FULL LENGTH ARTICLE

Fairness, Reliability and Validity Threats of the First University-based National Examination Management in Ethiopia:

Kinde Getachew[1], Fisseha Mikre[1], Frew Amsale[2], Ashenafi Belay[3], Netsanet Workeneh[4], Nega Jibat[5], Abel Worku[6], Jemal Abafita[7], Tekle Ferede[3], Getachew Tilahun[8] and Berhanu Nigussie[1]

[1]Jimma University, College of Education and Behavioural Sciences, Department of Psychology
[2]Jimma University, College of Education and Behavioural Sciences, Department of Educational Planning and Management
[3]Jimma University, College of Social Sciences and Humanities, Department of English Language and Literature
[4]Jimma University, Institute of Health, Faculty of Medical Science, Department of Pediatrics and Child Health
[5]Jimma University, College of Social Sciences and Humanities, Department of Sociology;
[6] Jimma University, College of Business and Economics, Department of Accounting and Finance
[7]Jimma University, College of Business and Economics, Department of Economics
[8] Jimma University, College of Social Sciences and Humanities, Department of Media and Communication Studies

***Corresponding author**:* kendie_get@yahoo.com

ABSTRACT
Academic dishonesty has become a serious educational challenge in Ethiopia. To address this challenge in the Ethiopian Secondary School Leaving Certificate Examination (ESSLCE), the MoE administered the examination for the first time in 2022/23 in public universities. Thus, this article presents research findings on the threats to fairness, reliability and validity of the ESSLCE actual administration. A concurrent mixed method design (Qual.+ Quan.) was used. Ten regions were purposely sampled, and one university was conveniently selected from each sampled region. Sixty-three (63) in-depth interviewees were conducted with various stakeholders and 938 students from Jimma Town took part in the survey. Thematic and descriptive analyses were the method of analysis. The results revealed that there were problems such as disturbing ambulance siren sounds, faulty item arrangements, shortages of exam booklets, ill-oriented exam chiefs and invigilators, and many students lacking the right concentrations and motivations to complete the examinations. Similarly, physical illnesses, exam anxiety, and home sicknesses were prevalent among students during the actual examination periods. Thus, it seems that the first university-based ESSLCE administration, which was characterized by adverse students' physical and psychosocial conditions, did not satisfy the criteria of good examinations. Therefore, it is recommended that the Ministry of Education (MoE) and the Educational Assessment and Examinations Service (EAES) have to seek for an alternative examination management which does not hamper the physical and psychosocial states of students but deters academic dishonesty and test mismanagement.

Keywords: Examination administration; Fairness; First university-based national examination management; Reliability; Validity

INTRODUCTION

Grade 12 examinations are mainly administered as evidences for certifying students for completing secondary education, and recruitment for higher education. Different factors, however, have challenged these purposes of the examination. Academic dishonesty, as one of the major challenges, introduces noises in the interpretation of student results (Wollack & Cizek, 2017; Zhu & Han, 2011). Moreover, adverse environmental, physical and psychological states of students during examination preclude students from demonstrating their true performances, and ultimately thwart the purposes of the examination intended to serve (Miller, Linn, & Gronlund, 2009).

Academic dishonesty, as one of the major threats to valid inference of examination results, is seen as a deliberate action performed by students or institutions mainly either before or during the examination administration to gain unfair advantages, or in some cases, to place a student or group of students at a disadvantaged position (Kellaghan & Greaney, 2020). Individual and institutional dishonest acts in schools or school districts are becoming worldwide problems that introduce systematic errors (Downing & Haladyna, 2009). Recently, a systematic review of high-stakes testing has shown that construct irrelevant variances such as academic dishonesty challenge the appropriateness of inferences drawn about content knowledge based on students' performance in an examination performance (Acosta et al., 2020).

In Ethiopia, academic dishonesty has become a serious challenge to educational outcomes. Various research findings have shown that rampant cheating behaviors are orchestrated by different actors, which include school principals, district educational leaders, parents and students (Badasa, Nuri, & Gizaw, 2019; Chala, 2021). To address the academic dishonesty in ESSLCE, the MoE has started to manage the administration of the examination at public universities located all over the country since 2022/23 academic year. The rationale for the Ministry's decision for switching the previous school-based ESSLCE administration to university-based one largely lies on the high prevalence rates of student-related and institutional dishonest academic acts (Nigussie, et al.,2023). Accordingly, all public universities hosted thousands of students from their catchment areas. More than 976,000 students sat for the 2022/23 grade 12 national examination (Ethiopian-Monitor, October 4, 2022).

The high incidences of physical illness and examination anxiety during the university-based ESSLCE management practices (Nigussie, et al.,2023) coupled with the inherent anxiety and stress in every national examination, may jeopardize valid interpretation of the examination results (Koramoah, Dzakadzie, & Danyoh, 2022; Miller et al., 2009). In relation to a change of examination settings, Becker and Bergstrom (2013) emphasize the need for cautious decisions before switching to new examination programs from the perspectives of validity, customer service, and cost. Moreover, research findings have shown that alteration of the test setting could significantly affect the psychometric properties of a test, thereby affecting students' test performance (Lee, Reynolds & Willson, 2003). Fairness, reliability and validity are essential criteria to evaluate any examination which are in turn dependent on many factors, among which factors related to pre- and while-examination administration are mentioned (Gronlund & Linn, 1990; Rukundo & Magambo., 2010).The objective of the research is, therefore, to analyze threats of fairness, reliability and validity in the first university-based national examination management in Ethiopia. To achieve this objective, we have used the qualitative approach of validity and reliability evidences (Kane, 1992, 2013). A brief notes on fairness, reliability and validity will follow to give readers some insights on the factors which could act as threats for fairness, reliability and validity.

Fairness of examinations has got much attention and effort in the design, development, and use of assessments. Although fairness does not have agreed upon definitions, American Educational Research Association (AERA), the American Psychological Association (APA), & the National Council on Measurement in Education (NCME) (2014) define it as "a responsiveness to individual characteristics and testing contexts so that test scores will yield valid interpretations for intended use" (p. 50). To make a test fair, the test developer and administrator should make sure that construct-irrelevant sources of variances are kept to the minimum for various groups of students (Zieky, 2013)

Moreover, American Educational Research Association et al. (2014) state that fairness is fair and equitable treatment of all test takers during the testing process. This pertains to fairness in measurement quality, fairness as the lack or absence of measurement bias, fairness as access to the constructs measured, and fairness as validity of individual test score interpretations for the intended use. In addition, test developers should treat all test takers equally, provide accommodations for people with disabilities, obtain diverse inputs, use an empirical indicator of item fairness, obtain validation information, and help score recipients use the test results appropriately (Zieky, 2013). Lam (1995) suggested a cautious approach to ensure fairness, i.e. equality and equity. The equality dimension emphasizes same administration, scoring, and interpretation procedure of an examination where as the equity aspect includes tailoring assessments to the individual student's instruction context and special background.

Classical Test Theory (CTT), also known as classical true score theory, states that the observed scores of a group of test takers contain random and systematic errors, no matter how carefully a test is designed and administered. Hence, the observed score (*Xo*), the information we get from the students in responses to questions, can be represented by the simple formula $Xo = Ts + E_{s,}$ where *Ts and* E_s represent a true score and the error term, respectively (Andrich & Marais, 2019).

Random and systematic errors negatively affect the interpretation of test scores, and they reduce the consistency of the test scores, i.e. their reliability. Reliability is defined as the consistency over replications of the testing procedure (American Educational Research Association et al., 2014). Systematic errors, which include various psychological and situational factors, cause scores to be inaccurate by introducing construct-irrelevant difficulty/easiness or construct-contamination in score interpretation (Haladyna & Downing, 2004). Henceforth, in any test designing and administration, it is essential to identify and minimize construct systematic errors, which either systematically increases or decrease test scores for a group of test takers or individual examinees (Kane, 2013).

American Educational Research Association et al. (2014) defines validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 11). Validity is the basis upon which all examinations are built with the aim of producing true estimates of students' ability that could help to make valid inferences from standardized examination or teacher made tests rather than the examination or exam scores (American Educational Research Association et al., 2014).

The contemporary views of validity guided by Kane (1992) and Messick (1994) characterize validity as a unitary concept which exists in a continuum that requires multiple sources of scientific evidences to support or refute the interpretation associated with assessment data, as opposed to the claim of earlier literature which assumes distinct types of validity such as content or predictive validity of the measurement instrument itself (Kane, 2013; Peeters & Harpe, 2020). Evidences generated from the interpretive/use argument serve either to support the interpretation/use, or oppose the interpretation/use of the results (Kane, 1992, 2013).

There are about five sources or facets of validity evidences in the process of validity argument approach (VAP); namely, evidences based on test content, response processes, internal structure, relation to other variables, and consequences (American Educational Research Association et al., 2014). This classification could provide a systematic evidence for concrete and systematic structure for validities studies (Downing & Haladyna, 2009). The argument-based approach of validity provides evidences of the existence of a problem (e.g., inadequate coverage of content or the presence of some form of systematic error), thereby pinpoint areas for further improvement (Kane, 1992, 2013).

Similarly, different factors may act as a threat to validity evidences. Threats to validity are any factors in the examination processes that interfere with the students' demonstration of competence. Although there are different source to threat to validity, the most common ones are Construct Underrepresentation (CU) and Construct Irrelevant Variance (CIV) (American Educational Research Association et al., 2014; Reeves & Marbach-Ad, 2016). CU refers to lack of representation for the

domains to be tested by the assessment method while CIV is a random or systematic error introduced in the assessment results by variables unrelated to the construct being measured. Both CU and CIV decrease the accuracy of interpreting the examination results, and thus, minimize evidence for validity (Downing & Haladyna, 2009; Messick, 1994). Valid interpretation of test scores relies on the expectation that every test administration has been conducted under the same, standardized conditions of measurement. Standardization is crucial in test administrations because it reduces the chances for sources of CIV, especially for CIV that is external to test takers. Test administrators must, therefore, consider that any compromises in standardization are likely to reduce the validity of score interpretations. One of the factors identified as a threat to validity is construct irrelevant variance (Haladyna & Downing, 2004).

Statement of the problem

The sources of random or systematic error in actual exam administration could result from the variability of examination site, interruptions during examination and student-related factors such as exam anxiety, lack of motivation to perform on a test, fatigue and negative attitude towards the examination (American Educational Research Association et al., 2014; Haladyna & Downing, 2004). The following table summarizes the sources of errors for CIV related to group of students as well as individual student, which are related to administration practices along with illustrative examples.

Actual test administration-related construct –irrelevant variance (CIV) which threatens validity are conceived as either affecting group of student or the individual student. Altering test administration requirements, variability across testing sites, interruption during test sessions, and differences in the time of the day at which the test is administered are the factors that affect group of students. On the other hand, test anxiety, cheating, prior exposure to test items; students motivation and students' fatigue (cognitive, emotional and physical) are sources of validity threats related to individual students. (American Educational Research Association et al., 2014; Haladyna & Downing, 2004)

The sources of group-level CIV from exam administration practices ultimately impact scores in one of two ways: students' scores are either higher or lower than they would be if the sources of CIV were eliminated. At some point in time, virtually all students experience test environments that are not conducive to effective test performance. Students' specific CIV is a systematic error that can overestimate or underestimate an individual examinee's score (American Educational Research Association et al., 2014; Haladyna & Downing, 2004). Compared with others sources of CIV, students potentially provide the most serious CIV threat to validity (Haladyna & Downing, 2004).

In line with this, the university-based exam management which was accompanied by several novel practices, and put students in severe adverse physical and psychosocial states may run against the principles of sound examination management (fairness, reliability and validity) and the introduction of random or systematic errors might be prevalent.. On this account, the exam management may largely contribute to the risk of false negative results (students got failed marks in fact when they passed or vice versa). This in turn may jeopardize the trustworthiness of the exam management (Miller et al., 2009).

This article, therefore, conveys research findings regarding the first university-based ESSLCE management focusing on the factors which threaten fairness, reliability and validity. We adopted the argument-based approach of validity evidences as it may provide evidences of the existence of a problem (e.g., the presence of some form of random or systematic error) in ESSLCE management, which in turn enable the responsible bodies to improve the procedures of the examination (Fechter et al., 2021; Kane, 1992, 2013).

Accordingly, attempts were made indicate the sources of examination administration errors that could contribute negatively to valid interpretation of the ESSLCE results. The findings reported in this article could provide valuable insights for the MoE and other stakeholders so that they could work together to improve the prospective ESSLCE managements in general. Particularly, the MoE and EAES can be able to consider and address the different sources of random or systematic error. A shared responsibility, honest examination and management of random or systematic errors are likely to

improve the effectiveness of the examination and increase the validity evidences for the future ESSLCE and other similar high-stake examinations. Specifically, the article targets the following research questions:

- What threats of fairness were observed in the exam management of the first University-based National Examination Management in Ethiopia?
- What threats of validity were prevalent in the exam management of the first University-based National Examination Management in Ethiopia?
- What threats of reliability were observed in the exam management of the first University-based National Examination Management in Ethiopia?

Conceptual framework of the study

To realize a smooth administration of examination, the period before, during and after the test should be carefully managed (Rukundo & Magambo., 2010); so as to minimize the sources of errors. Some research findings indicated the importance of test administration procedures as major sources of errors in students' academic measurement if not handled properly. Haladyna, Nolen, and Haas (1991) indicated several irrelevant factors in relation to examination administration practices which could pollute students' academic results.

As indicated in the figure below (the conceptual framework of the study being reported in this article), different factors constituted to the before, during and after exam administration in the first university-based ESSLCE management. Factors related to before and during periods of exam administration are particularly important and influence the interpretation of examination results (Haladyna & Downing, 2004).

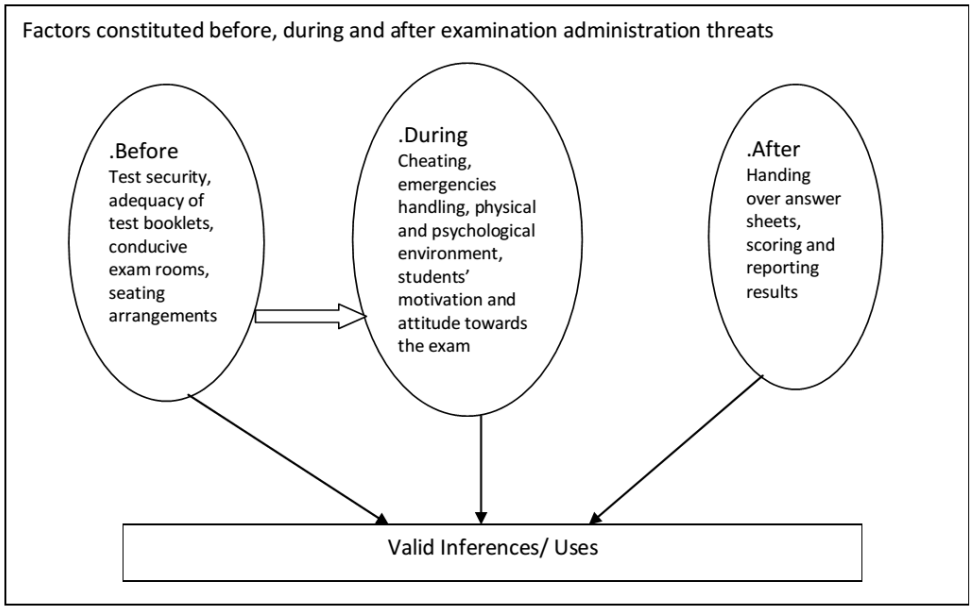

Figure 1: Factors constituted in the before, during and after examination administration threats

The before exam administration time includes tasks related to assembling and readying the exam booklets, securing the exam booklets, preparing conducive exam rooms, arranging the seating, recruiting and providing orientation for chiefs, supervisors and invigilators. Assembling and readying a test for administration require close attention to test validity. A typographical error can ruin an item or even a whole test (Campion & Miller, 2006). In a similar vein, Keeves (1994) and American

Educational Research Association et al. (2014) state that exam invigilators or supervisors should not be the same person who may be held accountable for the performance of students. Test administration guideline and manual should be prepared, and adequate orientation should be provided for invigilators to uphold the security and ensure the standardization of the examination (American Educational Research Association et al., 2014; Haladyna & Downing, 2004).

Location of the testing site and age of students are also one sources of error. Haladyna and Downing (2004) state that displacing students for large scale testing could be a source of error mentioning the classroom in which students had an experience with is the most natural environment. They also indicate younger students may perform better in their own classrooms. Cue dependent theory also posits that students are more likely to recall information during a test when they take a test in the same place as they were when learning (Tulving, 1974).

Studies have shown that high exam anxiety during actual examination time generally lowers students' examination performances. A meta-analysis conducted by Hembree (1988) indicated the deleterious effect of examination anxiety on student psychological state. Recently, Heissel, Adam, Doleac, Figlio, and Meer (2021) have shown that chronic stress which was brought about by violence, poverty, or family instability can affect how individuals' bodies respond to stressors in general, including the stress of standardized testing and questioned the validity of the high-stake examination as valid measure of students' actual ability. Heightened academic stress in the final years of schooling is a common concern, too (Wuthrich, Jagiello, & Azzi, 2020).

Due to environmental changes, students can potentially experience different types of stress that can affect their mental health, social health and academic achievement. Environmental stressors associated with senior school and high-stakes examinations may interact for biologically vulnerable students to exacerbate or trigger underlying stress vulnerability. These factors might be school-based (e.g. increased learning requirements, pressure to perform), home-based (e.g. pressure to perform, diet, sleep routines), peer group-based (e.g. social evaluation concerns, social contagion of stress) and the like (Wuthrich et al., 2020).

Students' demotivation during the test is also a powerful source of error regardless of the performance level of the student. Students may seriously underperform, make random marks on the answer sheets, omit answers, or not finish the test. The frequency of omitted responses and items not reached are signals of low motivation and non-compliance (Haladyna & Downing, 2004). Fatigue is another source of error. Young students may be more susceptible to fatigue in long testing situations than older students, and the conditions for test administration may interact with different types of students (American Educational Research Association et al., 2014).

METHODS AND MATERIALS

Design: A concurrent mixed method design (Qual.+ Quan.) was used. This method is helpful to secure detailed information from multiple sources, and to ensure triangulation (Bryman, 2012). The qualitative part mainly addressed experiences of different stakeholders in the pre and during examination management while the quantitative part addressed students' physical and the psychosocial states during the examination administration.

Participants and sampling procedures: We purposely sampled ten regions and one university located in the region conveniently (Table 1) as this approach ensures representation of the university's located in the country. Twelve university officials, ten university clinic health professionals, ten regional education bureau officials, ten regional assessment experts, six current invigilators, three current supervisors, three previous invigilators/supervisors, three students, three school directors, two EAES experts; and four MoE and EAES' documents were selected purposely as data sources for the qualitative part of the study. Nine hundred and thirty-nine (938) students who took the 2022/23 ESSLCE at Jimma University were surveyed in January 2022/23 to elicit data on their experiences of the ESSLCE management. The students were selected using available sampling method from schools located around Jimma Town, and 938 students were deemed a sufficient sample for the total population of 976, 000 exam takers based on Krejcie and Morgan (1970) sample size determination

method. The Krejcie and Morgan sample size determination method was used as it is relevant for surveying large number of participants.

Table 1: Sampled Regions, and Universities

Regions/Regional educational bureaus	Universities
Addis Ababa City Administration	Addis Ababa University
Benishangul Gumuz Regional State	Assossa University
Amhara Regional State	Bahir Dar University
South Western Ethiopian Peoples Regional State	Bonga University
Dire Dawa City Administration and Harari Regional State	Dire Dawa University
Gambella Regional State	Gambella University
SNNPR and Sidama Regional State	Hawassa University
Somali Regional State	Jigjiga University
Oromiya Regional State	Jimma University
Afar Regional State	Semera University

Instruments: Interview guides were used to collect data through interviews while standardized adjustment and exam anxiety scales, along with some open-ended items, were administered in the survey. The Examination Anxiety Scale (EAS) was adapted from the one developed by Abbasi and Ghosh (2020). The scale contains 21 items which measure four major components/subscales of examination anxiety. The subscales are emotional, behavioral, physical and cognitive anxiety. The psychological adjustment scale which was developed by Pennebaker (2013) contains 19 items that are categorized under three sub-scales: positive affect, negative affect and home sickness. The instrument was commented by experts, and it was modified by removing some items as they had little relevance to the study's context. A pilot-test was conducted in Seto High and Jimma Comprehensive High Schools with 90 students who sat in grade 12 examinations, and it was found the feasibility of the instrument was sound. The reliability estimates for the various subscales, along with the modifications made, is reported in Table 2. The behavioral, and positive-affect subscales of examination anxiety, and adjustment scales were excluded as they had very low reliability estimate, .585 and .524, respectively.

Table 2: Psychometric properties of the exam, and adjustment scales

Sub-scales	No. of items	Deleted	Deleted items	Cronbach's alpha
Emotional	5	1	When I sat for ESSLCE, I felt thrilled	.762
Cognitive	6	1	I thought about current events during ESSLCE	.753
Physical	5			.830
Negative effect	5	1	Felt angry	.638
Homesickness	6	3	Felt lonely, felt anxious, and liked being away from my parents	.700

Procedures: A team of trained data collectors and supervisors were dispatched to the MoE, EAES, universities, regional education bureaus and other sites to conduct the interviews and administer the questionnaire. The study participants were briefed about the nature of the research, and they provided their verbal consent for the interview. The interviews were tape-recorded and transcribed verbatim in the English language by the data collectors.

Data analysis: The data analysis was guided by the research questions stated in the introductory section of the article. For the qualitative analysis, coding, recoding, categorizing and then theme identification were undertaken. Put differently, the qualitative data gathered through in-depth

interviews were analyzed thematically. Frequency counts and percentages were used to analyze the data on the emotional anxiety and adjustment.

Ethical considerations: After the tool development was finalized and the proposal was ethically approved, the research team secured a letter of support from the Vice President for Research and Community Services (VPRCS) of Jimma University (JU). Participation in the study was voluntary, and participants and settings of the study were reported anonymously using codes. Only the research team accessed the recorded information for confidentiality, and the study ensured protection of participants from any type of psychological or physical harm.

RESULTS

Socio-demographic characteristics of respondents

Nine hundred thirty-four students (452, 44.0%) and (575, 56.0%) males and females, took part in the survey. The age range was from 17 to 29 years with the mean and median ages of 19.62 and 19.00, respectively, and standard deviation of 1.38. Four hundred seventeen (40.6%) and 610 (59.4%) students were from social and natural science streams, respectively. Eight hundred seventy three (873, 84.9%) attended public school while (155, 15.1%) attended private schools.

Threats to Fairness: Students traveling from distant areas and adjustment problems

The question of fairness abounds on account of collecting students to university compounds with which they are little familiarity, and the travel. In this regard, one participant commented:

> The university-based ESSLCE management put pressure on students. It had big impact on students. Let alone being moved to unfamiliar place, taking the exam being in one's own school is difficult as the exam is a "matric". We knew what we were feeling when we were examined. It has pedagogical implication. This is the last option. It is not advisable pedagogically under a normal situation. (UO, 9)

Students' traveling to university, especially from distant places, is also a source for questioning the exam for fairness. Students travelling to their respective universities were exhausting to the extent of putting them to extreme fatigue. In this regard, traveling long distances to take exams was very much tiresome for the students (students' responses to open-ended items of the questionnaire). A number of students encountered car accidents on their ways to the universities they were assigned to (UO, 6).

Threats for Reliability and Validity: Assignment of exam personnel, assembling and readying test booklets and students movement to university centers

Introduction of errors and thus invalid inferences of test results could occur in a situation where chiefs, supervisors and invigilators are ill-oriented. Shortages of exam booklets and answer sheets and poorly prepared exam booklets were some of the irregularities. Some participants reported that there were some problems with respect to chiefs, supervisors' and invigilators' knowledge of exam administration and behaviors during the actual examination. For instance, one participant claimed:

> *The invigilators who were assigned to the universities in our region were university teachers and examination experts in other regions. And, there appeared the issue of mistreating some students. The students have indeed been mistreated.* A p*rofessional sent from the MoE as exam center manager must be capable … the person who was sent during social examination round had no knowledge about how to handle the examination process. Even he said, 'The MoE sent me only because of the lack of human resources.'* (UO,06)

Assembling and readying a test for administration is a perquisite for valid inferences from test results. However, there were reports of some irregularities in items of arrangement and organization of exam booklets, which might have distracted the smooth execution of the new ESSLCE management. In this instance, it was reported that, "…. the presence of double or repeated questions in some of the exam booklets, and there were also shortages of exam booklets." (Region Education Assessment Experts, REAE, 28) and "not providing exam booklets for a student at Wachamo University…" (MoE implementation report).

Threats of Reliability and Validity: Magnitudes of exam cheating

The Ministry of Education reported that stealing/theft of exam booklets prior to the actual examination administration, which was rampant before, was eradicated during the new ESSLCE management. However, there were some reports of minor exam irregularities during the actual examination period across all universities. The commonly observed problems were that some students were found hiding their phones in their shoes, different parts of their bodies, their clothes and food ('kolo') and their hairs. Throwing the mobile phones to university compound, keeping a sim cards with them, was also observed. Few students were also caught red-handed using telegram app for sending messages during exam administration (MoE report).

Although almost all of the respondents in the in-depth interviews agreed that the new ESSLCE management helped to reduce the magnitude of students' cheating during actual examination administration close to zero, there were some instances of actual cheating or attempted cheating as reported by the exam invigilators, and students self-reported cheating indicated. For instance, an interviewed supervisor said, "There was a great expectation from students to copy answers from each other." (Supervisor 1). In addition, other interviewed supervisor said, "some students mindset was dependent on others so that they wanted to copy from other students." (Supervisor 2)

It was also reported that students used or tried to use various methods of cheating during the new ESSLCE administration. Accordingly, copying or trying to copy, snatching exam answers from other students by force, and using crib sheets were observed. The interviewed invigilators and supervisors reported that these problems were observed on both the private and government school students. On the contrary, majority of the interviewed invigilators and supervisors reported that they had not come across cheating in the form of using mobiles, paper exchanges, social media usages and exchange of exam sheets.

There were also reports of group cheating or intimidation of invigilators by students in some universities like Arbaminch University (Nech Sar Campus), Bahirdar University (Tibebe-Ghion Campus, Gish Abay Campus), Dilla University (Hasideli Campus), Wollo University (Kombolcha Campus), and Jigjiga University (MoE report). Moreover, it was reported that some students left the examination center as they were not willing to start or continue the exam. Hence, 1712 students (691 males and 1021 females) from Mekdella Amba University and 7356 students from Debre Tabor University left the examination centers for they were not willing to continue the exam (MoE report) (see summary in Table 3 below):

Table 3: Number of students who did not start and those who discontinued the examination

Exam Center		Registered	Showed up	Did not show up	Sat for the exam	Left the campus
Mekdela Amba university (Mekaneselam Center)	Male	1077			386	691
	Female	1563			542	1021
	Total	2640			928	1712
Debre Tabor 1		9,027	8,040	987	3,933	4,107
Debre Tabor 2		8,089	7,401	688	4,207	3,249
	Total	17,116	15,441	1,675	8,140	7,356

Source: MoE Implementation Report, October 2022

The quantitative finding also confirmed some form of exam irregularities. Likewise, it was found that 110(11.8%) students reported attempted cheating; 267(28.5%) students witnessed that they saw/heard of students who cheated; 24(2.6%) and 14(1.5%) reported that they saw invigilators and supervisors who unlawfully helped students, respectively (Table 4).

Table 4: Incidents of academic cheating

Variables	Yes	No
	f (%)	f (%)
Did you attempt to cheat in the ESSLCE held at university?	110(11.8%)	825(88.2%)
Did you see/hear students who cheat in ESSLCE?	267(28.5%)	669(71.5%)
Did you see invigilators who unlawfully help students in ESSLCE?	24(2.6%)	911(97.4%)
Did you see supervisors who unlawfully help students in ESSLCE?	14(1.5%)	920(98.5%)

Different strategies were used to prevent cheating during the new ESSLCE management. The coordination of the security with the exam supervisors and invigilators; invigilators' commitment for shouldering the exam administration responsibility at most professional level and integrity, the prohibition of mobile phones, the use of several coding for the exam booklets, the well-worked out seating arrangement, and the little familiarity between students and invigilators were some of the strategies. In this regard, one interviewed participant commented on the new exam management, in comparison to the previous ones as, "We were allowed to keep our phones with us in the previous ESSLCE examination…, and I had witnessed invigilators who played with their phones instead of proctoring students carefully, which had contributed to students cheating in the previous ESSLCE exams" (Supervisor 1).

Similarly, the participants unanimously agreed that the decision to bring the ESSLCE management to the university compound took the lion share for the low level of academic cheating as opposed to the school-based exam management. In this instance, one interviewed supervisor told the experiences comparing the previous and the new ESSLCE management. The supervisor said:

> *In the previous exam administration, everybody motivated students to cheat. The security personnel and the education personnel wanted students to cheat during the exam. When we reported on somebody who snatched a test booklet and run away to the nearby forest, no security personnel responded appropriately. The security personnel ran to the left while the culprit headed to the right. So, there was a temptation from all sides to encourage cheating. But, in the current exam administration, first the absence of cell phone had greatly reduced cheating. Secondly, since the exam environment and the invigilators were new for students, no incidences of academic cheating were observed. Students did not get an opportunity to cheat; so, they marked only what they knew; no cheating. If it continues like this, the burden for universities receiving unqualified students or students who join universities as a result of cheating will be greatly reduced, and this will bring quality of education eventually. I believe I supervised the current exam free from cheating.* (Supervisor 1).

There were also recommendations from the participants for the upcoming ESSLCE management based on the experiences they had for students attempted and the actual cheating observed during the new ESSLCE management. One interviewed participant recommended:

> *If we want the exam to effective in the future, it is better if students' seats are arranged in alphabetical order rather than arranging them to sit in the exam room by schools. If it is in this way, the students are not known to each other, and hence the invigilators will not also be challenged on controlling.* (UO, 9)

Threats to Reliability and validity: Psychosocial states of students and the testing environment

Students need to be put in a sound psychological state to properly display their maximum performances in any examinations. However, the new ESSLCE management affected most of students' psychological states owing to the new environment they were exposed to, couple with the inherent anxiety and stress any examination induces, let alone the ESSLCE.

In this regard, almost all of the interviewed participants disclosed that most of the students were in negative psychological states. Students experienced instability, test anxiety, stress, depression, expectation crisis. All these in turn led them to crying aloud, and some of them were fainted out while taking the examinations. One of the participants said, "It seemed that the students did not have the expectations for this kind of exam administration… the majority of the students seemed not very well-prepared for the exam... and were in stressful conditions." (Current Invigilator 5). Similarly, one participant noted, "…. There were many students who were in stress and fainted out before, during and after the exam; they also suffered from other psychological problems related to exam anxiety…." (UO, 5). One participant reported on the psychological state of students and the underlying factors as follows:

> *Several students were stressed. The first thing they were not allowed to take mobile phones with them, most students were separated from their family for the first time, as a result they had a feeling of homesickness, there were fear among some students because they were not able to communicate to their family as their phones were not with them. The environment was new for them; the food and everything were new. There were students who were stressed as a result. I witnessed students who were stressed because they came from far place. As a result, when they were alone without their phone, they were stressed as they felt they were left in the middle of nowhere. The environment was new to them, they were not allowed to go out of the campus, I think these all things hurt them a lot as they come from distant areas.* (Current Supervisor 2)

There were also incidences of students' illnesses during the actual exam administration which seemed to be the signs and symptoms of negative emotional states. One participant narrates the situation as follows, comparing the current incidences with the previous ESSLCE managements.

> *We took several tensed students from the exam room by hugging. They said they were sick. I had been invigilating or supervising for more than 15 years. I have not seen such problems in the past. In the room I entered to supervise, I came across some students who fainted out; even we questioned whether it was real or fake, or one of the techniques to cheat. The whole environment was tense. When we entered the center; we got checked. These things might induce fear; the fear is not good for students. The atmosphere was not conducive*. (Current Invigilators 2)

Similarly, the quantitative finding disclosed high levels of students' emotional and adjustment problems (Figure 2). One in two students exhibited high level of emotional anxiety; and one in three students reported high levels of physical, and cognitive anxieties. Adjustment problem related to homesickness was also highly prevalent (73.10 %), and nearly half of the students exhibited negative affect (45%).

There were also reports of discontinuing the exam among some students due to various illnesses. One student experienced still birth while the another one had gastritis (Jinka University); one student fainted out (Bahir Dar University, Zenzelima Camps); students shouted in groups, complaining that they were sick (Debark University); two students continually shouted and fainted out due to stress (Borena University), death of a student's due to heart failure was reported (Ambo University, Ambo Campus), two students were shouting during examinations and left the exam room (Weldiya University); five students were unable to continue due to stress (Addis Ababa University, 5 Kilo Campus); twenty students were shouting (Jimma University); eight student were on and off to complete the exam due to stress (Arsi University, Asseli Health Campus); ten students experienced stress and nine students completed the exam on and off but one student was unable to continue the exam (Jimma Technology Campus). (MOE Implementation Report, October 2022).

Data from some university clinics revealed the seriousness of students' physical illnesses. In university 09, the total number of students who received health support was 11,718(41%). Out of this, 7,507 (65%) registered in the first round (Social Science) and 4,211(35.93%) in second round (Natural Science) (UCHP 09).

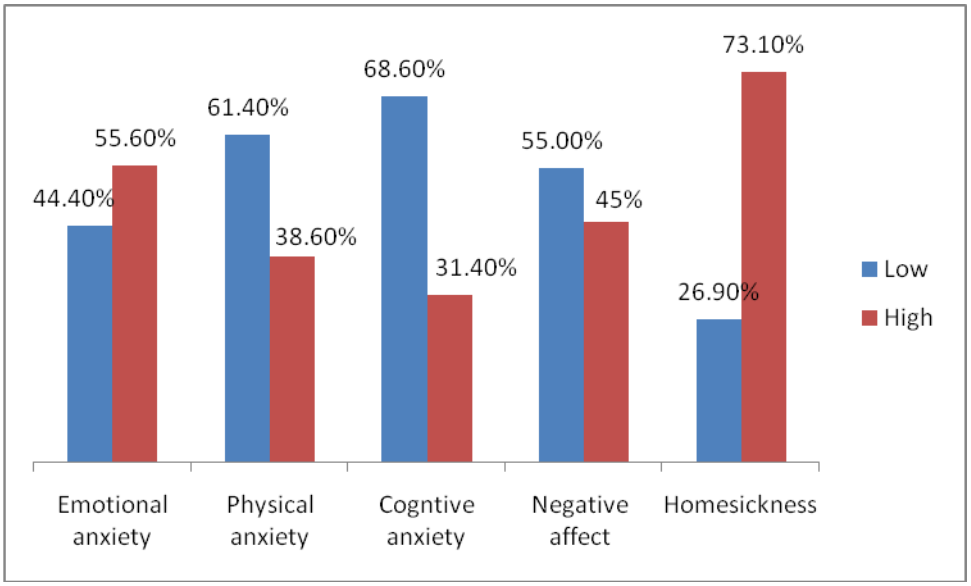

Figure 2: Percentages of various forms of examination anxiety, and psychological adjustment

Similarly, 6972 students sought medical services in university 08. In this university, 2018(28.94%) attended OPD; of this 1238(61.34%) were males while 780(38.65%) were females. On the other hand, 4954(71.06%) attended emergency, of which 1240 (25.03%) and 3714(74.97%) were males and females, respectively (Table 5).

Table 5: Students who were admitted to a university clinic

	Male	**Female**	**Total**
OPD	1238 (61.34%)	780(38.65%)	2018(28.94%)
Emergency	1240 (25.03%)	3714(74.97%)	4954(71.06%)
Total	2478 (35.55%)	4494(64.45%)	6972(100%)

Source: UCHP, 08

With respect to students' motivation to complete the examination, the majority of the participants reported that most of the students sat idle, showing little concentration on the examination and waiting for answers to come for them. Some also filled the answer sheet within 15 to 20 minutes, without reading the test booklets while some others did the exam by merely guessing and left the rooms. In this regard, one interviewed participant mentioned the psychological readiness and concentration of students and the underlying reasons as follows:

> *There were only few students (not more than 5%) who had psychological readiness and demonstrated concentrations on the exam. Others were sitting without working on the exam. In my view, the majority did not have preparations for the exam. The majority of the students sat idle without reading the exam sheet. Some of them randomly filled out the answers, without reading, and left the exam room very early, i.e., before 30 minutes... I asked some of the students concerning why they did not work on the exam. They responded that they had no enough educational resources in their schools, um, such as books…. They told me that their schools were empty and that they didn't have well-prepared and experienced teachers; as a result, they lacked good preparations for the national exam.* (Current Invigilator 3)

Some participants ascribed the students' poor concentrations on the exam to negative psychological states, and exhaustion, among the students, which resulted from travelling long distances to reach for the exam sites. For instance, one of the participants said, "In my observation, students got tired because

of travelling long distances. It would be good if they were assigned to nearby places. This affected their concentration and adjustment to work on the exam." (Current Iinvigilator 6).

With respect to the physical environment, the participants indicated the environmental conditions ranging from cleanliness of the exam classes, students seating arrangements, and classroom lightings were in very good conditions for proper exam administration. However, there were some problems in some examination centers. For example, one participant reported it as, "...most of the facilities were good but at the beginning of exams, there were deafening ambulance sirens while they took students who fell ill in exam rooms, but after few days, it was resolved" (Student 2). Another participant also added, "Invigilators were disturbing us within the class room" (a student's response to an open-ended item of the questionnaire). On the other hand, a third participant reported the following:

> *We supervised students in clean environment, spacious room with adequate lighting, and comfortable chairs, maintaining the distance from one student to another. I observed no problems with regard to classrooms and seating arrangements. In the previous exam management, students had been invigilated in muddy and suffocated rooms. But, at the university, there were standard classrooms which had adequate lighting. There were no student congestions, and we arranged the seating … comfortably to smoothly conduct the exam administration, maintaining appropriate sitting distance. I could say it was good in the current administration". (Supervisors 2)*

On the other hand, some students reported the poor quality of the physical environment. One student reported the quality of his/her exam room as, "The roof of the class where we took the exam had a hole that drops water. As a result, there was a nasty smell in the class…." (a student's response to an open-ended item of the questionnaire). Besides, the heavy security presence during the examination period induced fear and nervousness in the examinees. Over stated roles attributed to the security personnel compared to other relevant actors also inhibited free movement of invigilators, supervisors and supporting staff during the exam administration. In this regard, one student reported: "…in the exam room, the security personnel come to the exam room with their big gun. In the moment, I felt very nervous…." (A student's response to an open-ended item of the questionnaire).

DISCUSSIONS

Any sources of errors induced during examination administration thwart inferences of examination results. In the 2022/23 university-based ESSLCE management, several issues were evident before and during the actual exam administration which stand against fairness, reliability and validity of the examination results.

To minimize the systematic error and increase the validity of test results, careful recruitment of exam chiefs, supervisors and invigilators are vital, and proper orientation should be provided for these exam personnel, along with providing them with clear working guidelines (American Educational Research Association et al., 2014; Haladyna & Downing, 2004; Keeves, 1994). However, it was found that some invigilators mistreated students; it was also indicated that some exam chiefs did not have good knowledge of exam administration. There were also some problems pertaining to items arrangements and assembling of test booklets. Such irrelevant construct errors may interfere with the concentration of students and limit their abilities to demonstrate their full potential in the examination. In this respect, research finding indicated the need for working carefully to avoid typographical errors in test items and careful assembling and readying of examination booklets for valid examination results (Campion & Miller, 2006).

Cheating has been proved to be a serious challenge to test administration and interpretation of test results. The new university-based ESSLCE exam management deters academic dishonesty significantly compared to the previous ones. However, there were some instances of cheating during this ESSLCE management. This also could be one source of error that hampers valid test inferences. Several scholars established that any forms of academic dishonesty introduce systematic errors that

counter valid inferences of students examination performances (Haladyna & Downing, 2004). A systematic review of high-stakes testing has shown that construct irrelevant variances such as academic dishonest acts (i.e. systematic error that systematically affect test performance) challenge the appropriateness of inferences drawn about content knowledge based on students' examination performances (Acosta et al., 2020).

Students' physical and psychological conditions are also important elements in validity inferences. A research has shown that students need to be in good states of health and mind during examination as these influence their academic performances (Dawood, Al Ghadeer, Mitsu, Almutary, & Alenezi, 2016). Seen from this perspective, the 2022/23 ESSLCE introduced several errors. The majority of the students were not in good physical conditions. They encountered various forms of illnesses which hindered them from fully focusing on the examination, and even in sometimes forced them to discontinue it. It is obvious that students should be in a good health so as to focus on the examination and complete the exam, displaying their full potentials and with optimum motivation (Hamilton, Freche, Zhang, Zeller, & Carroll, 2021 ; Heissel et al., 2021). Otherwise, the students' health problems might be sources of errors that could jeopardize efforts made to make valid inferences from the examination results.

Another serious flaw in the 2022/23 ESSLCE management was students' negative psychological state. Students experienced exam anxiety and adjustment problems in the university. Higher level of anxiety threatens students' mental and physical health, and results in negative consequences on examination performance. A meta-analysis conducted by Hembree (1988) indicated the deleterious effect of examination anxiety on students' psychological state; the evaluative threat of examination can increase examination anxiety (Hill & Wigfield, 1984). Hancock (2001)) also provided evidences that disposition to anxiety and the high-stakes situation contribute to exam anxiety which in turn contributes to powerful sources of error. Recently, Heissel et al. (2021) have shown that anxiety can affect how individuals' bodies respond to stressors in general, including the stress of standardized testing; they questioned the validity of the high-stake examination as a valid measure of students' actual ability.

Students' level of motivation during the test can also be a powerful source of error. It was indicated that the majority of students exhibited low motivation to complete the examinations during 2022/23 ESSLCE management which can be inferred from their lack of concentration on the exams, random marking on the answer sheets, sitting quietly without reading the exam booklet, and not finishing the test. Such behaviors during examinations are signals of low motivation (Haladyna & Downing, 2004). The low motivation and concentration of students during the examination may be linked to cortisol responses for higher exam anxiety level (Heissel et al., 2021).

The testing environment can be responsible for heightened anxiety and adjustment problems among test-takers. Environmental stressors associated with high-stakes examinations may interact for biologically vulnerable students to exacerbate or trigger underlying stress vulnerability (Wuthrich et al., 2020). The heavy security presence and the university environment could be cited as sources of environmental anxiety. High exam anxiety, coupled with poor sleep pattern, impair academic performance, and the reciprocal relationship of sleep and anxiety may even influence the overall health of students (Hamilton et al., 2021).

Fatigues is another source of error (Haladyna & Downing, 2004). In 2022/23 ESSLCE management, students had to travel several kilometers to reach to the university compounds, and the receptions in universities took long times due to large numbers of students requiring the service. The cafeteria, accommodation, and security services of the universities did not meet expectations, and students had difficulties in sleeping due to disturbances around the dormitories. Research evidences have indicated that fewer hours of sleep is associated with higher risk of failing in examinations; conversely, adequate and quality sleep is correlated with better grades (Guadiana & Okashima, 2021; Okano, Kaczmarzyk, Dave, Gabrieli & Grossman, 2019; Vedaa, Erevik, Hysing, Hayley & Sivertsen, 2019).

CONCLUSIONS, IMPLICATIONS, AND RECOMMENDATIONS

Conclusions

The objective of the study was to examine threats to fairness, reliability and validity of the ESSLCE management, particularly focusing on the actual examination administration. The study was cross-sectional in nature, which needs cautious interpretations of the cause-effect relationship of variables. Moreover, the survey was administered two months after the ESSLCE ended, and this also might be resulted in recalling bias. Moreover, the investigation focused only on some of the before and during factors of exam administration due to lack of access to data on post-exam administration variables. A reader should take into note the limitation along with the following conclusions:

- It was evidenced that there were some problems in item arrangements in some subjects and shortage of exam booklets, which most likely disturbed the smooth flow of the exam administration and created stress among students and invigilators.
- There were some surrounding noises, like ambulance's sirens, which interfered with the concentration of students.
- There were still reports of actual cheating or attempts of cheating. However, in general, it is undeniable that the new ESSLCE management significantly reduced exam cheating. However, the question is at what cost the result was achieved, especially from the perspective of students' physical and psychological trauma.
- For students to properly demonstrate their maximum performances in any examinations, they need to be in positive psychological and physical states. The new ESSLCE management, in this regard, suffered from several limitations. Many students lacked the right concentrations and motivations to complete the exams. Similarly, physical illnesses, exam anxiety and home sicknesses were prevalent during the actual exam administration. Such negative physical and emotional states among students are likely to affect their exam performances, and in effect, the reliability and validity of the examination results could be questionable. Student who were ill and coming out of the exam classrooms, and allowed to continue the exam when they recovered. Such practices could have greatly compromised the validity and reliability of the ESSLCE management. Besides, how one can continue the examination in this state of mind and body conditions remains unanswered.
- In sum, the new ESSLCE administration, which was characterized by gross students' physical and psychological conditions, seem to fail satisfying the criteria for a good test in general. Although the MoE and EAES were greatly concerned about putting the new ESSLCE into place to fight off cheating, this ESSLCE management could have been measured or checked against the criteria of sound assessment principles equally, i.e., fairness, reliability and validity, prior to its implementation. Hence, many decisions await the MoE and EAES ahead to provide or experiment with an optimum ESSLCE management for the upcoming ESSLCE examination from the perspectives of validity, reliability and fairness, and even for ethical reasons.

Implications

In the light of the analysis and interpretation of the data from various sources, the potential effects, or consequences of the findings are forwarded based on the following pertinent questions related to assessment principles. The first question focuses on fairness, the second and the third are for reliability and validity.

1) Did the new ESSLCE management provide equal opportunities for all students?

Any examination management should provide equal opportunity for students. In other words, it should not work in favor or against any students. One of the principles of measurement is, thus, fairness. Fairness denotes creating suitable environment for all students, and providing equal playing field for all students in examinations. It could be seen from equality and equity perspectives. The new ESSLCE management seemed fair at it face value (provide equal field) for all students as it was administered under uniform conditions which were free from major irregularities. Nevertheless, deep inside, the question of fairness surrounding this examination is clearly evident. Grade 12 students, with different

exposures, and gathered from all corners of the country at universities to sit for the examination characterized with little time for adaptation to the environment, the university life and food types, are likely to face varying treatments, at least to some degree.

Most importantly, the majority of the students who were from rural and distant places travelled long miles, and as a result, suffered from physical and psychological exhaustion as a result of which their examination performances could be negatively affected. On the other hand, students who came from areas close to the center university, accustomed to city life and food provided in the university, seemed to be more advantageous than the students who were far from a university, rural areas and those who were less accustomed to the meals usually served at universities. In this regard, the new exam management seemed to violate the issues of fairness (equity was at risk). It was in favor of urban students, students who were from the surrounding areas, and students whose home diet matched with the food served in a university

2) Are students' ESSLCE results accurate and consistent?
This question focuses on validity and reliability. For an examination to be psychometrically sound, environmental, physical and psychological obstacles should be kept to a minimum. Otherwise, reliability and validity of the examination results would be at risk.

The supra environment in which the exam is being administered includes the experiences students had from the moment they left their villages on the way to the university to their whole stay at the university. During this time, students passed through several negative experiences. There were shortages of transportation, and consequently, students had to wait for several hours; there were also reports of congestion, and even traffic accidents. Moreover, in some places, students had to travel for days to reach their destination universities. Even after arrival at universities, there was long waiting hours, in the blazing hot sun, to get registered and obtain security clearance. These negative experiences undoubtedly hamper students physically and psychologically, and may negatively affect their concentrations and motivations during the actual exam administration.

Similarly, the heavy security presences, which are new experiences for most students, were sources of problems. There were also plenty of complaints from students concerning the quality of food services, and disturbances around the dormitory, which make sleeping difficult. Due to large number of students, it was necessary to get up early for breakfast, at 4:00 am at some universities. Fewer hours of sleep and poor quality of sleep are related, directly, to poor concentrations and low motivation to complete the ESSLCE exams, and eventually may affect the performance of students negatively.

Moreover, students' physical and psychosocial health plays vital roles in examination performances. Many students experienced different forms of illnesses during the examination. Some dropped out, and many students completed the exam struggling with their illnesses. There were also students who discontinued the examinations for a brief period due to illnesses but allowed to resume taking the exam. Most of the students encountered adjustment problems (such as homesickness) and had developed examination anxiety. The students' negative physical and psychosocial states during the examination could interfere with their performances and limit their abilities to display their capacities to the fullest, thereby ruining the valid interpretation of the examinations results.

In sum, it was in these situations, i.e., the adverse environmental, physical and psychosocial conditions that the students were forced to take the ESSLCE examinations. Therefore, we need to interpret students' ESSLCE results cautiously taking into account the adverse environmental, physical and psychosocial conditions the examinees were in.

3) How much of the variation in the ESSLCE results attributable to students' real performance, and how much to random and systematic errors?

According to Classical True score Theory (CTT), any observed score (o_s) is composed of a true score (t_s) and error score (e_s Es), mathematically represented as, $o_s = t_s + e_s$. To have accurate representations of students' performance (o_s) in any examination, adverse environmental conditions

and negative physical and psychosocial factors in students, which contribute to negative exam performances, should be controlled as much as possible. As a result, the error score (e_s) would be minimum.

When we evaluate the students' results (o_s) derived from the new ESSLCE management based on the theory of CTT, it seemed that students' observed scores (o_s) were contaminated with several irrelevant factors or errors, i.e., adverse environmental conditions plus adverse students' physical and psychosocial states, which in turn may have contributed to higher percentage of error variances (σ_e^2). Therefore, it seemed fair to conclude that the proportions of students' observed variance (σ_o^2) attributable for error variance (σ_e^2) was high as a result of adverse environmental conditions as well as adverse students' physical and psychosocial states. Had students been examined in favorable environments with proper physical and psychosocial states, some of the share of the variances attributed to error variances (σ_e^2) brought about by errors (adverse environmental conditions, adverse students' physical and psychosocial factors) would have been minimized, making the observed score (o_s) approach the true score (t_s), which could increase the reliability, and then the validity inferences.

Recommendations

In the light of the major findings of the findings on the new ESSLCE management, the following recommendations are forwarded.

- To implement exam administration effectively, the MoE and the EAES should put in place careful strategies to ensure proper procedures of item organizations in a test booklet, and meticulous handling of exam booklets. Moreover, careful selections of experienced exam chiefs and invigilators and providing them with ample proper orientations are of paramount importance.
- If the new ESSLCE management continues, it is recommended that arranging students' sitting in alphabetical order within the university, as opposed to arranging them from the same school in the same exam rooms, is vital. This helps to control students who are well informed about the previous seating arrangement and accordingly prepare for cheating in future examinations.
- The MoE and EAES should have a strict workable guideline on how to deal with students who become ill. The observed practices of allowing ill students to re-enter the exam room and continue the examination should be abolished. Students who are ill and leave the room during examination time should get a second chance as the Ministry did for some students, rather than allowing them to re-enter and continue the exam.
- The new ESSLCE should be modified taking into account the physical and psychological state. In addition, the MoE and EAES had better devise an alternative exam management that addresses the well-being of students, along with deterring examination malpractices.
- Further research is also needed to better determine the psychometric properties of the first university-based ESSLCE management and document its validity, reliability and item characteristics.

ACKNOWLEDGEMENTS

The authors are grateful to Jimma University for funding the research. The participants of the study are also acknowledged for providing the data required for the study.

REFERENCES

Abbasi, N., & Ghosh, S. (2020). Construction and standardization of examination anxiety scale for adolescent students. International Journal of Assessment Tools in Education, 7(4), 522-534. doi:10.21449/ijate.793084

Acosta, S., Garza, T., Hsu, H. Y., Goodson, P., Padrón, Y., Goltz, H. H., & Johnston, A. (2020). The accountability culture: A systematic review of high-stakes testing and English learners in the United States during No Child Left Behind. Educational Psychology Review, 32(2), 327-352.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.

Andrich, D., & Marais, I. (2019). A Course in Rasch Measurement Theory (Measuring in the Educational, Social and Health Sciences) || Reliability and Validity in Classical Test Theory. 41–53. doi:10.1007/978-981-13-7496-8_4 doi:10.1007/978-981-13-7496-8_4

Badasa, G. G., Nuri, R. A., & Gizaw, M. G. (2019). National Examination Misconduct: Magnitude and Perceived Determinants in Secondary and Preparatory Schools of Bale and West Arsi Zones. *Journal of Education and Practice, 10*(31). doi:10.7176/JEP/10-31-02

Becker, K. A., & Bergstrom, B. A. (2013). Test administration models. Practical Assessment, Research & Evaluation, 18(14). Available online: http://pareonline.net/getvn.asp?v=18&n=14. doi:10.7275/pntr-yz21

Bryman, A. (2012). *Social Research Methods, 4th Edition-Oxford University Press*. https://www.academia.edu/38228560/Alan_Bryman_Social_Research_Methods_4th_Edition_Oxford_University_Press_2012_pdf

Campion, D., & Miller, S. (2006). Test Production Effects on Validity. In S. M. Downing & T. M. Haladyna (Eds.), Handbook of test development (pp. 599-623). Mahwah, NJ: Erlbaum.

Chala, W. D. (2021). Perceived seriousness of academic cheating behaviors among undergraduate students: an Ethiopian experience. International Journal for Educational Integrity, 17(1), 1-15. doi:10.1007/s40979-020-00069-z

Dawood, E., Al Ghadeer, H., Mitsu, R., Almutary, N., & Alenezi, B. (2016). Relationship between Test Anxiety and Academic Achievement among Undergraduate Nursing Students. Journal of Education and practice, 7(2), 57-65.

Downing, S. M., & Haladyna, T. M. (2009). Validity and its threats. In S. M. Downing, & R. Yudkowsky (Eds.), Assessment in health professions education (pp. 21-55). London: Routledge.

Ethiopian-Monitor. (October 4, 2022). Daily News from Ethiopia, .

Fechter, T., Dai, T., Cromley, J. G., Nelson, F. E., Van Boekel, M., & Du, Y. (2021). Developing a validity argument for an inference-making and reasoning measure for use in higher education. In Frontiers in Education (Vol. 6, p. 727539). Frontiers Media SA. Doi: 10.3389/feduc.2021.727539. doi:10.3389/feduc.2021.727539

Gronlund, N. E., & Linn, R. L. (1990). Measurement and Evaluation in Teaching. McMillan Company, New York.

Guadiana, N., & Okashima, T. L. (2021). The Effects of Sleep Deprivation on College Students). Nursing | Senior Theses. 30. https://doi.org/10.33015/dominican.edu/2021.NURS.ST.09. doi:10.33015/dominican.edu/2021.NURS.ST.09

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. Educational Measurement: Issues and Practice, 23(1), 17-27.

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. Educational Researcher, 20(5), 2-7.DOI: 10.3102/0013189X020005002. doi:10.3102/0013189X020005002

Hamilton, N., Freche, R., Zhang, Y., Zeller, G., & Carroll, I. (2021). Test anxiety and poor sleep: A vicious cycle. International Journal of Behavioral Medicine, 28, 250-258. doi:10.1007/s12529-021-09973-1

Hancock, D. R. (2001). Effects of test anxiety and evaluative threat on students' achievement and motivation. The Journal of educational research, 94(5), 284-290. doi:10.1080/00220670109598764

Heissel, J. A., Adam, E. K., Doleac, J. L., Figlio, D. N., & Meer, J. (2021). Testing, stress, and performance: How students respond physiologically to high-stakes testing.

Hembree, R. (1988). Correlates, causes, and treatment of test anxiety. Review of Educational Research, 58, 47–7.

Hill, K. T., & Wigfield, A. (1984). Test anxiety: A major educational problem and what can be done about it. The elementary school journal, 85(1), 105-126. doi:10.1086/461395

Kane, M. T. (1992). An Argument-Based Approach to Validity.” Psychological Bulletin 112(3):527–35. doi: 10.1037/0033-2909.112.3.527.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1-73.

Keeves, J. P. (1994). National examinations: design, procedures and reporting.

Kellaghan, T., & Greaney, V. (2020). *Public Examinations Examined*

Koramoah, L. A., Dzakadzie, Y., & Danyoh, J. D. (2022). EXAMINATION ANXIETY AMONG FINAL-YEAR JUNIOR HIGH STUDENTS IN DANSOMAN, GHANA. *International Journal of Academic Research and Reflection*, *10*(1).

Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. Educational and psychological measurement, 30(3), 607-610.

Lam, T. (1995). Fairness in performance assessment. ERIC Digest. ED391982. Retrieved from http://www.ericdigests.org/1996-4/fairness.htm

Lee, D., Reynolds, C. R., & Willson, V. L. (2003). Standardized test administration: Why bother?Journal of Forensic Neuropsychology, Vol. 3(3) 2003

http://www.haworthpress.com/store/product.asp?sku=J151. doi:10.1300/J151v03n03_04

Messick, S. (1994). Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. Research Report RR-94-45.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). Measurement and Assessment in Teaching. In: Pearson.

Nigussie, B., Getachew, K., Mikre, F., Amsale, F., Belay, A., Workeneh, N., ... & , & Tilahun, G. (2023). A Move towards Choosing a Workable National Examination Management Approach for Ethiopia. The Ethiopian Journal of Social Sciences and Language Studies (EJSSLS), 10(1), 3-12.

Okano, K., Kaczmarzyk, J. R., Dave, N., Gabrieli, J. D., & Grossman, J. C. (2019). . Sleep quality, duration, and consistency are associated with better academic performance in college students. NPJ science of learning, 4(1), 16. doi:10.1038/s41539-019-0055-z

Peeters, M. J., & Harpe, S. E. (2020). Updating conceptions of validity and reliability. Research in Social and Administrative Pharmacy, 16(8), 1127-1130. https://doi.org/10.1016/j.sapharm.2019.11.017. doi:10.1016/j.sapharm.2019.11.017

Pennebaker, J. W. (2013). The college adjustment test (CAT). Measurement Instrument Database for the Social Science.

Reeves, T. D., & Marbach-Ad, G. (2016). Contemporary test validity in theory and practice: A primer for discipline-based education researchers. CBE—Life Sciences Education, 15(1), rm1. doi:10.1187/cbe.15-08-0183

Rukundo, A., & Magambo., J. (2010). Effective Test Administration in Schools: Principles and Good Practices for Test Administrators." African Journal of Teacher Education 1(1). doi: 10.21083/ajote.v1i1.1589. doi:10.21083/ajote.v1i1.1589

Tulving, E. (1974). "Cue-Dependent Forgetting: When We Forget Something We Once Knew, It Does Not Necessarily Mean That the Memory Trace Has Been Lost; It May Only Be Inaccessible." American Scientist 62(1):74–82.

Vedaa, Ø., Erevik, E. K., Hysing, M., Hayley, A. C., & Sivertsen, B. (2019). Insomnia, sleep duration and academic performance: a national survey of Norwegian college and university students. Sleep Medicine: X, 1, 100005. doi:10.1016/j.sleepx.2019.100005

Wollack, J. A., & Cizek, G. J. (2017). *The future of quantitative methods for detecting cheating: conclusions, cautions, and recommendations. In G. J. Cizek & J. A. Wollack (Eds.), Handbook of quantitative methods for detecting cheating on tests (pp. 390–399).* . New York: : Rutledge.

Wuthrich, V. M., Jagiello, T., & Azzi, V. (2020). Academic stress in the final years of school: A systematic literature review. Child Psychiatry & Human Development, 51, 986-1015. doi:10.1007/s10578-020-00981-y

Zhu, J., & Han, L. (2011). Analysis on the main factors affecting the reliability of test papers. Journal of Language Teaching and Research, 2(1), 236-238 doi:10.4304/jltr.2.1.236-238

Zieky, M. J. (2013). Fairness review in assessment. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 1. Test theory and testing and assessment in industrial and organizational psychology* (pp. 293–302). American Psychological Association. https://doi.org/10.1037/14047-017