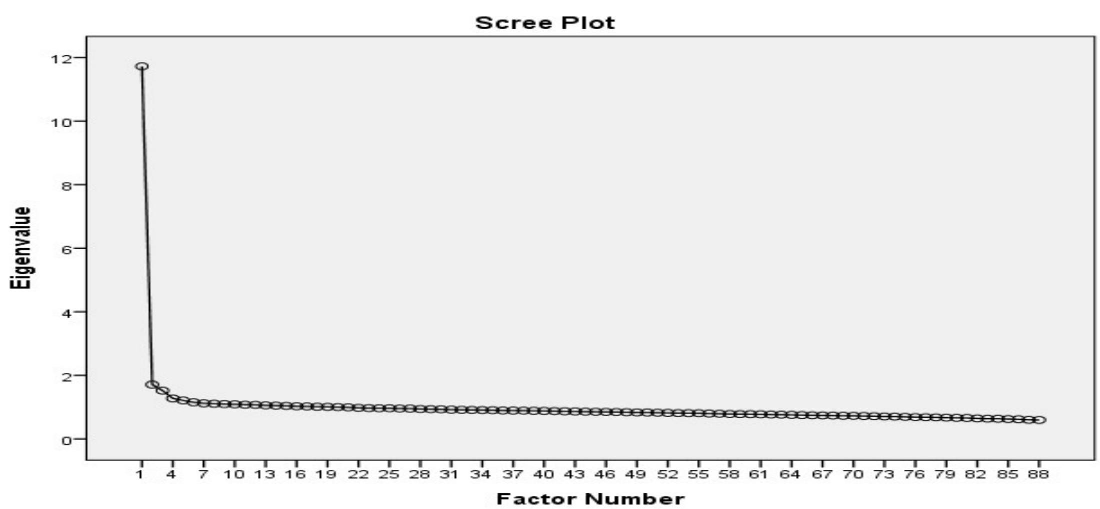

Figure 1. Scree plot of The SAAT test.

Parallel Analysis

After conducting a parallel analysis, and making a comparison between the parallel analysis eigenvalues and the eigenvalues from the sample data, results revealed four factors. The eigenvalue estimate for factor 1 was 11.725 versus 1.26 from simulated data and for factor 2 was 1.710 versus 1.24 (Table 5). The eigenvalues for the first two factors were greater than the parallel analysis estimates from simulated data. But only two interpretable factors were retained. Specifically, when the four factors were tested in the EFA, it showed that factor 4 had only one item that loaded substantially on the factor. Because of that factor 4 was deleted, and the items loading on the other three factors were examined. Running the three factors yielded 31 items with loadings exceeding .30 out of 88. However, two factors yielded 40 items loading above .30 and 48 items were removed. Therefore, a two-factor solution was considered to be most interpretable.

Table 5

Comparison between the Parallel Analysis Estimate and Eigenvalues for the Sample Data.

Factor	EFA Eigenvalue	Parallel Analysis Eigenvalue	Factor supported by EFA and parallel analysis (yes or no)
1	11.725	1.263	Yes
2	1.710	1.240	Yes
3	1.521	1.227	No
4	1.274	1.220	No
5	1.211	1.211	No
6	1.153	1.276	No
7	1.120	1.263	No
8	1.107	1.240	No
9	1.096	1.227	No
10	1.089	1.220	No
11	1.077	1.211	No
12	1.072	1.197	No
13	1.056	1.192	No
14	1.049	1.189	No
15	1.041	1.174	No
16	1.027	1.168	No
17	1.023	1.163	No
18	1.013	1.154	No
19	1.009	1.149	No
20	1.002	1.144	No

Eigenvalues and Variance

Using EFA with varimax rotation, two factors were initially identified, and accounted for 15.27% of the total variance explained for the SAAT test (Table 6).

Specifically, factor 1 explained 13.32 % of the variance and had an eigenvalue of 11.73. Factor 2, explained 1.94 % of the variance and had an eigenvalue of 1.71. Note that the two factors had eigenvalues (a measure of explained variance) greater than 1.0, which is a common criterion for the factors to be useful (Tabachnick & Fidell, 2013). In short, eigenvalues, percent variance, and parallel analysis results supported the possibility of two factors.

Table 6

Total Variance Explained After Rotation.

Component	Total	Percentage of Variance	Percentage of Cumulative
1	11.725	13.324	13.324
2	1.710	1.943	15.267

Factor Loadings

Factor number 1

A cut-off value of .30 for item loadings was used (Tabachnick & Fidell, 2013). Factor loadings for factor 1 ranged from .49 (fair) to .30 (poor) (Comrey & Lee, 1992). The first factor had a total of 28 items. Specifically, this factor had 10 items intended to assess achievement in biology, 5 items in chemistry, 10 items in physics, and 3 in math. To examine cross-loading, a 0.20 difference between factor loadings was considered (Tabachnick, & Fidell, 2013). There were four items that had cross-loadings with a second factor (items 8, 26, 61, and 40). The results showed that the majority of items loading on the first factor measured students' knowledge in biology as well as physics.

After deleting the cross-loading items, the range of the factor loadings did not change. Now, the factor had a total of 24 items, and these items were as follows: 9 items in biology, 4 items in chemistry, 8 item in physics, and 3 in math. With this in mind, factor 1 measured primarily students' knowledge in biology (Table 7).

Table 7

Items by Content Area for Factor One After Deleting Cross-Loading Items

SAAT Test Sections	Biology	Chemistry	Physics	Math	Total of Items	Deleted cross-loading-items
Number of Items	9	4	8	3	24	4

Factor number 2

Factor loadings for factor 2 ranged from .46 (fair) to .30 (poor) (Comrey & Lee, 1992). The second factor had a total of 19 items. Specifically, this factor had 4 items intended to assess achievement in biology, 9 items in chemistry, and 6 items in math. There were no items intended to assess physics achievement. There were two items that had cross-loadings with the first factor (items 9 and 86). The results showed that the majority of items loading on the second factor measured students' knowledge in chemistry.

After deleting the cross-loading items, the range of the factor loadings did not change. Now, the factor had a total of 16 items, and these items were as follows: 3 items in biology, 8 items in chemistry, and 3 in math. There were no items intended to assess physics achievement. Note that item 27 in chemistry did not load, therefore the item was deleted from the factor itself. With this in mind, factor 2 measured primarily students' knowledge in chemistry (Table 8).

Table 8

Number Items on Factor Two After Deleting Cross-Loading Items

SAAT Test Sections	Biology	Chemistry	Physics	Math	Total of Items	Deleted cross-loading Items
Number of Items	3	8	0	5	16	2

Factor Retention

Two factors were retained from this analysis – factor 1 (Biology) and factor 2 (Chemistry).

Confirmatory Factor Analysis (CFA)

The default estimator of the confirmatory factor analysis (CFA) is maximum likelihood (ML), but this estimator requires multivariate normality for the observed items (indicators). This assumption in the majority of the cases is not fulfilled when the items are ordered or nominal (categorical). In other words, ML is used when the data is continuous (Brown, 2006). When the data are not continuous, Brown (2006) suggested that the robust weighted least squares, WLSMV, should be used. The data in this study were binary. Therefore, the WLSMV approach was used.

Question One

Test of the Structure of the SAAT

To answer question one **"**Does the four-factor model of the SAAT adequately fit the data? And if not, what alternative model has adequate fit?" A CFA four-factor model (this is the theoretical model suggested by the NCA) was tested to see if the model has a good fit. Figure 2 shows the four-factor model.

The results showed that the four-factor model (Biology, Chemistry, Physics, and Mathematics) fit the data very well compared to the two-factor model (this model was built based on the EFA structure) because it had the following statistics: [(χ^2 (8648.710) = 3734, *p*< .001), CFI = .0.971, TLI = 0.970, RMSEA = 0.015, WRMR = 1.44; SRMR = 0.029 (Table 9). Specifically, the four-factor model had CFI > 0.95 and TLI > 0.95, RMSEA < 0.05, WRMR > 1, and SRMR < 0.05, which means that the criteria for a good fitting model was met by the model (Vehkalahti, 2011). In this study, the chi-square test was significant for the model, indicating that the observed covariance matrix did not equal the estimated covariance. Given the large sample size in this study, this finding was expected because the chi-square test is sensitive to sample size (Brown, 2006).

In addition, Cronbach's alpha was estimated for each factor to examine the level of internal consistency. The results for factor one gave us a Cronbach's Alpha of 0.79, factor two was 0.75, factor three was 0.68, and factor four was 0.76 . Note that all of the factors had an acceptable level of internal consistency (Lance, Butts, & Michels, 2006).

Based on these statistics, measurement invariance and ANOVA were determined for the four-factor model.

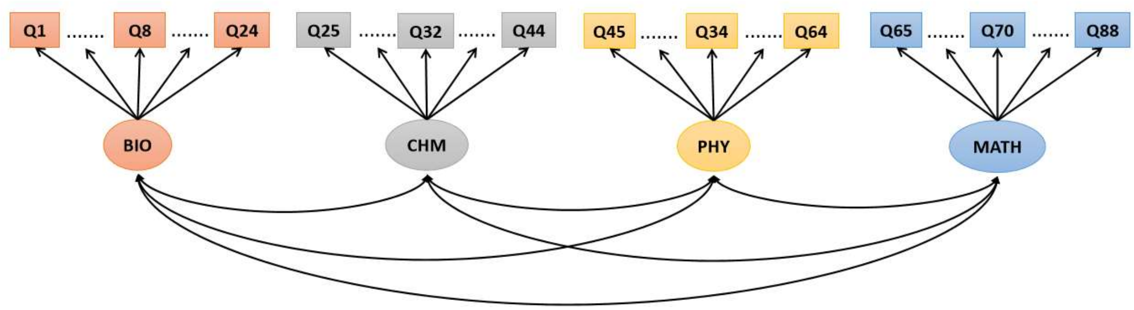

Figure 2. Four Factor Model

Table 9

Goodness of Fit comparison between Two-Factor and Four-Factor Models

Model	χ2	df	p	RMSEA	CFI	TLI	WRMR	SRMR
Four-factor	8648.710	3734	<0.001	0.015	0.971	0.970	1.44	0.029
Two-factor	4439.995	739	<0.001	0.029	0.947	0.944	2.13	0.042

Question *2*

To answer question two "Does the test factor structure differ for public and private schools?" tests of measurement invariance of the SAAT across school type were performed in the following hierarchical ordered of nested models: configural invariance, metric invariance, and scalar invariance (Brown, 2006). In this study, the configural invariance model, all parameters (factor variance, factor covariance, factor loading, indicator error variance) were free to vary across public and private groups. In the metric invariance, all factor loadings were equal across groups, where factor variance, factor covariance, and indicator error variance were free to vary across public and private groups. In scalar invariance, the loadings and intercepts were constrained to be equal across groups, where factor variance, factor covariance, and indicator error variance were free to vary across groups. The results for each invariance test are provided in Table 10, where all the invariance types of the SAAT across school type showed a good fit. The significance level of chi-square test differences between the three measurement invariance tests were examined to determine which model fit the sample in this study best (Brown, 2006).

Test of Configural Invariance

To test whether the four-factor model showed configural invariance across school type, its model fit indices were evaluated. The results showed that the four-factor model had a good fit [χ^2 (11614.295) = 7468, $p < .001$), CFI = .0.974, TLI = 0.973, RMSEA = 0.013, WRMR = 1.683; SRMR = 0.034] (Table 10). This means that the SAAT factor structure holds consistent for public and private school students. Therefore, configural invariance was achieved.

Test of Metric Invariance

A metric invariance test was examined to see if the factor loadings were equal across schools' types (Brown, 2006). The results showed that the four-factor model had good fit [χ^2 (11587.410) = 7552, $p < .001$), CFI = .0.974, TLI = 0.977, RMSEA = 0.013, WRMR = 1.796; SRMR = 0.036] (Table 13). This means that public and private schools had similar factor loadings. Therefore, metric invariance was achieved. Based on chi square difference test [(χ^2 26.885) = 84, $p >$0 .05], this researcher determined that the configural invariance test fit the data better than metric (Table 10). Note that the degree of freedom of the configural invariance model was less than metric because it was more complex. Specifically, the metric invariance model had the loading constrained across groups. On the other hand, configural had all the parameters (factor variance, factor covariance, factor loading, indicator error variance) free to vary across public and private groups. In addition, configural invariance model showed that both groups (public and private schools) have statistically significant estimates ($p < .001$). The variance explained by the items in the public group for the different items ranged between about 48.3% and about 2.9%. On the other hand, the variance explained by the private group ranged between 52% and 1.4%. Specifically, item 86 explained the highest variability in math for both the public (48.3%) and private (52%) groups.

Test of Scalar Invariance

A test of scalar invariance was examined to see if the factor loadings and intercepts were equal across school types (Brown, 2006). The result showed that the four-factor model had good fit [χ^2 (12098.282) = 7548, $p < .001$), CFI = .0.971, TLI = 0.971, RMSEA = 0.014, WRMR = 1.732; SRMR = 0.034] (Table 13). Therefore, scalar invariance was achieved. In addition, a comparison between scalar and metric invariance tests showed that metric invariance fit the data better than scalar [(χ^2 (510.872) = 4, $p < .0001$)]. This implies that the metric model was best suited for the sample (Table 10). Note that the degree of freedom of the metric invariance model was more than scalar because the model itself was less complex than scalar. Also, it was found that the metric model had chi square that was less than the other invariance models. Specifically, the metric invariance model had the loading constrained across groups. On the other hand, scalar had loadings and intercepts equal across school types. In addition, metric invariance showed that both groups (public and private schools) have statistically significant estimates (p < .001). The variance explained by the items in the public group for the different items ranged between about 47.5% and about 2%. On the other hand, the variance explained by the private group ranged between 53.2% and 2.4%. Specifically, item 86 explained the highest variability in math for both the public (47.5%) and private (53.2%) groups.

Table 10

Tests of Measurement Invariance for the SAAT test across School Type

Model	**$\chi 2$**	**df**	***P*-value**	**$\Delta\chi 2$**	**Δ*df***	**P-value**	**CFI**	**TLI**	**RMSEA**	**WRMR**	**SRMR**
Configural	11614.295	7468	——	——	—	—	0.974	0.973	0.013	1.683	0.034
Metric	11587.410	7552	<0.001	26.885	84	>0 .05	0.974	0.974	0.013	1.796	0.036
Scalar	12098.282	7548	<0.001	510.872	4	<0.001	0.971	0.971	0.014	1.732	0.034

Question 3

For question 3 which is **"**Does the test factor structure differ for gender," a test of measurement invariance of the SAAT across gender were performed in the same way as for question 2, where configural invariance, metric invariance, and scalar invariance were examined. Note that in this study, the configural invariance model, all parameters (factor variance, factor covariance, factor loading, indicator error variance) were free to vary across gender groups. In the metric invariance, all factor loadings were equal across groups, where factor variance, factor covariance, and indicator error variance were free to vary across groups. In scalar invariance, the loadings and intercepts were constrained to be equal across groups, where factor variance, factor covariance, and indicator error variance were free to vary across groups.

Test of Configural Invariance

To test whether the four-factor model showed configural invariance across gender, a model fit test was conducted. The result showed that the four-factor model had a good fit [χ^2 (11130.101) = 7468, $p < .001$), CFI = .0.974, TLI = 0.973, RMSEA = 0.013, WRMR = 1.647; SRMR = 0.034]. This means that gender had a similar factor structure. Therefore, configural invariance was achieved (Table 11).

Test of Metric Invariance

A test of metric Invariance was examined to see if the factor loadings were equal across gender (Brown, 2006). The result showed that the four-factor model had good fit to the data [χ^2 (13292.588) = 7552, $p < .001$), CFI = .0.959, TLI = 0.958, RMSEA = 0.016, WRMR = 1.961; SRMR = 0.040] (Table 14). This means that gender groups had similar factor loadings. Therefore, metric invariance was achieved. Based on chi square difference test [(χ^2 (2162.4)= 84, $p < .0001$, this researcher determined that the metric invariance test fit the data better than configural (Table 11). Note that the degree of freedom of the configural invariance model was less than metric because it was more complex. Metric model showed that both groups (male and female) had statistically significant estimates (p < .001). The variance explained by the items in the female group for the different items ranged between about 44% and about 1.3%. On the other hand, the variance explained by the male group ranged between 51% and 1.7%. Specifically, item 86 explained the highest variability in math for both the female (44%) and male (51%) groups.

Test of Scalar Invariance

A test of scalar invariance was examined to see if the factor loadings and intercepts were equal across gender (Brown, 2006). The result showed that the four-factor model had good fit [χ^2 (12983.917) = 7548, $p < .001$), CFI = .0.961, TLI = 0.961, RMSEA = 0.015, WRMR = 1.761; SRMR = 0.034] (Table 14). Therefore, scalar invariance was achieved. In addition, a comparison between scalar and metric invariance tests showed that metric invariance fit the data batter than scalar [(χ^(308.671) = 4, $p <$ .0001)]. This implies for this sample, a metric model showed better fit (Table 11). Note that the metric invariance model is less complex than scalar.

Table 11

Tests of Measurement Invariance for the SAAT Test across Gender

Model	$\chi2$	df	*P*-value	$\Delta\chi2$	Δdf	P-value	CFI	TLI	RMSEA	WRMR	SRMR
Configural	11130.101	7468	<0.001	----	----	----	0.974	0.973	0.013	1.647	0.034
Metric	13292.588	7552	<0.001	2162.487	84	<0.001	0.959	0.958	0.016	1.961	0.040
Scalar	12983.917	7548	<0.001	308.671	4	<0.001	0.961	0.961	0.015	1.761	0.034

Test of Latent Mean Differences

Two-way ANOVAs were used to determine if there were statistically significant differences between means by gender, school type, and their interaction in terms of the four-latent constructs (biology, chemistry, physics, and math). The results of the analysis for each factor are discussed below.

Biology

The latent means on the biology section on the SAAT test showed that there was a statistically significant difference for gender, $F(1,6105) = 603.04$, $p < 0.001$, and school type, $F(1,6105) = 6.73$, $p < 0.001$. Also, the interaction was statistically significant, $F(1,6105) = 14.96$, $p < 0.001$ (Table 12 and Figure 3). Therefore, multiple comparisons with a *Tukey* correction for inflation of Type I error was used for simple follow-up tests. Results showed that female students had higher mean scores on average than male students in private and public schools (Table 13). There was no difference between the female students in private and public schools ($p > 0.05$) (Table 14). On the other hand, the male students in private school had higher scores on average than males in public school.

Table 12

Two-way ANOVAs for Biology

Source	Sum of Squares	*df*	*F*	*p*	η^2
Gender	151.205	1	603.04	<0.001	0.0896
Schooltype	1.69	1	6.73	0.0095	0.0010
Gender*Schooltype	3.75	1	14.96	<0.001	0.0022
Residuals	1530.76	6105	NA	NA	NA

Table 13

*Gender*Schooltype Means for Biology with 95% Confidence Interval*

Gender	Schooltype	lsmean	SE	*df*	Lower.CL -	Upper.CL
Male	Public	-.180	.012	6105	-.211	-.149
Male	Private	-.099	.013	6105	-.131	-.067
Female	Private	.157	.017	6105	.115	.198
Female	Public	.180	.011	6105	.153	.207

Table 14

Multiple Comparisons with Tukey Correction for Biology

Contrast	Estimate	SE	*df*	*t*	*p*	Cohen's d
Male,public - Female,public	-.36	.017	6105	-21.69	<0.001	0.754
Male,public - Male,private	-.08	.018	6105	-4.51	<0.001	0.154
Male,public - Female,private	-.34	.021	6105	-16.18	<0.001	0.685
Female,public - Male,private	.28	.017	6105	16.50	<0.001	0.551
Female,public - Female,private	.02	.020	6105	1.16	0.654	0.049
Male,private - Female,private	-.26	.021	6105	-12.15	<0.001	0.480

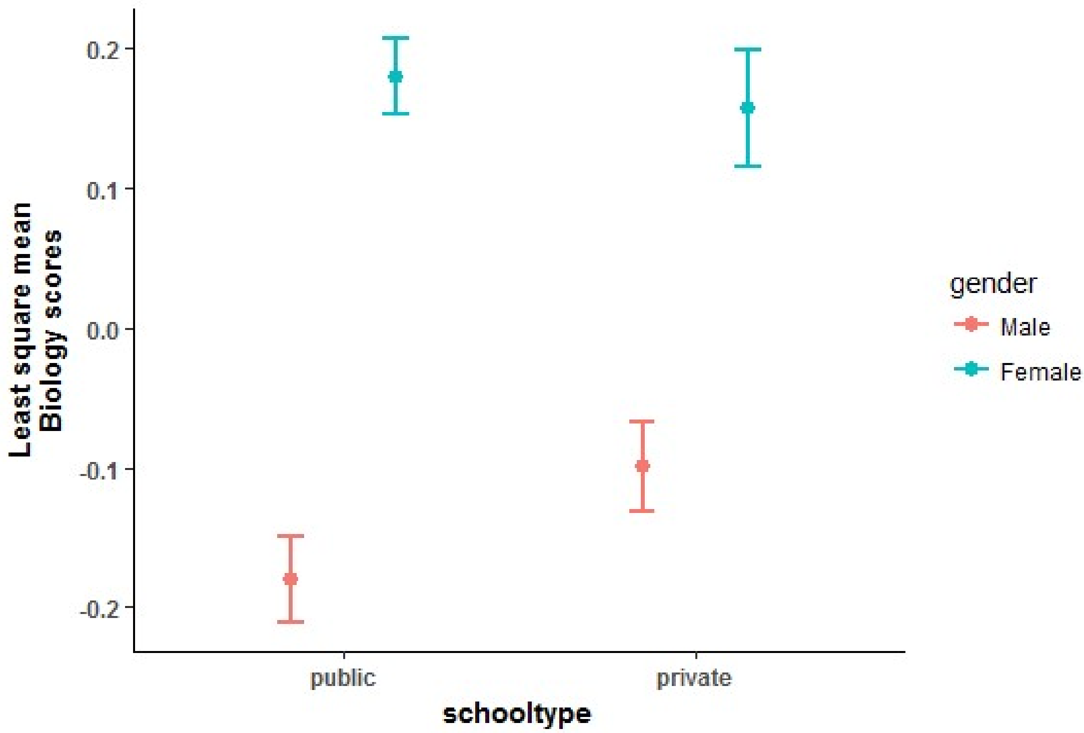

*Figure 3. Interaction between Gender*Schooltype for Biology*

Chemistry

The latent means on the chemistry section on the SAAT test showed that there was a statistically significant main effect of gender, $F(1, 6105) = 517.36$, $p < 0.001$, and school types $F(1, 6105) = 4.70$, $p < 0.05$. Also, the interaction was statistically significant, $F(1,6105) = 15.61$, $p < 0.001$ (Table 15 and Figure 4). Therefore, multiple comparisons with a *Tukey* correction for inflation of Type I error was used for follow-up tests. Based on group means the results showed that female students had higher mean scores on average than males in both private and public schools (Table 16). But there was no difference between the female students in private and public schools ($p > 0.05$) (Table 17). On the other hand, male students in private school had higher mean scores than the males in public school.

Table 15

Two-way ANOVAs for Chemistry

Source	Sum of Square	*df*	*F*	*P*	η^2
Gender	55.947	1	517.36	<0.001	0.0779
Schooltype	0.508	1	4.70	0.0302	0.0007
Gender*Schooltype	1.688	1	15.61	<0.001	0.0024
Residuals	660.195	6105	NA	NA	NA

Table 16

*Gender*Schooltype Means for Chemistry with 95% Confidence Interval*

Gender	Schooltype	lsmean	SE	df	Lower.CL -	Upper.CL
Male	Public	-.109	.008	6105	-.129	-.089
Male	Private	-.059	.008	6105	-.080	-.038
Female	Private	.093	.011	6105	.065	.120
Female	Public	.113	.007	6105	.095	.131

Table 17

Multiple comparisons with Tukey Correction for Chemistry

Contrast	Estimate	SE	*df*	*t*	*p*	Cohen's d
Male,public - Female,public	-.221	.011	6105	20.315	<0.001	0.708
Male,public - Male,private	-.050	.012	6105	-4.249	<0.001	0.145
Male,public - Female,private	-.202	.014	6105	14.752	<0.001	0.624
Female,public - Male,private	.172	.011	6105	15.436	<0.001	0.515
Female,public Female,private	.020	.013	6105	1.503	0.4358	0.063
Male,private - Female,private	-.152	.014	6105	10.967	<0.001	0.432

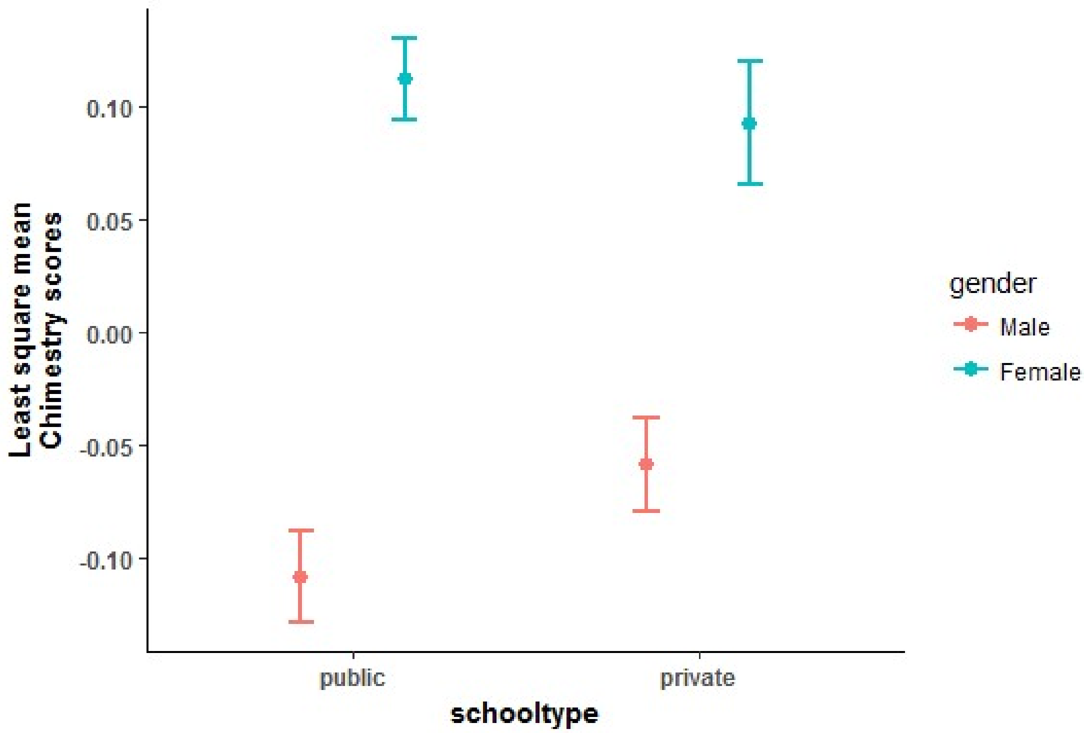

*Figure 4. Interaction Gender*Schooltype for Chemistry*

Physics

The latent means on the physics section on the SAAT test showed that there was a statistically significant main effect of gender $F(1, 6105) = 462.79$, $p < 0.001$ and school type $F(1,6105) = 8.16$, $p < 0.05$. Also, the interaction was statistically significant, $F(1, 6105) = 13.62$, $p < 0.001$ (Table 18 and Figure 5). In the interaction, the results showed that female students had higher scores on average than males in both private and public schools (Table 19). But there was no difference between the female students in private and public schools ($p > 0.05$) (Table 20). On the other hand, male students in private schools had higher mean scores than the males in public schools.

Table 18

Two-way ANOVAs for Physics

Source	**Sum of Square**	***df***	***F***	***P***	**η^2**
Fender	57.899	1	462.79	<0.001	0.0703
Schooltype	1.021	1	8.16	0.0043	0.0012
Gender*Schooltype	1.703	1	13.62	<0.001	0.0021
Residuals	763.783	6105	NA	NA	NA

Table 19

*Gender*Schooltype Means for Section with 95% Confidence Interval*

Gender	**Schooltype**	**lsmean**	**SE**	**df**	**Lower.CL -**	**Upper.CL**
Male	Public	-.113	.009	6105	-.135	-.092
Male	Private	-.056	.009	6105	-.078	-.033
Female	Private	.099	.012	6105	.070	.129
Female	Public	.111	.008	6105	.092	.131

Table 20

Multiple Comparisons with Tukey Correction for Physics

Contrast	**Estimate**	**SE**	***df***	***t***	***p***	**Cohen ’s d**
Male,public - Female,public	-.225	.012	6105	-19.185	<0.001	0.668
Male,public - Male,private	-.058	.013	6105	-4.589	<0.001	0.156
Male,public - Female,private	-.213	.015	6105	-14.479	<0.001	0.612
Female,public - Male,private	.167	.012	6105	13.969	<0.001	0.467
Female,public-Female,private	.012	.014	6105	.849	0.8307	0.036
Male,private - Female,private	-.155	.015	6105	-10.409	<0.001	0.410

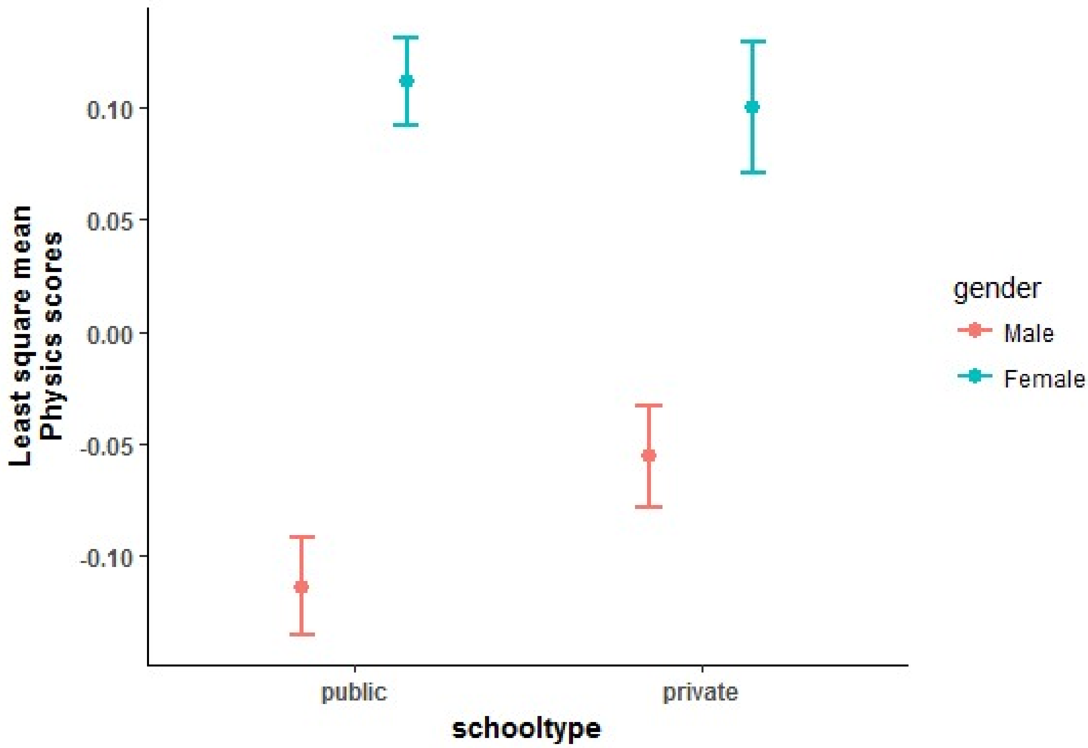

*Figure 5. Interaction Gender*Schooltype for Physics*

Math

The latent means on math section on the SAAT test showed that there was a significant main effect of gender $F(1,6105) = 380.97$, $p < 0.001$ and school type $F(1,6105) = 8.36$, $p < 0.05$. Also, the interaction was statistically significant, $F(1,6105) = 11.780$, $p < 0.001$ (Table 21 and Figure 6). Results showed that female students had higher mean scores than males in both private and public schools (Table 22). But there was no difference between the female students in private and public schools ($p > 0.05$) (Table 23). On the other hand, male students in private school had higher scores on average than the males in public school.

Table 21

Two-way ANOVAs for Math

Source	Sum of Square	*df*	*F*	*P*	η^2
Gender	51.529	1	380.97	<0.001	0.0586
Schooltype	1.131	1	8.36	0.0039	0.0013
Gender*Schooltype	1.593	1	11.78	<0.001	0.0018
Residuals	825.750	6105	NA	NA	NA

Table 22

*Gender*Schooltype Interaction for Math with 95% Confidence Interval*

Gender	Schooltype	lsmean	SE	df	Lower.CL -	Upper.CL
Male	Public	-.107	.009	6105	-.130	-.084
Male	Private	-.049	.009	6105	-.072	-.025
Female	Private	.096	.012	6105	.066	.127
Female	Public	.106	.008	6105	.086	.126

Table 23

Multiple Comparisons with Tukey Correction for Math

Contrast	Estimate	SE	*df*	*t*	*p*	Cohen 's d
Male,public - Female,public	-.213	.012	6105	-17.459	<0.001	0.608
Male,public - Male,private	-.058	.013	6105	-4.443	<0.001	0.151
Male,public - Female,private	-.204	.015	6105	-13.310	<0.001	0.563
Female,public - Male,private	.155	.012	6105	12.430	<0.001	0.415
Female,public-Female,private	.009	.015	6105	.634	0.9209	0.027
Male,private - Female,private	-.145	.016	6105	-9.378	<0.001	0.369

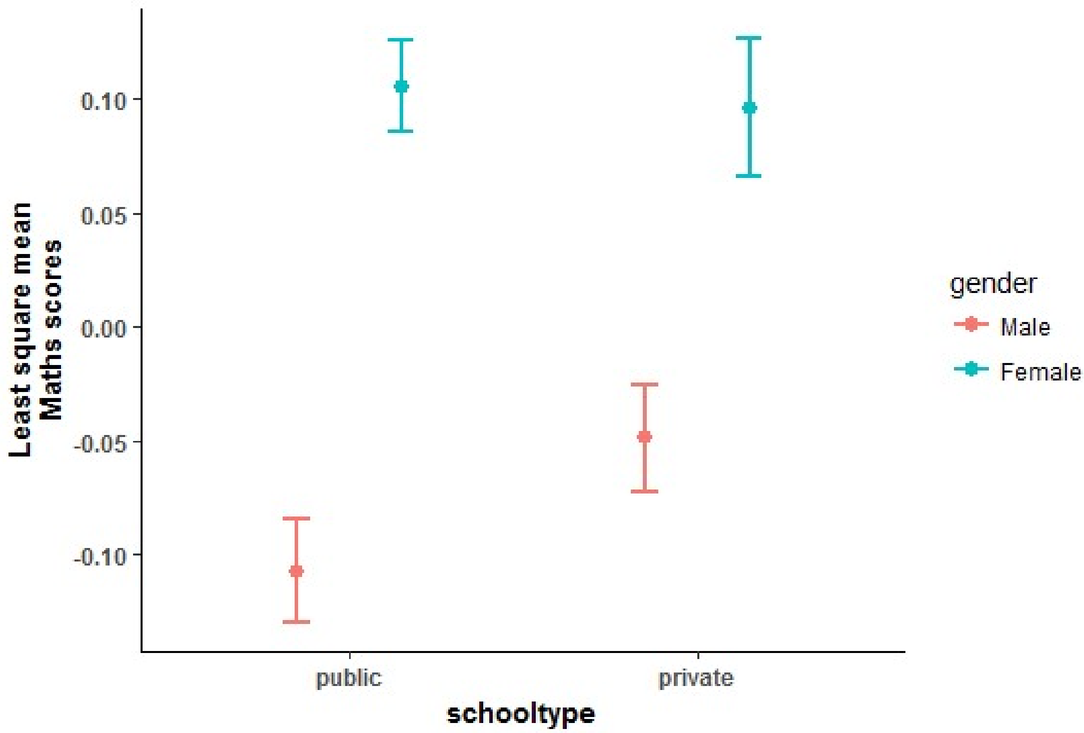

*Figure 6. Interaction Gender*Schooltype for Math*

Discussion

Comparison between the EFA and CFA Theoretical Models

It was found that there were no similarities between the EFA structure and the four-factor model. Specifically, the EFA structure showed that the SAAT test is measuring two factors. The first factor mostly measured students' knowledge in biology, whereas the second factor measured students' knowledge in chemistry. Note that the word "mostly" was used, which means that there are still other items from different sections that were in biology (factor 1) and chemistry (factor 2). On the other hand, the four-factor model of the SAAT that was suggested by the NCA had four factors as follows: biology, chemistry, physics, and math. In addition, the number of items that loaded on both factors in EFA was only 40. On the other hand, the four-factor model of the SAAT had 88 items. Therefore, the EFA result did not support the hypothesized CFA test structure. In other words, the EFA was not useful for this analysis because the two EFA factors cannot be directly compared with the CFA four-factor model and the EFA resulted in deletion of over half the items.

Even though the EFA analysis did not support the hypothesized CFA test structure, a question might be asked "why did only 40 out of 88 items (less than half) load adequately on a factor in the EFA? "Therefore, a recommendation to the NCA is to create a study where they measure the psychometric quality and validity of the SAAT test to make sure that the test is measuring what it is supposed to measure.

SAAT Test Structure

According to Vehkalahti (2011), a fitting CFA model has to meet particular criteria. These criteria are as follows: Root Mean Square Error of Approximation (RMSEA) must be less than .05, the Comparative Fit Index (CFI) and Tuker-Lewis Index (TLI) must be equal to or higher than .90, and the Standardized Root Mean Square Residual (SRMR) has to be less than .05. In question one, results showed that the CFA four-factor model of the SAAT fit the data better than the two-factor model. Specifically, the four-factor model met all the criteria above.

In addition, in the Tsaousis (2015) study, it was found that the four-factor model of the SAAT test fit the data, so results here can be considered as additional evidence that the four-factor model fit the data well for the SAAT test.

Multi-Group CFA

Three measurement invariance tests–configural invariance (equal factors), metric invariance (equal loadings), and scalar invariance (equal loadings and intercept)– across school types and gender were conducted. It showed that both groups (school type and gender) achieved the criteria above of having good model fit (Vehkalahti, 2011). Therefore, all three levels of measurement invariance were achieved. In school type, configural invariance fit better than metric, while metric was better than scalar. This indicated that the researcher will decide on which model fits best depending on research question. Based on our research questions, we focused on the metric invariance model results. In addition, based on the metric invariance model, it appeared that item 86 explained 47.5% in mathematics for public schools and 53.2% for private schools. More variability on math students' knowledge in private schools was accounted for by item 86 than that of students in public schools. In gender, when the metric invariance test was compared with configural, it showed that the data fit better with metric. However, when metric was compared with scalar, it showed that metric fit better. Also, it showed that item 86 explained the highest variability in math for both the female (44%) and male (51%) groups. In other words, we can say that item 86 measured the male students' knowledge of mathematics better than it did for females.

In addition, the prior study that was conducted at the NCA did not use the chi square test as a comparison of significance between the three tests (Tsaousis, 2015). Instead they followed Chen's (2007) guidelines, where they compared ΔCFI, ΔRMSEA, and ΔSRMR. They found good fit making these comparisons gender groups.

Latent Means Differences

Using the latent mean difference tests, it was found that for all four factors (Biology, Chemistry, Physics, and Math) female high school students scored better than male students. Specifically, by using Tukey comparisons, no differences between female high school students in private or public schools were found. On the other hand, males in private schools did better than public school male students on the SAAT test. With this in mind, it might help parents better decide what school system (private or public) is best for their children. However, note that the Cohen's effect size of male public and private high school students on all of the four SAAT test sections ranged from 0.156 to 0.145. This indicates that the difference of the effect size between male public and private schools is considered small because it less than 0.2 (Sawilowsky, 2009). This brings us to the following question "Why did female students in high school do better than males in all the four sections?" To answer this questions, an evaluative study is suggested to have a closer look at the problem, and find ways to explain these results. Also, other variables should be involved, such as age, grade level, and geographical location to better understand the significant differences that might show when we have a comparison between public and private schools.

Finally, it is worth mentioning that in Tsaousis's (2015) study, it was found that females did better than males in biology and chemistry, but males did better than females in physics and math. Note that this study only investigated gender not school types. This study differs from the Tsaousis's (2015) because female high school students achieved better than males in all of the SAAT test sections. With this in mind, another a question might be asked, "If there was a true decrease in achievement during the two-year period of 2015 and 2017 among males, what might be the causes?" Once again, an evaluative study might answer this question.

Limitations

This study could be expanded to include an analysis of the questions in the items themselves on the SAAT test. Specifically, the data for this study had only the items and the students' answers. It would be very helpful to have the question content, so the researcher has a better idea of what kinds of questions the students were being asked. For example, it would help the researcher to understand why only 40 out of 88 items loaded substantially on either of the two factors. It could be that a question listed as math dealt with physics more than math. In addition, it would be useful to know which questions cross-loaded in terms of content. Another limitation is that there was no qualitative component to this research, which could explain why females achieved better in both public and private schools compared to males.

Summary

This study found that there was no match between the EFA two-factor and theoretical four-factor models. The two-factor model had only 40 items, where the four-factor model had 88 items. Also, the test structure of the EFA analysis showed that the SAAT test only measured two factors (biology and chemistry). Therefore, the EFA results did not support the test structure hypothesized by the NCA. In addition, the CFA four-factor model showed a good fit in the CFA. Measurement invariance and ANOVA were also examined on the model. Note that the test itself showed that all measurement invariance steps had good fit across gender and school types. Also, it showed that each of the four factors had an acceptable level of internal consistency.

In addition, results showed that female students in high school did better than males on all four sections of the SAAT test. On the other hand, male students in public school did not achieve well on the test compared to females in both schools and males in private high school. Note that the effect size between male in public and private high schools was small. With this in mind, an evaluative study is suggested to take a closer look in order explain why females did better than males. Finally, it is worth mentioning that in Tsaousis's (2015) study, it was found that females did better than males in biology and chemistry, but males did better than females in physics and math. Note that the Tsaousis study only investigated gender not school types. With this in mind, another a question might be asked, "What caused an apparent decrease in achievement during the two-year period of 2015 and 2017 among males?" Once again, an evaluative study might answer this question.

References

Al-Algpary, M., & Al-Shapani, A. (2008). The decision of canceling the National Education Organization high school test is giving teachers the confidence they need to do their job. *Riyadh Online Newspaper*. Retrieved from http://www.alriyadh.com/350715

Abrqawie, A. (2015, January 29). Private schools are the reason why students do poorly on the attainment test. *Sabq Online Newspaper*. Retrieved from http://sabq.org/Gcxgde.

Albert, J., & Rizzo, M. (2012). R by Example : Concepts to Code. (Use R!).

Al-Falih, A. (2010). The achievement test for science colleges and the three years of the high school system. *Riyadh Online Newspaper*. Retrieved from http://www.alriyadh.com/548352

Al-Garieb, D. (2013). A study of why parents enter their children in private and international schools in Saudi Arabia. Retrieved from http://lisyanbu.blogspot.com/2013/09/blog-post.html

Al-Ghamdi, A. (2012). Female students believe that the private schools system is the reason behind their lower scores on GAP and ATFSC tests. *Riyadh Online Newspaper*. Retrieved from http://www.alriyadh.com/743687

Al-Hagbani, A. R. (2013, December 16). Half a million Saudis studying in private schools. *Aletisadiah*. Retrieved from 8http://www.aleqt.com/2013/12/16/article_808101.html

Ali, A, Ali, Z., & Naz, R. (2012). Study habits and education planning: A case study of comparison of private and public sector schools. *Dialogue, 7*(3), 309-318. ISSN: 1819-6470.

Al-Maliki, A. (2015). The Saudi Arabian National Center for Assessment in Higher Education test is a nightmare for high school students. *Riyadh Online Newspaper*. Retrieved from http://www.alriyadh.com/739273

Brown, T. (2006). *Confirmatory factor analysis for applied research* (Methodology in the social sciences). New York: Guilford Press.

Chen, F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464-504.

Deraney, P., & Abdelsalam, H. (2012). Private vs. public school education as a predictor for success for female students at private university in Saudi Arabia. *Journal of Arts, Science & Commerce, 3*(3_1), 1-7. Retrieved from http://www.researchersworld.com/vol3/issue3/vol3_issue3_1/Paper_01.pdf.

Khan, H., Chandio, J., & Farooqi, M. (2012). Comparison of performance appraisal system in public and private schools. *Pakistan Journal of Commerce and Social Sciences, 8*(1), 272-278.

King Saudi University. (2015). Application requirements: Achievement tests and GPA scores. Available from http://ksu.edu.sa/sites/KSUArabic/Students/ProsStds/AddmisionReq/Pages/Bachelors.aspx

Iqbal, M. (2012). Public versus private secondary schools: A qualitative comparison. *Journal of Research & Reflections in Education*, *6*(1), 40 -49.

Lance, C., Butts, M., & Michels, L. (2006). The Sources of Four Commonly Reported Cutoff Criteria. Organizational Research Methods, 9(2), 202-220.

Mishari, A. (2012). A deep look at secondary education and its development over the years. Retrieved from http://www.qiyas.sa/MAndAssesment/Researches/Pages/Researches.aspx

Newhouse, D., & Beegle, K. (2006). The effect of school type on academic achievement: Evidence from Indonesia. *The Journal of Human Resources, 41*(3), 529-557. doi: 10.3368/jhr.XLI.3.529

Saudi Arabia. National Center for Assessment in Higher Education. (2016). Achievement Test for Science Colleges: Males. Available from http://www.qiyas.sa/Sites/English/Tests/EducationalTests/Pages/Achievement-test-for-Science-Colleges---Males.aspx

Saudi Arabia. National Center for Assessment in Higher Education. (2015). The ATFSC standard achievement admission test booklet questions, samples for male students. Available from http://www.qiyas.sa/Sites/English/Tests/EducationalTests/Pages/Achievement-test-for-Science-Colleges---Males.aspx

Sawilowsky, S. S. (2009). New effect size rules of thumb. Journal of Modern Applied Statistical Methods, 8(2), 597-599.

Sideridis, G., & Tsaousis, I. (2014). *Identifying content and cognitive dimensions of the Standardized Test of English Proficiency (STEP)*. Retrieved from http://www.qiyas.sa/MAndAssesment/Researches/Pages/Researches.aspx

Tabachnick, B., & Fidell, L. (2013). Using multivariate statistics (Sixth ed).

Tsaousis, I. (2014). *Factorial structure of GAT and its measurement invariance across school types*. Retrieved from http://www.qiyas.sa/MAndAssesment/Researches/Pages/Researches.aspx

Tsaousis, I. (2015). *Factorial invariance and latent mean differences of scores on SAAT across gender.* Retrieved from http://www.qiyas.sa/MAndAssesment/Researches/Pages/Researches.aspx

Vehkalahti, K. (2011). A Beginner's Guide to Structural Equation Modeling, Third Edition by Schumacker, R. E., & Lomax, R. G. *International Statistical Review, 79*(2), 286-287.

Appendix A

Sample Items from the SAAT test

Below are some sample items from the ATFSC test for males:

Biology sample question:

- Photosynthesis is a main pathway in all the following groups expect:

A- fungi. B- algae.

C- cyanobacteria D- bryophyte

Chemistry sample question:

- Oxidation of primary alcohol produce:

A - Ketone. B- Organic acid.

C- Aldehyde D- Ester

Physics sample question:

- An object is placed at 20 cm from a convex lens with a focal length of 15 cm. How far in cm is the image from the lens?

A -60 B- 9

C- 9 D- 60000

Math Sample question:

- Let f be a function defined on the real numbers as $f(x) = x. | x |$, then f is:

A- increasing. B- decreasing.

C- concave up. D- concave down