EDConference Proceedings 2022 ISBN: 978-86-87043-81-7

Using Mantel-Haenszel, Distractor Response, and Logical Data Analyses in Detecting Differential Item Functioning in a Senior Science High School Entrance Examination

Diana Lou E. Sipalay
University of the Philippines Diliman, College of Education
Diliman, Quezon City, NCR, Philippines
desipalay@up.edu.ph

Jose Q. Pedrajita
University of the Philippines Diliman, College of Education
Diliman, Quezon City, NCR, Philippines
jqpedrajita@up.edu.ph

ABSTRACT

This study detected Differential Item Functioning (DIF) and its causes (gender-based, curriculum-based and school-based) in a Senior Science High School Entrance Examination using Mantel Haenszel, Distractor Response and Logical Data Analyses employing focus group discussions, semi-structured interviews and questionnaires. The DIF detecting methods revealed the presence of gender, curriculum and school biases across subtests. The causes of DIF were difference in the span of focus and interest, test sophistication, use of jargon, availability of materials in school, activity exposure, curriculum difference and teacher quality. Items with DIF, regardless of the group membership of examinees, with unreasonable difficulty that unfavorably affected the test performance of the students, were recommended for replacement or revision; for students to be assessed properly using tests with reviewed, evaluated and improved items.

KEYWORDS: Mantel-Haenszel Procedure, Distractor Response Analysis, Logical Data Analysis, Entrance Examination, Science High School, Senior High School

1 INTRODUCTION

The surging issue of detecting differential item functioning in measurement has come to the forefront in the field of testing. Unfair assessment arises when a student's test performance is imprecise because it disadvantages the student because of his/her group membership. It is important to understand that it occurs when it is not the student's ability that causes low performance, but the student's group membership does. Unfairly penalizing a group of test takers due to their gender, race, ethnicity, socioeconomic status, religion or other such group defining characteristics adds to the unfairness of a particular assessment tool. Naturally, bias items display differential item functioning that underestimate or overestimate the value of variables the items are designed to measure (Salubayba, 2013) [1].

Questions of test bias are intertwined with the questions on the validity and reliability of a particular instrument. The first critical element to an effective assessment item is validity. It is important to detect biased items because they may result in a dubious differential performance for test takers of the same ability.

The removal or revision of potentially and identified biased items is required in all tests, may it be a wide scale, high stakes or a simple teacher made test used in the classroom. It is vital that before the administration of any assessment instrument, the biased items must be detected first and then either eliminated or revised (Pedrajita, 2007) [2].

Differential item functioning (DIF) analysis is one method to investigate biases in written tests at the item level. DIF is evident in a test item when despite the effort to control for overall test performance, test-takers from different majority and minority groups have a different probability of answering an item correctly or when test takers from two subgroups with the same trait level have different expected scores on the same item (Camilli & Sheppard, 1994) [3]. DIF Detection involves comparison of the performance of matched majority (or reference) and minority (or focal) group examinees. Thus, an item that exhibits DIF may or may not be biased for or against any group (Kanjee, 2007) [4].

Entrance Examinations are conducted in educational institutions to select and sift prospective students for admission. Given that test fairness is related to the interpretations and uses of test scores as well as the claims made from those interpretations and uses; it is critical to obtain and weigh validity evidence to support or refute the score interpretations, their uses, and the potential socio-political consequences in order to evaluate fairness (Banerjee, 2016) [5].

The study generally aimed to produce a fairer – more reliable and valid Entrance Examination for Senior High Students of a selected Science High School. Specifically, it aimed to detect the presence of large DIF in the items of the examination and determined the causes of DIF.

The use and misuse of high stakes tests are a controversial topic in public education, especially in the Philippines, where they have become especially popular in recent years because of its use to not only assess students but attempts to increase the quality of education through fair testing (Maca & Morris, 2014) [6].

Bias comes in many forms which can be evident in gender, cultural, ethnic, religious, or class. The origin or cause of item bias can be identified using Logical Data Analysis (LDA). Curriculum bias is affected by the content of the curriculum that is taught in science and non-science high schools. The extent to which curriculum may be biased for or against examinees from science or non-science institutions has little to never been the focus of empirical researches.

Furthermore, potential sources of such bias might be group differences in examinees, such as differences in subject inclination, subjects taken during prior education and schema being trained as a science or a non-science high school student (Kunnan, 2004) [7].

The type of school from where the test takers are from significantly affects the knowledge of the examinees in terms of content. Therefore, school bias with regards to school type (public versus private) occurs when the knowledge on a specific content of a test of the examinee is comparatively more for one group of students than for others.

Gender biased tests are those that favor one sex or one gender group. This type of test is specifically prohibited because of the discrimination it produces among males and females (Einarsdóttir and Rounds, 2009) [8].

Various aspects of fairness including fairness with respect to standardization, test score use and item bias have been the center of attention. However, the DIF concept (Perrone, 2006) [9] and detection developed by the Educational Testing Service (ETS) in 1986, is the standard of psychometric bias analysis.

Accordingly, DIF which may reflect measurement bias has received a great deal of attention in educational measurement. There is no single best method of DIF analysis which is effective and useful for all purposes (Millsap & Everson, 1993; as cited in Van den Noortgate and Boeck, 2005) [10].

2 RESULTS

The validity and reliability of the test were established before the detection of DIF. Validity ranged from slightly adequate to adequate in the six main disciplines represented by each subtest found in the test.

The result of the Cronbach coefficient alpha indicated that the Senior High School Entrance Examination is reliable in measuring the examinees' achievement and academic knowledge on topics needed for senior high school STEM strand.

The DIF detecting methods revealed the presence of gender, curriculum and school bias across subtests. The causes of DIF were noted as difference in the span of focus and interest, test sophistication, use of jargon, availability of materials in school, activity exposure, curriculum difference and teacher quality.

The results of both statistical procedures and the logical data analysis showed that there are more items in the examination that displayed gender-based DIF against female examinees, curriculum-based DIF against Non-Science high school examinees and school-based DIF against Private school examinees.

Items with unreasonable difficulty that unfavorably affected the test performance of the students were recommended for replacement or revision to the school test committee to ensure that the students would be assessed properly using only a test with reviewed, evaluated and improved items.

3 DISCUSSION

Item bias is attributable to the degree of item validity. Test bias is a major threat against validity and therefore test bias analyses should be made before the administration of the test itself. If a test has poor validity, there is no justification for using the test results for their intended purpose (Fraenkel and Wallen, 1994) [11].

Content validity is a logical process where connections between the test items and the subject matter related tasks are established. Bias in content validity is evidenced either by items that ask information that disadvantaged students have not had equal opportunity to learn or by wording of the question is unfamiliar and a disadvantaged student who may "know" the answer is unable to respond because he/she did not understand the question.

The test's validity was estimated by gathering a group of subject matter experts (SMEs) to review the test items. The content validity of the Senior High School Entrance

Examination was based on whether the items adequately and properly sampled the universe of items that probe understanding of the content domains. The scale below, adapted from the study of Pedrajita (2007) was used in quantifying the adequateness of the items in the subtest.

Table 1 Scale for Quantifying Adequateness of Items for Content Validation

Value	Implication
86%-100%	Adequate
71%-85%	Moderately Adequate
56%-70%	Slightly Adequate
41%-55%	Slightly Inadequate
26%-40%	Moderately Inadequate
Below 25%	Inadequate

Internal consistency is typically a measure based on the correlations between different items on the same test. It measures whether several items that propose to measure the same general construct produce similar scores. It is usually measured with Cronbach's alpha coefficient. Nunnally and Bernstein (1994) [12] published the interpretation of the Cronbach's Alpha value, supported by the work of Cronbach in 1951.

Table 2 Interpretation of the Cronbach's Alpha Value

Cronbach's Alpha	Internal Consistency
$\alpha \geq 0.9$	Excellent
$0.9 > \alpha \geq 0.8$	Good
$0.8 > \alpha \geq 0.7$	Acceptable
$0.7 > \alpha \geq 0.6$	Questionable
$0.6 > \alpha \geq 0.5$	Poor
$0.5 > \alpha$	Unacceptable

The test was considered a valid measurement of achievement in the six main disciplines. In terms of content validity, the Senior High School Entrance Examination was moderately adequate in representing the universe of content standards needed to probe understanding of the content domain for Senior High School STEM strand.

Table 3 Overall Validity of the Senior High School Entrance Examination

Subtest	Number of Basic Content Standards	Actual Number of Basic Content Standards in the Examination
Language Proficiency	9	6
Mathematics Proficiency	10	7
Science Process Skills	5	3
Science Proficiency		
Biology	5	5
Chemistry	9	8
Physics	19	13
Earth Science	3	3
Mechanical and Spatial Skills	1	1
ICT Skills	1	1
Total	**62**	**47**
Remarks	**76% - moderately adequate**	

The Cronbach's Alpha coefficient values of Language and Mathematics Proficiency showed acceptable reliability for the two subtests while the alpha for the Science Process Skills subtest meant a questionable reliability. Mechanical and Spatial subtest and ICT subtest got alpha values that connoted poor reliability while all subparts under the Science

Proficiency subtest got unacceptable reliability. The overall Cronbach's Alpha Coefficient that was computed for the Entrance Examination is 0.899, indicating a good reliability.

Table 4 Test for Reliability – Cronbach Alpha Coefficient Values per Subtest

Subtest	Number of Items (N)	Cronbach's Alpha Coefficient
Language Proficiency	30	0.711
Mathematics Proficiency	40	0.723
Science Process Skills	30	0.682
Science Proficiency		
Biology	15	0.368
Chemistry	15	0.435
Physics	15	0.327
Earth Science	15	0.403
Mechanical and Spatial Skills	20	0.596
ICT Skills	20	0.589

The Mantel-Haenszel procedure identified a total of 25 items showing severe gender-based DIF. Six items showed large DIF in favor of female examinees and 19 items in favor of male examinees. This indicates that the items mentioned were more difficult to answer if an examinee belongs to the disadvantaged group.

Table 5 Identified Potentially Biased Items Using Mantel-Haenszel Procedure

Subgroups (Reference and Focal)	Potentially Biased Items Against	Number of Items
Gender	Male Examinees	6
	Female Examinees	19
School	Public Examinees	6
	Private Examinees	68
Curriculum	Science HS Examinees	8
	Non-Science HS Examinees	92

The procedure also identified a total of 100 items showing severe curriculum-based DIF. Eight (8) items showed large DIF in favor of the Non-Science High School examinees and 92 items in favor of Science High School examinees. This indicates that the items mentioned were more difficult to answer if an examinee belongs to the disadvantaged group. There were 74 items identified showing severe school-based DIF. Six (6) items showed large DIF in favor of the Private School examinees and 68 items in favor of Public School examinees.

On the other hand, using Distractor Response Analysis, there were 21 recorded items in the whole examination that showed DIF between male and female examinees. All 21 items were found potentially biased against female examinees with Item 4 under Earth Science as an exception, having a distractor A, as potentially biased against male examinees.

Examining the curriculum-based, the results of DR analysis revealed that there were 78 items that were potentially biased against a particular subgroup and there were 59 items flagged with school-based DIF found in the Entrance Examination using the DRA procedure. Ten (10) items showed DIF against the reference group and 49 showed DIF in favor of the reference group.

Table 6 Identified Potentially Biased Items Using Distractor Response Analysis

Subgroups (Reference and Focal)	Potentially Biased Items Against	Number of Items
Gender	Male Examinees	1
	Female Examinees	21
School	Public Examinees	11
	Private Examinees	51
Curriculum	Science HS Examinees	15
	Non-Science HS Examinees	69

Review of the potentially biased items in this study was adapted from the practice stated by Reynolds (2006) [13] in his research where he stated that LDA involves determining whether the flagged item is (a) easier for the reference than the focal group; (b) easier for the focal than the reference group or (c) of equal difficulty among the comparison groups. Only the DIF displaying items were tackled in the interviews and focus group discussions and not the DIF-free or items with moderate to negligible DIF. Part of the procedure was to know if there are items that are construct irrelevant and was pinpointed by the subject expert and item writers in the given questionnaire.

All items identified as displaying DIF under MH and DRA procedures were considered potentially biased and were discussed in the FGDs, unstructured interviews and pinpointed in the questionnaires given to the item writers. The examinees pinpointed the difficulties they encountered and the techniques they often use to answer the item/s despite the difficulties. They were also asked as to why they found a particular item, specifically the DIF items, difficult for them to answer.

Table 7 Themed Causes of Item Bias based on Logical Data Analysis Procedure

Evaluated Biases	Causes of Item Bias
Gender-based DIF	Span of Focus
	Interest
Curriculum-based DIF	Test Sophistication Training
	Use of Unfamiliar Words
	Use of Available Materials
	Curriculum
	Type of Activities
School-based DIF	Level of Exposure and Experiences
	Teacher Factor

The LDA revealed that in the gender-based DIF, the common reasons for the biases are the difference in the span of focus between male and female students and the difference in their interests. On the other hand, the common reasons for curriculum-based DIF were gathered as their readiness for any kind of test or test sophistication, use of unfamiliar words,

use of available materials in the school, types of activities they are exposed to and lastly, the curriculum difference itself. The causes of DIF in school type bias were the level of exposure and experiences of the students and the quality of teachers that they have in school.

4 METHODOLOGY

Descriptive research design was employed to determine items exhibiting DIF, to know the underlying causes of DIF in the entrance examination. The researcher considered non manipulative variables and establishes a formal procedure, in this study, to detect DIF and determine the causes of DIF. The reliability and validity of the entrance examination was computed then three matched groups in terms of curriculum, school and gender were used for Mantel-Haenszel statistic, Distractor Response Analysis and Logical Data Analysis for comparison. The groups were compared in terms of their probability of success on each item on all the subtests in the entrance examination. The comparison groups were as follows: science high school and non-science high school completer-examinees, public and private completer-examinees and male and female completer-examinees.

Table 8 Descriptive Research Design

Matched Groups	Reference Group	Focal Group	Intervening Variables	Dependent Variables
Curriculum	Science HS	Non-Science HS	1. DIF Analysis -Mantel-Haenszel Statistic - Distractor Response Analysis 2. Logical Data Analysis 3. Content Validity 4. Internal Consistency Reliability	Curriculum-based DIF
School	Public	Private		School-based DIF
Gender	Male	Female		Gender-based DIF

The 200-item Entrance Examination was subdivided into six main skills that an incoming Senior High School that is STEM inclined. No field/pilot testing was conducted and the reliability and validity of the test were not established prior to the administration of the examination. There is no table of specifications for the test.

Table 9 Six Subtests Under the Senior Science High School Entrance Examination

Subtest	Number of Items
Language Proficiency	30
Math Proficiency	40
Science Process Skills	30
Science Proficiency	60
Mechanical and Spatial	20
Information, Communication and Technology	20
Total	**200**

The researcher conducted a cognitive interview with the available item writers to examine the content of the examination against the prescribed syllabi for the subtests found in the examination. Focus group discussion and interview with the stakeholders – the administration of the school, the testing committee, the test item writers and the examinees who took the test were conducted.

The scores of the examinees in the test served as the primary data for the study. The test papers and results were gathered through the registrar and admission committee of the school for the quantitative study.

Content validity and internal consistency reliability were established with the help of subject matter experts (SMEs) and Cronbach's Alpha Coefficient, respectively. DIF detecting procedures were employed to detect gender-based DIF, curriculum-based DIF and school-based DIF. Mantel-Haenszel Statistic and Distractor Response Analysis were the methods used to detect items exhibiting DIF. In the Mantel-Haenszel procedure, the basis for flagging the item as potentially bias was the value of delta-MH. The classification used by the Educational Testing Service (ETS) was used in tagging the items as negligible, moderate and exhibiting large DIF. Only the items with large DIF were flagged as potentially bias to a specific matched group. In the Distractor Response Analysis, the chi square for each distractor was computed and significant values were evaluated and flagged as potentially bias against a specific matched group.

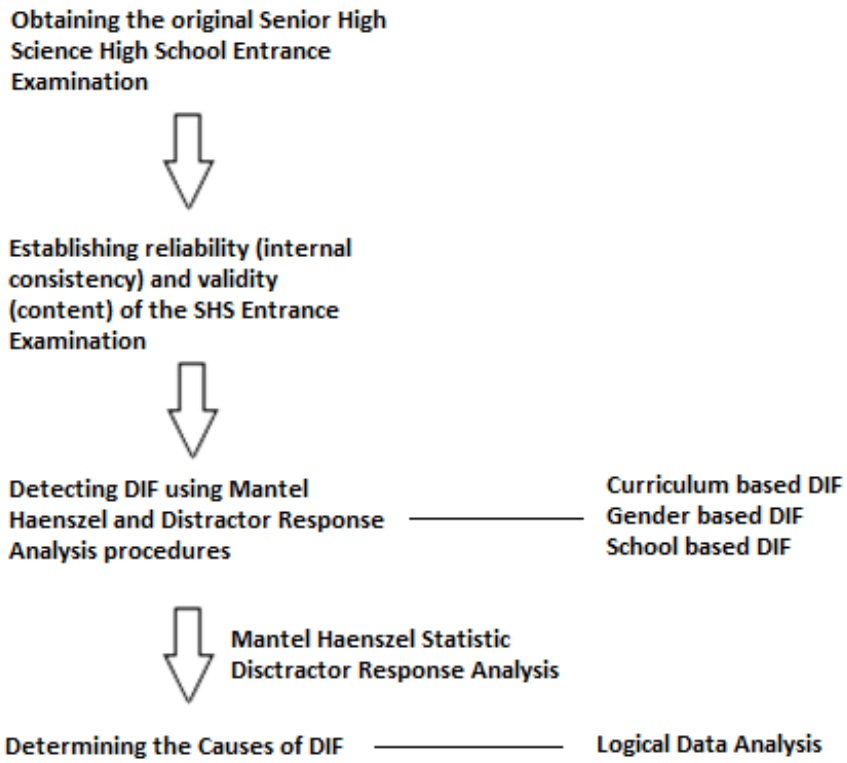

Figure 1: Methodological Flowchart

For the qualitative part, which is the Logical Data Analysis, the subject matter experts and item writers were interviewed and focus group discussions were conducted in accordance with the comparison groups to know the causes of DIF.

The analysis of data included (1) validity and reliability procedures, specifically, content validity and internal consistency reliability; (2) DIF analysis procedures such as Mantel-Haenszel statistic and Distractor Response Analysis for the quantitative part of the study; and (3) knowing the causes why the items are exhibiting DIF using the qualitative method, Logical Data Analysis.

Distractor analysis involves the calculation of bivariate frequency distributions for unscored items as categorical variables. The chi square for each distractor was computed and significant values were evaluated and items containing DIF exhibiting distractors were flagged as potentially biased items against a specific matched group.

Another DIF analysis procedure is Mantel-Haenszel statistic. The MH DIF statistic defines DIF in terms of the ratio of the odds of a correct response on the studied item for the reference group to the odds of a correct response for the focal group.

The Mantel-Haenszel Statistic computed for each item in all the subtests was further classified into items with negligible, moderate and large DIF according to the classification used by the ETS to specify the degree of DIF. In this study, only the items which fell on "C" category or large DIF, were evaluated.

Table 10 Summary of Selected Item Bias Detection Models

Item Bias Model	Focus of Analysis	Measure of Bias
Distractor Response Analysis	Difference in proportions selecting distractors	Significance of Chi-square
Mantel-Haenszel	Perform statistical test for evaluating the amount of DIF	Significance of Chi-square and large DIF Effect (C items)

Items that are flagged with DIF were not automatically revised or discarded. Being flagged with DIF means the item is potentially biased but this empirical evidence of differential performance between groups is not enough to reject the item/s out of the Entrance Examination. Instead, further acquisition of logical reasons as to why it displayed DIF was further investigated using Logical Data Analysis.

5 CONCLUSION

In terms of content validity, the Senior High School Entrance Examination items were moderately adequate (76%) in representing the universe of content standards needed to probe understanding of the content domain for Senior High School STEM strand. The alpha value indicated that the entrance Examination as a whole has good reliability. After careful considerations, all of the null hypotheses in the MH procedure and DRA were rejected, in favor of the alternative hypotheses, in all subtest.

The causes of item bias based on Logical Data Analysis (LDA) procedure revealed that in the gender-based DIF, the common reasons for the biases are the difference in the span of focus between male and female students and the difference in their interests. On the other hand, the common reasons for curriculum-based DIF were gathered as their readiness for any kind of test or test sophistication, use of unfamiliar words, use of available materials in the school, types of activities they are exposed to and lastly, the curriculum difference itself. The cause of DIF in school type bias were the level of exposure and experiences of the students and the quality of teachers that they have in school.

The results of the DIF detecting procedures, after logical analysis, revealed that there are more items in the examination that displayed gender-based DIF against female examinees, curriculum-based DIF against Non-Science high school examinees and school-based DIF against Private school examinees.

The presence of DIF displaying items directly affected the validity of the examination. The three subtests with the highest number of items flagged as potentially bias across matched groups were Language Proficiency, Mathematics Proficiency and Science Process Skills. The validity of the three subtests were all slightly adequate with 60% to 70% accuracy. The reliability of Language Proficiency and Mathematics Proficiency were both acceptable except Science Process Skills with a questionable reliability value.

The high adequacy of the subtests was supported by the results attained by the DIF detection procedures. The causes of DIF were more on the difference of interest between males and females, subject matter schema between Science and Non-Science high school examinees and availability of materials for ICT between Public and Private school examinees.

The examination's overall reliability was computed to be good but majority of the subtests have unacceptable to questionable reliability and only two subtests got an acceptable reliability value. Although the overall alpha computed was good, following the test development procedure and increasing the number of items can actually improve the examination itself.

The overall validity of the examination was computed to be moderately adequate. Out of 62 learning competencies that the examination should contain, 47 were actually embedded in the items of the test. A moderately adequate validity is already acceptable in statistical terms but aiming for a higher validity can increase the truthfulness of the assessment in evaluating and assessing the test takers' ability. Detecting items that are potentially biased against a particular subgroup and finding the causes of DIF through logical analysis could help in eliminating biased items and increasing the validity of the test.

FUNDING: This research received no external funding.

CONFLICT OF INTEREST: Authors declare no conflict of interest.

REFERENCES

[1] Salubayba, T. (2013). Determining Differential Item Functioning in an Achievement Test Using Mantel-Haenszel, Item Response Theory and Logical Data Analysis, pp. 1-3.

[2] Pedrajita, J. (2007). Identifying Biased Test Items by Differential Item Functioning Analysis Using Contingency Table Approaches: A Comparative Study. Education Quarterly. 67 (1).

[3] Camilli, G. & Sheppard, L. (1994). Methods for Identifying Biased Test Items. London: SAGE

[4] Kanjee, A. (2007). Using logistic regression to detect bias when multiple groups are tested. South African Journal of Psychology, 37(1), 47-61.

[5] Banerjee, H. L. (2016). Test fairness in second language assessment. Studies in Applied Linguistics and TESOL, 16(1).

[6] Maca, M. & Morris, P. (2014). Education, National Identity and State Formation in the Modern Philippines. Constructing Modern Asian Citizenship, 5, 125.

[7] Kunnan, A. J. (2004). Test fairness. European language testing in a global context, 27-48.

[8] Einarsdóttir, S. & Rounds, J. (2009). Gender bias and construct validity in vocational interest measurement: Differential item functioning in the Strong Interest Inventory

[9] Perrone, M. (2006). Differential item functioning and item bias: Critical considerations in test fairness. Studies in Applied Linguistics and TESOL, 6(2).

[10] Van den Noortgate, W., & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. Journal of Educational and Behavioral Statistics

[11] Fraenkel, J. and Wallen, N. (1994). How to Design and Evaluate Research in Education. 2nd Edition, McGraw-Hill Inc.

[12] Nunnally, J. C., & Bernstein, I. H. (1994). Psychological Theory.

[13] Reynolds, R.C., et.al. (2006). Measurement and Assessment in Education. Boston: Pearson