Effect of Items Position Change on Students' Achievement in the Ethiopian University Entrance Examinations (EUEEs)

Robel Getachew[1], Mekbib Alemu[2], Misrak Getahun[3]*

[1] *Doctoral Candidate, Department of Physics, Hawassa University, Ethiopia;*
[2] *Associate Professor, Department of Teacher Education, Addis Ababa University, Ethiopia;*
[3] *Associate Professor, Department of Physics, Hawassa University*
** Corresponding Author: mekbib.alemu@aau.edu.et*

DOI: https://doi.org/10.63990/ejtel.v2i2.12444
Received: June 21, 2024; Revised: Dec 26, 2024; Accepted: July 28, 2025

Abstract:

To combat exam malpractice in crowded examination rooms, the Ethiopian University Entrance Examinations have been administered in four coded booklets of different reshuffling of item orders. However, research has revealed that systematic item position changes have significant effects in achievement scores. The main purpose of this study was to find out if random item order reshuffling would also have mean achievement score differences depending on which of the exam booklets test-takers were tested with. To address this purpose, the Entrance Examination 5 subjects (English, Mathematics, and 3-sciences) for 6-years for 21 sample public schools (11,376 grade 12 students) was received from National Education Assessment and Examinations Agency. In addition to the usual descriptive statistics, the data was analyzed with Spearman's rank-order correlation to determine if the item distributions in the four booklets of the same exam significantly differ with each other. Besides, one-way ANOVA was used to determine if there are statistically significant differences in students' achievement mean scores by booklets. The Spearman's rank-order analysis shows weak to moderate item position order differences among booklets. In contrast to this, statistically significant mean achievement differences were found in 66.67% of the exams, which put at a serious disadvantage up to 16.64% of test-takers due to which exam booklets they were tested with. Hence, it was recommended that all stakeholders: test developers, exam booklet developers, result publishers and decision makers be aware of the unfairness of the current practice with item reordering and therefore take appropriate compensatory measures.

Keywords: Achievement scores, Exam Malpractice, Item Position, Test Fairness, Test Anxiety

Introduction

Background of the Study

Large-scale assessments play a critical role in enhancing educational systems by offering a structured mechanism to monitor and evaluate student performance at various levels (American Educational Research Association [AERA] et al., 2014). These assessments serve multiple purposes, ranging from measuring student achievement and certifying attainment to informing policy decisions that affect educational planning and outcomes (Kellaghan & Greaney, 2019). Furthermore, they provide valuable data that support decisions related to student placement, admission to higher education institutions, and workforce entry, thus aligning educational outcomes with societal and economic needs (Ollennu & Etsey, 2015).

However, the challenge of exam malpractice is panic in the sector of test administration. One of the challenges of exam malpractice is to administer examinations in crowded examination rooms, and the convenience of using several alternative forms of a test to reduce the possibility of exam cheating (Anastasi, 1976; Carlson & Ostrosky, 1992). But questions are frequently raised on these alternative forms of examinations such as: are these alternative item position booklets equivalent in their total scores, even though the contents are identical? The concern about whether alternative forms of examinations maintain equivalence in total scores, despite identical content, is a significant issue in educational assessment. This question arises from the need to ensure fairness, validity, and reliability when using alternative test formats, such as rearranged item booklets. The concept of equivalence in alternative test forms revolves around the idea that different versions of the same test should yield similar outcomes in terms of total scores, reflecting consistent measurement of the same constructs (Brennan, 2013). When item positions are altered, the test should still measure students' abilities accurately, without introducing biases due to differences in the cognitive load or strategies triggered by the sequence of questions (Frey, 2018).

Normally it is advised by test experts and researchers to order examination items from easy-to-hard ordering systems for the measurement and psychological advantages stated in common measurement texts (Anastasi & Urbina, 1997; Cronbach, 1990; Mehrens & Lehmann, 1991; Plake, 1980). If the position of items changes and administered in different alternate forms of tests to different students, then attention should be on the nature and treatment of psychometric properties (AERA et al., 2014; Colwell, 2013; Ollennu & Etsey, 2015; Wu et al., 2016). However, Pettijohn II and Sacco (2007) reported that many of test developers mix up the order of test questions in alternate test forms without thinking of the consequences. This might also be true for testing organizations. The consequence may have an impact on students' examination performance, stress, perceptions, test reliability, and expectations of students' easy-to-hard ordering of items (Opara & Ogbanu, 2023; Plake, 1980). Wu et al. (2016), pointed out that if either test-takers' performance or items' characteristics are affected due to item position changes, then the validity of the test in interpretation will also be threatened.

National examination in the young Ethiopian modern education system which suffered disruption from 1936 to 1941, began in 1946 after liberation from Italian occupation. Soon after, the national examination at grades 8 and 12 started in 1950 with tests coming from London, Grate Britain (Alamneh, 2017; Mamaru et al., 2023). Mamaru and coauthors (2023) who studied the history of the national exam in Ethiopia from 1946 to 2023, noted that the constructed-response essay type London-based General Certificate of Education (GCE) exam was overtaken by locally developed and administered Ethiopian School Living Certificate Examination (ESLCE) by 1955. The ESLCE retained the GCE exam format for some time till marking and scoring by those subject experts who developed the exam became difficult due to

increasing number of examinees. While at the beginning the GCE for grade 12 consisted of five subjects: English Language, Mathematics, General Science, Geography, and Ethiopian History and Civics, the ESLCE came by 1955 with additional subjects and included: Amharic, English, Mathematics, Biology, General Science, Chemistry, Physics, Geography, History, French, Geez, Economics, and Accounting (Chala & Agago, 2022 and Mamaru et al., 2023). The list of exam subjects changed further slightly in 1975 and 1991 following government and ideology changes (Alamneh, 2017 and Mamaru et al., 2023).

Though the exam format started to change by including partly multiple-choice items in some of the ESLCE subjects till 1966, it remained including the constructed-response items in all subjects until 1974 (Alamneh, 2017). To resolve the marking problem that was gaining weight due to the ever-increasing test takers population, the ESLCE abolished the partly essay type exam format in favour of total multiple-choice format in all subjects since 1977 (Alamneh, 2017). The reliance of the National exams on just the multiple-choice exam formats, despite resolving the marking difficulty by making the speedy and reliable machine marking possible, came up with additional exam malpractices in the examination halls. To manage this (while still maintaining the benefits of the objective type curriculum-based examinations) the exam development and administration body first introduced the parallel exam forms (4-booklets) approach in 1996 and shifted the exam centers from secondary high schools to university premises since 2022 (Alamneh, 2017; Chala & Agago, 2022; Mikre et al, 2023).

The Ethiopian University Entrance Examination (EUEE), as mentioned above, is a high-stakes test with multiple-choice item format where its score determines the future of students' academic life. So, it is exposed to multiple exam malpractices inside as well as outside of the examination room (Chala & Agago, 2022; Mikre et al., 2023). To combat some of the exam-room malpractices, the examination has been administered in four coded booklets of different reshuffling of item order. However, some research interventions inform that the position changes of items be considerate of such psychometric characteristics as item difficulty level (Anastasi & Urbina, 1997; Cronbach, 1990; Plake, 1980). But, in EUEE examinations are prepared by an independent body (namely, the Addis Ababa University's Institute of Educational Research) in line with subject structures. After the development process is completed, final examination booklets with four codes are made ready to be printed, published, administered, and scored under the responsibility of Educational Assessment and Examinations Service (EAES) (FDRE-Council of Ministers, 2012; Federal Democratic Republic Government of Ethiopia (FDRGE), 1994, sec. 3.3.7). Consequently, the item-position reshuffling to come up with a different exam booklet, shatters the arrangement based on item difficulty levels. Finally, the cutoff scores for test-takers to be admitted to higher education is decided by high stakeholders considering the universities' in-take capacities, gender, socio economic status of regions from where test-takers come, and disabilities. But when such decisions were made, no

reflections about item position effects and equivalence of scores with different exam booklets were made (AERA et al., 2014, European Federation of Psychologists Associations (EFPA) and European Association of Work and Organizational Psychologists (EAWOP), 2007). Colwell (2013) explains that when high-stakes decisions are based on the test scores obtained from such examinations, the issues item position must be addressed to ensure that tests provide fair representations of all students' abilities.

Despite such recommendations by testing experts and researchers (AERA et al., 2014), and while there are good practices in different countries (EFPA & EAWOP, 2007), in Ethiopia the University Entrance Examination results are understood as if there are no differences among test forms (Chala & Agago, 2022; Mikre et al., 2023). Even though these decisions are generally considered as fair, the judgmental fairness must be empirically questioned. In the case of the EUEE, which determines the future of hundreds of thousands of students every year, decisions must be based on meticulous considerations of position effects. Maybe in these exams, item position effects are considered to be minimized due to the non-systematic distributions of items in different booklets. However, such high-stakes decisions should not be left to general assumptions, instead searching for empirical evidences to what extent item position affects students' results and how much those are affecting decisions must be conducted.

Review of Related Literature

One of the test development principles repeatedly appearing in textbooks and examination guidance documents is to arranging test items in a systematic pattern in ascending order of difficulty (Anastasi & Urbina, 1997; Ollennu & Etsey, 2015; Opara & Ogbanu, 2023). The idea behind this is that if candidates answer the easier questions first and are successful, it will build their confidence and give them a mental boost, which will stimulate them, lower their exam anxiety, and promote more successful answers to the following difficult questions (Holzknecht et al., 2021; Mehrens & Lehmann, 1991). However, candidates who encounter the more difficult items first (descending order), especially in a timed test, may spend a lot of time on one specific question and not finish the test with the simpler items done. Also, researchers argued that fatigue and pressure to finish could account for poorer performance on easy items when they appear later in the test than when they appear earlier in the test (Hambleton & Traub, 1974; Wu et al., 2016). On the other hand, Hambleton and Traub (1974) explain that when test items are arranged in reverse order, difficult-to-easy order of items, a student with experience and expectation of the common order of items from easy-to-difficult encounters difficulty. When faced with difficult items at the beginning, the test taker expects even more difficult items at later stages and more stressful test situation. This might make test takers more anxious with the likely result that test performance would be adversely affected.

Items may also be placed in an inconsistent order (mixed order); this method involves placing difficult items throughout the test at specified intervals, and then followed by subsequently easier ones. The idea behind this method is that an ascending order technique disappoints the candidate when they encounter and attempt too many difficult items in a row. Consequently, they end up with not answering these items at all, guessing, and cheating on them, and this can't show the candidates' true ability on that trait (Ekele, 2002 as cited in Opara & Ogbanu, 2023).

However, different researchers found different results about the achievement score differences and psychometric nature of items in the examination booklets produced by item position changes. Many research findings suggested and cited that item arrangements significantly influenced test performance, but this influence occurs when the examination is administered at a speed test2 rather than a power test3 (Hambleton & Traub, 1974; Opara & Ogbanu, 2023; Plake, 1980). However, power tests have no practical significance, but some researchers suggest that easy-to-hard ordering is still preferable, citing student expectation as the rationale (Flaugher et al., 1966; Monk & Stallings, 1970; Plake et al., 1982). In multiple-choice items, MacNicol (1960 as cited in Plake, 1980) investigated the effects of changing an "easy-to-hard" arrangement to either hard-to-easy or a random arrangement. He found out that the hard-to-easy arrangement was significantly more difficult than the original easy-to-hard order while the random arrangement was not significantly different in their scores. This finding was argued by different scholars (Anastasi, 1976; Ollennu & Etsey, 2015; Plake, 1980; Plake et al., 1982; Shepard, 1994).

However, some researchers found no significant difference in performance when items were arranged according to easy-to-hard, hard-to-easy arrangement, or random order (Gerow, 1980; Ollennu & Etsey, 2015; Soyemi, 1980). Also, researchers discovered that different arrangements of items could affect performance adversely or positively depending on the levels and subjects in question. For instance, Ollennu & Etsey (2015) worked on English, mathematics, and Science subjects of the Basic Education Certificate Examination in Ghana, found significant differences in the performance of each subject. Also, the mean score on a mathematics test of a high school grade 11 course with items arranged in the order difficult-to-easy was significantly lower than the mean score on a test with the same items arranged in the order easy-to-difficult (Hambleton & Traub, 1974). But Nagy et al. (2018) found weak differences with Science, Mathematics, and reading tests of PISA 2016 analysis where the strongest effect was observed in the reading sections.. Flaugher et al. (1966) indicated that moderate rearrangement of items on the College Entrance Examination Board, Scholastic Aptitude Test was associated with significantly different test scores in the Verbal portions of the test but not in the Mathematical portions (Monk

[2] Time-restricted test
[3] Time unrestricted test

& Stallings, 1970). Even Abdullahi et al. (2020) in a college-level experiment on the subject of Mathematical-Economics course of randomized distribution of items had significantly greater achievement mean scores than easy to hard items order and no significant differences were observed. Also, Satti et al. (2019) noted that of 5th-year medical graduate examinations composed of form-A ordered according to the content sequences, form-D prepared in the reverse order of form-A, and the remaining B and C prepared in a randomized order. There were no statistical differences among the mean scores of the different forms (A, B, C, and D).

In the preceding literature review it has been observed that the research on item position effects was going on over the decades with experimental studies and based on high stake exam records (Abdullahi et al., 2020; Hambleton & Traub, 1974; Plake et al., 1982; Soysal & Kogar, 2021). In most of these studies, either comparison between systematically ordered item arrangements (easy-to-difficult and difficult-to-easy) or between ordered and disordered item arrangements (easy-to-difficult and moderately disordered or different clustering) were considered. Few of these studies considered high-stake international examinations such as PISA and TIMSS, in which item-position effects are already recognized (Wu et al., 2016) and such measures as booklet design are used to curb the negative effects (Hartig & Buchholz, 2012; Soysal & Kogar, 2021). However, the study of item-position effects in totally random and non-clustered item arrangements such as in the case of the EUEE are rare.

Even though recognizing item position effects in alternate forms of exam booklets and booklet designs are trusted to limit one or the other form of position effects, studies still show that individual and group differences in final scores are persisting (AERA et al., 2014; Hartig & Buchholz, 2012). In the Standards for Educational and Psychological Testing, AERA et al. (2014) suggest the need for the final score equating to make final judgments and use of examination results fair and defensible. Further, by way of establishing evidences for equating scores from alternate forms, the "Standards" recommends to make an appropriate choice from four alternate measures. These measures are:

1. administering the forms to be equated to the same sample of examinees or to equivalent samples;
2. administering alternate forms to equivalent samples, usually through random assignments;
3. administering a common set of items, referred to as anchor items, to the samples taking each form; or
4. use an external anchor test in which the anchor items are administered in a separate section and do not contribute to the total score on the test.

(AERA et al., 2014, p. 97-98)

In the case of EUEE, neither during exam development nor in publishing and administering processes that the existence of item position effects is recognized (Chala & Agago, 2022; Mikre et al., 2023). At the same time, when the high-stake decisions are made based on exam results, there are no evidence of measures to equate scores from alternate forms (AERA et al., 2014; Gregory, 2011; Zelman, 2013). Therefore, it is critical to conduct research on item position effects in the EUEE and present to educational and examination stakeholders about the existence and extent of effects of item position changes. Such research will not only contribute to the practical judgmental validity of the EUEE, but it also contributes to the literature bases by illuminating the status of item position effects in alternate forms with random item distribution (AERA et al., 2014; Hartig & Buchholz, 2012; Wu et al., 2016).

Purpose of the study

The purpose of this Study was to explore the effect of item position changes on the achievement scores of the Ethiopian University Entrance Examinations (EUEE) across selected subjects. Research has consistently shown that test item positioning can influence student performance, with early or late placement of items potentially affecting cognitive load, fatigue, and anxiety levels (Haladyna & Downing, 2004; Schweizer et al., 2017). In high-stakes exams like the EUEE, understanding the effect of item position is crucial, as it has implications for fairness, validity, and reliability of the test scores, which in turn impact university admissions decisions.

More specifically, the study sought to:

1. Identify whether significant item position effects exist among different test booklets for each of the five EUEE subjects (English, Mathematics, Biology, Chemistry, and Physics) when items are randomly distributed across versions.
2. Determine the extent to which item position changes influence test-takers' scores based on the booklet versions they received during the examination (e.g., Booklets 01, 02, 03, or 04).

Methods

Research Design

This study is exploratory ex-post facto design research aiming at determining if there were achievement differences in grade 12 students mean scores in the EUEE based on item positioning in different booklets of the same exam. And further, if item position effects are observed, the Study aims to estimate the extent of the effect on test-takers scores. For these purposes, the EUEE scores for five subjects were categorized by year and booklet groups and analyzed by comparing mean scores by quartets.

Data Source

For various analysis at national level, the National Examination Agency samples 21 public preparatory Schools (recently named high schools) based on proportionate stratified random sampling from nine federal regions and two city administrations (National Education Assessment and Examinations Agency [NEAEA], 2017). The same was taken in this Study. This resulted in 6,498 (Biology-2020) to 11,376 (English-2015) number of test-takers (student population), where the variation is depending on the type of subject. Six years of exam records for years 2015-2020 from NEAEA (recently named as Educational Assessment and Examinations Service, EAES) record were taken with permission from the organization. This resulted in 30 examinations (5 subjects per year) and 120 (4 booklets per exam) booklets. The examination subjects sampled in this study were: English, Mathematics, Biology, Chemistry, and Physics. These subjects were selected because, until recently 70% of student population was from natural science stream required to take these subject-examinations (Japan International Cooperation Agency (JICA), 2017; NEAEA, 2017; Teferra et al., 2018) and that makes the research give picture of the larger proportion of test takers. Care was taken to keep anonymous individuals whose scores were used in the study. For this purpose, the data received from EAES was with name codes for individuals and schools. In the Excel sheet from EAES, test takers' names (TestTaker) were replaced with such codes as TestT0001, TestT0002, … and similarly, schools names (School) was coded as SchC01, SchC02, … However, as the goal of the study was not about test takers and their schools, data analysis was not affected at booklets groups level.

Data Analysis

The first step in the analysis was to filter out scores of each subject by booklet code with their respective year. Even though there were four booklets (coded with Code-01, 02, 03, and 04 for analysis purpose) per examination for every subject and for every year, the EAES codes were retained when the data was exported to the SPSS for further statistical analysis. This helped in presenting analysis results as relevant to specific exam booklets in the final report. Applying descriptive statistics of mean, median, and standard deviations was important to see the score distributions of each coded booklet.

Before making the mean score comparison among booklets, which was the focus of the objectives of this Study, Spearman's rank-order correlations were used to see to what extent item distributions in the four booklets of an exam are different /similar with each other. This was important because of the assumption that item position effects would be minimized in random item distribution cases (AERA et al., 2014; Soysal & Kogar, 2021), in the cases of comparison of alternate forms with easy-to-difficult and difficult-to-easy item distributions, there is the least similarity (rho ~ 0.0).

To determine if there are statistically significant differences in students mean scores by booklets, one-way ANOVA was used. After finding the exams with significant mean score differences for the examinations from 2015-to-2020, the post-hoc analysis was used for pair-wise analysis and locating the statistically significant difference between pairs of booklets of an exam.

Finally, to address the second purpose of this research, that is to estimate to what extent test-takers are affected by the item-position effects, the pooled standard deviation was calculated for each exam pairs with significant mean differences observed in the post-hoc analysis. By dividing the mean difference by the pooled standard deviation of the respective pairs, the absolute mean differences were calculated. This absolute mean difference (in unites of pooled standard deviation) was used to estimate the population within the range between the mean and the absolute mean difference in the standard normal distribution for each exam pairs (Peck et al., 2008). This resulted in estimation of the proportion of test-takers who were affected by the item position effects.

Results

Relative item distributions among exam booklets

Inspection of the different exam booklets revealed that test items in booklets were not ordered according to test development principles advise, from list difficult to most difficult. Thus, instead of the ideal order in terms of items psychometric characteristics, when the item positioning in the four booklets of the same exam was considered, the relative order with respect to each other was considered. Thus, the first observation was made with EUEE exams by comparing the relative randomness of distributions of items in four booklets of the same exam

Table 1
Mean Spearman's rho for list and most item order randomization difference among booklets by subject ($p<0.01$)

Test	Year	N	Mean	STD
English	2016	120	0.9870	0.00539
	2018	120	0.9935	0.00269
Biology	2016	100	0.9725	0.00568
	2018	100	0.9653	0.00844
Chemistry	2018	80	0.9942	0.00383
	2019	80	0.9215	0.02188
Mathematic	2016	65	0.9918	0.00217
	2017	65	0.9043	0.04021
Physics	2017	50	0.5915	0.09700
	2019	50	0.8292	0.04366

(same year and same subject) using Spearman's rank-order correlation. Here in Table 1, the

mean of Spearman's rho for randomly selected two exams for each of the five subjects are presented. Revealed

As can be seen from the data in the table, the Spearman's rho varied from ρ_{min} =0.592 (SD=0.097) among Physics 2017 exam booklets to ρ_{max} =0.994 (SD=0.004) among Chemistry 2018 exam booklets. In addition, it could be observed from the data in Table 1 that except for Physics examinations, the item order differences for all the other cases were minimal (rho in very strong range). However, Physics exams showed moderate to strong item order differences (Schober, 2018) in the range of ρ_{min} =0.592 (SD=0.097) to ρ_{min} =0.829 (SD=0.044). Furthermore, there appeared to exist correlations between number of items and Spearman's rho. The correlation coefficient between number of items and Spearman's rho was found to be r=0.530 (p=0.008) which is in moderately strong range. This is to be expected as Spearman's coefficient increases with sample size.

The maximum item order randomization difference was observed between Physics 2017 exam booklets Code 69 and Code 70, with Spearman's rho of 0.318 ($p<0.05$), next between booklets Code 67 and Code 70 of the same exam with Spearman's rho of 0.574 ($p<0.01$). In general, the item order difference among booklets of this exam were the highest of all the 30 exams with mean rho 0.592 ($p<0.01$). On the other hand, the minimum observed randomization difference was between Chemistry 2018 exam booklets of Code 38 & 39, Code 38 & 40, and Code 39 & 40, all with Spearman's rho of 0.998 ($p<0.01$). The next were booklets from English 2018 between Code 22 & 23 and Code 23 & 24, with rho of 0.997 ($p<0.01$). In general, the coefficient rho was randomly distributed between rho of 0.318 for Physics 2017 and 0.998 for Chemistry 2018 exams. The only exceptionally different item order differences, as described above were among Physics 2017 which were in the range of weak to strong correlation (Schober et al., 2018). This means, even if the exam items were reordered in the different booklets of the same exam, the order differences were not that strong in many of them to expect significant achievement difference among test-takers (students) due to item position effects (AERA et al., 2014; Soysal & Kogar, 2021).

In general, there were no strong differences in items distributions among booklets of the same exam. Booklets inspections revealed that the items are not redistributed individually in a complete random fashion. Item groups (blocks) containing random numbers of items (between 1 to 7) are picked randomly and put at random positions in the different booklets. Therefore, some of the items kept their relative positions with respect to some of the item group members. Probably this was the reason for low observed randomization difference (or high Spearman's rho).

Mean scores and standard deviations of students' scores from 2015 to 2020

The main purpose of this research was to find out if the item order randomization differences among exam booklets had effect on students' achievement. To see this, the crude data received from NEAES (National Education Assessment and Examinations Service) data center was classified based on the booklets' codes and mean achievement scores for each subject and year were calculated by booklets. The item numbers in the six years in EUEE (Ethiopian University Entrance Examination) differ by subjects from 45 to 120. As mean scores were to be compared specific for subjects and years or examination, the item number difference from exam to exam, and/or from year to year did not matter in addressing the research objectives. The analysis result was based on row scores of 120 items in English, 80 items in chemistry, and 100 items in Biology. In Mathematics exams 65 items were used except in 2020 (61 items) and similarly in physics, 50 items were counted except 45 items in 2017. In table 2 on the next page, a sample of 10 exams (2 for each subject) are presented to show variability in students' mean achievement score from booklet to booklet. The full-length data that is used for analysis is found in Appendix A.

Inspection of Table 2 (and also Appendix A) show that the mean achievement scores populated the lower half of the ideal mean (50%) in every subject with more than 30% variability (SD>0.30 of mean). Only in the case of 11 exams (out of 30), that students mean scores were at or barely above the ideal mean. While in all of the English and Physics exams students mean scores were totally below the 50% mark, there was one Mathematics, four Chemistry, and six (all) biology exams in which the mean scores were found to be at or barely above the ideal mean. Besides, it was only in Biology 2019 and Chemistry 2019 that in the sample schools the maximum possible score was achieved.

The other observation that could be made from the data in Table 2 is that in most of the examinations the mean scores showed very little variations from booklet to booklet. With each of the 30 examinations analyzed in this study, there are 4 different booklets that make 6 independent booklet pairs. That means, the total 30 examinations constituted 180 pairs of booklets showing mean score differences between each other. However, if we just count those with mean differences greater than one point, there would be 39 pairs of booklets. That means, 21.67% of the booklet pairs showed more than 1 point mean score differences. In this respect, Biology Exams showed the largest number of mean differences between pairs of booklets with 15 out 36 booklet pairs. The other four subjects exhibited 5 mean differences of more than 1-point between 5 or 6 booklet pairs.

Table 2
Last two years' sample of students' mean scores and standard deviations (M(SD)) by booklet for same examinations

	Code 01	Code 02	Code 03	Code 04	Total	Max	Test-Teker
English 2019	42.32 (13.74)	41.7 (13.45)	43.04 (13.71)	42.09 (13.65)	42.28 (13.64)	98	10577
English 2020	52.47 (19.75)	52.96 (18.79)	53.15 (18.67)	54.12 (18.15)	53.17 (18.86)	110	9612
Math 2019	21.86 (8.13)	21.77 (7,77)	22.65 (7.85)	21.85 (8.21)	22.03 (8.00)	59	7012
Math 2020	31.01 (8.82)	31.72 (8.00)	30.95 (8.93)	28.66 (10.95)	30.62 (39.63)	58	6503
Biology 2019	51.01 (17.35)	51.77 (16.69)	51.47 (16.83)	50.72 (17.28)	51.25 (17.04)	100	7012
Biology 2020	52.66 (14.40)	52.65 (13.24)	53.41 (15.07)	54.34 (13.96)	53.24 (14.19)	94	6498
Chemistry 2019	36.52 (11.04)	36.74 (10.85)	36.20 (11.84)	35.51 (11.58)	36.26 (11.33)	80	7006
Chemistry 2020	46.13 (11.75)	45.59 (12.20)	41.02 (12.02)	43.03 (10.56)	43.99 (11.84)	74	6501
Physics 2019	17.78 (5.79)	17.40 (5.40)	17.38 (5.53)	17.28 (5.71)	17.47 (5.61)	46	7006
Physics 2020	22.59 (7.09)	24.63 (6.93)	23.62 (7.36)	25.06 (7.03)	23.95 (17)	45	6500

The largest mean score difference between booklets was observed in Chemistry 2020 exam. A 5.11 difference was observed between Code 35 and Code 37 booklets in Chemistry 2020 exam. In the same exam the second largest mean score difference was also observed between Code 36 and Code 37 booklets in Chemistry with mean score difference of 4.57. Still the same exam exhibited the third largest mean score difference of 3.09 between booklet Code 35 and Code 38. Apart from that exhibited by Chemistry 2020 exam, Mathematics 2020 exam booklets of Code

48 and Code 50, Physics 2020 booklets of Code 39 and Code 42, Biology 2015 booklets of Code 31 and Code 32, and English 2018 booklets of Code 22 and Code 23, showed maximum of mean score differences of 3.066, 2.474, 2.332 and 1.991, respectively.

Significance of Mean score differences among booklets of exams from 2015 to 2020

From the descriptive analysis it was observed that 39 (21.67%) out of a total of 180 EUEE exam booklet pairs exhibited more than 1-point mean score differences. 1-point minimum was arbitrarily taken to make sense of the extent of difference observed among different booklet pairs. However, all booklet pairs exhibited mean score differences ranging from a minimum of 0.014 to 5.109. Now the question is are these observed differences statistically significant to claim that students' achievements were affected by item positioning in different booklets.

After checking and confirming that all the group of data (by subject, year of examination, and booklets) satisfy the assumptions for ANOVA analysis, one-way ANOVA was used for statistical significance of mean score differences. The one-way ANOVA analysis showed that 21 out of 30 of the exams (70%) had statistically significant differences among the respective 4 booklets. However, the post-hoc analysis resulted in dropping of Physics 2018 exam score as

Table 3

ANOVA table for exams exhibiting statistically significant mean score difference in students' achievement scores between pairs of exam booklets (p=0.05)

	Exam	Item n	Test-taker N	Mean	Std. Dev.	Mean Diff.	df	F	Sig.
1.	English 2018	120	9840	51.58	15.32	1.157	3	8.706	0.000
2.	English 2019	120	10577	42.28	13.64	0.706	3	4.478	0.004
3.	English 2020	120	9612	53.17	18.86	0.860	3	3.245	0.021
4.	Math 2017	65	7778	24.00	8.86	0.405	3	4.820	0.002
5.	Math 2018	65	6623	27.02	8.58	0.607	3	5.543	0.001
6.	Math 2019	65	7012	22.03	8.00	0.445	3	4.684	0.003
7.	Math 2020	61	6503	30.62	9.63	1.543	3	30.563	0.000
8.	Biology 2015	100	8203	62.29	16.45	1.293	3	8.873	0.000
9.	Biology 2018	100	6621	56.22	14.92	0.980	3	4.713	0.003
10.	Biology 2020	100	6498	53.24	14.19	0.970	3	5.094	0.002
11.	Chemistry 2015	80	8200	45.488	11.67	1.202	3	13.473	0.000
12.	Chemistry 2016	80	7665	37.25	9.93	0.644	3	5.284	0.001
13.	Chemistry 2018	80	6624	42.21	12.37	1.109	3	9.547	0.000
14.	Chemistry 2019	80	7006	36.26	11.33	0.667	3	3.816	0.010
15.	Chemistry 2020	80	6501	43.99	11.84	2.980	3	67.648	0.000
16.	Physics 2015	50	8218	19.25	6.29	0.315	3	3.148	0.024
17.	Physics 2016	50	7659	21.37	6.98	0.314	3	2.707	0.044
18.	Physics 2017	50	7787	20.04	6.10	0.336	3	3.896	0.009
19.	Physics 2019	50	7006	17.47	5.61	0.254	3	2.742	0.042
20.	Physics 2020	50	6500	23.95	7.17	1.406	3	39.626	0.000

there was no statistically significant pairwise mean score differences among the 4 booklets, even-if the F-value (F(3,6626)= 2.883; p=0.034) was statistically significant. Therefore, in Table 3 below, the relevant descriptive statistics and ANOVA analysis for 20 (66.67%) of the exams are presented.

As can be seen from the table, 15 of the examinations (50%) have statistically significant test-takers' achievement mean score differences at less than 0.01 while only 5 of them (16.67%) showed difference at 0.05 significance level. As noted in the descriptive analysis, Chemistry 2020 examination is with the highest mean difference of 2.98 points (25.17% of the mean standard deviation). Like Chemistry, Physics exhibited significant differences in most of the exams (5 out of 6 exams), even if the mean differences among those booklets are as low as 0.25 points (4.53% of the mean standard deviation). This means, even if the mean differences are very small, there are chances that those differences are statistically significant and occur due to the difference in item position order in the different booklets. In terms of the frequency of statistically significant difference among exam booklets, next to Chemistry and Physics exams, Mathematics exhibited 4 out of 6 times, and English and Biology 3 times out of 6 exams.

Table 4
Pairwise comparison of mean score difference between booklets of the same exam showing significance of minimum and maximum ($p<0.05$)

Minimum significant difference			Maximum significant difference		
Exam	Booklet pairs	Mean Diff.	Exam	Booklet pairs	Mean Diff.
English 2018	Code 24 & 22	1.299	English 2018	Code 23 & 22	1.991
Math 2015	Code 16 & 15	0.707	Math 2020	Code 48 & 50	3.066
Biology 2015	Code 34 & 32	1.433	Biology 2015	Code 31 & 32	2.332
Chemistry 2019	Code 27 & 30	1.004	Chemistry 2020	Code 35 & 37	5.109
Physics 2019	Code 23 & 26	0.500	Physics 2020	Code 42 & 39	2.474

To identify where the significant mean score differences lie, post-hoc analysis was conducted. Table 4 presents the sample of exam booklet pairs with the minimum and maximum significant mean score differences. For detail analysis, see to the data in Appendix B. The post-hoc analysis revealed that out of 180 independent pairs of booklets, there were 44 (24.44%) pairs with statistically significant mean differences. As observed in the descriptive analysis and the one-way ANOVA analysis, still Chemistry exams were the leading pairs in significant mean score difference with 15 (41.67%) of the booklet pairs. Mathematics and physics followed Chemistry each with 9 (25%), and Biology with 6 (16.67%) of booklet pairs. English examination booklet pairs were with the least (5 or 13.89%) number of booklet pairs to show statistically significant mean score differences. This means nearly in a quarter of EUEE there were statistically significant achievement mean score differences among students due to item position differences among booklets.

On the other hand, Physics and mathematics occupied the two least mean score differences with 0.50 and 0.71, respectively. These differences are 8.9% for Physics and 8.3% for Mathematics of their respective mean standard deviations. In another extreme, Chemistry and Mathematics occupied the top two positions of mean score differences with 5.11 and 3.07, respectively. These are 43.15% for Chemistry and 31.84% for Mathematics of their respective mean standard deviations. These are very large differences.

Estimation of Proportions of test takers significantly affected by Item Position Changes

The second measure purpose of this research was to estimate to what extent test-takers are affected by the item-position effects. To address this purpose, after identifying the exam pairs with significant mean differences, the mean differences were calculated in terms of the pooled standard deviation units. This absolute mean difference was used to estimate the population within the range between the mean and the absolute mean difference in the standard normal distribution for each exam pairs (N=44, 24.44%). Table 5 presents the maximum and minimum mean score differences by subject, and the mean population proportion with significant item position effects by subject and overall mean. The detail of the population proportion estimate is presented in Appendix B along with other relevant data.

Table 5
Estimation of the proportion of the test-taker population with minimum and maximum IP effects ($p<0.05$)

	Exam	Booklet pairs	Mean Diff.	Absolute Diff.	Population Proportion (%)	Sig. (p)
Minimum	English 2018	Code 22 & 24	1.299	0.0916	3.59	0.016
Maximum	English 2018	Code 22 & 23	1.991	0.1269	4.97	0.000
	Average	5 (13.89%)	1.656	0.1012	3.94	
Minimum	Math 2015	Code 15 & 16	0.707	0.0811	3.18	0.036
Maximum	Math 2020	Code 48 & 50	3.066	0.2996	11.6	0.000
	Average	9 (25.00%)	1.462	0.1581	6.16	
Minimum	Biology 2015	Code 32 & 34	1.433	0.0850	3.19	0.029
Maximum	Biology 2015	Code 31 & 32	2.332	0.1432	5.57	0.000
	Average	6 (16.67%)	1.866	0.1206	4.64	
Minimum	Chemistry 2019	Code 27 & 30	1.004	0.0909	3.58	0.044
Maximum	Chemistry 2020	Code 35 & 37	5.109	0.4300	16.64	0.000
	Average	15 (41.67%)	2.153	0.1843	7.22	
Minimum	Physics 2019	Code 23 & 26	0.500	0.0869	3.38	0.042
Maximum	Physics 2020	Code 39 & 42	2.474	0.3504	13.68	0.000
	Average	9 (25.00%)	1.1251	0.1641	6.36	
	Grand mean	44 (24.44%)	1.698	0.157	6.10	

As it can be observed from Table 5, there is a mean population proportion of 6.10% with a mean difference of 1.698 (15.7% of mean standard deviation). A simple statistical analysis (with normal distribution) shows that a one-standard deviation changes from the mean results in 34.1% change in the population and a half- standard deviation change from the mean results in 19.1% change (Peck et al., 2008). Further it can be observed from the data in table 5, that the item position effect is the strongest in Chemistry, with 16.64% affected population proportion for 0.43Std ($p<0.001$) absolute mean difference in mean score. This occurred in Chemistry 2020 examination for booklet pairs Code 35 and 37. This means, up to 16.64% of the test takers in EUEE could be either unfairly lost or unfairly advantaged just by the booklet they were examined with due to item position effects. In this case, Mathematics and Biology seem to be with the minimum population proportion to be affected with item position effects. The minimum proportion for Mathematics was 3.18% ($p=0.036$) and for Biology 3.19% ($p=0.029$). However, it is in English EUEE exams that the least mean population (3.94%) that was observed with minimal item position effect, while chemistry (with 15 out of 36 (41.67%) of the exam booklet pairs that demonstrated the maximum population proportion of exam takers at 7.22% mean population proportion.

Discussions

The main purpose of this research is to find out in EUEE (Ethiopian University Entrance Examination) if there are test-takers achievement mean score differences among different exam booklets. Furthermore, the study aimed at estimating the proportion of exam takers who would be affected by item position effects, if a significant achievement mean scores were observed. To address these objectives, there was a secondary question to raise: how significantly were the booklets of the same exam differing from each other? In EUEE, there were four booklets per exam containing the same items but in different orders. Therefore, to answer the question, the extent of difference in item positions were statistically determined. Unlike in other studies on the effect of exam item position on students (test-takers) performance (Ollennu & Etsey, 2015; Pettijohn & Sacco, 2007), the items in the EUEE were not ordered based on any assessment logic. By inspection of exam booklets from the sample of EUEE between 2015-to-2020, it was observed that the production of alternate booklets resulted in high similarity in item positioning. With spearman's correlations analysis it was found that all of the 180 booklet pairs, except one pair of Physics 2017 exam where those were with more than moderate similarity (Schober et al., 2018), all of the exams exhibited more than strong similarity of item distributions.

Due to the high degree of similarity between booklets and the lack of order based on test construction principles in the random item distribution case of EUEE, it was unlikely to find achievement differences among test-takers due to item position differences. Even if those differences appear, their magnitudes would not be as large as observed in other similar studies

(for example in Ollennu & Etsey, 2015 and Opara & Ogbanu, 2023). Added to that, researchers suggested that random item distribution may minimize the item position effects (AERA et al., 2014; Soysal & Kogar, 2021; Wu et al., 2016), implied that the likelihood of finding significant mean score differences based on booklet differences is minimized. However, the results in this research contradicted the consensus among researchers in the likelihood of occurrence and in significance of the difference.

After observing that for majority of the EUEE examinations from 2015 to 2020 were not exhibiting strong difference of item positioning patterns, the data was analysed to address the main research purpose. It was observed that 66.67% (20 out of 30) of the exams showed statistically significant mean score difference based on booklet differences in item arrangement. In contrast to the mean score differences observed in other previous researches (for example Alakayleh, 2017; Ollennu & Etsey, 2015), the magnitudes of the mean difference observed in the current study appeared small. As the number of items per exam differs from subject to subject (from about 50 to 120), the small mean differences could not be compared meaningfully, even though those are statistically significant. However, by using the standard deviation as a unit of measure of the mean difference, it was observed that the absolute mean difference varies between 0.0811 (for Mathematics 2015 exam) to 0.4300 (for Chemistry 2020 exam). Therefore, from this result we can see that the item position effect is contributing to such strong differences in students (test-takers) achievement scores in the EUEE.

This is significant not only to the local practical context, but the result in a way confirms the general findings in other researches (Alakayleh, 2017; Hartig & Buchholz, 2012; Ollennu & Etsey, 2015; Soysal & Kogar, 2021; Wu et al., 2016). Even though many of the previous studies on item-position effects were by comparing alternate exam forms (booklets) with systematic item arrangements (for example by Abdullahi et al., 2020 and Alakayleh, 2017), and sometimes with a different purpose such as analysis of performance persistence of test-takers throughout a test (for example by Soysal & Kogar, 2021, Wu et al., 2016), all asserted that test-takers performances are significantly affected by item positioning. The result in this Study further illuminates the case of exam booklets with non-systematic (or random) item arrangements in which case research is scares and the assumption is item position effects to be minimized by random item arrangements (AERA et al., 2014).

The difference in mean score of students is not unique to specific subject instead it was observed in 20 out of 30 examinations investigated in this research with all the five subjects. As discussed above, this was unexpected to occur among booklets which have so high degree of similarity in item ordering. In similar studies so far (Alakayleh, 2017, Ollennu & Etsey, 2015, Opara & Ogbanu, 2023) the comparisons were between a random and either easy-to-difficult or difficult to easy item sequencing. Therefore, even if the number of items in general are less than the

number of items in any of the EUEE exams, the difference in item ordering were very high (either rho= –1 or very close to zero). But, in the current Study, it was found that none of the booklets were sequenced according to psychometric characteristics and yet the random sequencing in the different booklets did not result in significant differences among them. In contrast to this, statistically significant achievement differences were found in 66.67% of the exams.

The disparity observed between the mean score differences among exam booklets and the high degree of similarity among the booklets is a fundamental finding. So far research on this issue was focused on the difference in item ordering (Alakayleh, 2017, Baffoe, 2021, Ollennu & Etsey, 2015, Opara & Ogbanu, 2023, Pettijohn, & Sacco, 2007). However, the finding in the current study is suggestive of the existence of a more profound factor other than mere reordering of items in exam booklets resulting in a significant difference in students (test-takers) achievement. This requires probably a more complex analysis of the data at item levels.

In addition to investigating the existence of item position effects in the EUEE, this Study also attempted to estimate the population proportion of exam takers affected by the position effects. It was found that in between 3.1 up to 16.64% of test takers were affected by item position effects with examinations in different years and subjects. Furthermore, it was observed that position effects were varying from year to year by subject. While the effect was minimal but significant for English examinations with 3.94%, it was Chemistry with the highest affected population size with mean of 7.22%. In many of the item-position effects studies, the magnitude of the impact on test-takers were not reported except the consensus that it has significant effect on students' performance (Hartig & Buchholz, 2012; Ollennu & Etsey, 2015; Wu et al., 2016). However, fairness is one of the major principles of assessment (AERA et al., 2014; EFPA & EAWOP, 2007). Therefore, at least in reporting high-stake exam results, the item position effects and those who would be affected by it should be reported (AERA et al., 2014; Soysal & Kogar, 2021; Wu et al., 2016).

Except in chemistry exams, the achievement mean score differences among booklets do not look that much alarming. For many of the examinations the mean score differences were by about 1 and 2 points. However, when these differences were converted to absolute difference (by comparing with the standard deviation) those results become more meaningful. The alarming face of the item position effects became apparent when it is translated to victims count. In this study it was observed that overall, 6.10% of test-takers were disadvantaged by taking one of the exam booklets and not the other one. In the latest Educational Statistics Annual Abstract (Federal Ministry of Education, 2023) the total number of students who registered for the EUEE were 845,099 out of which 356,878 were from Natural Science stream. Natural Science stream students take these five subjects in the EUEE. Therefore, the mean effect obtained in this study

(6.10%) means well above 21,700 students (test-takers) are affected by item position effects. When this is compared with the number of students who have scored more than 50% in EUEE and said to have possessed to higher education last year, that is 22,974 (6.8%) (Federal Ministry of Education, 2023), judgement based on EUEE is unfair to such large proportion of students and put the judgemental validity in question (AERA et al., 2014; EFPA & EAWOP, 2007; Wilson, 2023).

Conclusions and Recommendations

Conclusions

The main purpose of this exploratory ex-post facto research was to find out if there were mean achievement score differences between test-takers depending on which of the exam booklet (01, 02, 03, or 04) they were tested within the subjects: English, Mathematics, Biology, Chemistry, and Physics. The results are that while direct inspection of a set of booklets for an exam showed item position differences, these differences were not statistically significant; majority of the exams in all of the five subjects exhibited significant achievement mean score differences regardless of insignificant differences in item ordering among exam booklets; and significant number of students are affected by the booklet based mean score differences. Therefore, it has been concluded that the observed differences are strong enough to raise questions about the fairness of the EUEE, which puts at a serious disadvantage on average up to 6.10% of test-takers (in tenths of thousands) just due to which exam booklets they were tested with.

Recommendations

In this study it was found that item reordering is consequential to test-takers while assessment authorities are warning exam developers against malpractices with respect to assessment fairness (AERA et al., 2014; EFPA & EAWOP, 2007; Wilson, 2023). In order to maintain exam fairness, the Standards for Educational and Psychological Testing (AERA, et al., 2014), recommends that "Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses" (pp. 85). Furthermore, the authors of the Standards emphasize that "Fairness is a fundamental validity issue and requires attention throughout all stages of test development and use." (pp. 49). Mark Wilson (2023), in relation to test items, declared that "The fundamental fairness requirement of the items design is that, across important subgroups, items function in a similar way for respondents who are at the same [ability] location" (pp. 239). Despite the fact that random item distribution is implied as a factor minimizing item-position effects by such assessment authorities as AERA et al, (2014), strong effect has been observed in this study. This anomalous finding may suggest further and deeper investigations into the nature and behavior of the items in such high-stake tests and test-takers behaviors (Hartig & Buchholz 2012; Soysal & Kogar,

2021; Wu et al., 2016). Thus, for high-stake examinations such as EUEE, in order to address the requirements of the fundamental principle of validity and fairness:

1. Examination (test) development experts (such as those in EAES) be aware of the existence of test takers performance variations due to construct irrelevant factors such as item position changes as observed in this and many other studies. There are many alternative ways of addressing the issue recommended by researchers and assessment authorities to minimize the effect (AERA et al., 2014; Wu et al., 2016). Thus it is professional responsibility of test developers to get familiar with nature of the issue and the suggested measures so that appropriate decisions will be taken to make the EUEE fair and valid.

2. Exam developers, responsible for production of alternate exam booklets, either have to look for other ways of producing alternate tests or practice great caution, should they still use the option of item reordering to curb the exam room malpractice (Mehrens & Lehmann, 1991; Ollennu & Etsey, 2015). Exam development experts (as those in EAES) should be aware of the findings in this study showed that even the insignificant reordering in different booklets put a significant number of test-takers at a serious disadvantage.

3. In the current understanding of validity, validity is in the final argument based on assessment products. Among other things, in high-stake examinations individuals' fates are determined based on test-takers scores. Faulty interpretation of assessment results, lead to unfair judgments by exam result users (AERA, et al., 2014, Colwell, 2013) and victimizes individuals. Therefore, exam result publishers and users should identify the existing exam inconsistencies and apply compensatory approach for the disadvantaged groups by such effects as item positions. Furthermore, result publishers should organize and document evidences about the fairness, reliability and validity of exam results to support decision processes about individuals and systems.

4. While majority of research in this area of high-stake assessment is focused on the effects of item reordering (item positioning) on various students learning outcomes, the causes why item position effects occur were not studied. What is in the nature or characteristics of items that is resulting in change in their functioning in different test forms need to be studied to come up with effective measures against item position effects. However, in the current research indicative result about the seriousness of the problem beyond mere item reordering was found. Therefore, further research should be continued in determining the hidden variables and the extent of their effects in test takers achievements results due to alternate tests with or without item positioning.

5. Many of the studies referred to in this study revealed that item-position effects are not the only factors contributing to individual performance differences in such high-stake examinations. Therefore, research in this area should be extended to determining the particular contribution of item-position effects among other factors.

Limitation of the study

Having noted the extent to which item position effects can be, the result here has to be moderated by paying attention to the limitations of the study. Item position researchers acknowledge several factors contributing to achievement differences beside item position effects (Hartig & Buchholz 2012; Soysal & Kogar, 2021, Wu et al., 2016). However, in this study only item position effect was considered as factor behind the observed achievement mean score differences between test takers with different booklets. Thus, the present result should be taken as strong evidence of the existence of item position effects to the extent to challenge the fairness of test results but, the exact determination of the extent has to be further researched by taking into account other variables impacting achievement differences (AERA et al., 2014; Soysal & Kogar, 2021, Wu et al., 2016).

Even though we observed the existence of achievement differences based on exam booklets over all of test subjects in this study, we did not study the relationship between the pattern of achievement mean score difference and subject matter. This was partly due to the limited number of years of examination data we secured and exclusion of those subjects other than Natural Science fields in the high schools. Besides, other than working on the data, the EAES experts responsible for the EUEE preparation, administration, marking and publishing were not contacted to secure critical information about the development of the exam booklets and nature and process of high-stake decisions based on students' achievement scores in EUEE.

Acknowledgment

The researchers in this study are grateful to the officials of the EAES for having trust on us, independent researchers, and allow us to use the national data. We are also indebted to them for preparing and delivering the data based on the sampling areas presented. Also, our thanks go to members and experts of the EAES, giving supporting documents for analysis of this research.

References:

Abdullahi, G. S., Vincent, P., & Akwashiki, A. G. (2020). Effect of changes in item-sequence on students' academic achievement in multiple-choice test of mathematical-economics in colleges of education, Lagos State Nigeria. World Journal of Innovative Research, 9(4), 69–76. https://www.wjir.org/download_data/WJIR0904032.pdf

Alakayleh, A. S. (2017). Could different Items arrangements affect 10th grade students ' performance in multiple-choice tests in Maths , Science , and English language final exams. *Journal of Education and Practice*, *8*(17), 180–187. https://www.iiste.org/Journals/index.php/JEP/article/view/37481

Alamneh, A. (2017). National/Public Examination, Assessment in Focus, Bi-annual Publication of the Testing Center, 7(1), 5–10. http://repository.smuc.edu.et/bitstream/123456789/3047/1/assessement%20in%20focus.pdf

American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education. (2014). *STANDARDS for Educational and Psychological Testing*. Joint Committee on Standards for Educational and Psychological Testing (U.S.). American Educational Research Association. https://www.testingstandards.net/uploads/7/6/6/4/76643089/9780935302356.pdf

Anastasi, A. (1976). *Psychological Testing* (4th ed.). Macmillan publishing Co., Inc.

Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (7th ed.). Prentice-Hall, Inc.

Baffoe, J. (2021). *Effects of Item Sequencing in Multiple-Choice Tests on Senior High School Students' Academic Performance In The Kumasi Metropolis: The Moderating Role of Gender*. (Unpublished MPhil thesis, University of Cape Coast). https://ir.ucc.edu.gh/xmlui/handle/123456789/7345

Brennan, R. L. (2013). Commentary on "Validating the interpretations and uses of test scores." *Journal of Educational Measurement*, *50*(1), 74–83. https://doi.org/10.1111/jedm.12001

Chala, L., & Agago, M. (2022). Exploring national examination malpractice mechanisms and countermeasures: An Ethiopian perspective. *International Journal of Instruction*, *15*(3), 413–428. https://doi.org/10.29333/iji.2022.15323a

Carlson, J. L., & Ostrosky, A. L. (1992). Item sequence and student performance on multiple-choice exams: Further evidence. *Journal of Economic Education*, *23*(3), 232–235. https://doi.org/10.2307/1183225

Colwell, N. M. (2013). Test anxiety, computer-adaptive testing and the Common Core. *Journal of Education and Training Studies*, *1*(2), 50–60. https://doi.org/10.11114/jets.v1i2.101

Cronbach, L. J. (1990). *Essentials of Psychological Testing* (5th ed.). Harper Collins Publishers, Inc.

European Federation of Psychologists' Associations [EFPA], & European Association of Work and Organizational Psychologists [EAWOP]. (2007). *European test user standards for test use in work and organizational settings* (Version 1.92). http://www.eawop.org/uploads/datas/10/original/European-test-user-standards-v1-92.pdf

FDRE-Council of Ministers. (2012, January 26). National Educational Assessment and Examination Agency Establishment, Council of Ministers Regulation No.260/2012. *Federal Negarit Gazette*, *15*, 6289–6294.

Federal Democratic Republic Government of Ethiopia (FDRGE). (1994). *Education and Training Policy* (1st ed.). St. George printing Press.

Federal Ministry of Education (2023). Education Statistics Annual Abstract 2022/23 (2015E.C.). Education Management Information System (EMIS) and ICT Executive Office, MOE. https://moe.gov.et/storage/Books/ESAA-2022/23_2015E.C.pdf

Flaugher, R. L., Melton, R. S., & Myers, C. T. (1966). *A study of the effects of item rearrangement* (Research Bulletin No. RB-66-31). Educational Testing Service.

Frey, B. B. (Ed.). (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation* (Vols. 1–4). SAGE Publications. https://us.sagepub.com/en-us/nam/the-sage-encyclopedia-of-educational-research-measurement-and-evaluation/book245469

Gregory, K. (2011). *National examinations: General Education Quality Improvement Program Ethiopia, 2011* [Unpublished report]. Ministry of Education, Ethiopia.

Gerow, J. R. (1980). Performance on achievement tests as a function of the order of item difficulty.

Teaching of Psychology, 7, 93 – 96.

Haladyna, T. M., & Downing, S. M. (2004). Constructing multiple-choice items to measure higher-order thinking. *Assessment & Evaluation in Higher Education, 29*(2), 153–163. https://doi.org/10.1080/0260293042000188455

Hambleton, R. K., & Traub, R. E. (1974). The effects of item order on test performance and stress. *The Journal of Experimental Education*, *43*(1), 40–46. https://doi.org/10.1080/00220973.1974.10806302

Hartig, J., & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling*, *54*(4), 418–431. https://www.psychologie-aktuell.com/fileadmin/download/ptam/4-2012_20121224/06_Hartig.pdf

Holzknecht, F., McCray, G., Eberharter, K., Kremmel, B., Zehentner, M., Spiby, R., & Dunlea, J. (2021). The effect of response order on candidate viewing behaviour and item difficulty in a multiple-choice listening test. *Language Testing*, *38*(1), 41–61. https://doi.org/10.1177/0265532220917316

Japan International Cooperation Agency (JICA). (2017). Project for Capacity Development for Improving Learning Achievement in Mathematics and Science Education in the Federal Democratic Republic of Ethiopia (LAMS).

Kellaghan, T., & Greaney, V. (2019). *Using assessment to improve the quality of education*. UNESCO International Institute for Educational Planning. https://unesdoc.unesco.org/ark:/48223/pf0000126231

Mamaru, A., Getachew, E., & Tafese, D. (2023). Historical Development of National Examination in Ethiopia - 1946 to 2023. J*ournal of educational assessment and examinations,* 1(1), 84 – 122.

Mehrens, W. A., & Lehmann, I. J. L. (1991). *Measurement and Evaluation in Education and Psychology* (4th ed.). Wadsworth/Thomson Learning.

Mikre, F., Getachew, K., Amsale, F., Belay, A., Ferede,T., Workeneh, N., Jibat, N., Worku, A., Abafita7, J., Tilahun, G., & Nigussie, B. (2023). First University-based Grade 12 National Examination Management in Ethiopia: Challenges Encountered. *Ethiopan Journal of Education & Science, 19*(1), 50-64. https://journals.ju.edu.et/index.php/ejes/article/view/5278

Monk, J. J., & Stallings, W. M. (1970). Effects of item order on test scores. *The Journal of Educational Research*, *63*(10), 463–465. https://doi.org/10.1080/00220671.1970.10884067

Nagy, G., Nagengast, B., Frey, A., Becker, M., & Rose, N. (2019). A multilevel study of position effects in PISA achievement tests: Student- and school-level predictors in the German tracked school system. *Assessment in Education: Principles, Policy & Practice*, *26*(4), 422–443. https://doi.org/10.1080/0969594X.2018.1449100

National Education Assessment and Examinations Agency [NEAEA]. (2017). *Ethiopian Third National Learning Assessment of Grade 10 and 12 Students Achievement (ETNLA).* http://www.oecd.org/pisa/pisaproducts/44455820.pdf

Ollennu, S. N. N., & Etsey, Y. K. A. (2015). The impact of item position in multiple-choice test on student performance at the Basic Education Certificate Examination (BECE) level. *Universal Journal of Educational Research*, *3*(10), 718–723. https://doi.org/10.13189/ujer.2015.031009

Opara, I. M., & Ogbanu, G. I. (2023). Effect of item order on the reliability of mathematics test among secondary school students in Rivers State. *Journal of Advances in Education and Philosophy*, *7*(11), 460–466. https://doi.org/10.36348/jaep.2023.v07i11.003

Peck, R., Olsen, C., & Devore, J. (2008). *Introduction to statistics and data analysis* (3rd ed.). Thomson Brooks/Cole.

Pettijohn II, T. F., & Sacco, M. F. (2007). Multiple-choice exam question order influences on student

performance, completion time and perceptions. *Journal of Instructional Psychology*, *34*(3), 142–149. https://psycnet.apa.org/record/2007-15457-010

Plake, B. S. (1980). Item Arrangement and Knowledge of Arrangement on Test Scores. *The Journal of Experimental Education*, *49*(1), 56–58. https://eric.ed.gov/?id=EJ239594

Plake, B. S., Ansorge, C. J., Parker, C. S., & Lowry, S. R. (1982). Effects of item arrangement, knowledge of arrangement, test anxiety, and sex on test performance. *Journal of Educational Measurement*, *19*(1), 49–57. https://doi.org/10.1111/j.1745-3984.1982.tb00114.x

Satti, I., Hassan, B., Alamri, A., Khan, M. A., & Patel, A. (2019). The effect of scrambling test item on students' performance and difficulty level of MCQs test in a College of Medicine, KKU. *Creative Education*, *10*(8), 1813–1818. https://doi.org/10.4236/ce.2019.108130

Schober, P., Boer, MM., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia*, *126*(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

Schweizer, K., Troche, S. J., & Rammsayer, T. H. (2017). Item sequencing in tests of cognitive abilities: Impact on test scores and psychometric properties. *Psychological Test and Assessment Modeling*, 59(2), 157-175.

Shepard, L. A. (1994). *The Challenges of Assessing Young Children Appropriately*. *76*(3), 206–212. https://eric.ed.gov/?id=EJ492843

Soyemi, M. O. (1980). Effect of item position on performance on multiple-choice tests. Unpublished M.Ed. dissertation, University of Jos.

Soysal, S., & Koğar, E. Y. (2021). An investigation of item position effects by means of IRT-based differential item functioning methods. *International Journal of Assessment Tools in Education*, *8*(2), 239–256. https://doi.org/10.21449/ijate.779963

Teferra, T., Asgedom, A., Oumer, J., Tassew, W., Aklilu, D., & Berhannu, A. (2018). Ethiopian Education Development Roadmap (2018-30): An Integrated Executive Summary. Ministry of Education, Education Strategy Center.

Wilson, M. (2023). *Constructing measures: an item response modeling approach*, (2nd Ed.). Routledge. https://doi.org/10.4324/9781003286929

Wu, Q., Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2019). Predictors of individual performance changes related to item positions in PISA assessments. *Large-scale Assessments in Education*, *7*(5), 1–21. https://doi.org/10.1186/s40536-019-0073-6

Zelman, M. (2013). Scoping Study for Item Bank Development: General Education Quality Improvement Program Study. In *General Education Quality Improvement Program*.

Appendices

Appendix A

Students' mean scores and standard deviations (M(SD)) by booklet for same examinations

	Code 01	Code 02	Code 03	Code 04	Total	Max	Test-Teker
English 2015	53.12 (16.87)	53.72 (15.97)	52.68 (15.81)	53.5 (15.98)	53.26 (16.17)	105	11376
English 2016	49.78 (15.51)	49.9 (15.53)	50.1 (15.41)	49.75 (15.69)	49.88 (15.53)	97	11183
English 2017	46.01 (15.08)	45.36 (14.44)	45.57 (14.81)	45.33 (13.73)	45.57 (14.53)	96	11337
English 2018	51.01 (15.39)	50.68 (13.96)	52.67 (17.26)	51.98 (14.41)	51.58 (15.32)	105	9840
English 2019	42.32 (13.74)	41.7 (13.45)	43.04 (13.71)	42.09 (13.65)	42.28 (13.64)	98	10577
English 2020	52.47 (19.75)	52.96 (18.79)	53.15 (18.67)	54.12 (18.15)	53.17 (18.86)	110	9612
Math 2015	25.12 (8.91)	25.83 (8.51)	25.42 (8.37)	25.5 (8.31)	25.46 (8.54)	54	8212
Math 2016	29.68 (1o.87)	30.00 (11.04)	30.23 (11.15)	29.81 (11.26)	29.93 (11.08)	61	7667
Math 2017	24.48 (8.74)	24.12 (8.86)	23.44 (9.02)	23.9 (8.79)	24.00 (8.86)	59	7778
Math 2018	26.95 (8.68)	27.67 (8.68)	26.47 (8.27)	26.98 (8.72)	27.02 (8.58)	61	6623
Math 2019	21.86 (8.13)	21.77 (7,77)	22.65 (7.85)	21.85 (8.21)	22.03 (8.00)	59	7012
Math 2020	31.01 (8.82)	31.72 (8.00)	30.95 (8.93)	28.66 (10.95)	30.62 (39.63)	58	6503
Biology 2015	63.13 (15.08)	60.8 (17.43)	62.99 (16.93)	62.23 (16.22)	62.29 (16.45)	96	8203
Biology 2016	59.91 (15.98)	60.54 (15.46)	60.76 (15.54)	60.57 (15.54)	60.44 (15.63)	94	7662
Biology 2017	60.28 (17.28)	59.77 (18.10)	59.99 (17.99)	59.19 (16.86)	59.82 (17.57)	94	7773
Biology 2018	57.27 (15.5)	55.41 (14.27)	56.23 (14.83)	55.94 (14.99)	56.22 (14.92)	89	6621
Biology 2019	51.01 (17.35)	51.77 (16.69)	51.47 (16.83)	50.72 (17.28)	51.25 (17.04)	100	7012
Biology 2020	52.66 (14.40)	52.65 (13.24)	53.41 (15.07)	54.34 (13.96)	53.24 (14.19)	94	6498
Chemistry 2015	44.54 (11.68)	45.70 (10.82)	46.71 (12.29)	45.02 (11.77)	45.49 (11.67)	74	8200
Chemistry 2016	36.59 (10.19)	37.26 (9.63)	37.83 (9.90)	37.38 (9.95)	37.25 (9.93)	67	7665
Chemistry 2017	40.99 (12.26)	41.26 (13.03)	41.59 (12.54)	41.60 (12.10)	41.35 (12.49)	70	7776
Chemistry 2018	41.50 (12.11)	41.72 (12.80)	43.59 (12.46)	42.11 (12.01)	42.21 (12.37)	71	6624
Chemistry 2019	36.52 (11.04)	36.74 (10.85)	36.20 (11.84)	35.51 (11.58)	36.26 (11.33)	80	7006
Chemistry 2020	46.13 (11.75)	45.59 (12.20)	41.02 (12.02)	43.03 (10.56)	43.99 (11.84)	74	6501
Physics 2015	19.4 (6.43)	19.11 (6.19)	18.98 (6.14)	19.51 (6.39)	19.25 (6.29)	43	8218
Physics 2016	21.02 (7.15)	21.57 (6.86)	21.57 (7.04)	21.33 (6.84)	21.37 (6.98)	45	7659
Physics 2017	19.82 (6.01)	20.35 (6.26)	20.18 (6.19)	19.8 (5.92)	20.04 (6.10)	41	7787
Physics 2018	20.86 (6.39)	20.90 (6.27)	20.87 (6.04)	20.35 (6.07)	20.75 (6.20)	42	6630
Physics 2019	17.78 (5.79)	17.40 (5.40)	17.38 (5.53)	17.28 (5.71)	17.47 (5.61)	46	7006
Physics 2020	22.59 (7.09)	24.63 (6.93)	23.62 (7.36)	25.06 (7.03)	23.95 (17)	45	6500

Appendix B

Estimation of proportion of test-takers to be affected by Item Position Changes

Subject	Year	Booklet Code (I)	Booklet Code (J)	Mean Diff. (I-J)	Std. Error	Sig.	Pooled Std	Absolute Diff.	Populatio n Diff.	Test-Taker		Standard Deviation	
										N(I)	N(J)	SD(I)	SD(J)
English	2018	Code 21	Code 23	1.664*	0.435	0.001	16.34	0.1018	3.98	2513	2441	15.39	17.26
5(13.89%)	2018	Code 22	Code 23	1.991*	0.437	0.000	15.69	0.1269	4.97	2472	2441	13.96	17.26
	2018	Code 22	Code 24	1.299*	0.438	0.016	14.18	0.0916	3.59	2472	2414	13.96	14.41
	2019	Code 12	Code 13	1.337*	0.375	0.002	13.58	0.0985	3.78	2674	2619	13.45	13.71
	2020	Code 51	Code 54	1.656*	0.544	0.012	18.98	0.0872	3.38	2443	2363	19.75	18.15
Mathematics	2015	Code 15	Code 16	0.707*	0.262	0.036	8.72	0.0811	3.18	2162	2074	8.91	8.51
9(25.00%)	2017	Code 59	Code 61	1.048*	0.282	0.001	8.88	0.1181	4.57	2036	1921	8.74	9.02
	2018	Code 26	Code 27	1.204*	0.298	0.000	8.44	0.1427	5.57	1682	1630	8.6	8.27
	2019	Code 15	Code 17	0.791*	0.268	0.017	8.00	0.0989	3.78	1826	1722	8.13	7.85
	2019	Code 16	Code 17	0.885*	0.27	0.006	7.81	0.1133	4.38	1778	1722	7.77	7.85
	2019	Code 17	Code 18	0.804*	0.274	0.018	8.03	0.1001	3.98	1722	1686	7.85	8.21
	2020	Code 47	Code 50	2.358*	0.337	0.000	9.89	0.2384	9.29	1700	1540	8.82	10.95
	2020	Code 48	Code 50	3.066*	0.339	0.000	10.23	0.2996	11.6	1655	1540	9.52	10.95
	2020	Code 49	Code 50	2.297*	0.341	0.000	9.97	0.2304	9.09	1608	1540	8.93	10.95
Physics	2015	Code 25	Code 26	0.533*	0.199	0.037	6.26	0.0851	3.38	2049	1958	6.14	6.39
9(25.00%)	2017	Code 67	Code 68	0.537*	0.194	0.028	6.13	0.0875	3.39	2020	1953	6.01	6.26
	2017	Code 68	Code 70	0.551*	0.197	0.027	6.10	0.0904	3.58	1953	1878	6.26	5.92
	2019	Code 23	Code 26	0.500*	0.19	0.042	5.75	0.0869	3.38	1835	1661	5.79	5.71
	2020	Code 39	Code 40	2.042*	0.245	0.000	7.01	0.2913	11.41	1693	1667	7.09	6.93
	2020	Code 39	Code 41	1.026*	0.248	0.000	7.22	0.1421	5.57	1693	1598	7.09	7.36
	2020	Code 39	Code 42	2.474*	0.25	0.000	7.06	0.3504	13.68	1693	1542	7.09	7.03
	2020	Code 40	Code 41	1.015*	0.249	0.000	7.14	0.1421	5.57	1667	1598	6.93	7.36
	2020	Code 41	Code 42	1.448*	0.254	0.000	7.20	0.2011	7.26	1598	1542	7.36	7.03
Chemistry	2015	Code 27	Code 28	1.166*	0.3591	0.006	11.26	0.1035	3.98	2120	2087	11.68	10.82
15(41.67%)	2015	Code 27	Code 29	2.177*	0.3619	0.000	11.98	0.1817	7.142	2120	2025	11.68	12.29
	2015	Code 28	Code 29	1.011*	0.3633	0.028	11.57	0.0874	3.39	2087	2025	10.82	12.29

Appendix B

Estimation of proportion of test-takers to be affected by Item Position Changes

Subject	Year	Booklet Code (I)	Booklet Code (J)	Mean Diff. (I-J)	Std. Error	Sig.	Pooled Std	Absolute Diff.	Populatio n Diff.	Test-Taker		Standard Deviation	
										N(I)	N(J)	SD(I)	SD(J)
	2015	Code 29	Code 30	1.691*	0.3686	0.000	12.04	0.1405	5.56	2025	1968	12.29	11.77
	2016	Code 71	Code 73	1.246*	0.318	0.001	10.05	0.1240	4.77	2021	1878	10.19	9.9
	2018	Code 37	Code 39	2.089*	0.427	0.000	12.28	0.1701	6.74	1726	1620	12.11	12.46
	2018	Code 38	Code 39	1.865*	0.429	0.000	12.63	0.1476	6.35	1692	1620	12.8	12.46
	2018	Code 39	Code 40	1.480*	0.436	0.004	12.24	0.1209	4.77	1620	1586	12.46	12.01
	2019	Code 27	Code 30	1.004*	0.384	0.044	11.05	0.0909	3.58	1829	35.51	11.04	11.58
	2019	Code 28	Code 30	1.228*	0.386	0.008	10.86	0.1130	4.38	1782	35.51	10.85	11.58
	2020	Code 35	Code 37	5.109*	0.405	0.000	11.88	0.4300	16.64	1699	1613	11.75	12.02
	2020	Code 35	Code 38	3.091*	0.41	0.000	11.20	0.2760	10.83	1699	1543	11.75	10.56
	2020	Code 36	Code 37	4.570*	0.409	0.000	12.11	0.3773	14.61	1646	1613	12.2	12.02
	2020	Code 36	Code 38	2.552*	0.413	0.000	11.44	0.2232	8.7	1646	1543	12.2	10.56
	2020	Code 37	Code 38	2.018*	0.415	0.000	11.33	0.1781	6.94	1613	1543	12.02	10.56
Biology	2015	Code 31	Code 32	2.332*	0.505	0.000	16.28	0.1432	5.57	2144	2099	15.08	17.43
6(16.67%)	2015	Code 32	Code 33	2.193*	0.512	0.000	17.19	0.1276	4.97	2099	2022	17.43	16.93
	2015	Code 32	Code 34	1.433*	0.518	0.029	16.86	0.0850	3.19	2099	1938	17.43	16.22
	2018	Code 41	Code 42	1.863*	0.511	0.002	14.91	0.1250	4.77	1726	1685	15.5	14.27
	2020	Code 31	Code 34	1.687*	0.499	0.004	14.19	0.1189	4.57	1694	1545	14.4	13.96
	2020	Code 32	Code 34	1.689*	0.501	0.004	13.59	0.1243	4.77	1657	1545	13.24	13.96

1. **Absolute Difference** is calculated as the ratio of the mean difference to the pooled standard deviation.
2. **Population Difference** is determined using standard distribution table as percentage of the population between the mean and the absolute difference. In order not to exaggerate the size of the population affected with Item-position effects rounding up in calculations was not made.