European Journal of Education Studies
ISSN: 2501 - 1111
ISSN-L: 2501 - 1111
Available online at: www.oapub.org/edu

DOI: 10.46827/ejes.v11i3.5233 Volume 11 | Issue 3 | 2024

ASSESSMENT FAIRNESS AT THE UNIVERSITY OF CAPE COAST: EVALUATING THE FACTOR STRUCTURE AND MEASUREMENT INVARIANCE ACROSS GENDER

Samuel Oppong[1i],
Nathaniel Quansah[2],
Regina Mawusi Nugba[3],
Vera Rosemary Ankoma-Sey[4],
Samuel Yaw Ampofo[5],
Gabriel Essilfie[6]
[1,2]Student,
Department of Education and Psychology,
University of Cape Coast,
Ghana
[3]Lecturer,
Department of Education and Psychology,
University of Cape Coast,
Ghana
[4,5,6]Senior Lecturer,
College of Distance Education,
University of Cape Coast,
Ghana

Abstract:
The specific purpose of the study was to evaluate the factor structure of statistics assessment fairness inventory (SAFI) which was adapted for the Ghanaian context and further examine measurement invariance across gender. SAFI factor structure and construct validity were assessed with a sample of 735 Ghanaian university students through exploratory and confirmatory factor analysis. Through an exploratory factor analysis, five dimensions were obtained which include: learning materials and practices, test design, opportunity to demonstrate learning test administration and grading, and offering feedback. A confirmatory factor analysis was then conducted to examine the factor loadings of the items. Further analysis revealed a partial measurement invariance of the fairness construct with equal means and equal variances in relation to gender. This suggested that the SAFI is valid, reliable, and can equally work across gender. It is recommended that SAFI has important implications for decisions regarding the conduct of especially, high-stakes examinations.

[i] Correspondence: email soppong005@stu.ucc.edu.gh

Keywords: educational assessment, assessment fairness, measurement invariance, statistics assessment fairness

1. Introduction

The concept of fairness holds significant importance in the educational and assessment experiences of students (Laing, Mazzoli Smith & Todd, 2018; Tierney, 2013). Research has indicated a correlation between students' academic achievement, engagement, and motivation for learning with their perception of classroom fairness (Holmgren & Bolkan, 2014; Berti, Molinari, & Speltini, 2010; Chory-Assad, 2002). On the contrary, the observation of inequitable treatment in the classroom has been linked to student absenteeism (Ishak & Fin, 2013), academic dishonesty (Murdock, Miller, and Goetzinger, 2007), and adverse conduct such as hostility and aggression (Chory-Assad & Paulsel, 2004b). Despite the robust empirical associations, the notion of fairness remains a concept that lacks clarity within the assessment literature. This is due to the existence of diverse conceptions and definitions that are often based on a psychometric rationale (Tierney, 2013).

Fairness has been a prevalent ideal throughout the history of education in democratic societies. This can be observed from the implementation of merit-based systems many centuries ago to the structuring of modern-day classrooms (Tierney, 2013). The attribute of fairness is widely regarded as a desirable characteristic in various educational evaluations, ranging from identifying the learning requirements of individual students to conducting extensive global assessments of academic accomplishments.

Green *et al.* (2007) asserted in their study on assessment ethics that the concept of fairness is widely accepted as a fundamental principle. Tierny (2013) is of the view that although statistical evidence is important when evaluating test fairness, its determination is not only a matter of statistics. The issue of ensuring fairness in educational assessment is multifaceted, and no single approach can guarantee its attainment. Various methods can be employed to attain a more equitable assessment, with certain circumstances or approaches carrying greater significance in certain contexts relative to others, contingent upon the assessment's objective and the individual assessed (Tierney, 2016). It is imperative to proactively consider conditions and strategies for ensuring fairness during the design and development of assessment tools and tasks. Additionally, it is crucial to maintain this consideration throughout assessment interactions and retrospectively during the review of the assessment process.

While the importance of fairness in assessments as a factor affecting students' academic performance is acknowledged, there has been limited research focused on evaluating the concept of assessment fairness in developing nations. This gap in research persists mainly because the majority of studies exploring assessment fairness have taken place outside of Ghana and the African continent. Specifically, most of the assessment

fairness research has been done in Europe (Murillo & Hidalgo, 2020; Tierney, 2016; Wallace, 2018; Bazvand and Rasooli, 2022; Kunnan, 2004; Palmer, 2010; Kane & Burns, 2013).

This study aimed to address this research gap by concentrating on the validation of the Statistics Assessment Fairness Inventory (SAFI), which was developed based on the Assessment Fairness Questionnaire (AFQ; Rezai, 2022) and modified for the Ghanaian context. Measurement instruments, such as the AFQ, play a crucial role in social science research, and the level of concern regarding the data quality produced by these instruments has been highlighted by Cizek (2012). The process of adapting assessment instruments is a multifaceted task that involves careful consideration of various factors, including content maintenance, psychometric qualities, validity for the target population, and cultural suitability for the target community. These considerations demand a rigorous methodological approach (Borsa, Dam´asio, & Bandeira, 2012). The current study emphasizes the implementation of these optimal methodologies to enhance the level of confidence in the results derived from the SAFI.

2. Literature Review

2.1 Conceptual Review

Fairness in assessment, as described by Educational Testing Service (2014), encompasses three key components: the absence of bias, equitable treatment of all individuals taking the test, and equal access to preparation materials for achievement exams. Furthermore, McNamara and Ryan (2011) assert that fairness encompasses the degree to which a test's psychometric quality guarantees procedural equity for both individual test-takers and subgroups, as well as the sufficiency of the construct's representation in the test materials and procedures. Ensuring the fairness of an assessment is a significant concern for educators. In order for a test to be considered fair, several criteria must be met. Firstly, any disparities in performance between different groups should be attributed to factors that are directly related to the content and construct of the test (ETS, 2014). Secondly, the administration processes should be consistent for all students, as emphasised by Kunnan (2018) and McNamara & Ryan (2011).

Additionally, it is important for teachers and students to engage in communication that is characterised by professionalism and dignity. Lastly, the score obtained on the test should accurately and impartially reflect the individual's performance (Wallace, 2018). This implies that the concept of fairness in assessment holds significant importance within the field of social and psychological sciences, hence necessitating comprehensive investigation worldwide and locally.

In the context of this study, fairness has been viewed as the extent to which students perceive the assessment process as impartial or just treatment of all test takers. This means that fairness is not related to the psychometric property of the test but to how students view the whole process as fair and the extent to which students' perceptions of

fairness construct are the same across genders and different programmes of study. This can be seen in the perception students have in relation to getting access to learning materials and practices, perception about the test design, administration, grading, offering feedback, and opportunities to demonstrate learning (Rezai, 2022). Students also perceive the assessment process to be equal if it calls for the same administration, content, scoring, and interpretation of results to achieve objectivity (Murillo & Hidalgo, 2020). This is what Nisbet and Shaw (2019) classified as equality assessment. Pepper and Pathak (2008) conducted a study on the perceptions held by university students regarding fair assessment practices at Southwestern University. The researchers discovered that students deemed assessment to be fair when there was a clear and explicit approach to the administration of assessments and grading criteria, regular provision of feedback, and proactive engagement in the assessment process. In a study conducted by Murillo and Hidalgo (2017), the researchers discovered that students' judgements regarding fair assessment were linked to concepts such as equality, objectivity, transparency, and evaluation of course material. In a comparable investigation conducted by Wallace (2018), the administration of the test was characterised by a notable degree of interactional fairness and procedural fairness.

2.2 Structural Equation Modeling (SEM)

SEM, which was employed for the study can also be referred to as means and covariance structure analysis, causal modeling, latent variable path analysis, and latent variable modeling. The term "SEM" refers to a broad range of statistical techniques utilised for assessing the empirical validity of substantive theories (Lei & Wu, 2007). From a statistical perspective, it can be observed that it constitutes an expansion of the general linear model (GLM) techniques, which encompass methodologies such as ANOVA and multiple regression analysis. SEM integrates a measurement model into its analysis to account for potential errors that may arise from the measurement of a given variable (Owolabi, Ayandele & Olaoye, 2020).

This statistical method has become widely used because it allows for the measurement of both direct and indirect correlations among causative variables using a single model (Meydan & Sesen, 2011). SEM allows for the analysis of research hypotheses through a single procedure by simulating intricate interactions among various observable and latent variables. The process of examining the connections between manifest and unobservable constructs necessitates the use of SEM. Latent variables are constructs that cannot be directly measured and are instead inferred from observed variables that are believed to be indicators of the construct.

Civelek, Ince, and Karabulut (2016) assert that apart from SEM, the majority of statistical techniques are designed to determine correlations within a given data set. According to Karagoz (2016), it can be argued that SEM is a more appropriate approach for hypothesis testing compared to other methods. In the context of this study, SEM

allows us to build complex structures of assessment fairness and model it to determine the factor loadings on the latent variables through the observed indicators.

3. Materials and Methods

3.1 Participants

The current study sampled 735 students who offered educational statistics within the first semester 2022/2023 academic year. Specifically, a census approach was employed to select all students who offered educational statistics that semester. The sample was selected over six programmes of study; Arts, Social Science, Home Economics, Computer Science, Fine Arts, and Communication Design. The selection of all students was driven by the objective of the present study, which aimed to examine the temporal stability of gender invariance. The inclusion of a large sample size was deemed necessary for this purpose.

3.2 Procedures

Before the researchers embarked on the data collection, an introduction letter was obtained from the head of Education and Psychology Department. The letter explained the purpose of the study, the need for individual participation as well as the confidentiality and anonymity of participants' responses. This letter was sent to the sampled departments. This in a way to ensure that the lecturers involved are pre-informed about the data collection. Ethical clearance was also sought from the Ethical Review Board of the College of Education Studies, University of Cape Coast with Ref No. CES/ERB/UCC/EDU/08-23/25.

Based on ethical issues highlighted by Oppong, Nugba, Asamoah, Quansah, and Ankoma-Sey (2023), the authors ensured that none of them were violated. After all permissions were sorted and granted, the researchers followed up to arrange for the time and date on which the data was convenient to be collected. After this, the researchers trained four research assistants who helped in the data collection process. The research assistants underwent training in effective communication strategies with respondents, including techniques for explaining complex concepts and ensuring consistent information gathering. A duration of fourteen (14) days was allocated for the distribution and retrieval of the questionnaire.

3.3 Statistics Assessment Fairness Inventory (SAFI)

The Scale, Statistics Assessment Fairness Inventory (SAFI) was derived from a collection of questionnaires initially created by Rezai (2022) with the purpose of assessing overall fairness in the context of classroom assessments. The initial scale was developed by Rezai (2022) as a 110-item questionnaire with 10 subscales. Participants were required to rate their responses on a 4-point Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree). In the process of developing SAFI, a total of 46 fairness indicators were

incorporated, organised into five distinct subscales. These subscales include learning materials and practises, test design, opportunity to demonstrate learning, test administration, and grading and offering feedback. The aforementioned indicators were later subjected to exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) procedures.

3.4 Analysis

In order to assess the factor structure of the revised SAFI, an EFA was done. The examination of factor structures in the Ghanaian context was deemed necessary due to the original discovery of eleven-dimensional component structures by the developers. This was particularly essential as the scale had been first designed and tested only with youth in the United States, lacking any validation with Ghanaian youth. The varimax rotation method was employed for all EFAs. CFA was further employed in order to validate the results obtained from the EFA, using the validation sample. The components were extracted using the Varimax with Kaiser Normalisation method, specifically, the maximum likelihood estimation.

According to Hau and Marsh (2004), the maximum likelihood estimation method with resilient standard errors demonstrates robustness in its ability to accommodate minor deviations from the assumption of normality. According to Costello and Osborne (2005), factor loadings that were equal to or more than .50 were considered to be satisfactory. Furthermore, we conducted an analysis to evaluate the applicability of the SAFI in the specific context of Ghana, while also investigating the extent to which the scale remains consistent across different genders. The invariance tests were conducted to evaluate the statistical equivalence of the parameters in the assessment fairness measurement model between male and female participants. The analysis of the data involved the application of the partial least squares path modelling (PLSPM) technique within the framework of SEM specifically employing the stages of measurement invariance of composite models (MICOM) approach.

The evaluation of the model constructed at each level involved an assessment of many fit indices, including the requirement for equal indicators in both groups, equivalent data treatment, consistent PLSPM algorithm settings, correlation between the composite scores of both groups and permutation's p-values exceeding a threshold of .05. The study conducted an analysis using three nested hierarchical models of Multiple Group Confirmatory Factor Analysis (MGCFA). These models comprised the configural model, the composite model, and the equal means and variance model.

4. Results

4.1 Ensuring Validity and Checking Assumptions Underlying EFA

The process of adapting and developing the SAFI was meticulously executed in order to guarantee the accuracy and reliability of the obtained responses. Careful measures were taken to guarantee that the questions were appropriately modified and formulated to embody the principle of fairness in the evaluation of statistical knowledge. Following the compilation of the items, the instrument underwent a validation process, which involved a thorough evaluation by professionals specialising in the field of Measurement and Evaluation, specifically PhD students. This approach aligns with Anim's (2005) argument that expert judgement plays a crucial role in determining the content and construct validity.

Pallant (2007) suggested that a sample size of 150 is deemed appropriate for conducting factor analysis. Similarly, Tabachnick and Fidell (2019) were of the view that it is comforting to have at least 300 cases for factor analysis. Based on these suggestions, a sample size of 735 fitted for accurate estimation of parameters in the EFA. Further, other two statistical measures were generated using SPSS software to measure the sampling adequacy of the dataset. Bartlett's test of sphericity (Bartlett, 1954) was first inspected and it was found to be significant at $\chi^2(1035) = 11592$, $p<.05$. Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was also found to be appropriate with a value of .925 which is greater than the cut-off value of .60 (Pallant, 2007). Additionally, the factorability of the correlation matrix was checked and a correlation of at least $r= .3$ or greater was found among some of the variables. Table 3 provides a summary description of the assumption checks. The researchers did an exploratory factor analysis (EFA) using the principal component analysis (PCA) method to identify the elements underlying the scale. The scree plot was employed to ascertain the number of factors.

The preliminary research unveiled a total of five factors, as depicted in Figure 1.

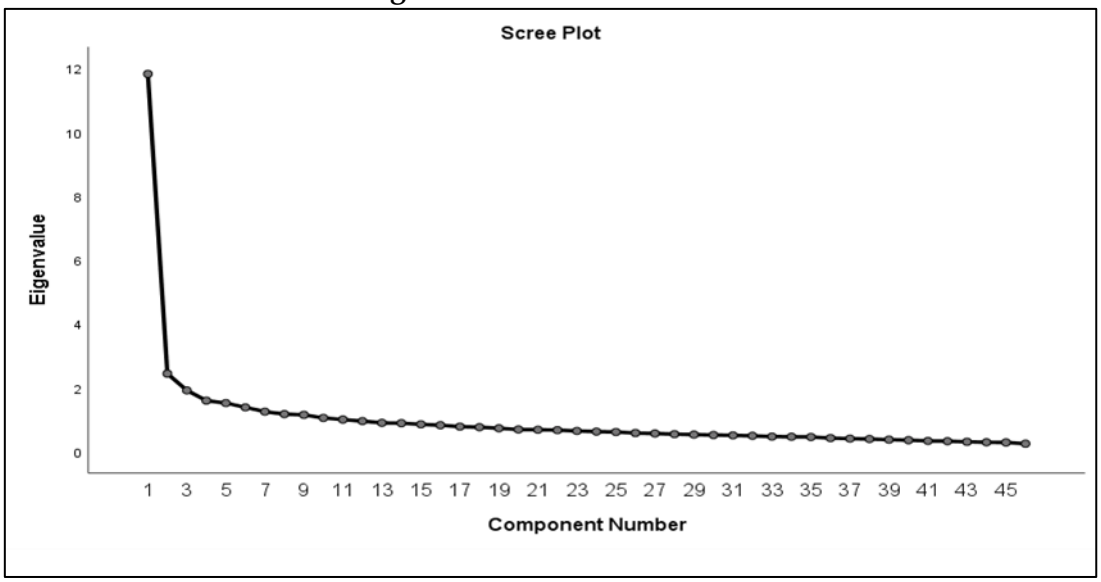
Table 1: KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.925
Bartlett's Test of Sphericity	**Approx. Chi-Square**	11592.403
	df	1035
	Sig.	.000

4.2 Factor Extraction and Model Fit of the Inventory

Factor extraction is the process of identifying the most concise set of factors that effectively capture the relationships among a given set of variables. Tabachnick and Fidell (2007) suggested that the best approach to extracting the various components in factor analysis is the use of PCA. Also, to ensure a good and accurate judgment of the model fit and to ensure the number of components, a scree plot test was used.

Figure 1: Result of Scree Plot



A confirmatory analysis was undertaken using the Maximum Likelihood Method to examine the factor loadings of each item. Items that had a factor loading < .5 were excluded from the analysis, as indicated in Table 2. Following the completion of the factor analysis, a total of 42 items were found to have been retained, and rejecting four items due to low factor loadings. The five components identified in the confirmatory factor analysis were assigned the following labels: learning materials and practices, test design, opportunity to demonstrate learning, test administration, and grading, and offering feedback.

Table 2: Factor Rotation

Indicators	**Component**				
	1	**2**	**3**	**4**	**5**
I am sufficiently aware of the educational calendar and its activities provided by the university.	**.306***				
I have sufficient access to qualified capable educational statistics lecturers to teach the course content.	.582				
My lecturers provide a reasonable and adequate explanation of the learning objectives for the semester.	.660				
Course outlines are being shared from the beginning of the course and it is being followed through to the end.	.619				
I have sufficient access to quality appropriate facilities to learn the course content.	.602				
The lectures adequately prepare us well before test administration.	.711				
Considering my learning style and abilities, I have enough opportunities to concentrate in the classroom.	.622				
My lecturer uses an activity-oriented method to teach in the classroom for a better understanding.	.684				
The pedagogical skills adopted by my statistics lecturers are appropriate.	.690				

The teaching of statistics in the classroom is more of practical (students-centred).	.571				
Enough opportunities are created to work and learn with my classmates in group activities.	.508				
Most of the quizzes I write are made up of familiar items		**.317***			
My SHS background influences my performance in my statistics.		**.456***			
Tests have good design; appearance, proper font and size, and enough space to answer item questions.		.692			
Tests are designed and tailored to the course outline given.		.671			
The time and venue for statistics tests are agreed upon by lecturers and students.		.637			
My views and concerns regarding the design and preparation of classroom tests are considered by lecturers.		.627			
Test items are logically ordered from simple to difficult.		.647			
End-of-semester exams are a combination of objective questions (e.g., multiple choice and fill-in the blank) and essay questions.		.590			
Items formulated for quizzes are mostly made up of calculations.		.512			
Tests measure exactly what they intend to assess (e.g., statistics tests only assess statistics competence).		.608			
My lecturers provide me with multiple assessment opportunities.			.655		
I have enough time to show my abilities during quiz administration.			.757		
There are enough facilities for students with exceptional physical problems (e.g., blindness, deafness, etc.) to show their abilities.			.672		
There are good conditions such as comfortable chairs, a quiet place during test administrations.			.673		
My lecturers provide adequate direction/instruction on how to answer test items.				.594	
Adequate time is given to enable me to respond to statistics test items.				.643	
The instructions of tests items are easy to understand.				.642	
I am able to read and understand statistics test items when assessed.				.635	
The conditions of tests administration are the same for all students.				.555	
Announcements are made about the time at regular intervals during test administration.				.684	
My lecturers resolved problems encountered by me during test administration answer.				.605	
My lecturers ensure that the testing environment is free from distractions				.656	
My statistics lecturers interact well with me when effecting corrections after test administration				.666	
My lecturers try as much as possible to prevent students from cheating during statistics test administration.				.570	
Difficulty test items are brought before the easier ones and this put me off.					**.383***
My lecturers provide sufficient and valid information about grading criteria at the beginning of the semester.					.624
There is a consistency in grading criteria.					.664

My lecturers provide a reasonable and adequate explanation of grading criteria.					.705
There is transparency in announcing test results (It is clear when, how, and by whom test results are announced).					.735
My grades are kept confidential in the classroom.					.678
My lecturers announce test results and grades on time.					.712
The feedback that lecturers offer on my performance is clear to me.					.765
The feedback that lecturers offer on my performance is in line with the educational materials taught during the semester.					.751
The feedback that lecturers offer on my performance is reasonable and adequate.					.739
My lecturers listen carefully to my requests for a review of my grade.					.657

Note: 1 - learning materials and practices; 2 - test design; 3 - opportunity to demonstrate learning; 4 - test administration; 5 - grading and offering feedback.
*Item rejected.

4.3 Estimating Reliability

Estimating the reliability of items cannot be overlooked because every investigator considers it necessary to gather objective and accurate information. There is the need, therefore, to estimate the reliability of responses to the construct of interest (Quansah, 2017). The reliability of the instrument was estimated using the Cronbach's Alpha Reliability Method. The reliability estimates for each sub-scale as well as the whole instrument were investigated (See Table 3). The overall reliability estimate of the instrument was .85. This reliability coefficient is sufficient enough to ensure reliable responses as indicated by Pallant (2010) and a reliability coefficient (alpha) of .70 or higher is considered appropriate.

Table 3: Reliability Estimate for Sub-scales

Subscale	No. of Items	Coefficients
Learning Materials and Practices	11	.826
Test Design	7	.758
Opportunity to Demonstrate Learning	4	.712
Test Administration	10	.828
Grading and Offering Feedback	10	.886

Source: Field Data.

4.4 Measurement Invariance of Statistics Assessment Fairness Inventory based on Gender

Assumptions that necessitate the analysis of MGCFA were tested. Among the assumptions included missing values, multicollinearity, and variance inflation values (VIF). The initial step before the conduct of the analysis involved the examination of missing values. However, there were no missing values identified in the data set. Multicollinearity was then checked to identify the relationships among the items and none was found. Tabachnick and Fidell (2013) indicated that a correlation exceeding .90

among the items gives rise to the issue of multicollinearity because a high correlation signifies that the items assess similar properties. Also, all the subscales of the inventory had a VIF lower than 10. After all these assumptions were fulfilled measurement invariance of composite model (MICOM) was performed.

4.5 Test for Configural Invariance

In ensuring that configural invariance exists among groups, we did a qualitative assessment of the composites' specification across all the groups by ensuring that three criteria were met. Firstly, identical indicators for each of the measurement model was ensured across the groups. The researcher also had already checked for face and expert validity of the instrument (hair *et. al*, 2017) to ensure the same set of indicators across the groups. Secondly, identical data treatment was ensured across the gender group. In doing this, outliers and missing values were checked and none was found in both groups. Lastly, identical algorithm settings or optimization criteria through variance-based model estimation methods were ensured so that differences in the group-specific model estimations do not result from dissimilar algorithm settings. After establishing configural invariance for both groups, compositional invariance was tested.

4.6 Compositional Invariance of the Instrument Based on Gender

The compositional invariance in other measurement invariance literature is also known as metric invariance. To test for compositional invariance, permutation multigroup analysis of MICOM was used. In order to ensure a compositional invariance, the original correlation between composite scores obtained from group 1 (females) and group 2 (males) should be greater than or equal to the 5% quantile of the empirical distribution of the correlations between the composite scores of group 1 (females) and group 2 (males).

This result is summarised in Table 4.

Table 4: Compositional Invariance Assessment

Construct	Original Correlation	5.00%	Permutation p-value	Partial Measurement Invariance
Grading and Offering Feedback	0.999	0.998	0.676	Established
Learning Materials and Practices	0.995	0.996	0.046	Not Established
Opportunity to Demonstrate Learning	0.998	0.987	0.747	Established
Test Administration	0.997	0.995	0.242	Established
Test Design	0.998	0.994	0.64	Established

Source: Field Data.

From Table 4, it was observed that the permutation p-values were all greater than .05 except for the construct, learning materials and practices. Not all the latent factor structures can be compared across the two groups thereby establishing a partial

measurement invariance. After establishing this, the mean and variance of each construct were examined to ensure equality of composite mean values and variances.

These results are presented in Table 5 and 6 respectively.

Table 5: Equality of Means

Construct	Mean Original Difference	Confidence Interval	Permutation p-value	Equality of Means
Grading and Offering Feedback	-0.03	[-0.151, 0.142]	0.712	Equal
Learning Materials and Practices	-0.099	[-0.143, 0.152]	0.181	Equal
Opportunity to Demonstrate Learning	0.087	[-0.138, 0.144]	0.237	Equal
Test Administration	-0.123	[-0.14, 0.155]	0.094	Equal
Test Design	-0.123	[-0.151, 0.15]	0.113	Equal

Source: Field Data.

From Table 5, it was revealed that all the permutation p-values were greater than .05 and this suggested that the mean of each construct in both groups is not significantly different from each other.

The variance of each construct was further examined and the result is presented in Table 6.

Table 6: Equality of Variances

Construct	Variance Original Difference	Confidence Interval	Permutation p-value	Equality of Variances
Grading and Offering Feedback	-0.185	[-0.232, 0.232]	0.128	Equal
Learning Materials and Practices	-0.082	[-0.228, 0.234]	0.483	Equal
Opportunity to Demonstrate Learning	-0.116	[-0.229, 0.216]	0.323	Equal
Test Administration	0.031	[-0.227, 0.232]	0.793	Equal
Test Design	0.035	[-0.217, 0.243]	0.784	Equal

Source: Field Data.

From Table 6, it was revealed that all the permutation p-values were greater than .05 and this suggested that the variance of each construct in both groups is not significantly different from each other. However, because of the partial measurement invariance at the compositional level, full measurement invariance cannot be concluded. After ensuring partial measurement invariance, the final step was assessed which is estimating the path coefficient using the pooled data.

Group group-specific model was estimated separately and the results are presented in Table 7.

Table 7: Path Coefficient

Path Coefficient	Path Coefficient (females)	p - value (females)	Path Coefficient (males)	p - value (males)	Invariant
Grading and Offering Feedback -> Learning Materials and Practice	**0.27**	**0.00**	**0.30**	**0.00**	**Yes**
Opportunity to Demonstrate Learning -> Grading and Offering Feedback	0.48	0.00	0.42	0.00	Yes
Test Administration -> Learning Materials and Practice	0.43	0.00	0.35	0.00	Yes
Test Administration -> Test Design	0.53	0.00	0.62	0.00	Yes
Test Design -> Opportunity to Demonstrate Learning	0.35	0.00	0.49	0.00	Yes

Source: Field Data.

From the path coefficient in Table 7, it was seen that the p-value for each construct in both female and male groups was less than .05 and this means that invariance has been achieved. To further confirm that differences do not exist between the two groups, the Mann-Whitney U test was conducted.

The result is summarised in Table 8.

Table 8: Mann Whitney U Test

Gender	N	Mean Rank	Mann-Whitney U	df	Sig. Value
Females	426	348.84	66739.00	1	.096
Males	292	375.06			

Source: Field Data.

A Mann-Whitney U test was performed to evaluate whether classroom assessment fairness differed by gender. The results indicated that there was no significant difference between assessment fairness perceptions of female students and male students where U = 66739.00, p=.096. The null hypothesis is therefore failed to be rejected.

5. Discussion

The primary objective of this study was to examine the construct validity of the SAFI within the Ghanaian setting, as well as assess the measurement equivalence of the scale across gender. The results of this study additionally offered empirical evidence to substantiate the applicability of the scale to both male and female individuals at a specific point in time. The examination of the factor structure of the assessment fairness construct was one of the primary objectives of this work. The maximum likelihood estimation approach of PCA was employed to accomplish this task. A total of 41 scale items were included in the final model, distributed across five subscales. Due to the inability of certain items to fulfil the predetermined factor loadings cutoff criteria, which

required a minimum loading of .50, they were removed. The ramifications of this study's findings extend to the future examination of the SAFI and the modification of other scales to accommodate culturally different environments, such as the growing number of diverse schools in Western contexts. The inclusion of a diverse range of response options enabled participants to express their perceptions regarding the level of fairness they attributed to the assessment process.

The SAFI items encompass five discrete subdomains, namely:

a) learning materials and practises,
b) test design,
c) opportunity to demonstrate learning,
d) test administration, and
e) grading and offering feedback.

These subdomains collectively pertain to the concept of fairness in the evaluation of statistics. The inclusion of fairness assessments pertaining to arrays in the inventory provides additional evidence supporting the inventory's validity as a measure of assessment fairness within the academic context.

In addition, the study found no significant difference in relation to how male and female students perceive the fairness construct. The results from the measurement invariance test established a partial measurement invariance of the construct across gender. This implies that both male and female students offering educational statistics at the University of Cape Coast can be meaningfully compared along the fairness construct. Hence lecturers can make meaningful comparisons between genders. Establishing measurement invariance indicates that educational statistics tests and exams are equitable for all students regardless of their demographic characteristics (gender). The findings pertaining to partial measurement invariance indicate that the degree of non-invariance and discrepancies in factor loadings have an impact on the goodness of fit indices. On the other hand, the magnitude of the factor loading exerts a greater influence on parameter estimations and bias. Hence, it is imperative for educational researchers to assess the assumption of invariance prior to forming evaluations. The study was unable to achieve complete measurement invariance, which has implications related to the size of the model. The impact of model size on goodness of fit indices was shown to be significant, leading Donahue (2006) to argue that model size should be carefully considered when examining partial measurement invariance.

In the study of Donahue (2006) on the effect of partial measurement invariance on prediction, it was revealed that if certain model parameters, such as factor loadings, factor covariances, and error variances, are discovered to be noninvariant between groups, they are permitted to vary, while other parameters remain invariant. This allows for continuing testing. This approach enables the utilisation of a measurement system that accommodates potential variations among the groups while maintaining the significance of the overall comparison. This is the reason why although some model parameters of the assessment fairness instrument were seen to be non-invariant, the researchers still

continued to established partial measurement invariance as a result of the equal means and variances obtained. Similarly, Yu & Hudders (2022) conducted a study on measurement invariance of the modified brand luxury index scale across gender, age, and countries. A subsequent examination revealed a partial measurement invariance across the United States, China, and India in support of the above, Steinmetz, Schmidt, Tina-Booh, Wieczorek & Schwartz (2009) tested measurement invariance using multigroup CFA to examine the differences between educational groups in human values measurement. The results of the analysis indicated that there was a partial invariance seen for the majority of the 10 values and parameters.

Moreover, the methodologies employed in the adaptation of the SAFI to suit the Ghanaian context provide valuable perspectives on the procedures that may be utilised to standardize and evaluate the applicability of various psychosocial measures in multiple sociocultural settings. Furthermore, the results obtained from this research have significant implications for the application of the validated scale within the context of educational training and development practice. The development of site-specific instruments aimed at ensuring the reliability and validity of statistics assessment fairness is a complex task that necessitates specialised technical expertise, ample resources, and a significant investment of time. This undertaking surpasses the capabilities typically possessed by education professionals, including educational development specialists, practitioners, and administrators. Therefore, various education stakeholders will have the opportunity to utilize the validated SAFI to acquire dependable and accurate data regarding the fairness of statistics assessments in educational institutions in Ghana, as well as in other developing nations that exhibit comparable characteristics to Ghana.

6. Conclusions and Recommendations

There is statistical evidence to conclude that the assessment fairness instrument can be compared across gender. The construct is understood equally among male and female students. This indicates that all males and females offering educational statistics at the University of Cape Coast understand the construct of fairness equally and know the implications of a fair assessment process. The SAFI instrument can therefore be adopted in varying contexts to examine assessment fairness. It is further recommended that every study should provide evidence of validity for the instrument in terms of its internal structure and measurement invariance from the perspective of equality of measurement. This can help investigate specific items or constructs that are driving the differences and consider modifying or removing problematic items. In this regard, constructs can be measured meaningfully across groups of students.

Conflict of Interest

The authors confirm the originality of this research and declare that no conflict of interest.

About the Author(s)

Samuel Oppong, Department of Education and Psychology, University of Cape Coast, Ghana. Research interests: educational measurement, instrument validation, classical test theory, structural equation modelling, assessment in education.
ORCID ID: https://orcid.org/0009-0005-0851-9186
Web of Science: https://www.webofscience.com/wos/author/search
Wiley: https://wiley.atyponrex.com/dashboard?siteName=emip

Nathaniel Quansah, Department of Education and Psychology, University of Cape Coast, Ghana**.** Research interests: validation and validity theory, statistical modelling, item response theory, high-stakes assessment, and program evaluation.
ORCID ID: https://orcid.org/0009-0001-8274-1480

Regina Mawusi Nugba, Lecturer**,** Department of Education and Psychology, University of Cape Coast, Ghana**.** Research interests: educational measurement, research methods in education, assessment in education, test development, curriculum development in education
ORCID ID: https://orcid.org/0000-0002-3537-645X

Vera Rosemary Ankoma-Sey, Senior Lecturer**,** College of Distance Education**,** University of Cape Coast, Ghana. Research interests: educational administration educational planning educational leadership educational assessment, gender and administrative management
ORCID ID: https://orcid.org/0000-0002-3254-0680

Samuel Yaw Ampofo, Senior Lecturer**,** College of Distance Education**,** University of Cape Coast, Ghana. Research interests: educational administration, distance education, teacher education, and development.
ORCID ID: https://orcid.org/0000-0001-6510-2858

Gabriel Essilfie, Senior Lecturer**,** College of Distance Education**,** University of Cape Coast, Ghana. Research Interests: distance education and teacher preparation, educational leadership, teacher attitude in educational management.
ORCID ID: https://orcid.org/0000-0002-7807-7732

References

Bazvand, A. D., & Rasooli, A. "Students' experiences of fairness in summative assessment: A study in a higher education context". Studies in Educational Evaluation, 2022, (72), 101118.

Berti, C., Molinari, L., & Speltini, G. "Classroom justice and psychological engagement: Students' and teachers' representations". Social Psychology of Education, 2010, (13), 541-556. doi:10.1007/s11218-010-9128-9

Chory-Assad, R. "Classroom justice: Perceptions of fairness as a predictor of student motivation, learning, and aggression". Communication Quarterly, 2002 (50), 58-77. doi:10.1080/01463370209385646

Chory-Assad, R., & Paulsel, M. "Antisocial classroom communication: Instructor influence and interactional justice as predictors of student aggression". Communication Quarterly, 2004b, (52), 98-114. doi:10.1080/01463370409370184

Civelek ME, İnce H, Karabulut AT. "The mediator roles of attitude toward the web site and user satisfaction on the effect of system quality on net benefit: A structural equation model on web site success". European Scientific Journal, 2016.

Cizek, G. J. "Defining and distinguishing validity: interpretations of score meaning and justifications of test use". Psychological methods, 2012 17(1), 31.

Donahue, B. H. "The effect of partial measurement invariance on prediction". University of Georgia, 2006 (Doctoral dissertation).

Green, S., Johnson, R., Kim, D., & Pope, N. "Ethics in classroom assessment practices: Issues and attitudes". Teaching and teacher education, 2007, (23), 999-1011. doi:10.1016/j.tate.2006.04.042

Hau K, T, Marsh H, W. "The use of item parcels in structural equation modelling: Non-normal data and small sample sizes". British Journal of Mathematical and Statistical Psychology. 2004, 57(2):327-51.

Holmgren, J., & Bolkan, S. "Instructor responses to rhetorical dissent: Student perceptions of justice and classroom outcomes". Communication Education, 2014, (63), 17-40. doi:10.1080/03634523.2013.833644

Ishak, Z., & Fin, L. "Truants' and teachers' behaviors in the classroom". Procedia-Social and Behavioral Sciences, 2013, (103), 1228-1237. doi:10.1016/j.sbspro.2013.10.451

Laing, K., Mazzoli Smith, L., & Todd, L. "The impact agenda and critical social research in education: Hitting the target but missing the spot"? Policy Futures in Education, 2018, 16(2), 169-184.

Lei PW, Wu Q. "Introduction to structural equation modeling: Issues and practical considerations". Educational Measurement: issues and practice, 2007, 26(3):33-43.

Lunenburg, F. "Self-efficacy in the workplace: implications for motivation and performance". International Journal of Management, Business, and Administration, 2011, 14, 1–7.

McNamara T, Ryan K. "Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test". Language Assessment Quarterly, 2011 Apr 1;8(2):161-78.

Murdock, T., Miller, A., & Goetzinger, A. "Effects of classroom context on university students' judgments about cheating: Mediating and moderating processes". Social Psychology of Education, 2007, 10, 141-169. doi:10.1007/s11218-007-9015-1

Murillo, F. J., & Hidalgo, N. "Students' conceptions about a fair assessment of their learning". Studies in Educational Evaluation, 2017, (53), 10-16.

Nisbet I, Shaw SD. Fair assessment viewed through the lenses of measurement theory. Assessment in Education: Principles, policy & practice, 2019, 26(5):612-29.

Nitko A, J. "Conceptual frameworks to accommodate the validation of rapidly changing requirements for assessments". Curriculum and assessment. 2001, (1):143-63.

Oppong, S., Nugba, R. M., Asamoah, E., Quansah, N., & Ankoma-Sey, V. R. Teachers Confidence of Classroom Assessment Practices: A Case of Basic Schools in Upper Denkyira West District, Ghana. European Journal of Education Studies, 10(11). October 2023.

Owolabi HO, Ayandele J K, Olaoye D D. "A Systematic Review of Structural Equation Model (SEM)". Open Journal of Educational Development (ISSN: 2734-2050), 2020, 1(2):27-39.

Pallant, J. "SPSS survival manual—A step-by-step guide to data analysis using SPSS for windows". 2007, Maidenhead: Open University Press.

Pepper MB, Pathak S. "Classroom contribution: What do students perceive as fair assessment"? Journal of Education for Business, 2008, 83(6):360-8.

Quansah, F. "The use of Cronbach alpha reliability estimates in research among students in public universities in Ghana". Africa Journal of Teacher Education, 2017 (1);7822.

Rezai, A. "Fairness in classroom assessment: development and validation of a questionnaire". Language Testing in Asia, 2022, 12(1), 17.

Steinmetz H, Schmidt P, Tina-Booh A, Wieczorek S, Schwartz S, H. "Testing measurement invariance using multigroup CFA: Differences between educational groups in human values measurement". Quality & Quantity, 2009, (43), 599-616.

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. "Using multivariate statistics" Boston, MA: Pearson, 2013, Vol. 6, pp. 497-516.

Tierney, R. "Fairness in classroom assessment". In J. H. McMillan (Ed.), SAGE handbook of research on classroom assessment, 2013, (pp. 125–144). Thousand Oaks, CA: SAGE Publications

Tierney, R. D. "Fairness in educational assessment. In M. A. Peters (Ed.), Encyclopedia of Educational Philosophy and Theory. 2016, DOI: 10.1007/978-981-287-532-7_400-1

Wallace, M. P. "Fairness and justice in L2 classroom assessment: Perceptions from test takers". Journal of Asia TEFL, 2018, 15(4), 1051.

Yu S, Hudders L. "Measurement invariance of the modified brand luxury index scale across gender, age and countries". Journal of Fashion Marketing and Management: An International Journal. 2022, 26(5):870-89.