Title page

Post exam analysis: Implication for intervention

By

Tsegaye, Kindu Nibret, Biology unit, Gondar CTE, Gondar Ethiopia

Corresponding author: Tsegaye, Kindu Nibret, e-mail: kindutsegaye21@gmail.com

Abstract: ***The difficulty index, discrimination and distracter efficiency of college level exam paper was analyzed as an input for taking actions in future test developments. The exam papers of 176 first-year regular pre-service diploma students at Gondar CTE were analyzed using descriptive analysis. Difficulty indices and distracter efficiencies were calculated using Microsoft Excel 2007. Other test statistics such as mean, bi-serial correlations and reliability coefficients were computed using SPSS version 20. Results showed that the mean test score, out of 31, was 17.23 ± 3.85. Average difficulty and discrimination indices were 0.56 (SD 0.20) and 0.16 (SD 0.28), respectively. The mean distracter efficiency was 92.1% (SD 17.2%). The reliability of the test was 0.58. Out of 31, 13 (41.9%) items were either too easy or too difficult. Only two items fell into good or excellent discrimination power. Inconsistency in option formats and inappropriate stems were observed in the exam paper. Based on the results the college level exam paper has acceptable level of difficulty index and distracter efficiency. However, the average discrimination power of exam was very low (0.16, acceptable ≥ 0.4). The internal consistency reliability was also less than the acceptable level (0.58, acceptable ≥ 0.7). Thus, future test development interventions should give due emphasis on item reliability, discrimination coefficients and item face validity.***

1. Background

The rapid expansion of higher education institutions in Ethiopia seem to compromised quality of education in the country (4). According to Arega Yirdaw (2016), problems related to the teaching-learning process stood first as a key factor that determine quality education in private higher institutions in Ethiopia. To this end, an effective assessment tool, among others, had to be in place to see if the required outcomes could be achieved.

Higher education institutes need to combine different approaches and instruments for assessing students (5). This is because students' assessment and evaluation are an integral part of the teaching - learning process (2). The assessments should be relevant while tracking each student's performance in a given course. Considering this, instructors at higher institutions must be aware of the quality and reliability of tests. Otherwise, the final results may be influenced by the test itself, which could lead to a biased assessment (5). Usually instructors receive little or no training on assessment quality. If training given, it doesn't focus on strategies to construct test or item-writing rules but only on large-scale test administration and standardized test score interpretation (2). Tavakol and Dennick (2011) pointed out the importance of post- exam item analysis to improve the quality and reliability of assessments.

Item-analysis is the process of collecting, summarizing and using information from students' responses to assess the quality of test items (21). It allows teachers to identify: too difficult or too easy items, items that do not discriminate high and low able students or items that have implausible distracters (2, 3). In these cases, teachers can remove too easy or too difficult items or non-discriminating items. Item analysis also help teachers modify instruction to correct any misunderstanding about the content or adjust the way they teach (2).

A number of reports on Ethiopian education quality indicated that there was a serious problem in quality of education (4, 19). Assessment for grading students' achievement in the Ethiopian school system is mainly exercised through the administration of teacher made classroom tests and national examinations (20). It is thought that the exclusive reliance on multiple-choice questions for school and national examinations may be causing a negative back-wash effect on

55 education quality (20). Objective examination results can be analyzed to improve the validity and
56 reliability of assessments. Post-exam analysis is one intervention to improve the quality and
57 reliability of assessments (17). As far as the knowledge of the researcher is concerned, no item
58 analysis was conducted at Gondar CTE. Hence the objective of this study was to improve the
59 skills of college instructors to systematically use standardized and validated student assessments
60 with the autonomy of the department.
61

62 2. Methods

63 2.1 Research Design

64 The validity and reliability of a summative test in basic natural science course was assessed
65 using descriptive analytical method. Item difficulties, discrimination coefficient, reliability, face
66 validity and distracter efficiency for multiple choice questions (MCQ) were calculated.
67

68 2.2 Study population

69 All regular first year diploma students at Gondar CTE during 2017/18 academic year were taken
70 as the study population.
71

72 2.3 Sample size and sampling technique

73 A total of 176 (33.5%) students were selected using stratified random sampling. A stratified
74 sampling technique was employed to include representative samples from each stream. The
75 sample exam papers were collected from science instructors in the department, Natural Science,
76 Gondar CTE.
77

2.4 Instrument and scoring

The summative exam paper in the course 'basic natural science' was used as the research instrument. The exam paper comprised of 31 objective items of which 21 are multiple choice, 7 true/false and 3 matching items. All the objective test items were considered for analysis. For item analysis, results of all papers were coded as 1 for right and 0 for wrong responses. The maximum mark possible to score was 31 and minimum zero, with no negative marking.

2.5 Face validity

The test paper was reviewed for the following face validity parameters.

- typing and punctuation errors
- inappropriate/incomplete stems
- Inappropriate options formats/alternatives format for MCQs.

2.6 Internal consistency reliability

The internal consistency reliability of the exam paper was determined as it was considered to be the most relevant and accurate method for determining test reliability. The acceptable range of value for test reliability in most literatures is $\alpha \geq 0.7$. KR-20 was recommended to determine the internal consistency of a dichotomous item (17). Objective test items can dichotomously be scored as either right or wrong (17). In this study, the Kuder-Richardson method was employed to estimate test reliability. A KR-20 value of 0.7 or greater was considered as reliable.

2.7 Item Difficulty Index (p)

The item difficulty statistic is an appropriate choice for achievement tests when the items are scored dichotomously (i.e. correct vs. incorrect). Thus, it can be calculated for true-false,

multiple choice and matching items. Difficulty index was computed simply by dividing the number of test takers who answered the item correctly by the total number of students who answered the item (correct + incorrect). Its value ranges from 0 - 1; the higher the value, the easier the item and vice versa. The recommended range of difficulty level is between 0.3 – 0.7 (1 and 6). Items having *p*-values below 30% and above 70% are considered too difficult and too easy respectively (1).

Item Difficulty Index (*p*)	Item Evaluation
p > 0.7	Too easy
p = 0.3 – 0.7	Acceptable
p < 0.30	Too difficult

Source: Instructional Assessment Resources (IAR 2011) in (1)

2.8 Discrimination coefficient (*r*)

The item discrimination index is a value of how well a question is able to differentiate between students who are high performing and those who are not (17). It can be calculated either by extreme group method or point bi-serial correlation coefficient (*r*). The extreme group method considers only 54% of the respondents (top 27% and bottom 27%). On the other hand, the point bi-serial correlation coefficient indicates the relationship between a particular question (correct or incorrect) on a test and the total tests score (12, 17). For this reason, the point bi-serial correlation coefficient was calculated in this study. The point bi-serial correlation is computed to determine the relationship between student's performance in each item and their overall exam scores (12). It was computed using SPSS version 20 and its value ranges between -1 and 1; a

119 higher value indicates a powerful discrimination power of the item. The test items in this study
120 were classified as in the table below.

r value	Quality	Recommendations
≥ 0.4	Excellent	Retain
0.3 – 0.39	Good	Possibilities for improvement
0.2 – 0.29	Average	Need to check/review
0.0 – 0.19	Poor	Discard or review in depth
< - 0.01	Worst	Definitely discard

121 Adopted from (7)

122

123 ## 2.9 Distracter efficiency (DE)

124 Distracters are classified as the incorrect answers in a multiple-choice question. Student
125 performance in an exam is very much influenced by the quality of the given distracters. Hence, it
126 is necessary to determine the effectiveness of the distracters in a given MCQ. Distracter
127 effectiveness for the option indicates percentage of students choosing the option of item as an
128 answer. It was calculated based on the number of non-functional distracters (NFDs) per item. An
129 NFD was defined as an incorrect option in MCQ selected by less than 5% of students. The DE
130 was considered to be 0%, 33.3%, 66.7% or 100% if an item had three, two, one or zero NFDs,
131 respectively.

132

133 ## 3 Data analysis

134 Data was stored and analyzed using SPSS version-20 for discrimination coefficient and
135 reliability and Microsoft Excel 2007 for difficulty index and DE. Mean and standard deviations

were also computed for difficulty index, discrimination and reliability measures. Count and percentages for distracter efficiency were displayed in tables. Face validity was qualitatively described and graphs and tables were used to display results.

4 Results

Descriptive statistics: The scores of 176 students ranged from 5 to 27 out of 31. The mean test score was 17.23 ± 3.85 (Table 1). The median score is 17.0, slightly lower than the mean score. The skewness and kurtosis values for the scores were − 0.020 and − 0.137, respectively. The values for asymmetry and kurtosis between −2 and +2 are considered acceptable in order to prove normal univariate distribution. Figure 1 illustrates a fairly symmetrical distribution of the total score.

Table 1: Item descriptive statistics

Mean	**17.23**
Std. Error of Mean	0.29
Median	17.00
Mode	**15**.00[a]
Std. Deviation	3.85
Skewness	-0.020
Std. Error of Skewness	0.183
Kurtosis	-0.137
Std. Error of Kurtosis	0.364
Range	22.00
Minimum	5.00
Maximum	27.00
a. Multiple modes exist. The smallest value is shown	

Face validity: Results from face validity revealed the following findings.

- Punctuation missing (full stop & question mark missing) in question No.5, 6, 7, 11 and 15.
- Option format inconsistency in MCQs were observed in question No from 26 to 31 (option letters changed).
- Inappropriate stems (stem not meaningful or incomplete) were observed in question No. 12, 13, 14, 27 and 28.
- Inappropriate options/alternatives (all of the above, A and B) – occurred in question No. 12, 13, 15, 22, 27 and 28.
- No negatively phrased stems (*not* or *except*).
- Absolute terms, in question number 4 (most).

Internal consistency reliability

The performance of the test as a whole was evaluated based on the internal consistency reliability. The computed KR-20 of the comprehensive test was 0.58. This value is less than the recommended range in many literatures (≥ 0.7).

Difficulty index: Appendix A shows the distribution of difficulty indices (p) for each item. One item (q.19) has the highest p-value (0.82) and item number **18** has the lowest (0.15). Eighteen (58.1%) items have moderate difficulty level (p-value between 0.3 – 0.7). Twelve items (38.7%) have excellent difficulty level (p-values between 0.4 – 0.6). Thirteen (41.9%) items lie outside the moderate difficulty range (0.3 – 0.7) i.e. three items were too difficult (p-value < 0.3) and ten

(32.3%) items too easy (*p*-value > 0.7). The mean difficulty index was 56% that is $p = 0.56$, SD. 0.20. A summary of difficulty index was illustrated in Table 2.

Table 2: Difficulty index summary

p value	Interpretation	# of items	Action
< 0.3	Difficult	3 (9.6%)	Discard
0.3 – 0.7	Moderate	18 (58.1%)	Accept
> 0.70	Easy	10 (32.3%)	Reject

Item Difficulty Index Mean 0.56, SD. 0.20

Discrimination coefficients: Appendix B shows the result of the point bi-serial correlation coefficient for each item. Three items (**q.2**, **q.18** and **q.20**) have negative discrimination powers (*r* - worst). Only a single item (**q.31**) has excellent discrimination power ($r > 0.4$). Seventeen items (54.8%) were categorized as poor ($r < 2.0$) and nine items (29%) as average ($r = 2.0 – 2.9$) (Table 3). Question number **7** is an ideal item in terms of difficulty level ($p = 0.54$, Appendix A), but good in terms of discrimination ($r = 0.39$, Appendix B). The mean discrimination power is **0.16** (SD. 0.28). In many literatures, the acceptable mean *r* value is ≥ 0.4.

Table 3: Distribution of items in terms of level of discrimination

Point bi-serial Correlation (*r*)	**No. of Items**	**%**	**Action**
Excellent (above 0.40)	1	3.23%	Retain
Good (0.30 – 0.39)	1	3.23%	Possibilities for improvement
Average (0.20-0.29)	9	29.03%	Usually need and subject to improvement

Poor (Below 0.19)	17	54.84	Discard or review in depth
Worst (< - 0.01)	3	9.67%	Definitely discard

Item Discrimination Coefficient Mean 0.16, SD, 0.28.

Table 4 displays a combination of the two indices i.e. item difficulty and discrimination. According to table 4, two (6.5%) items (**q.7** and **q.31**) have good *p*-values ranging from 0.3 to 0.7 and $r \geq 0.3$. However, if only items with excellent *p*-value (0.4 – 0.6) and excellent *r* (≥0.4) considered, there is no any single item which could be labeled as "**excellent**" (Table 4). Furthermore, easy items ($p > 0.7$) such as **q.3**, **q.5**, **q.10** and **q.24** have poor discrimination ($r < 0.3$). Difficult items ($p < 0.3$) such as **q.12**, **q.14** and **q.18** also have very low discriminating power ($r < 0.2$). The difficulty level for **q.2** was ideal ($p = 0.51$) but its discrimination power was negative ($r = -0.004$). Ideal questions with *p*-values from 0.4 to 0.6 (q.6, q.16, q.21, q.23 and q.26) have poor *r*- values (< 0.2). There is no statistically significant correlation between difficulty index and discrimination coefficient (Pearson correlation, Sig. (2-tailed) p = 0.279).

Table 4: Item difficulty (*p*) and discrimination (*d*) indices

item	*p*	*d*	item	*p*	*d*
q1	0.71	0.208	q17	0.64	0.017
q2	0.51	-0.004**	q18	0.15*	-0.046**
q3	0.74	0.082	q19	0.82	0.202
q4	0.67	0.211	q20	0.3	-0.02**
q5	0.78	0.16	q21	0.56	0.166
q6	0.44	0.166	q22	0.8	0.207

q7	0.54	0.393	q23	0.53	0.186
q8	0.71	0.12	q24	0.81*	0.143*
q9	0.35	0.013	q25	0.7	0.061
q10	0.78	0.186	q26	0.51	0.065
q11	0.6	0.253	q27	0.53	0.234
q12	0.19*	0.184*	q28	0.75	0.159
q13	0.53	0.237	q29	0.74	0.162
q14	0.2*	0.025*	q30	0.45	0.253
q15	0.38	0.203	q31	0.34	0.404
q16	0.47	0.173			

Figure 2 shows the graphical representation of difficulty index and discrimination power. The scatter plot allows identification of appropriate (valid and reliable) questions at the center of the plan. Moreover the representation is useful to notice immediately questions that are too easy or too difficult. According to figure 2, *r* increases up to a point where *p* approaches to 0.4, then after declines.

Distracter efficiency: The distracter analysis shows that four (12.9%) items (**q.12**, **q**.**13**, **q**.**19** and **q**.**28**) have five NFDs, with a choice frequency of < **5**%. All other items do not have any NFDs (Appendix C). In addition, four items (**q.12**, **q**.**14**, **q**.**18** and **q**.**20**) have distracters selected by more students than the (key) correct answer. There are no items with three NFDs. Only one item (**q.13**) has two NFDs (Table 5). The overall mean of DE was 92.1% with minimum 33.3% and maximum 100%.

212 **Table 5: Distracter analysis (DE) summary**

Number of Items	21
Total Distracters	63
Functional distracters	58 (92.1%)
Non functional distracters (NFDs)	5 (7.9%)
Items with 3 NFDs (DE=0%)	0
Items with 2 NFDs (DE=33.3%)	1
Items with 1 NFD (DE=66.7%)	3
Items with 0 NFD (DE=100%)	16
Items with over distracters	4
Overall mean DE	92.1±17.2%

213

214 Difficult items such as **q.12, q.14** and **q.18** have DE between 66.7% and 100%. Similar result
215 was recorded for easy questions such as **q.18, q.22** and **q.24**. Some items with poor or good *p*-
216 values have similar **DE** values (Table 6). Only a single item (**q.31**) satisfies all the three
217 parameters of ideal questions ($p > 0.3$, $r > 0.4$ and DE = 100%) (Table7).

218 **Table 6: Comparison of item difficulty with distracter efficiency**

Item	*p*	DE (%)
q11	0.6	100
q12	0.19*	66.7
q13	0.53	33.3
q14	0.2*	100

q15	0.38	100
q16	0.47	100
q17	0.64	100
q18	0.15*	100
q19	0.82	66.7
q20	0.3	100
q21	0.56	100
q22	0.8	100
q23	0.53	100
q24	0.81*	100
q25	0.7	100
q26	0.51	100
q27	0.53	100
q28	0.75	66.7
q29	0.74	100
q30	0.45	100
q31	0.34	100

219 **Table 7: Comparison of *p*, *d* and DE**

item	*p*	*d*	DE
q11	**0.6**	**0.253**	**100**
q12	**0.19***	**0.184***	**66.7**
q13	**0.53**	**0.237**	**33.3**
q14	**0.2***	**0.025***	**100**

q15	0.38	0.203	100
q16	0.47	0.173	100
q17	0.64	0.017	100
q18	0.15*	-0.046**	100
q19	0.82	0.202	66.7
q20	0.3	-0.02**	100
q21	0.56	0.166	100
q22	0.8	0.207	100
q23	0.53	0.186	100
q24	0.81*	0.143*	100
q25	0.7	0.061	100
q26	0.51	0.065	100
q27	0.53	0.234	100
q28	0.75	0.159	66.7
q29	0.74	0.162	100
q30	0.45	0.253	100
q31	0.34	0.404	100

220

221 # Discussion

222 By analyzing summative assessments, it is possible to modify future test development techniques
223 or modify classroom instructions. This was the intention of the current study. The findings of this
224 study might pinpoint areas where interventions are required.

225

The internal reliability calculated in this summative test was 0.58. This value is a beat less than the expected range in most standardized assessments ($\alpha \geq 0.7$). According to (8), a Cronbach alpha of 0.71 was obtained in a standardized Italian case study. Reliability analysis could be categorized as: excellent if $\alpha > 0.9$, very good if between 0.8-0.9 and good if between 0.6-0.7 (1). If the reliability value lies within 0.5 - 0.6, revision is required. It will be questionable if reliability falls below 0.5 (1).

According to (1), the summative test administered in this study requires revision (KR-20 = 0.58). This might also imply that college educators need to validate their assessment tools through item analysis. According to Fraenkel and Wallen in (12), one should attempt to generate KR-20 reliability of 0.70 and above to acquire reliable test instruments.

According to Table 1, 58.1% (18) of the items in the summative test have average difficulty (p = 0.3 - 0.7). An exam item is considered to be good item if its p-value lies in the moderate range (17). In this study, a little higher than half of the exam items could be considered as good questions. This revealed that there is some gap in the preparation of good questions. A similar finding was reported in many other literatures (1, 6 and 9).

Questions which are too easy or too difficult for a student contribute little information regarding student's ability (17). Data in this study showed that 32.3% of the test items were too easy (recommended – 10%-20%) and 9.6% were too difficult (recommended-20%). Though it is advisable to include easy and difficult items in a given test (10), it would be better if the recommended limits were met. Hence as per the recommendations, there were more easy items

and fewer difficult items in the current study. A difficult item could mean either the topic is difficult for students to grasp (10, 11) or not taught well (10) or mis-keyed (12) or poor preparation of students.

According to (12), the discriminatory power of individual items can be computed either by discrimination index, biserial correlation coefficient, point biserial coefficient or phi coefficient. In this study, the discrimination power of every item was calculated by using point-biserial coefficient. The point-bi-serial coefficient result (Table 3) showed that only one item was considered as 'Excellent" ($r > 0.4$) and another one reasonably good ($r = 0.393$). All other items in this summative test need revision or subjected to improvement ($r < 0.3$). Similar study was reported by (3) that there was no a single item having discrimination index greater than or equal to 0.30. Contrary to this study, 46.67% of items in one study (15) were classified as good to excellent in their discrimination power ($r \geq 0.3$).

Large and positive values are required for the point bi-serial correlation as it indicates that students who get an item right tend to obtain high scores on the overall test and vise-versa (8). An item with negative and/or low discriminating power needed to be considered in subsequent test development phases. In this study, three items have negative discrimination. This could be due to the fact that low ability students guessed the item right and high ability students suspicious of any clue less successful to guess (16). Items with negative discrimination decrease the validity of the test and should be removed from the collection of questions (10, 12, 13, 14 and 15).

Difficulty and discrimination indices are often reciprocally related. However, this may not always be true. Questions having high *p*-value discriminate poorly; conversely, questions with low *p*-value may discriminate well (17). This variation could be as a result of students who make a guess when selecting the correct responses (12). Data (Table 4) showed that guessing has occurred in this study. According to (1), moderately difficult items should have the maximal discriminative ability. The findings of this study contradict with (1). This may reflect that some extent of guessing occurred during test administration.

Distracter analysis was conducted to determine the relative usefulness of distracters in each item. Seventeen (81%) items have no NFDs (DE = 100%), three items (**q.12**, **q.19** and **q.28**) have 1 NFDs (DE = 66.7%) and one item (**q.13**) has 2 NFDs (DE = 33.3%) (Table 5). There is no item with 3 NFDs (DE = 0). On the other hand, seven distracters (11%) (12 - A, 14 - A, B, 18- B, C and 20-A) were selected by more students than the correct answer. This may indicate that the items were confusing (12). An overall DE mean of 92.1% (considered as ideal/acceptable) was obtained in this study. Similar finding was reported by (10) in an internal microbiology examination in India.

Non-functional distracters make an item easy and reduce its discrimination (10). Question number 31 (Table 7) has moderate difficulty and excellent discrimination power. Probably this could be due to absence of NFDs. However, this doesn't work for other test items probably due to random guessing or some flaws in item writing (10).

5.1 Conclusion and Recommendations

Post exam item analysis is a simple but effective method to assess the validity and reliability of a test. It detects specific technical flaws and provides information for further test improvement. An item with average difficulty (p = 0.3 – 0.7), high discrimination ($r \geq 0.4$) and higher DE value (>70%) is considered as an ideal item. In this study, the summative test as a whole has moderate difficulty (mean = 0.56) and good distracter efficiency (mean = 92.1%). But it poorly discriminates between high and low achieving students. The test as a whole needs revision as its reliability was not reasonably good. Some flaws in item writing were also observed.

According to Xu and Liu (2009) in (1), teachers' knowledge in assessment and evaluation is not static but rather a complex, dynamic and ongoing activity. Therefore, it is plausible to suggest that teachers or instructors should have some in-service seminars on test developments. Since most of the summative tests constructed within the college are objective types, item analysis is recommended for instructors at some points in their teaching life. It is also suggested that there might be a specific unit responsible for testing and the analysis of the items after exam administration.

Competing interests

The author declares that there is no competing interest.

Acknowledgment

I would like to thank lecturers at Department of Natural Science, Gondar CTE, for providing exam papers for the study. I would like to extend my appreciation to Mr. Awoke Debebe for his assistance in data entry and critical review of the manuscript.

References

1. Zia-ul-Islam, Usmani A. (2017). Psychometric analysis of anatomy MCQs in modular examination. *Pak. J. Med. Sci*. 2017; 33(5).
2. Anna Siri and Freddano M. (2011). The use of item analysis for the improvement of objective examinations. *Procedia - Social and Behavioral Sciences* 29 (2011) 188 – 197.
3. Deshpande S. and Prajapati R.K. (2018). Item analysis of mid-trimester test paper and its implications. *Int. J. Manag. and App. Sci.*, Volume-4, Issue-2, Feb.-2018.
4. Deena Kheyami, Ahmed Jaradat, Tareq Al-Shibani and Fuad A. Ali (2018). Item analysis of multiple choice questions at the department of paediatrics, Arabian Gulf University, Manama, Bahrain. *SQU Medi. J*., Volume 18, Issue 1, pp. e68–74.
5. Towns Marcy H. (2014). Guide to developing high-quality, reliable, and valid multiple-choice ssessments. *J. Chem. Educ*. (910) 1426−1431.
6. Chauhan P., Chauhan G.R., Chauhan B.R., Vaza J.V and Rathod S.P. (2015). Relationship between difficulty index and distracter effectiveness in single best-answer stem type multiple choice questions. *Int. J. Anat. Res*. 3(4):1607-10.
7. Backhoff, E., Larrazolo, N., & Rosas, M. (2000). The level of difficulty and discrimination power of the basic knowledge and skills examination. *Revista Electrónica de Investigación Educativa*, 2 (1).

8. Gnaldi P., Matteucci M., Mignani S. and Falocci N. (2015). Methods of item analysis in standardized student assessment: an Application to an Italian case study.
9. Kennedy Quaigrain & Ato Kwamina Arhin (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education* (2017), 4: 1301013.
10. Ardra Ravindranathan Menon and Prithi Nair Kannambra (2017). Item analysis to identify quality multiple choice questions. *National J. Lab.y Medi*. 6(2): MO07-MO10.
11. Shenoy P., Sayeli V. and Rao R.R. (2016). Item-analysis of multiple choice questions: A pilot attempt to analyze formative assessment in pharmacology. *Res. J. Phar. Bio. Chem. Sci*. 7(2): 1683.
12. Shafizan Sabri S. (2013). Item analysis of student comprehensive test for research in teaching beginner string ensemble using model based teaching among music students in public universities. *Int. J. Edu. Res*. 1(12).
13. Mitra N. K, Nagaraja H. S, Ponnudurai G, Judson J. P. (2009). The levels of difficulty and discrimination indices in type-A multiple choice questions of pre-clinical semester-1 multidisciplinary summative tests. *IeJSME* 2009: 3 (1): 2-7.
14. Sim S.M. and Rasiah R.I. (2006). Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. *Ann. Acad. Med*. (35): 67-71.
15. Mukherjee P. and Lahiri S.K. (2015). Analysis of multiple choice questions: Item and test statistics from an assessment in a medical college of Kolkata, West Bengal. *J. Den. Medi. Sci.* 14(12): 47-52.

16. Kolte V. (2015). Item analysis of multiple choice questions in physiology examination. *Indian J. Basic App. Medi*. Res.:4 (4): 320-326.

17. Tavakol M. and Dennick R. (2011). Post-examination analysis of objective tests. *Medical Teacher*; 33: 447–458.

18. Arega Yirdaw (2016). Quality of education in private higher institutions in Ethiopia: The role of governance. *SAGE open*, pp. 1–12.

19. Fekede Tuli (2012). Examining quality issues in primary schools in Ethiopia: Implications for the attainment of the education for the all goals. *ECPS J*. 5/2012.

20. Ministry of Education, Ethiopia. (2008). General education quality improvement package (GEQIP). November, 2008.

21. Adhi M. I. and Aly S. M. (2018). Student perception and post-exam analysis of one best MCQ and one correct MCQs: A comparative study. *J. Pak. Med. Assoc*. 68 (4).

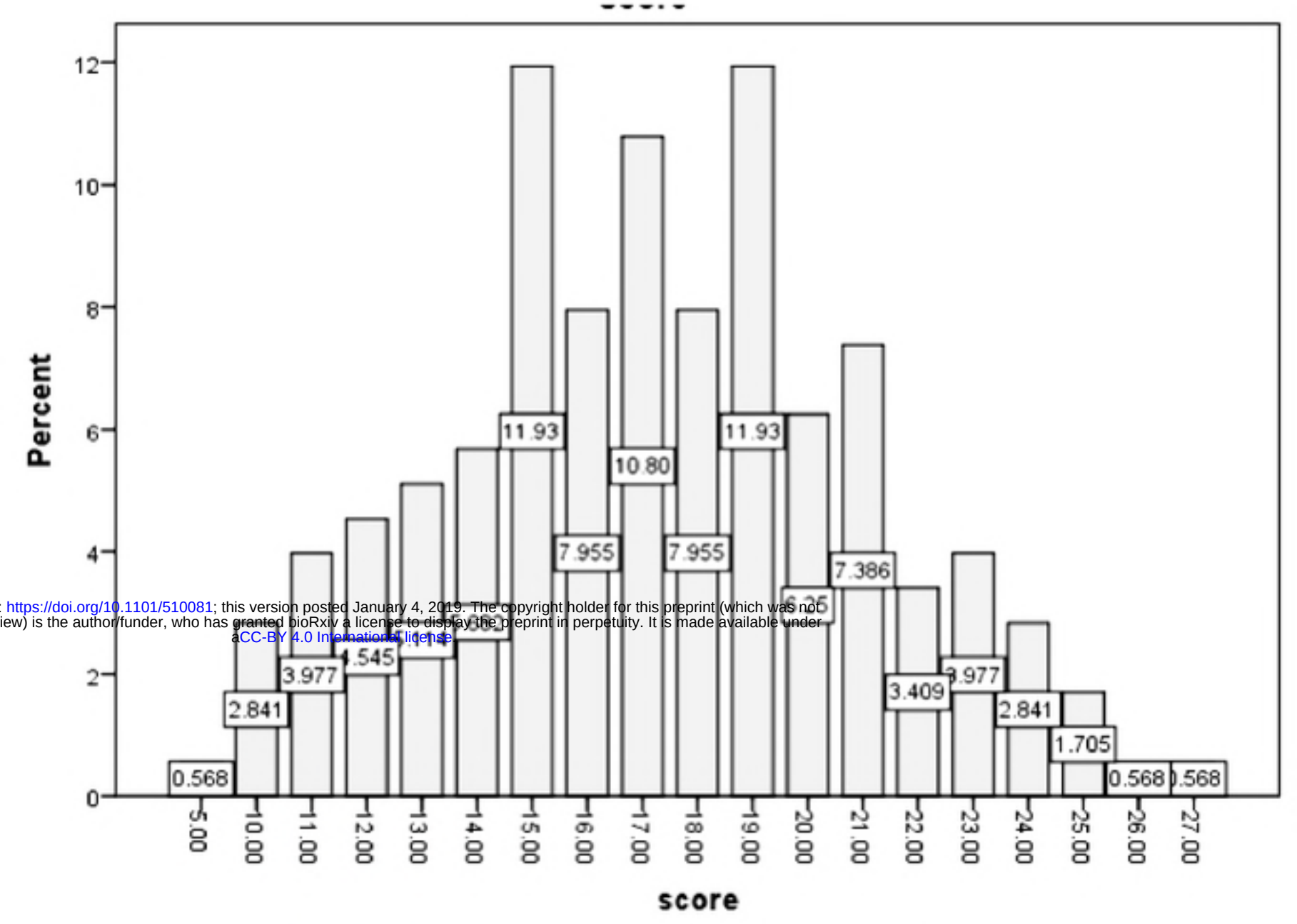

Figure 1: Percentage distribution of total test score

Figure