International Journal of Humanities Social Science and Management (IJHSSM)
Volume 4, Issue 3, May.-June, 2024, pp: 941-951 www.ijhssm.org

Examining Effect of Item Positions in Multiple-Choice Physics Test on Gender Achievement Gap.

Fagbenro W. Ayoola (PhD.)
Department of Science Education
Federal University Wukari.

Shamsu Aliyu
Department of Science Education
Federal University Wukari

Date of Submission: 14-05-2024 Date of Acceptance: 29-05-2024

Abstract

This study explored the effect of changes in item sequence on student's achievement in multiple-choice physics tests in Senior Secondary School II in Taraba State. The study adopted repeated measures two-group within-subject experimental research design. The research collected data in order to answer two research questions and test two hypotheses. The sample comprised 450 senior secondary II Physics students $(male = 303\ and\ female\ = 147\ )$ drawn from population of Physics students in Taraba State. Multi stage sampling technique was employed to randomly select twelve schools from three Local Government Areas of the three Senatorial Districts of Taraba State, and an intact arm of SS II from each of the sampled schools was used. Two parallel 40-items Physics Achievement Test developed by the researcher were used for data collection. The resulting data were collated and analysed using descriptive statistics and t-test. The finding revealed that there is significant difference between the students mean achievement score in the format A and Format B of the physics achievement test $(t = 4.409\ , df\ = 898, p < 0.05, two - tailed)$. There is a significant difference in the mean score of male students between format A and format B. The mean difference between conditions for the female students was also significant. Change in format accounted for 36% of variance in mean score of the female students between conditions while it accounted for 26% mean score difference of the male students. Sequels to the findings, it is concluded that students will perform better in physics achievement test if the test items are arranged randomly than in descending order of difficulty. Presenting items in descending other of difficulty will adversely affect the female students. It's therefore recommended that test developers should endeavour to arrange physics test items randomly and all test taker be administered with the same test form so that the desired trait can be adequately measured.

I. Introduction

The enrolment ratio between female and male, known as gender parity index (GPI), has been on the increase. This shows that more females enrolled than males, reporting an increase in drop rate among males (Nishimura, 2017; Belal, 2009). Also, the GPI indicates rise of gender-sensitive policies resulting in increase of gender gaps. In spite of the swell in the attendance of females, women still need to be as well educated as men in physics (Evans et al., 2020). As captured by UNESCO (2000), gender equity means fairness of treatment for both women and men, with respect to their needs. This may include equal treatment or treatment that is different but which is considered equivalent in terms of rights, benefits, obligations, and opportunities. From all the studies carried out in relation to gender in education sector, it is crystal clear that there are gender related issues surrounding students. The issues vary from differences among the teaching strategies employed by teachers and among the teachers (Ananga, 2021; Barnett-Cooper, 2012; Belal, 2009; Lee, 2021; Toraman & Ozen, 2019), effect of gender on students' achievement (Ananga, 2021; Barnett-Cooper, 2012; Lee, 2021; Meinck & Brese, 2019), to the context of school, curricula and policies (Dickey, 2014; Kollmayer et al., 2020; Wigati, 2019).

Science education stand out as a major contributor to nation's prosperity, welfare, and security, among subjects offered in schools (Abraha et al., 2019; Abraha et al., 2021). Many gender equity based studies have been carried out in science education. These studies investigated varied gender-related topics such as gender-biased career guidances, science curricula, and messages in science textbooks and stereotypical gender images in science textbooks. There is absence of notable studies that buttress the connection between the construction of scientific knowledge and gender (Hearn & Husu, 2011). Students are assured of equal rights to quality science education through educational equity (Jalak & Nasri, 2019). Researches among students enrolled in STEM have shown gender differences. For instance, girls prefer trades in the social sciences in contrast to boys, who are often seen in profession related to STEM (Meinck & Brese, 2019). While male students prefer chemistry and physics, female students are fascinated by Biology (Kang et al., 2019). The masculinity associated with Physics, Mathematics and chemistry poses significant challenges among female students (Makarova et al., 2019). Girls' underachievement and under representation in science at various levels of schooling, mostly starting with the secondary school level have been documented by researches worldwide (Lundberg, 2020; McDool & Morris, 2022; UNESCO, 2017). According to Alexakos and Antoine (2003), there is a consistent gap in participation, interest, and achievement in primary and secondary school classrooms, with the gap more pronounced in physics and chemistry. Allegrini (2015) reported that females are underrepresented in STEM, especially in computer science, physics, mathematics and engineering. It is reported that countries such as Finland, Sweden and Norway, characterized by gender equality have wider gaps in science than the countries that rank poorly in gender equality (Stoet & Geary 2018).

There is an urgent need to address the gender disparity in female participation, attainment and outcomes in physics. Most studies in science education concentrated on the preferences and perceptions of males and females regarding STEM subjects (Kang et al., 2019; Makarova et al., 2019) and their career aspirations (Meinck & Brese, 2019). Tytler& Osborner (2012) submitted that the quality of teaching is a key determinant of student engagement and success in science. Much effort has been geared towards promoting gender equality in schooling in terms of access and completion rate, while less attention has been paid to gender biased assessment practices that affect inclusion in the teaching and learning of science. There is an urgent need to pay attention to the quality of instrument used in assessing students in science. In terms of assessment, there is bias in a question "if a factor other than ability (in this case gender) affects the likelihood that a student will answer the question correctly" (Dietz, Pearson, Semak, & Willis, 2012).

Measuring Gender Achievement Gap

Measurement of students' scientific reasoning and knowledge processes is a complex job, yet vital to efficient teaching and learning (National Research Council 2001, 2007). To shed light on gender disparities in educational opportunity and to comprehend how gender norms and stereotypes shaped students' lives, test-based gender achievement gaps are often used. The male and female students' average total scores on an assessment can be used to estimate gender achievement gaps by comparing the scores. On an average, females do better than males on reading/English language Arts and males do better than females on math test (Chatterji, 2006; Fryer & Levitt, 2009; Lee, Moon, & Hegar, 2011; Penner & Paret, 2008;Robinson & Lubienski, 2011; Sohn, 2012). But the conclusions drawn may be sensitive to how gender achievement gaps are measured on standardized tests. For a test that assesses a uni-dimensional construct, this method is suitable. But if gender differences in achievement vary among the set of skills tested, whatever gender gap calculated from the test scores will depend on the mix of skills measured by the test. Prior researches suggested that there is evidence of association between gender achievement gaps and item format. Gaps are often more female bias on tests with more constructed response items and male bias on tests with more multiple-choice items. The differences may be the use of different item types to measure the different skills. Alternatively, gender differences in the ancillary, construct-irrelevant skills needed by the different item types (e.g., the handwriting skills required for essay questions) may be the cause of the differences or pattern. Whatever way it is, relationships exist between test item format and gender achievement gap (National Centre for Education Statistics, 2009a; 2009b). Research generally shows that male do better on multiple-choice items than female while female do better than male on constructed response items. This pattern may be due to gender differences in the ancillary construct irrelevant skills needed by the different item types (the handwriting skills required for essay) and or gender differences on the skills intended to be measured by the test (Taylor & Lee, 2012; Willingham & Cole, 2013).

Item Position Effect on Multiple Choice-Objective Question

Recent studies in science education using MC have documented item feature effects that considerably influence the extraction and measurement of student knowledge. For example, prior knowledge of the behavioural trait to be tested correlated with higher poise in response accuracy in physics education and better performance on MC assessments (e.g., Caleon & Subramaniam, 2010) and chemistry education (e.g., Rodrigues, Taylor, Cameron, Syme-Smith, & Fortuna, 2010). To make the results of MC test more reliable, one of the widely used methods is to place items in different positions or locations within the tests (Bulut et al., 2017). Thus, problems such as individuals memorizing items or copying answers of other examinees during the test application can be overcome (Bulut, 2015). This method solved the problem of exam malpractice that may affect the psychometric properties of the test, however, it leads to item position effects (Bulut, 2015). The consequence of item position effect on individuals' abilities is ignored in many testing situation. If it occurs, it is assumed to be the same for all persons and all items (Hahne, 2008 : Albano, 2013). In practice, individuals' test scores can vary according to item position (Albano,2013), that is, it poses threat to the validity of test score interpretations (Trendtel & Robitzsch, 2018). The positions of items in test forms created by item position manipulations may lead to differential item functioning (DIF) (Akayleh, 2018; Balta & Omur Sunbul, 2017; Debeer & Janssen, 2013; Erdem, 2015). While some studies posit that item position effect affect examinees achievement (Debeer & Janssen, 2013; Hartig & Buchholz, 2012; Ollennu & Etsey, 2015; The West African Examinations Council [WAEC], 1993), others have concluded that item position does not affect examinees achievement (Doğan Gül & Çokluk Bökeoğlu, 2018; Perlini et al., 1988; Tal et al., 2008). Other studies agreed that item position caused bias in item parameter estimates (Debeer & Janssen, 2013; Doğan Gül & Çokluk Bökeoğlu, 2018; Hecht et al., 2015; Meyers et al., 2009). Majority of the studies on IP effects are mostly based on Classical Test Theory (CTT), those based om Item Response Theory (IRT) framework are also available (Debeer & Janssen, 2013; Hahne, 2008; Hohensinn et al., 2008; Qian, 2014; Weirich et al., 2014).

Research Hypotheses

Hypothesis I**:** There is no significant difference in the students' mean scores when the item sequence of multiple-choice physics test is changed.

Hypothesis II**:** There is no significant difference in the mean scores of male and female students when the item sequence of multiple-choice physics test is changed.

Hypothesis III: There is no significant difference in the means scores of male students between different question formats.

Hypothesis IV: There is no significant difference in the mean scores of female students between different question formats.

Research Questions

Research Question 1: What is the effect size between conditions for both male and female?

II. Method

The study used the repeated measures two-group within-subject experimental research design. In this design, the subjects are their own controls because the model assesses how a subject responds to all of the treatments (Kerlinger & Lee, 2000). The treatment is the variable being manipulated whose effect is under investigation; in this case, the variable being manipulated is the item sequence. The subjects were selected into two groups and the groups randomly assigned to treatments. The first group answered format A of the Physics Achievement test first, while the second group answered the format B first. It followed that the first group answered format B during the second administration while the second group answered format A.

The target population for this study comprised all Senior Secondary School II Students in Taraba State, Nigeria. The estimated population of all the students in Senior Secondary School II in Taraba state is 9,528. The choice of SS II students as participants for the study emanated from the fact that SS II students are not preparing for any external examination. It is also assumed that the students would have covered enough Physics (electricity) content to be able to respond to any question given to them by the researcher. Table 1 shows the distribution of Senior Secondary School II Science Students in Taraba State by Council Area and Sex.

The sampling procedure adopted was a multistage random sampling technique. The twenty (20) local government areas in Taraba State were first stratified into the three senatorial districts. Random sampling technique was used to select one local government areas from each of the three senatorial districts. The secondary schools in each of the three randomly selected local government areas were stratified into private and public secondary schools. Two public and two private secondary schools were randomly selected from each local government areas. The selected arms in the schools were an intact class. The sample size is four hundred and five (405). Table 1 shows the distribution of Senior Secondary School II Science Students in Taraba State by Council Area and Sex.

Table 1: Sample Frame for Effect of Item Sequence on Physics Achievement

S/N	**Selected Local Government Area**	**No of School**	**Sample of SSII**	**Sample of SSII by Sex**	
				Male	**Female**
1	Wukari	4	154	108	46
2	Jalingo	4	193	134	59
3	Bali	4	103	61	42
Total		**12**	**450**	**303**	**147**

The instrument employed for data collection in this study is Physics Achievement Test. The initial draft of PAT consisted of 60 items. It was developed by the researcher. However, the physics syllabus prepared for SSCE by WAEC and NECO, as well as the Physics curriculum prepared by the Federal Ministry of Education, Abuja, Nigeria was taken into consideration. The items were developed from the content of the physics syllabus and physics curriculum for senior secondary one and senior secondary two. In addition, the items were written by following the pattern of WAEC and NECO. That is, each item was placed on four-option response mode of A, B, C, and D. The items covered one main theme in physics, this is electricity.

The decision to develop items from electricity was taken, because, in the first place, analysis of the scheme of work in all the schools that were sampled showed that all the physics teachers had taught electricity. Two, there are many formula and equations which the students have to master in electricity. As noted by the Physics Chief Examiners' (WAEC, 2012), many candidates have difficulties in the use of equations and formulas in test items.

Table 2: Table of Specification

S/N	Content	Knowledge	Comprehension	Application	**Total**
1	Electric charge	2	2		**4**
2	Current in a simple circuit		1	4	**5**
3	Potential difference	1	2	2	**5**
4	Resistance	2	2	2	**6**
5	Series circuit	1	1	3	**5**
6	Parallel circuit	1	1	3	**5**
7	Electric power	2	1	2	**5**
8	Electric energy	1	2	2	**5**
	Total	**10**	**12**	**18**	**40**

Test blue print, illustrated in table 2, was developed to ensure the content validity of the test. The thought processes were limited to knowledge, comprehension and application because of the age of the students and reduction of tedium. Also, opinions of panel of qualified experts in Physics Education and Education Evaluation were sought in deciding the appropriateness of the items to give logical validity index of 0.77.

The draft copy of PAT consisting of 60 items was first administered to 150 students to determined the difficulty index. The time allowed for the students to take the test was 60 minutes. On the average, it took the students about 50 minutes to finish the test.

Item analysis was carried out using CTT to select the final items. On the basis of the criteria set for the difficulty indices (i.e. $0.30 \geq p \leq 0.80$), items which failed to satisfy the conditions were deleted.

The format A of Physics Achievement Test consists of 40 items that were randomly arranged and the format B of Physics Achievement Test consists of 40 items arranged from "easy to hard” based on their difficulty. The test reliability was estimated using parallel form method. The two PAT tests yielded two sets of scores which were correlated and this gave coefficient of Equivalence of 0.736. Table 3 shows the reliability coefficient of the instrument used.

Table 3 : Correlation of total score of PATA and PATB

		TscoreA	TscoreB
Tscore A	Pearson corr.	1	0.736
	Sig. (2-tailed)		0.000
	N		50
Tscore B	Pearson corr.	0.736	1
	Sig. (2-tailed	0.000	
	N	50	

Hypothesis I: There is no significant difference in the students’ mean scores when the item sequence of multiple-choice physics test is changed

Table 4: Summary of the descriptive statistics of Achievement in Physics Scores by Format

	Question format	N	Mean	Std. Deviation	Std. Error Mean
Physics Achievement Test Score	Format A	450	22.2867	5.19466	0.24488
	Format B	450	20.7000	5.59480	0.26374

Table 5: Summary of the descriptive statistics of Achievement in Physics Score By Format and Sex

	Sex	N	Mean	Std. Deviation	Std. Error Mean
Physics Achievement Score format A	Male	303	22.6040	5.39943	0.31019
	Female	147	21.6327	4.69491	0.38723
Physics Achievement Score format B	Male	303	21.1254	5.70153	0.32754
	Female	147	19.8231	5.27945	0.43544

Hypothesis I: There is no significant difference in the students’ mean scores when the item sequence of multiple-choice physics test is changed.

Table 4 shows that the mean score in Physics Achievement Test Format A $(\bar{X} = 22.2867, S.D = 5.19)$ is more than the mean score in Physics Achievement Test Format B $(\bar{X} = 20.7000,\ S.D = 5.59)$. The mean difference between the two conditions was 1.5867.

Table 6: The Student’s t Table

	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
Physics Achievement Test Score	4.409	898	.000	1.58667	0.35990

Table 6 shows that the difference between conditions was significant $(t = 4.409, df = 898, p = 0.000, two - tailed)$.

The null hypothesis of no significant difference between the mean scores of the two conditions, when items are arranged randomly and when they are arranged in descending order of difficulty, is rejected. This implies that there is a significant difference in the mean achievement score when the item sequence of the multiple-choice physics achievement test is changed.

Hypothesis II: There is no significant difference in the mean scores of male and female students when the item sequence of multiple-choice physics test is changed.

As indicated in table 5, the male students achieved more than the female students in Format A of the Physics Achievement test. The mean difference is 0.97131. The male students achieved better than the female students on physics achievement test format A

Table 7: The Student's t Table

	t	df	Sig.	Mean Difference	Std. Error Difference
Physics objective test format A	1.865	448	0.063	0.97131	0.52070

From table 7, an independent t-test showed that the mean difference between male and female was not significant $(t = 1.865, df = 448, p = 0.063, two - tailed)$. This implies that there is no significant difference in the achievement of male and female students when the items on physics achievement test are arranged randomly.

Also from table 5, the mean score of the male students in Physics Achievement test Format B is 21.1254 and that of the female is 19.8231. The mean different of the two groups is 1.30228. The male students achieved better than female students on physics achievement test format B.

Table 8: The Student's t Table

	t	df	Sig.	Mean Difference	Std. Error Difference
Physics objective test format B	2.327	448	0.020	1.30228	0.55961

As shown in table 8, an independent t-test showed that the difference between male and female achievement was significant $(t = 2.327, df = 448, p = 0.020, two - tailed)$. This implies that there is a significant difference in the achievement of male and female students when the items on physics achievement test are arranged in descending order of items difficulty. It follows that arranging items of physics achievement test in descending order of item difficulty discriminate between male and female students.

Hypothesis III: There is no significant difference in the means scores of male students between different question formats.

Between conditions, the mean achievement score of male students in Format A of the Physics Achievement test (22.604) is greater than their mean achievement score in Format B (21.1254). The mean difference of the achievement score between conditions is 1.47855. The male students achieved better in format A of the physics achievement test than format B.

Table 9: The Student's t Table

	t	df	Sig.	Mean Difference	Std. Error Difference
Physics Achievement Test Score	3.278	604	0.001	1.47855	0.45111

From table 9, an independent t-test showed that the mean difference between the two conditions for male students was significant $(t = 3.278, df = 604, p = 0.001, two - tailed)$. This implies that the male students' achievement in the format A of the physics achievement test is significantly different from that of the format B. It follows that the manner in which the items on the physics achievement test are arranged affected the response of the male students. The hypothesis that there is no significant in the responses of the male when the sequence of the items is changed is rejected.

Hypothesis IV: There is no significant difference in the mean scores of female students between different question formats.

As illustrated in table 5, the mean achievement score of female students in Format A of the Physics Achievement test (21.6327) is higher than their mean achievement score in Format B (19.8231). The mean difference of the achievement between conditions is 1.80952. The female students achieved better in the format A of the physics achievement test than format B.

Table : The Student's t Table

	t	df	Sig.	Mean Difference	Std. Error Difference
Physics Achievement Test Score	3.105	292	0.002	1.80952	.58271

This implies that the difference in the mean achievement of female students between conditions is significant. It follows that the way the items on the physics achievement test are arranged affected the responses of the female students. The null hypothesis that the change in the sequence of items of multiple-choice physics test did not affect the responses of the female is rejected.

Research Questions

Research Question 1: What is the effect size between conditions for both male and female?

Between conditions, for male, the mean difference in achievement is 1.49 and that of the female is 1.81. The mean difference of the female between conditions is larger than of the male, meaning that the arrangement affected female more than the male. The effect size calculated for the males is 26.6 and that of the females is 36.3. Meaning that 36.3% of the variance in the achievement of the female students is explained by the manner in which the items in Physics Achievement test are arranged while 26.6% of the variance in the achievement of the male is explained by how the items are arranged.

III. Discussion

The study found out that the male students achieved better than their female counterpart in the two formats of the physics achievement tests. Also, male and female students achieved better in the format A of the physics achievement test than format B. An independent t-test showed that the mean difference between conditions for both male and female is significant. This implies that changing the sequence of items on physics achievement test from random arrangement to arrangement in descending order of difficulty affected both male and female students. Therefore, the hypothesis of no significant difference in the mean score of male and female students when the item sequence of multiple-choice physics test is changed is rejected. It follows that the manner in which the items on physics achievement test are arranged will affect both male and female students.

Observing the mean difference of both male (1.49) and female (1.81) in the two conditions, the mean difference of the female is larger than of the male, meaning that the arrangement affected female more than the male. The effect size calculated for female is 36.3 and that of the male is 26.6. Meaning that 36.3% of the variance in the achievement of the female students is explained by the manner in which the items in Physics Achievement test are arranged while 26.6% of the variance in the achievement of the male is explained by how the items are arranged. It follows that how items that composed physics achievement test is presented to the students affected the female students more than the female students.

Previous studies have submitted that the features of assessment questions can result in gender bias (Cassels & Johnstone, 1984). Halpern et al.(2007) argued that males outperform females on visual-spatial questions whereas females tend to perform better on more "verbal" tasks . More recently, a study of the impact of exam question structure on the performance of first-year physics undergraduates showed that while student performance improved with increased scaffolding of questions, the increase in average examination mark was greater for female students (13.4%) than for male students (9%) (Gibson, Jardine-Wright & Bateman, 2015). However, subsequent research indicated that scaffolding was not the dominant determinant of gender gaps in MCQs, but instead that questions with a high visual-spatial content (diagrams and multidimensional context) were stronger indicators of male bias (Dawkins, Hedgeland &Jordan, 2017). This result was consistent with work on the gender differences in performance over eight years in the Australian

Science Olympiad Exam for physics, which revealed that the gender gaps in achievement correlated with the question type, particularly with respect to the content, context, and presentation (Cassels & Johnstone, 1984). This result is also in agreement with Gladys, Furst, Holdsworth, & Dastoor (2023) that highlights gender bias in multiple-choice physics examinations based on question characteristics that could be a useful tool in understanding the presence and origin of gender gaps in student performance.

IV. Conclusion and Recommendations

This study investigated the effects of changes in item sequence on students achievement in multiple-choice physics tests in Taraba state of Nigeria. The findings revealed that arrangement of items of Physics achievement test is associated with Achievement in Physics. Arranging the items randomly make students achieve better than when they are arrange in descending order of difficulty. Though the descending order of difficulty arrangement of items on physics achievement test affected both male and female, the effect on female is more pronounce than that of female. The practice is no favourable to the female students and may increase the achievement gap between male and female. It is concluded that items should be presented to all the physics students in the same manner. The practice of presenting physics items to a student in one manner and to another in different manner should be stopped. All physics students should presented physics items in random arrangement rather than arrangement in descending order of item difficulty.

Sequels to the findings, recommendation were made thus: The items that made up the physics achievement test should be arranged randomly. The teacher should endeavour to present the same arrangement to all the students.

References

[1]. Abraha, M., Dagnew, A., & Seifu, A.(2019). Gender responsive pedagogy: Practices, challenges & opportunities - A case of secondary schools of North Wollo Zone, Ethiopia. Journal of Education, Society and Behavioural Science, 30(3), 1-17.

[2]. Abraha, M., Seifu, A., & Dagnew, A. (2021). Manifestation of the Fawe's gender responsive pedagogy in the Ethiopian general secondary school science teaching. International Journal of Management, 12(1), 1560-1571.

[3]. Akayleh, A. S. A. (2018). Precision of the estimations for some methods of the CTT and IRT as a base to display the differential item functions on the different item ordered test formats.https://bit.ly/3aJeFKx

[4]. Albano, A. D. (2013). Multilevel modeling of item position effects. Journal of Educational Measurement, 50(4), 408–426.

[5]. Alexakos, K., & Antoine, W. (2003). The gender gap in science education. Science Teacher (Normal, Ill.), 70(3), 30.

[6]. Allegrini, A. (2015). Gender, STEM studies and educational choices. Insights from feminist perspectives. In understanding student participation and choice in science and technology education (pp. 43-59). Springer, Dordrecht

[7]. Ananga, E. D. (2021). Gender responsive pedagogy for teaching and learning: The practice in Ghana's initial teacher education programme. Creative Education, 12, 848-864.

[8]. Balta, E., & Omur Sunbul, S. (2017). An investigation of ordering test items differently depending on their difficulty level by differential item functioning. Eurasian Journal of Educational Research, 72, 23-42.

[9]. Barnett-Cooper, D. (2012). A study of the impact of single-gender classes on middle school students in an urban setting [Dissertation, Rowan University]. Rowan Digital Works. Theses and Dissertations, 144. https://rdw.rowan.edu/etd/144

[10]. Belal, F. O. (2009). Gender equality in secondary education: A study of girls' educational access and participation in Jordan between 2000 and 2005 [Dissertation, Seton Hall University]. Seton Hall University Dissertations and Theses (ETDs). 439. https://scholarship.shu.edu/dissertations/439

[11]. Bulut, O. (2015). An empirical analysis of gender-based DIF due to test booklet effect. European Journal of Research on Education, 3(1), 7-16.

[12]. Bulut, O., Quo, Q., & Gierl, M. J. (2017). A structural equation modeling approach for examining position effects in large-scale assessments. Large-scale Assessments in Education, 5(1), 8. http://doi.org/10.1186/s40536-017-0042-x

[13]. Caleon, I.S. & Subramaniam, R. (2010). Do students know what they know and what they don’t know? Using a four-tier diagnostic test to assess the nature of students’ alternative conceptions. Research in Science Education, 40, 313-337.

[14]. Cassels, J. R. T. & Johnstone, A. H. (1984). The effect of language on student performance on multiple choice tests in chemistry, J Chem. Educ. 61, 613.

[15]. Chatterji, M. (2006). Reading achievement gaps, correlates, and moderators of early reading achievement: Evidence from the Early Childhood Longitudinal Study (ECLS) kindergarten to first grade sample. Journal of Educational Psychology, 98(3), 489.

[16]. Dawkins, H., Hedgeland, H. &Jordan, S. (2017). Impact of scaffolding and question structure on the gender gap, Phys. Rev. Phys. Educ. Res. 13, 020117.

[17]. Debeer, D., & Janssen, R. (2013). Modeling item‐position effects within an IRT framework. Journal of Educational Measurement, 50(2), 164-185.

[18]. Dickey, M. W. (2014). Gender-specific instructional strategies and student achievement in 5th grade classrooms [Dissertation, University of South Carolina]. University of South Carolina University Libraries Theses and Dissertations. https://scholarcommons.sc.edu/etd/2624

[19]. Dietz, R. D., Pearson, R. H., Semak, M. R. & Willis, C.W. (2012). Gender bias in the Force Concept Inventory? AIP Conf. Proc. 1413, 171.

[20]. Doğan Gül, Ç., & Çokluk Bökeoğlu, Ö. (2018). The comparison of academic success of students with low and high anxiety levels in tests varying in item difficulty. Inonu University Journal of the Faculty of Education, 19(3), 252-265. https://doi.org/10.17679 /inuefd.341477

[21]. Evans, D. K., Akmal, M., & Jakiela, P. (2020, January 17). Gender gaps in education: The long view. Center for Global Development. https://www.cgdev.org/publication/gender-gaps-education-long-view

[22]. Erdem, B. (2015). Investigation of Common Exams Used in Transition to High Schools in Terms of Differential Item Functioning Regarding Booklet Types with Different Methods Hacettepe University. Ankara. Int. J. Assess. Tools Educ., Vol. 8, No. 2, (2021) pp. 239–256

[23]. Fryer, R. G., Jr., & Levitt, S. D. (2009). An empirical analysis of the gender gap in mathematics. American Economic Journal: Applied Economics, 2(2), 210-240.

[24]. Gibson, V., Jardine-Wright, L., & Bateman, E. (2015). An investigation into the impact of question structure on the performance of first year physics undergraduate students at the University of Cambridge, Eur. J. Phys. 36, 045014

[25]. Gladys, M. J., Furst, J. E., Holdsworth, J. L. & Dastoor, P. C. (2023). Gender bias in first-year multiple-choice physics examinations, Physical Review Physics Education Research, 19, 020109.

[26]. Hahne, J. (2008). Analyzing position effects within reasoning items using the LLTM for structurally incomplete data. Psychology Science Quarterly, 50(3), 379–390.

[27]. 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.

[28]. Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S. & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics, Psychol. Sci. Publ.Interest 8, 1.

[29]. Hearn, J., & Husu, L. (2011). Understanding gender: Some implications for science and technology. Interdisciplinary Science Review, 36(2), 103-113. https://doi.org/10.1179/030801811x13013181961301

[30]. Hecht, M., Weirich, S., Siegle, T., & Frey, A. (2015). Effects of design properties on parameter estimation in large-scale assessments. Educational and Psychological Measurement, 75 (6), 1021-1044. https://doi.org/10.1177/0013164415573311

[31]. Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L., & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the linear logistic test model. Psychology Science Quarterly, 50, 391-402.

[32]. Jalak, J. T., & Nasri, N. M. (2019). Systematic review: The impact of pedagogy on equity in science education in rural schools. Creative Education, 10(12), 3243–3254. https://doi.org/10.4236/ce.2019.1012248

[33]. Kang, J., Hense, J., Scheersoi, A., & Keinonen, T. (2018). Gender study on the

relationships between science interest and future career perspectives. International Journal of Science Education, 41(1), 80–101. https://doi.org/10.1080/09500693.2018.1534021

[34]. Kerlinger, F.N. & H.B. Lee (2000). Foundations of Behavioural Research, 4th edition, Wadsworth, Thomson Learning Inc.

[35]. Kollmayer, M., Schultes, M.-T., Lüftenegger, M., Finsterwald, M., Spiel, C., & Schober, B. (2020). REFLECT – A teacher training program to promote gender equality in schools. Frontiers in Education, 5. https://doi.org/10.3389/feduc.2020.00136

[36]. Lee, S. M. (2021). Exploring gender-responsive pedagogy for STEM education. IISRR- International Journal of Research, 7(II). http://www.iisrr.in/mainsite/wp-content/uploads/2021/10/6.-Shok-Mee-LEE-Exploring-Gender-Responsive-Pedagogy-for-STEM-Education.pdf

[37]. Lee, J., Moon, S., & Hegar, R. L. (2011). Mathematics skills in early childhood: Exploring gender and ethnic patterns. Child Indicators Research, 4(3), 353-368.

[38]. Lundberg, S. (2020). Educational gender gaps. IZA DP No. 13630. www.iza.org

[39]. Makarova, E., Aeschlimann, B., & Herzog, W. (2019). The gender gap in STEM fields: The impact of the gender stereotype of math and science on secondary students' career aspirations. Frontiers in Education, 4(60). https://doi.org/10.3389/feduc.2019.00060

[40]. McDool, E., & Morris, D. (2022). Gender differences in science, technology, engineering and maths uptake and attainment in post-16 education. Manchester School, 90(5), 473–499. doi:10.1111/manc.12403

[41]. Meinck, S., & Brese, F. (2019). Trends in gender gaps: Using 20 years of evidence from TIMSS. Large-scale Assessments in Education, 7(1). https://doi.org/10.1186/s40536-019-0076-3

[42]. Meyers, J. L., Miller, G. E., & Way, W. D. (2009). Item position and item difficulty change in an IRT based common item equating design. Applied Measurement in Education, 22(1), 38-60. https://doi.org/10.1080/08957340802558342

[43]. National Center for Education Statistics. (2009a). The Nation's Report Card: 2009 Grade 4 Sample Questions for Mathematics, Reading, and Science. U.S. Department of Education, Institute of Education Sciences: Washington, DC. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/demo_booklet/09SQ-O-G04-MRS.pdf

[44]. National Center for Education Statistics. (2009b). The Nation's Report Card: 2009 Grade 8 Sample Questions for Civics, Geography, U.S. History, Mathematics, Reading and Science Probe. U.S. Department of Education, Institute of Education Sciences: Washington, DC. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/demo_booklet/09SQ-G08-MRS.pdf

[45]. National Research Council. (2001). Knowing what students know: The science and design of educational assessment. National Academy Press, Washington, D.C.

[46]. National Research Council. (2007). Taking science to school: Learning and teaching science in grades K-8. National Academy Press, Washington, D.C.

[47]. Nishimura, M. (2017). Effect of school factors on gender gaps in learning opportunities in rural Senegal: Does school governance matter? JICA Research Institute, 141. https://jicari.repo.nii.ac.jp/index.php?action=repository_action_common_download&item_id=803&item_no=1&attribute_id=9&file_no=1&page_id=13&block_id=21

[48]. 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.

[49]. Penner, A. M., & Paret, M. (2008). Gender differences in mathematics achievement: Exploring the early grades and the extremes. Social Science Research, 37(1), 239-253.

[50]. Perlini, A. H., Lind, D. L., & Zumbo, B. D. (1998). Context effects on examinations: The effects of time, item order and item difficulty. Canadian Psychology/Psychologie Canadienne, 39(4), 299-307. https://doi.org/10.1037/h0086821

[51]. Qian, J. (2014). An investigation of position effects in large-scale writing assessments. Applied Psychological Measurement, 38(7), 518-534. https://doi.org/10.1177/014662161453431

[52]. Robinson, J. P., & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during

elementary and middle school: examining direct cognitive assessments and teacher ratings. American Educational Research Journal, 48(2), 268-302.

[53]. Rodrigues,S., Taylor, N., Cameron, M., Syme-Smith, L. &Fortuna, C.(2010). Questioning Chemistry: The role of level, familiarity, language and taxonomy. Science Education International, 21(1), 31-46.

[54]. Sohn, K. (2012). A new insight into the gender gap in math. Bulletin of Economic Research, 64(1), 135-155.

[55]. Stoet, G., & Geary, D. C. (2018). The gender-equality paradox in Science, technology, engineering, and mathematics education. Psychological Science, 29(4), 581–593. doi:10.1177/0956797617741719 PMID:29442575

[56]. Taylor, C. S., & Lee, Y. (2012). Gender difference in reading and mathematics tests with mixed item formats. Applied Measurement in Education, 25(3), 246-280.

[57]. Tal, I. R., Akers, K. G. & Hodge, K. G. (2008). Effect of Paper color and question order on exam performance. Teaching of Psychology, 35(1), 26-28. https://doi.org/10.1080/0098 6280701818482

[58]. Toraman, C., & Ozen, F. (2019). An investigation of the effectiveness of the gender equality course with a specific focus on faculties of education. Educational Policy Analysis and Strategic Research, 14(2), 6–28. https://doi.org/10.29329/epasr.2019.201.1

[59]. Trendtel, M., & Robitzsch, A. (2018). Modeling item position effects with a Bayesian item response model applied to PISA 2009–2015 data. Psychological Test and Assessment Modeling, 60(2), 241-263.

[60]. Tytler, R., & Osborne, J. (2012). Student attitudes and aspirations towards Science. In B. J. Fraser, K. G. Tobin, & C. J. McRobbie (Eds.), Second International Handbook of Science Education (pp. 597–625). Springer. doi:10.1007/978-1-4020-9041-7_41

[61]. UNESCO. (2017). Cracking the code: Girls' and women's education in science, technology, engineering and mathematics (STEM). UNESCO. https://unesd oc.unesco.org/image s/0025/00253 4/25347 9E.pdf

[62]. WAEC (1993). The effects of item position on performance in multiple choice tests. Research Report, Research Division, WAEC, Lagos.

[63]. Weirich, S., Hecht, M., & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. Applied Psychological Measurement, 38, 535-548. https://doi.org/ 10.1177/0146621614534955

[64]. Wigati, I. (2019). The social aspects of gender-responsiveness in schools. Sawwa: Jurnal Studi Gender, 14(2), 147–162. https://doi.org/10.21580/sa.v14i2.4523

[65]. Willingham, W. W., & Cole, N. S. (2013). Gender and fair assessment. Routledge. Xi, X. (2010). How do we go about investigating test fairness? Language Testing, 27(2), 147-170.