



Quality Item Generation through Item Analysis: Exploring SPSS

Etobro, A. Benjamin & Taiwo, Olufunmi

Department of Educational Foundations and Counselling Psychology, Faculty of Education, Lagos State University, Ojo

Abstract

The study assessed the development of items with good qualities adopting item analysis to determine level of difficulty and discrimination index of items. The study was guided by two research questions. Fifty items were administered to 50 undergraduates in their third year. Item analysis was carried out on the responses. Findings showed that 16% of the entire test had difficulty index of less than 0.2. Furthermore, 18% of the items had a difficulty index of 0.20-0.39. This implies that 18% of the items were difficult. Twenty-three items (46%) were moderately easy (0.60-0.79). The item discrimination indices reveal that 36% of the items had discrimination level of 0.41 to 0.70; 28% had discrimination level of 0.21 to 0.40; 16% of the items had discrimination index between 0.00 and 0.20 indicating that they are poor items. In conclusion, the difficulty and discrimination powers of most of the test items were valid and reliable.

Key Words: item generation, item analysis, item discrimination, difficulty index

Introduction

Assessment and evaluation appears to be integral parts of the instructional process. One of the major aims of education is all round personality development of the child. But in the contemporary system of education, this wholistic development has centred only on the cognitive domain. Academic achievement of learners could be ascertained by employing different assessment techniques. Assessment outcomes can be used in measuring the attainment of learning behavioural objectives and determining teachers' effectiveness and efficiency. The importance of assessment cannot be overemphasized because through assessment according to Ukashatu et al., (2019), decisions about grades, placement, advancement, instructional needs, as well as about the curriculum innovations are made. Thus, for effective and meaningful assessment to take place there is need to utilize questions of good psychometric qualities. This implies that the 21st century instructors be well equipped with test construction skills in order to generate valid and reliable items.

The teacher during instructional process is constantly assessing learners' progress. Assessment plays a vital role in instructional process in the educational system because according to Crocker, (2019) assessment provides education stakeholders with ample data in decision making. Making appropriate and effective decisions from the

assessment data obtained, would require constructing valid and reliable test items that are void of ambiguities. The test constructor needs to observe due process in test construction by performing validation process that should include test item analysis.

Item analysis is performed to determining the item difficulty level, facility index, discriminating powers and distracter efficiency. It is expedient to explore the possibility of adopting item analysis with the ever-increasing students' population among university general compulsory courses. Item analysis according to Urbina (2014) refers to all the techniques adopted to assess the psychometric qualities usually employed during test development construction process. Item analysis is a process that examines student's responses to individual test items to assess the quality of the items and the test as a whole (Sharma, 2021).

Oftentimes examinees in a teacher-made test could be seen in examination halls requesting for explanations of some items. The clarifications may emanate from ambiguity of the test item, too broad an item or similar options in a question of multiple choice items. Some of these shortcomings in teacher-made tests would have been taken care of, if the items had passed through the item analysis process as part of standardization exercise. These inadequacies on the part of the instructors or test developers in item

construction perhaps may be due to lack of skills in establishing the validity and reliability of quality test items. Instructors or teachers can build their test construction skills by exploring the adoption of technology in the use of statistical tools and packages.

The rapid development of information and communication technologies has brought about the modernization of teaching and learning across the globe. In the contemporary instructional process, teaching and learning have been made easy through the use of computer technologies, electronic gadgets and other multi-media tools. Consequently, the adoption of electronic learning metamorphosed into electronic assessments or examinations (e-assessments and e-examinations). The adoption of e-assessment should require the construction of good items. Items of good quality could be generated through performing item analysis employing Statistical Package for Social Sciences (SPSS). The study determined item difficulty index and discriminating power exploring SPSS.

Two research questions guided the study. Fifty, Faculty of Education students in their third year of 2020/2021 academic session, who took EST 305: “Tests, Measurement and Educational Statistics” were randomly sampled adopting systematic sampling technique. An achievement test tagged “Tests, Measurement and Educational Statistics Test”

was adopted from a pool of lecturer-constructed items in the question bank. There is dearth in Literature on the practical determination of item qualities adopting SPSS. In order to fill this gap, the present study carried out item analysis of Tests, Measurement and Educational Statistics multiple choice questions obtained from a year three course. The study focused on analyzing the difficulty level and discrimination power of the test. Based on this, two research questions that guided the study are: What is the difficulty level of the multiple choice questions in Tests, Measurement and Educational Statistics? What is the discrimination power of the multiple choice questions in Tests, Measurement and Educational Statistics?

Results

Research Question 1

What is the item difficulty level of Test, Measurement and Evaluation objective questions among undergraduate students?

To answer this question, descriptive statistics of means (p-values) was employed. Score the items, ‘1’ for correct response and ‘0’ for incorrect response, the difficulty index of each item could be computed. Compute the item difficulty index, by clicking on “Analyze”; then click on “Descriptive Statistics”. This action was followed by clicking on “Descriptives” as shown in Figure 1.

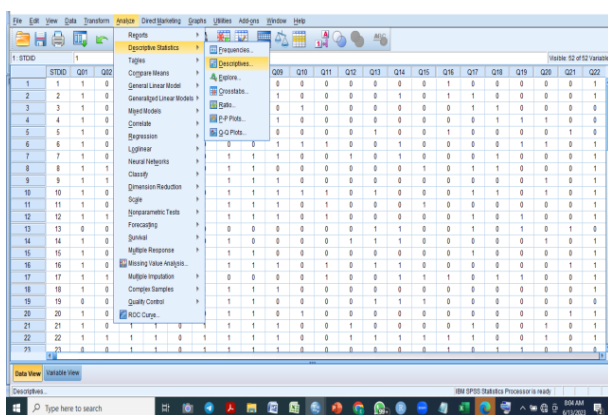


Figure 1: Performing Descriptive Statistics

Take all the items to the dialogue box labeled “Variable(s)” (Figure 2) and click on “Option”. From the “Descriptive options” interface that appears click on “Standard deviation, Maximum and Minimum”, leaving out only “Mean” and click on “Continue” as

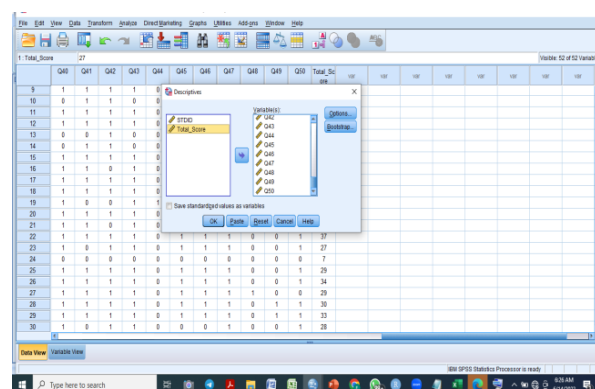


Figure 2: Performing Descriptive Statistics by clicking on “Descriptives”

shown in Figure 3. Then click on “OK”. SPSS “Descriptive statistics” output displays. The values in the “Mean column” in the SPSS “Descriptive statistics” output which, is the p-value of each item represents the “Difficulty index”, Figure 4.

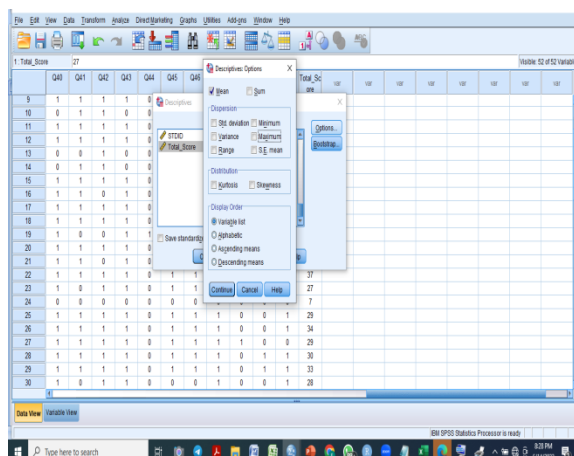


Figure 3: Performing Descriptive Statistics

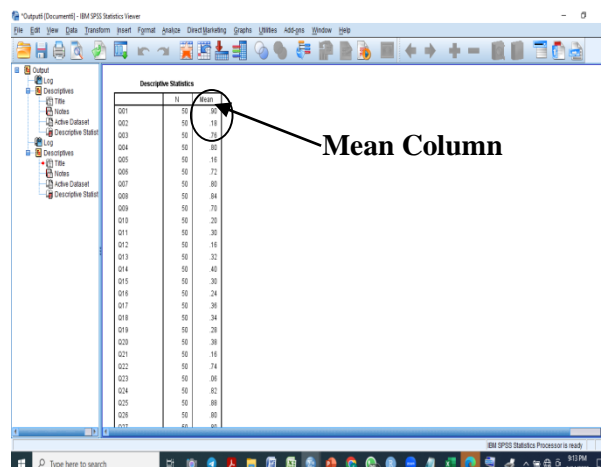


Figure 4: Difficulty index

Table 1: Difficulty Indices of Test, Measurement and Evaluation Objective Questions among Undergraduates

Difficulty Index	<i>f</i>	%	Items	Item Evaluation	Remarks
Less than 0.20	8	16	5, 21, 23, 29, 33, 38, 44, 48	Most difficult	Remove
0.20-0.39	9	18	2, 10, 11, 12, 13, 15, 16, 18, 19	Difficult	Retain
0.40-0.59	5	10	14, 17, 20, 36, 49	Moderately difficult	Retain
0.60-0.79	23	46	1, 3, 4, 6, 7, 9, 22, 24, 25, 26, 27, 28, 30, 31, 34, 35, 37, 40, 41, 42, 45, 46, 47	Moderately easy	Retain
0.80-0.89	5	10	8, 32, 39, 43, 50	Easy	Retain
Total	50	100			

The Difficulty indices of the achievement test are 0.04 to 0.89 as shown in Table 1. Specifically, Table 1 reveals that eight (8) items 5, 21, 23, 29, 33, 38, 44 and 48 representing 16% of the entire test fell in the most difficulty index of less than 0.2. There is need to delete the eight items. Nine (9) items, 2, 10, 11, 12, 13, 15, 16, 18 and 19 had difficulty indices of .20 to .39. This implies that nine questions were difficult. Furthermore, five items, 14, 17, 20, 36 and 49 were within the range of the difficulty index of

(0.40-0.59). It indicates that five questions were moderately difficult. Twenty-three (23) items, 1, 3, 4, 6, 7, 9, 22, 24, 25, 26, 27, 28, 30, 31, 34, 35, 37, 40, 41, 42, 45, 46 and 47 representing 46% of the entire test fell in the moderately easy index of (0.60-0.79). This shows that 23 items were moderately easy. However, only five items, (8, 32, 39, 43 and 50) had .80 to .89 difficulty indices indicating that the items were easy.

Research Question 2

What is the item discrimination index of “Tests, Measurement and Educational Statistics” test items?

To answer this question, score the items, ‘1’ for correct response and ‘0’ for incorrect response, the item discrimination index of each item was computed. To compute the item discrimination index, the total score is computed for each student. Compute

the sum of scores obtained by each student by ‘Clicking on Transform’ and then “Compute Variable”; then Type “Total Score”. In order to sum up the scores for all the items, all the items (Q01 to Q50) are entered into the dialogue box “Numeric Expression” one after the other. To get the Total Score for each student click “OK”, Figure 5.

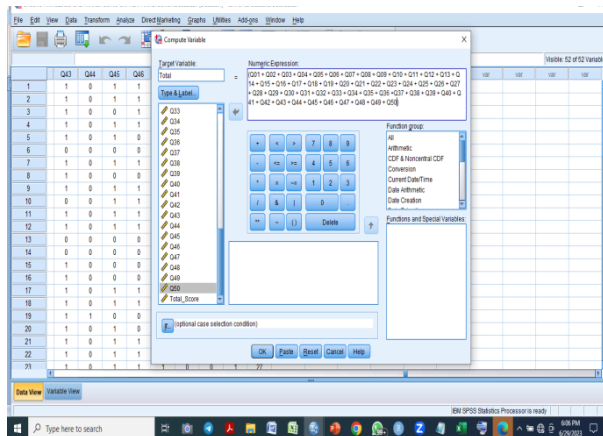


Figure 5: Computing total score

Computing the discrimination power of the items require arranging students' total score in descending order. Click on “Data” and then click “Sort Cases”. Move “Total score” into Dialogue Box “Sort by” and click on “OK” as in Figure 6. The next step is to rearrange the scores with student’s ID swapping with the question and between the columns and rows by clicking on “Transpose”. All the scores are moved

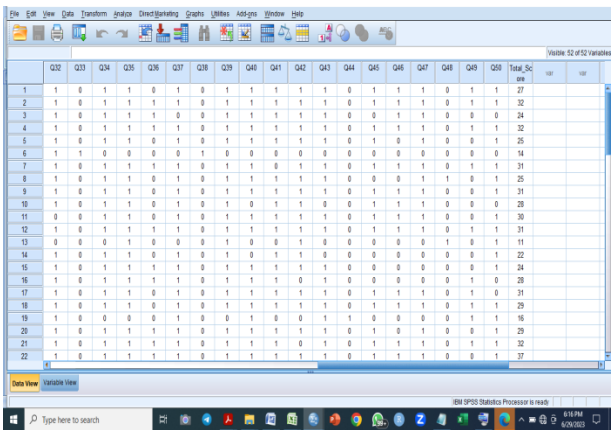


Figure 6: Total score

into the “Variable(s) box” and the student ID into "Name Variable". Then click on “OK”. This action transposes the questions into rows and students' ID into columns, as shown in Figure 7. Then group the students into upper, top 27% and lower, lower 27% resulting to 14 students in each group as indicated in Figure 8.

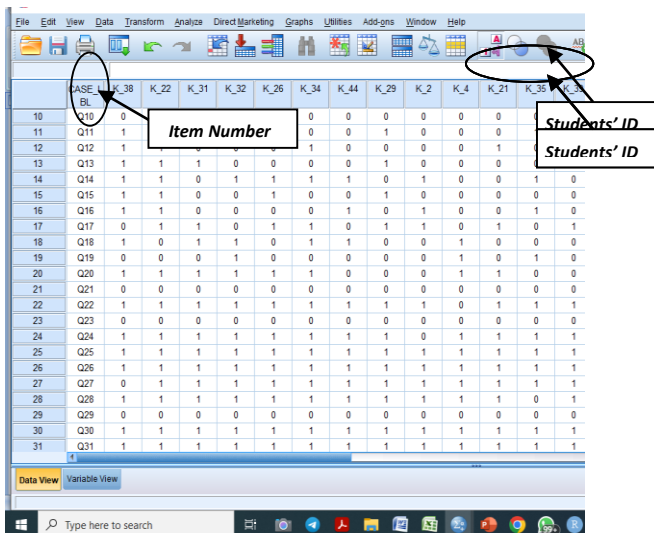


Figure 7: Transposed data

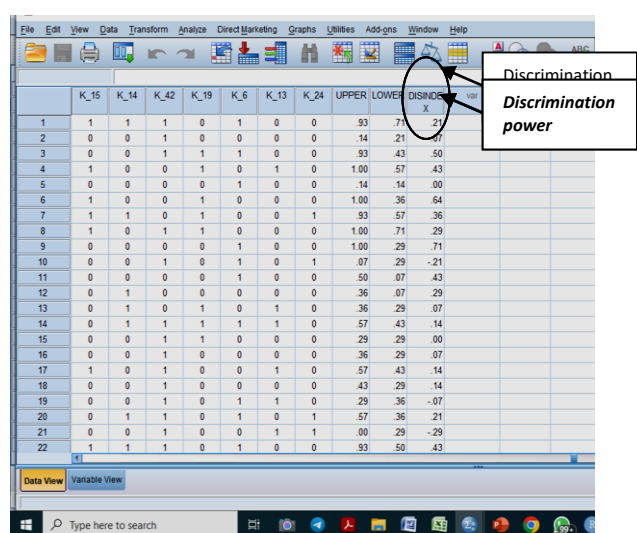


Figure 8: Transposed data sorted into Upper and Lower groups as well as item discrimination power

The proportion of students in both upper and lower groups who got an item right is computed by clicking on “Transform”. Then click on “Compute variables”. Compute variable dialogue box will appear. The variable to be computed is typed into the “Target variable” and moves the top 27% (the 14

upper achievers) one after the other into the “Numeric Expression” box and divides this by 14, then click “OK”. A similar action is performed for the lower achievers, Figure 8. Compute the discrimination index by subtracting the proportion of students in the lower group from that of upper group, Figure 8.

Table 2: Discrimination Indices of Test, Measurement and Evaluation Objective Questions among Undergraduates

Item Discriminating Index	<i>f</i>	Percent	Items	Remarks
.41 - .70	18	36	3, 4, 6, 9, 11, 22, 24, 25, 26, 27, 30, 34, 36, 41, 45, 46, 47 and 49	Good
0.21-0.40	14	28	1, 7, 8, 12, 20, 28, 31, 35, 37, 39, 40, 42, 43 and 50	Satisfactory
0.00-0.20	8	16	5, 13, 14, 15, 16, 17, 18 and 32	Poor
Negative	10	20	2, 10, 19, 21, 23, 29, 33, 38, 44 and 48	Rejected

Table 2 reveals that 18 (36%), 3, 4, 6, 9, 11, 22, 24, 25, 26, 27, 30, 34, 36, 41, 45, 46, 47 and 49 had discrimination level of 0.41-0.70. This implies that the items distinguished between students in the upper and lower groups. Furthermore, 14 (28%) items (1, 7, 8, 12, 20, 28, 31, 35, 37, 39, 40, 42, 43 and 50) had discriminating level of .21 to .40 and thus, were satisfactory since they distinguished between students in the upper and lower groups. Eight (16%) of the items are poor since they had indices of .20 and below. However, 10 items (20%) had negative discrimination indices; hence they were removed.

Discussion of Results

Results on the difficulty powers of Tests, Measurement and Educational Statistics multiple choice questions among undergraduate students revealed that only eight (8) items representing 16% of the entire 50 items fell in the most difficulty index of less than 0.2. Meanwhile, 42 items corresponding to 84% of the items had difficulty indices of .20 to .89. The findings in this study, is in tandem with that of Toksöz and Ertunç (2017) who reported that most of the items were of moderate difficulty level. Odukoya *et al.*, (2017) who conducted item analysis on the multiple choice items in a Nigerian private university reported that most of the items did not meet the acceptable level of difficulty and distractive indices. The findings in this study is also in agreement with a study carried out by Burud *et al.*, (2019) on item analysis, who reported that 29.16% of the items were very easy, 35% were excellent, 16.67 % were good while 4.17 % were very difficult items. Furthermore, findings on discrimination powers of the multiple choice questions of Tests, Measurement and Educational Statistics showed that 10 items

(20%) had negative discrimination indices. This implied that more of the lower achievers attempted the items than students in the upper group hence they were discarded from the pool of items. Eight (16%) of the items are poor since they had indices of .20 and below, while 10 items had negative discrimination power. However, 32 items were good for inclusion in the item pool since their discrimination power ranges from 0.21 to 0.70. These items with poor and negative item discrimination power could have been due to ambiguity in the item construction. These types of items would fail to assess testees’ mastery of target knowledge and may not adequately distinguish between upper and low achievers. Jannah *et al.*, (2021) argued that this suggests that such items with low and negative item discrimination power be revised or discarded. The finding in this study supported that of Hartanti and Yogi (2019) who reported in their study that most of the items were of satisfactory item discrimination power. Similarly, Danuwijaya (2018) reported that the majority of the test items had discrimination power that was satisfactory.

Conclusion

This study examined the process of item generation through item analysis by exploring SPSS. A survey research design was adopted. The study was guided by two research questions. Fifty items in Test, Measurement and Evaluation were administered to 50 undergraduates in their third year. The item analysis of the MCQs was practically demonstrated exploring SPSS. The item difficulty level and item discrimination power were presented with annotated SPSS displays. Results showed that some questions had difficulty power as low as less than 0.2. Such

items were recommended for deletion from the entire test because they are very difficult. The test has items with low discrimination index of between 0.00 and 0.20 need to be revised while the items with negative discrimination indices were suggested for deleting. Conclusively, the difficulty and discrimination powers of most of the test items were valid and reliable. Lecturers, teachers and other stakeholders involved in assessment should endeavour to perform item analysis to avoid items with ambiguity.

References

- Burud, I., Nagandla, K., & Agarwal, P. (2019). Impact of distractors in item analysis of multiple choice questions. *International Journal of Research in Medical Sciences*, 7 (4), 1136-1139.
- Danuwijaya, A. A. (2018). Item analysis of reading comprehension test for post-graduate students. *English Review: Journal of English Education*, 7(1), 29-40. <https://doi.org/10.25134/erjee.v7i1.1493>.
- Hartati, N., & Yogi, H. P. S. (2019). Item analysis for a better quality test. *English Language in Focus (ELIF)*, 2(1), 59-70. <https://doi.org/10.24853/elif.2.1.59-70>
- Jannah, R., Hidayat, D. N., Husna, N. & Khasbani, I (2021). An item analysis on multiple-choice questions: a case of a junior high school English try-out test in Indonesia. *Leksika*.Vol.15 (1), 9-17
- Mahjabeen, W., Alam, S., Hassan, U., Zafar, T., Butt, R., Konain, S., Rizvi, M. (2017). Difficulty index, discrimination index and distractor efficiency in multiple choice questions. *Annals of PIMS*, 310-315.
- Odukoya, J. A., Adekeye, O., Igbino, A. O., & Afolabi, A. (2017). Item analysis of university-wide multiple choice objective examinations : the experience of a Nigerian private university. *Quality & Quantity*, 1-15. <https://doi.org/10.1007/s11135-017-0499-2>
- Sharma, L. R. (2021). Analysis of Difficulty Index, Discrimination Index and Distractor Efficiency of Multiple Choice Questions of Speech Sounds of English. *International Research Journal of MMC (IRJMMC)*; Vol. 2 (1), pp 15-28.
- Taiwo O.A (2020). Effects of Test Score Equating Methods on Students Assessment Measures in Southwest; Unpublished Ph.D Thesis, Ekiti State University, Ado.
- Urbina, S. (2014). *Essentials of Psychological Testing* (2nd ed.). Hoboken, NJ: John Wiley & Sons.