Journal of Policy Research, 9(1), 497-512.
https://jprpk.com
https://doi.org/10.5281/zenodo.7477685

Development and Content Validity of Test for the HOTRTG (Higher Order Thinking Research Test for Graduates) in the Subject of Research Methods at the University level in Punjab, Pakistan

Ghous Bakhsh[1], Dr. Shahid Hussain Shahid[2], Dr. Jam Muhammad Zafar[3], Muhammad Kamran[4]

Abstract

Tests for Higher order thinking (HOT) skills are available around the globe. In Pakistan, the quality of tests used for assessing HOT is very rare. Specialized tests are required to measure the HOT of students for which this study was made. It aimed to 1) develop a multiple choice and open-ended question test for measuring the HOT of graduate students, and 2) ensure the content validity and reliability of the test. This was instrumental research consisting of three phases; 1) instrument development, 2) expert validation, and 3) pilot testing. The instrument was developed in Multiple Choice Questions (MCQs) and Open-Ended Questions (OEQs) format. The developed instrument consisted of 38 items (30 MCQs, 8 OEQs). The CVR value for the test ranged from 0.571 to 1.000 and the CVI value for the test was 0.774 to ensure the validity of the test. The tool was validated by expert opinion on the material, construction, and language aspects which showed that all 38 items (30 MCQs and 8 OEQs) were valid and of very good quality. At the pilot testing phase, it was conducted on 80 graduate students of the faculty of Humanities and Social Sciences, KFUEIT, R Y Khan, Punjab, Pakistan. The tested instrument was then analyzed to determine the level of difficulty, discrimination index, and reliability of the test. On the bases of the results of difficulty level and discrimination index 26 questions (20 MCQs and 6 Short Questions) were valid and 10 MCQ items and 2 OEQ items were found below the criteria which were dropped from the final draft of the test. The Alpha value of reliability of the test was 0.735 which indicates that the test is highly reliable. On the bases of the results, the developed instrument proved valid to measure the HOT skills of graduate students in the subject of research at the University level in Punjab, Pakistan.

Keywords: Higher Order Thinking Skills, Instrument Development, Validation, Graduate students

1. Introduction

An essential component of educational research and assessment is gauging higher order thinking abilities (*higher order thinking will be written as HOT from now on*). Researchers and educators have employed various tools to assess cognitive processes over the years. To assess students' critical thinking skills, Smith and Johnson created a thorough survey in 2017 (Smith & Johnson, 2017). The following year, Thompson and Brown conducted a study (Thompson & Brown, 2018) that used a performance-based evaluation to gauge HOT in a classroom context. Similarly, Martinez and Davis conducted a longitudinal study in 2019 utilizing a standardized test to monitor middle school pupils' growth in HOT abilities (Martinez & Davis, 2019).

New measurement techniques have been investigated with the rise of digital technologies. Jones and Lee (Jones & Lee, 2020) looked into the use of computerized adaptive testing to evaluate HOT abilities. In 2021, Chen and Wang focused their research on how artificial intelligence and machine learning algorithms could be used to evaluate HOT (Chen & Wang, 2021). In addition, Garcia and Hernandez's study from 2022 offered a unique framework for evaluating HOT utilizing game-based evaluations (Garcia & Hernandez, 2022).

These studies demonstrate the evolving scenery of measurement tools for HOT skills, ranging from traditional surveys and standardized tests to innovative approaches leveraging technology. Researchers continue to explore and refine these tools to gain deeper insights into students' cognitive abilities and promote effective teaching and learning strategies (Arooj et al., 2022; Iqbal et al., 2022; Kamran et al., 2017).

The educational reforms are based on learning taxonomies like Bloom's taxonomy. According to Bloom's Taxonomy creation of something new is the application of HOT. HOT includes the learning of multifaceted skills like critical thinking and issue-solving (Watson, 2019). It is the dire need of the 21st century that a student should have critical thinking and industrial skills, innovate and solve problems through teamwork, and have the capability to communicate efficiently (UNESCO, 2013; Scotts, 2015; Pretorius et al., 2017).

[1] Ph.D. Scholar Education, Department of Education, KFUEIT, Rahim Yar Khan, Pakistan, ghousdteryk@gmail.com

[2] Assistant Professor, Institute of Humanities and Arts, KFUEIT, Rahim Yar Khan, Pakistan, shahid.hussain@kfueit.edu.pk

[3] Assistant Professor, Department of Education, KFUEIT, Rahim Yar Khan, Pakistan, dr.zafar@kfueit.edu.pk

[4] Assistant professor in the Department of Education, University of Loralai, Balochistan, Pakistan, Muhammad.kamran@uoli.edu.pk

HOT refers to cognitive processes that demand people analyze, evaluate, and synthesize information in addition to simple memory and understanding. It is widely acknowledged as an important element of education and plays a significant part in the development of critical thinking abilities. The complexity of the modern world and the velocity of change in recent years (2017–2022) have made the need for HOT more critical than ever.

HOT is important in preparing students for the issues they will encounter in the future, according to experts in 2017 (Author, 2017). Critical and creative thinking skills are increasingly important (Andleeb et al., 2022; Kamran et al., 2022) as technology develops and automation replaces conventional tasks. These abilities are required for people to handle complex situations and reach wise decisions. By encouraging students to think deeply and critically, they develop a desire to explore new ideas and creativity (Kamran et al., 2021a; Kamran et al., 2021b).

In order to adapt to new circumstances and come up with creative solutions, HOT is crucial, as the global pandemic of 2020 demonstrated (Smith, 2020). In order to navigate the challenges presented by the crisis, individuals and communities must possess the capacity to engage in creativity (Arooj et al., 2021; Kamran et al., 2017), adaptability, and problem-solving. In 2021, researchers underscored the connection between HOT and creativity, emphasizing that it enables individuals to generate novel ideas and solutions (Johnson, 2021). In order to evaluate the quality of research in education, the use of appropriate data collection tools is very important (Huma Naz et al., 2023).

2. Theoretical Framework

The HOT concept originated from Bloom's Taxonomy (1956) which was based on the constructivist theory of education. This Taxonomy was revised by Anderson and Krathwohl in 2001, according to them thinking skills are grouped into six levels such as remembering, understanding, applying, analyzing, evaluating, and creating (Anderson et al. 2001). The last three skills (analyzing, evaluating, and creating) are considered the HOT skills, and the first three skills (remember, understand, and apply) are considered to be lower order thinking skills. HOT skills are considered more complicated and challenging to achieve than lower order thinking skills. There is no agreement on what comprises HOT skills, however, according to Lee and Choi (2017) majority of researchers viewed that HOT skills involve 'complex cognitive actions such as developing arguments formulating hypotheses, making comparisons and inferences, elaborating, interpreting and analyzing information, applying multiple criteria, integrating and synthesizing information, and sorting multiple solutions. HOT skills consist of such thinking processes that go beyond knowledge and understanding. According to another definition, HOT skills include skills such as critical thinking, evaluative and problem-solving skills (Gorin et al., 2011).

3. Methodology

3.1. Design of the study

This study aims to develop and validate HOT Research Test for Graduates (HOTRTG). The design of the study is an instrumental design because the researchers have to develop and validate the instrument/test (Huma Naz et al., 2023). The process of development consisted of three phases; (1) the Designing phase, (2) the Expert Validation phase, and (3) Pilot testing validity.

a. Phase I: Designing Phase-Development of Higher-order thinking-based Research Test for Graduates (HOTRTG)

The test for graduate students was developed to measure the students' HOT. It was developed in a mixed format (Close-Ended and open- Ended), based on four higher levels (applying, analyzing, evaluating, and creating) of revised Bloom's Taxonomy. The common topics of three schemes of studies, i.e., M Phil education, MS English, and MS Islamic studies were selected covering the three (3) main themes, i.e., Concept of Research, Research Methodology, and Data Analysis, and eight (8) sub-themes, i.e., the introduction of research, types of research, development of objectives/hypothesis/research questions, comparison of research designs, sampling methods, development of research tools, data analysis and data presenting of the subject of "Research Methods" as contents for the intervention of the study. The questions were developed about the four top levels; (applying, analyzing, evaluating, and creating) of revised Bloom's Taxonomy. The questions were based on situations, concepts, procedures, graphic organizers, and images that require HOT to answer. Initially, fifty (50) MCQs were developed on different topics of research covering the three levels; applying, analyzing, and evaluating, and eight (8) open-ended questions on the 6^{th} level of the cognitive domain of Bloom's revised Taxonomy. Test items, scoring key for MCQs, and scoring rubrics for open ended questions were discussed with the supervisor,

Professors, and M.Phil. and Ph.D. scholars of the education department. In light of the discussion and recommendation of the supervisor, thirty (30) MCQs and eight (8) OEQs were finalized. Topic-wise and cognitive level-wise detail of contents, question distribution, and marks
distribution are given in the tables below.

Table 1: Topic wise & Cognitive Level wise Question Numbers Distribution

Main Topics	Sub-Topics	Topic wise & Cognitive Domain wise Question Numbers					
		Apply	Analyze	Evaluate	Create	Total	Topic wise weightage
Concept of Research	Introduction of Research		2	2	1	5	13%
	Types of Research		3	2		5	13%
	Development of objectives/hypothesis/research questions	1	1	2	1	5	13%
Research Meth	Comparison of Research designs	1	1		1	3	7%
	Sampling Methods		3	1	2	6	16%
	Development of Research tools	1	1	1	2	5	13%
	Data Analysis	1	2	2	1	6	16%
Data Analysis	Data Presenting	1	1	1		3	7%
Total Questions		**5**	**14**	**11**	**8**	**38**	**100%**
Weightage of Questions		**13%**	**37%**	**29%**	**21%**	**100%**	**100%**

Source: Fictitious data, for illustration purposes only

Table 1 indicates that the content of the test is comprised of three main themes "Concept of Research", "Research Methodology" and "Data Analysis" having eight (8) sub themes of the subject of Research Methodology of M.Phil. programs of Social Sciences. The main theme wise distribution of question is as 15 items from 1st theme, 14 items from 2nd theme, and 9 items from 3rd theme, and on the bases of levels of cognitive domain wise distribution of marks is as; 5 items from 3rd level "Apply", 14 items from 4th level "Analyze", 11 items from 5th level "Evaluate", and 8 items from 6th level "Create".

Figure 1 shows that three main themes of contents for the development of test and questions are distributed according to the main theme as; 15 questions are from "Concept of Research|", 14 questions from "Research Methodology" and 9 questions from 'Data Analysis" are developed.

Figure 2 indicates that questions of the test are developed covering the four top levels of the cognitive domain of revised Bloom taxonomy; 5 items from the 3rd level "Apply", 14 items from the 4th level "Analyze", 11 items from 5th level "Evaluate", and 8 items from 6th level "Create" are developed.

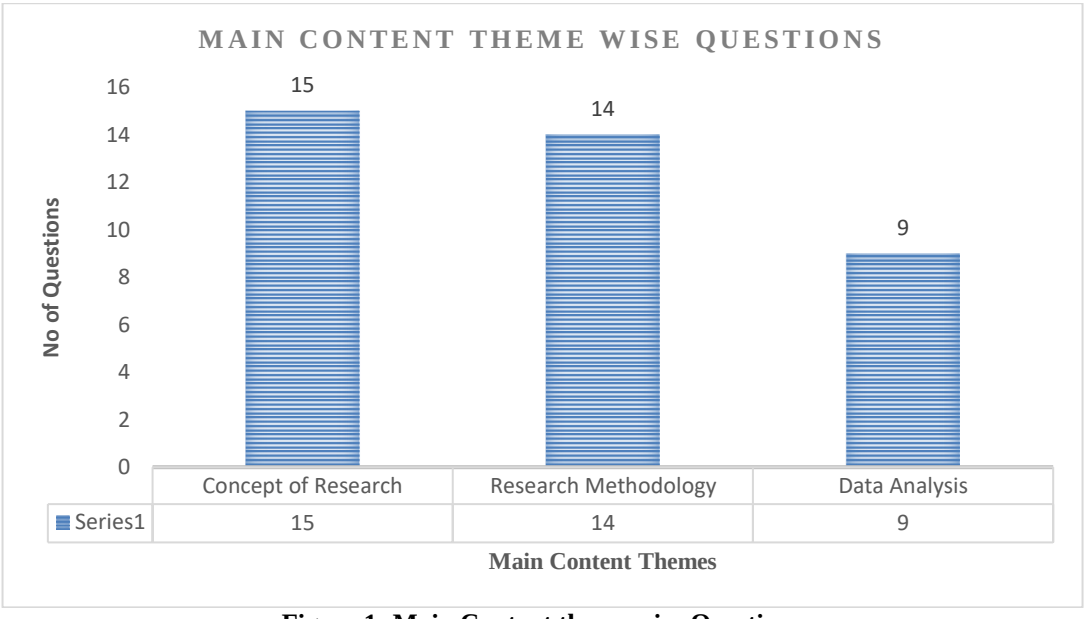


Figure 1: Main Content theme wise Questions

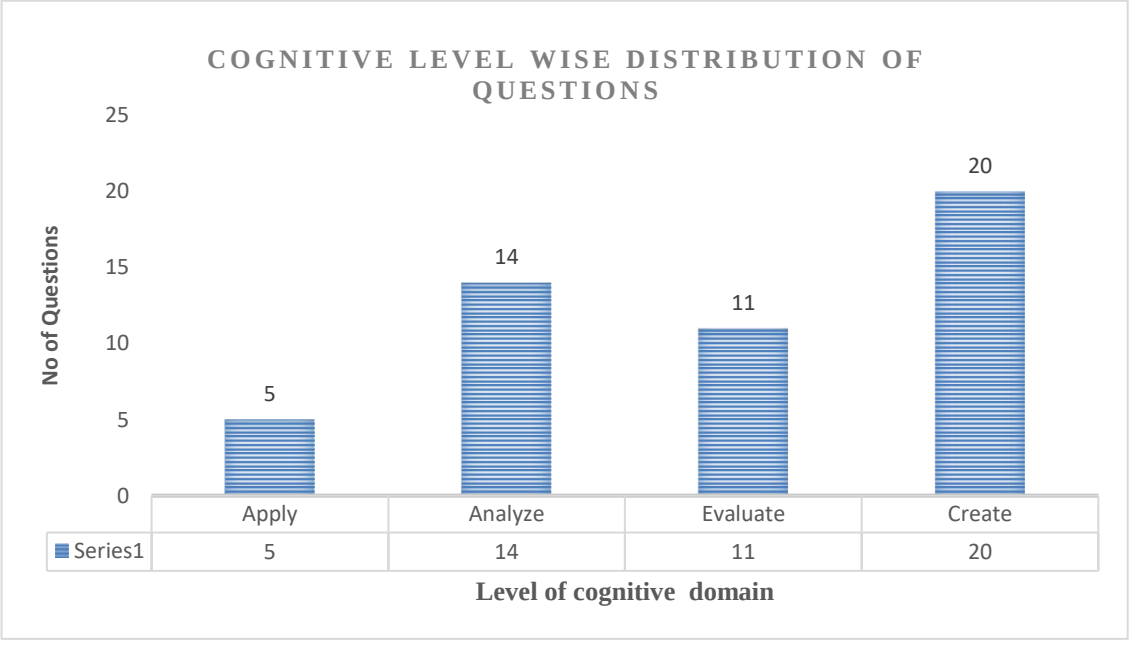


Figure 2: Main theme wise distribution of questions

Table 2 represents the distribution of marks on the bases of Main Content Themes and on the bases of levels of the cognitive domain. Main Content Themes-wise distribution of marks is given as; 20 marks are allocated to 'Concept of Research", 21 marks are allocated to "Research Methodology" and 9 marks are allocated to "Data Analysis". On the bases of levels of the cognitive domain, the distribution of marks is given as; 5 marks for 3rd level "Apply", 14 marks for 4th level "Analyze", 11 marks for 5th level "Evaluate" and 20 marks for 6th level "Create" for the development of the test.

Table 2: Topic wise & Cognitive Level wise Marks Distribution

Main Topics	Sub-Topics	Topic wise & Cognitive Domain wise Question Numbers					
		Apply	Analyze	Evaluate	Create	Total	Topic wise weightage
Concept of Research	Introduction of Research		2	2	5	9	18%
	Types of Research		3	2		5	10%
	Development of objectives/hypothesis/ research questions	1	1	2	2	6	12%
Research Methodology	Comparison of Research designs	1	1		3	5	10%
	Sampling Methods		3	1	4	8	16%
	Development of Research tools	1	1	1	3	6	12%
Data Analysis	Data Analysis	1	2	2	3	8	16%
	Data Presenting	1	1	1		3	6%
Total Questions		**5**	**14**	**11**	**20**	**50**	**100%**
Weightage of Questions		**10%**	**28%**	**22%**	**40%**	**100%**	**100%**

Source: **Fictitious data, for illustration purposes only**

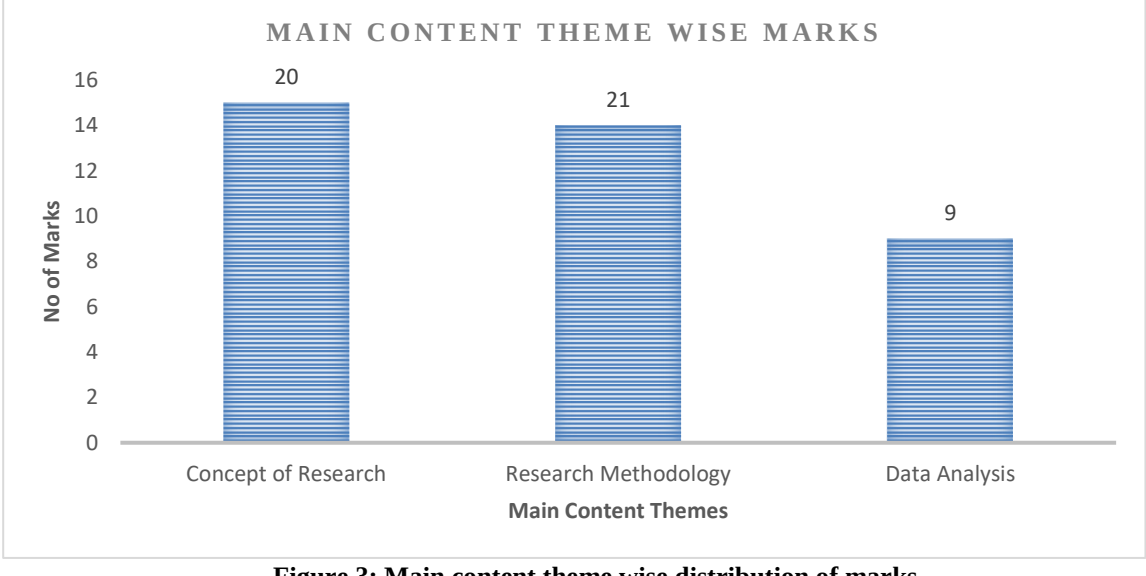


Figure 3: Main content theme wise distribution of marks

Figure 3 shows that the marks of the test are distributed according to the three main themes of contents. It indicates that 20 marks are allocated to 1st theme "Concept of Research", 21 marks are allocated to 2nd theme "Research Methodology, and 9 marks are allocated to 3rd theme "Data Analysis" in the development of the test.

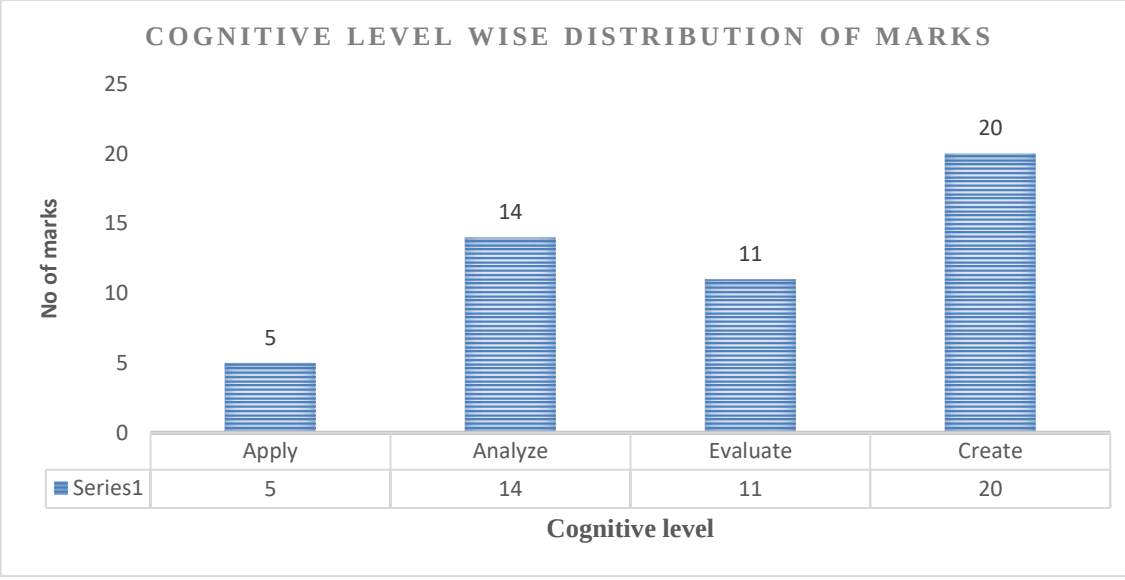


Figure 4: Cognitive level wise distribution of marks

Figure 3 shows that marks of the test are distributed into the top four levels of the cognitive domain of the revised taxonomy of Bloom. Level-wise distribution of marks is given as; 5 marks for 3rd level "Apply", 14 marks for 4th level "Analyze", 11 marks for 5th level "Evaluate", and 20 marks for 6th level "Create" for the development of the test.

b. Phase II-Validity of Higher Order Thinking Research Test for Graduates (HOTRTG)

The finalized draft of test was distributed to twenty (20) national and international subject experts, item developers and language experts out of which fourteen (14) experts returned with their expert opinions. The validity of the Higher Order Thinking Research Test for Graduates (HOTRTG) is supported by the following evidences.

i. Content validity

The developed test was validated by a panel of experts. The experts reviewed the items in the test to ensure that they were relevant to the construct of HOT and research. The content validity of the test was ensured by calculating the Content Validity Ratio (CVR) and Content Validity Index (CVI). CVR was first introduced by Lawshe (1975). Calculating the CVR and CVI makes the researcher ensure the selection of the most important and best content statistically. The panel of experts is asked to review each item for CVR based on the 3-point Likert scale: 1. Necessary 2. Useful but Not Necessary 3. Not Necessary

The CVI for each item and the overall Content Validity Index (CVI) of the questionnaire were calculated to improve the quality and ensure its validity.

The panel of experts is asked to review each item based on the 4-point Likert scale: 1- Not Relevant 2- Somewhat relevant 3- Relevant 4- Completely Relevant

Lawshe (1975) concluded that a CVR value greater than 0.51 is acceptable for fourteen (14) experts. Similarly, a CVI value of more than 0.70 is an acceptable value. Table1.3 describes that the CVR value for the test ranged from 0.571 to 1.000 and the CVI value for the test 0.774 was computed to ensure the validity of the test. Hence the results of CVR and CVI ensure that the test is highly valid and recommended to use for measuring HOT skills.

Table 3: CVR and CVI of Higher Order Thinking Research Test for Graduates (HOTRTG)

Item #	No. of Experts	No. of Necessary Items	CVR	CVI	Decision
1	14	12	0.714	0.857	Accepted
2	14	11	0.571	0.786	Accepted
3	14	11	0.571	0.786	Accepted
4	14	12	0.714	0.857	Accepted
5	14	13	0.857	0.929	Accepted
6	14	12	0.714	0.857	Accepted
7	14	11	0.571	0.786	Accepted
8	14	13	0.857	0.929	Accepted
9	14	14	1.000	1.000	Accepted
10	14	13	0.857	0.929	Accepted
11	14	12	0.714	0.857	Accepted
12	14	12	0.714	0.857	Accepted
13	14	11	0.571	0.786	Accepted
14	14	12	0.714	0.857	Accepted
15	14	11	0.571	0.786	Accepted
16	14	13	0.857	0.929	Accepted
17	14	12	0.714	0.857	Accepted
18	14	12	0.714	0.857	Accepted
19	14	11	0.571	0.786	Accepted
20	14	13	0.857	0.929	Accepted
21	14	13	0.857	0.929	Accepted
22	14	11	0.571	0.786	Accepted
23	14	12	0.714	0.857	Accepted
24	14	12	0.714	0.857	Accepted
25	14	11	0.571	0.786	Accepted
26	14	12	0.714	0.857	Accepted
27	14	13	0.857	0.929	Accepted
28	14	14	1.000	1.000	Accepted
29	14	13	0.857	0.929	Accepted
30	14	13	0.857	0.929	Accepted

c. Phase III-Pilot Testing of Higher-order thinking-based Research Test for Graduates (HOTRTG)

Initially, this test was comprised of thirty (30) MCQs and eight (8) OEQs. The pilot study was conducted on 80 graduate students of the faculty of Humanities and Social Sciences, Khwaja Fareed University of Engineering and Information Technology, R Y. Khan, Punjab, Pakistan.

i. Level of Difficulty (p)

The difficulty value of an item is defined as the proportion or percentage of the examinees who have answered the item correctly". The level of Difficulty (p) is a measure of how difficult a test item is. It is calculated as the proportion of students who answered the item correctly. For example, if 50% of students answered an item correctly, then the level of difficulty would be 0.50. The level of difficulty is an important factor in test construction. It is used to ensure that the test is challenging enough for all students, but not so difficult that it is impossible to pass. The level of difficulty is also used to compare the difficulty of different test items.

ii. Criteria for Level of Difficulty

J.P. Guilford (1956) described the criteria for the level of difficulty as:

Correct %	Item Difficulty Designation
0 – 20	Very difficult
21 – 60	Difficult
61 – 90	Moderately difficult
91 – 100	Easy

A low difficulty value index means, that item is the high-difficulty one. For example, D.V=0.20 » 20% only

answered correctly for that item. So that item is too difficult. Conversely, a high difficulty value index means, that item is an easy one. For Example, D.V=0.80 » 80% answered correctly for that item. So that item is a too-easy one

iv. Power of Discrimination (D)

According to Blood and Budd (1972), the Index of discrimination is the ability of an item on the basis of which the discrimination is made between superiors and inferiors". The distraction power of a test is concerned with the wrong options, and primarily concerned with the ability of the wrong options to attract those who do not know and fail to attract those that know. Power of Discrimination (D) differentiates between the high achiever and low achiever on the bases of responded right answers (Aulia et al., 2014; Zaidi et al., 2018). Suparman (2011) recognized that the Discrimination index is the most important method to measure the Power of Discrimination (D).

Table 4: Discrimination Index-based Item Quality Standards

Excellence	Variety	Discrimination Status	Comments
Discrimination Index	Below zero	Poor	Dropout
	Zero	No	Dropout
	0.00-0.19	Average	Retain
	0.20-0.34	Good	Retain
	Above 0.35	Excellent	Retain

Source: Kolte (2015)

v. Reliability (α)

The alpha value indicates the test reliability in terms of low, average, and highly reliable tests (Khurram et al., 2021).

Table 5: Criteria for measuring the reliability

Quality	Alpha Value	Status of Test	Remarks
Reliability of test Item (α)	0.000-0.400	Low	Not sufficient Reliable
	0.401-0.700	Average	Sufficient Reliable
	0.701-1.000	High	Good

On the bases of Table 5, the calculated Alpha value of the reliability of the test was 0.735 which indicates that the test is highly reliable and recommended for the use of measuring the HOT of graduates.

d. Item Analysis of Higher Order Thinking Research Test for Graduates (HOTRTG)

Different values like the Difficulty level of items, the Discriminating Index, and the alpha values were calculated using the ITEMAN software. ITEMAN is a statistical software program that is used to analyze test statistics.

Table 6 shows that on the bases of the calculated statistics of test items; level of difficulty and discrimination index, ten (33%) of items are below the criteria and twenty (20) 67% of items fulfill the criteria for retention. So, Item numbers 1,2,5,7,13,15,19,22,23, and 26 i.e., ten (33%) of items were dropped from the list of items of the test, and item numbers 3, 4, 6, 8, 9, 10, 11, 12, 14, 16, 17, 18, 20, 21, 24, 25, 27, 28, 29, and 30 i.e., twenty (67%) of items were finally selected for the final draft of the test.

Figure 5 indicated that on the bases of the results of level of difficulty (p) value and discrimination index (D.I), 33% of test items were below the criteria and dropped from the test and 67% of items meet the criteria, and are included in the final test.

Figure 6 indicates that 4 MCQ items were easy and 11 items were moderately difficult and 9 items were Difficult and 6 items were Very Difficult. According to the criteria for the level of difficulty of the item, a total of 10 MCQ items (4 Easy, 6 very Difficult) items were dropped from the test and 20 MCQ items were retained for the final draft of the test.

Figure 7 represents that 6 MCQ items were Poor, 2 items were No, 2 items were Acceptable, 14 items were Good and 6 items were Excellent. According to the criteria of the Discrimination Index Poor, No, and Acceptable items were below the criteria.

On the bases of the results of item analysis statistics, 10 items were dropped from the test and 20 MCQs Good and Excellent items were finally retained for the test. Initially, there were 8 OEQ items developed to measure

the highest level (6th level "Create") of revised Bloom's Taxonomy. On the bases of qualitative analysis of OEQs, 2 items were dropped and finally, 6 OEQs were retained for the final draft of the test.

Table 6: Item analysis based on difficulty level and Discrimination Index

			Item Status		
Item#	**Correct Prop**	**Disc. Index**	**Correct Prop Based**	**Disc. Index Based**	**Remarks**
1	0.91	-0.20	Easy	Poor	Rejected
2	0.93	-0.22	Easy	Poor	Rejected
3	0.67	0.28	Moderate Difficult	Good	Retain
4	0.55	0.32	Difficult	Good	Retain
5	0.15	0.00	Very Difficult	No	Rejected
6	0.80	0.30	Moderate Difficult	Good	Retain
7	0.17	-0.09	Very Difficult	Poor	Rejected
8	0.43	0.44	Difficult	Excellent	Retain
9	0.63	0.56	Moderate Difficult	Excellent	Retain
10	0.70	0.65	Moderate Difficult	Excellent	Retain
11	0.51	0.48	Difficult	Excellent	Retain
12	0.70	0.82	Moderate Difficult	Excellent	Retain
13	0.92	0.08	Easy	Acceptable	Rejected
14	0.76	0.62	Moderate Difficult	Excellent	Retain
15	0.12	0.00	Very Difficult	No	Rejected
16	0.73	0.33	Moderate Difficult	Good	Retain
17	0.50	0.27	Difficult	Good	Retain
18	0.67	0.29	Moderate Difficult	Good	Retain
19	0.14	-0.150	Very Difficult	Poor	Rejected
20	0.63	0.45	Moderate Difficult	Excellent	Retain
21	0.48	0.38	Difficult	Excellent	Retain
22	0.14	-0.24	Very Difficult	Poor	Rejected
23	0.10	0.05	Very Difficult	Acceptable	Rejected
24	0.68	0.51	Moderate Difficult	Excellent	Retain
25	0.74	0.40	Moderate Difficult	Excellent	Retain
26	0.92	-0.55	Easy	Poor	Rejected
27	0.36	0.44	Difficult	Excellent	Retain
28	0.40	0.66	Difficult	Excellent	Retain
29	0.54	0.42	Difficult	Excellent	Retain
30	0.55	0.43	Difficult	Excellent	Retain

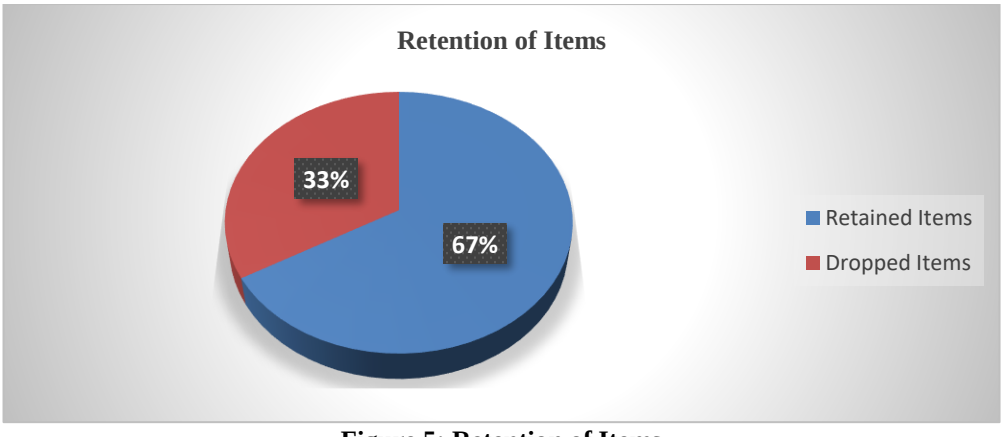


Figure 5: Retention of Items

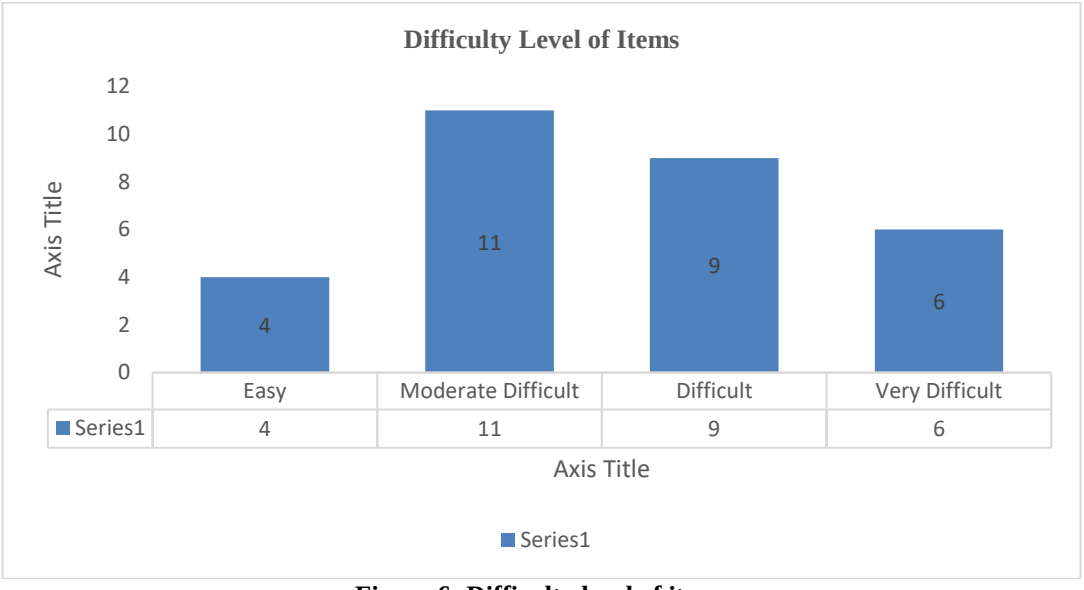


Figure 6: Difficulty level of items

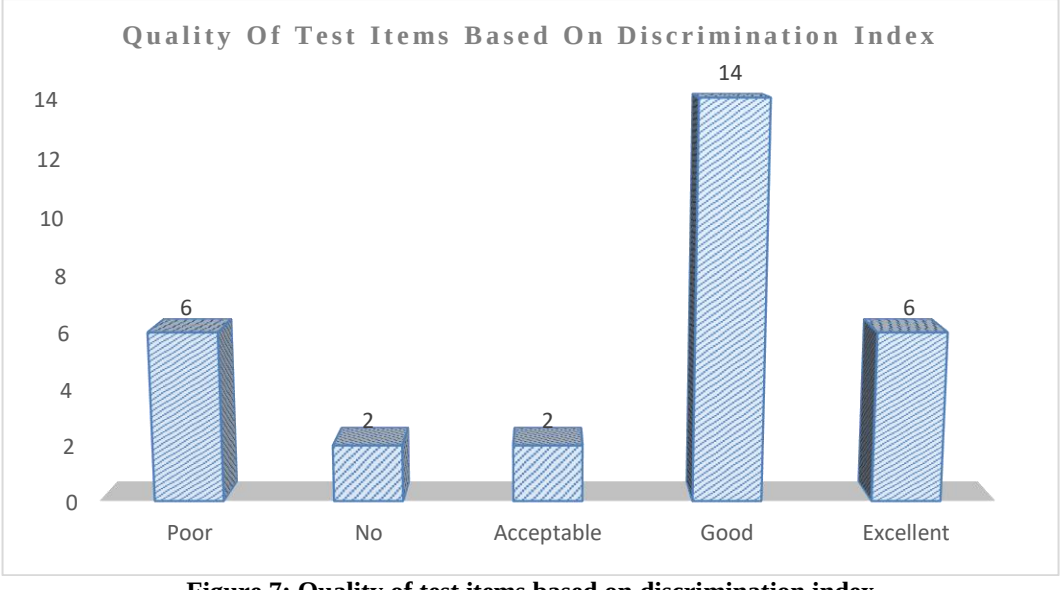


Figure 7: Quality of test items based on discrimination index

Figure 8 represents that on the bases of the main content theme wise, there are 9 items from "Concept of Research", 10 items from "Research Methodology" and 7 items from the "Data Analysis" were retained for the final draft of the test.

Hence after the validation and pilot testing process, finally 20 MCQ items and 6 OEQ items were recommended for the final draft for administration to measure the HOT of graduate students regarding the subject of Research Methods which is attached in Appendix-A.

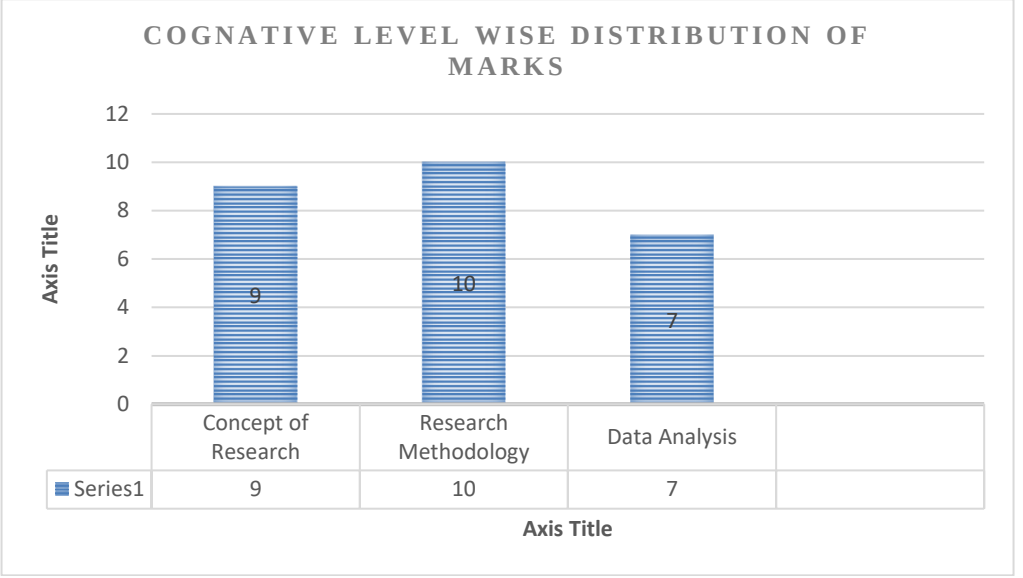


Figure 8: Cognitive level wise distribution of marks

4. Discussion

Looking toward the future, HOT will undoubtedly remain important. HOT is a fundamental skill for success in the 21st century. HOT is important in preparing students for the issues they will encounter in the future. Critical and creative thinking skills (Arooj et al., 2021; Kamran et al., 2017) are increasingly important as technology develops and automation replaces conventional tasks. HOT has a good effect on problem-solving and decision-making skills. In Pakistan, tests regarding the measurement of HOT are very hardly found in the literature. We did not come across any suitable test to measure the HOT of the students. Since specialized tests are required to measure the HOT of students, therefore, this study is conducted with the aim to develop a validated and standardized test for measuring the higher order thinking of graduate students in the subject of Research Methods at a University level in Punjab.

5. Conclusion and Implications

On the bases of the results, it is concluded that HOTRTG is ready to be used as a measurement tool for measuring the HOT of graduate students in the subject of Research Methods at a University level. The data analysis of the contents according to the levels of the cognitive domain shows that it was well articulated in the light of the scheme of studies of M.Phil. programs of social sciences (Education, English, and Islamic Studies). From the results of expert validation on each item and overall test, it is concluded that on the bases of CVR and CVI values items are eligible and the test is used as a measurement tool for measuring the HOT of graduates. The results of the level of difficulty, discrimination index, and reliability values showed that all the items of HOTRTG are at an excellent level of application.

On the bases of the results, it can have implications for the development of a curriculum for promoting HOT skills at the university level. The results of this study can be implemented for the development of standardized tests focusing on the higher levels of Bloom's taxonomy. This study will help in developing the students' learning outcomes. The results will be implemented in developing the training module for designing the process of test development for other related subjects. It is recommended that at least at the graduate level, it should be ensured that teachers are well aware of the concept of HOT skills, strategies, methods, practices, and activities that are helpful in promoting the HOT of graduates. On the bases of this study, it is recommended to conduct research on the development of training modules on teaching and measuring the HOT of graduate students.

References

Anderson, Lorin W., and David R Krathwohl. 2001. A Taxonomy for learning, teaching, and assessing. Boston: Allyn and Bacon.

Andleeb, N., Kamran, M., & Akram, H. (2022). Examination of the Demographic Variables in Promoting Creativity in Pakistan: A Follow-Up Study. *International Journal of Business and Management Sciences*, *3*(2), 35-47.

Arooj, T., Parveen, S., Iqbal, M., & Kamran, M. (2021). Proposing Creativity Inclusion in the Primary Education of Pakistan: Analysis of Educational Policy Documents and Curricula of Multiple Countries to Draw the Framework. *Turkish Online Journal of Qualitative Inquiry*, *12*(9), 3088-3095.

Arooj, T., Ameer, I., & Kamran, M. (2022). Pre-service Teachers' Cognitive or Non-Cognitive Preferences: Variance in the Learning Strategies from the Lenses of Gender Category in Balochistan, Pakistan. *Journal of Social Sciences Review*, *2*(4), 207-212.

Aulia, I. F., Sukirlan, M., & Sudirman, S. (2014). Analysis of the Quality of Teacher- made Reading Comprehension Test Items Using Iteman. U-JET, 3(4).

Gorin, Joanna S., and Svetina Dubravka. 2011. Test design with higher order cognition in mind. In Assessment of higher order thinking skills, edited by Gregory Schraw and Daniel R Robinson, 121-149, Charlotte: Information Age Publishing.

Huma Naz, L., Muhamamd Zafar, J., Farooq Ahmad Khurram, A., & Kamran, M. (2023). Analysis Of External Monitoring And Evaluation System To Propose A Rationalized Model (Instrument) For The School Education Department In Punjab-Exploratory Factor Analysis. *Pakistan Journal of Society, Education and Language*, *9*(2), 297–303.

Iqbal, M., Faizi, W. U. N., & Kamran, M. (2022). Exploring the students'most and least preferred learning strategies from the university of loralai, balochistan perspective. *Pakistan Journal of Social Research*, *4*(1), 366-372.

Kamran, M., Shah, S. A., & Rao, C. (2017). Secondary Science Teachers'views About The Placement Of Creativity In Secondary Classes: A Qualitative Study. *European Journal of Education Studies*, *3*(8), 838-851.

Kamran, M., Arooj, T., & Amjid, M. (2021a). Examination Of The Perception Level Of Teachers About The Promoters To Creativity In Pakistan: Seeing Through The Demographic Differences Of Gender, Area And Marital Status. *Pakistan Journal of Humanities and Social Sciences Research*, *4*(2), 77-88.

Kamran, M., Hayat, F., & Khan, M. (2021b). Shaping A Contextualized Theory For The Definitions Of Creativity: A Case Of Pakistani Secondary Science Teachers. *PalArch's Journal of Archaeology of Egypt/Egyptology*, *18*(10), 2306-2317.

Kamran, M., Zafar, N., & Bashir, M. (2022). Teachers' Creativity Promotion with Respect to Teaching Experience, Teaching Level, Sector and Gender vise Schooling System, and Subject Teachers Teach: Evidence from a Broader Pakistani Context. *Journal of Peace, Development and Communication*, *6*(3), 30–51.

Kolte, V. (2015). Item analysis of multiple choice questions in physiology examination. *Indian Journal of Basic and Applied Medical Research*, *4*(4), 320-326.

Lee, Jihyun, and Hyoseon Choi. 2017. What affects learner's higherorder thinking in technology-enhanced learning environments? The effects of learner factors. *Computers & Education*, 115, 143-52.

Watson, S (2019). A novel 3D in vitro model of gliobl astoma reveals resistance to temozolomide which was potentiated by hypoxia. *Journal of Neuro-oncology*, 142(2), 231-240.

Suparman, U. (2011). The implementation of iteman to improve the quality of English test items as a foreign language: An assessment analysis. *AKSARA-Jurnal Bahasa, Seni, dan Pengajarannya*, 12(1), 86-96.

United Nations Educational, Scientific and Cultural Organization (UNESCO) (2013). Shaping the Education of Tomorrow: Full Length Fontenoy

Zaidi, N. L. B., Grob, K. L., Monrad, S. U., Holman, E. S., Gruppen, L. D., & Santen, S. A. (2018). Item Quality Improvement: What Determines a Good Question? Guidelines for Interpreting Item Analysis Reports. *Medical Science Educator*, 28(1), 13-17.

APPENDIX-A

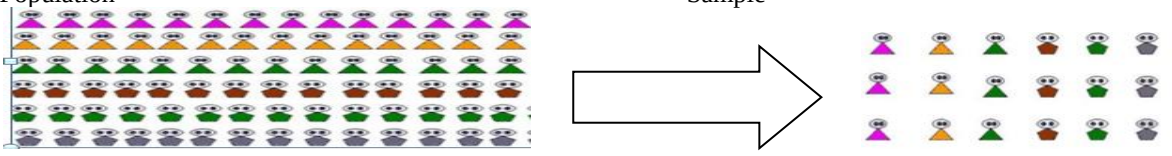
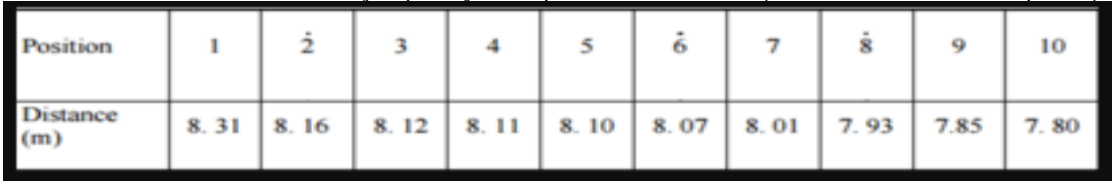
HOTRTG (*Higher Order Thinking Research Test for Graduates*)

Scholar: …… Subject: …….
Program: ……Department: …….
Institute: …….
Date: …… Time Allowed: 1:20 Hour Total Marks: 40

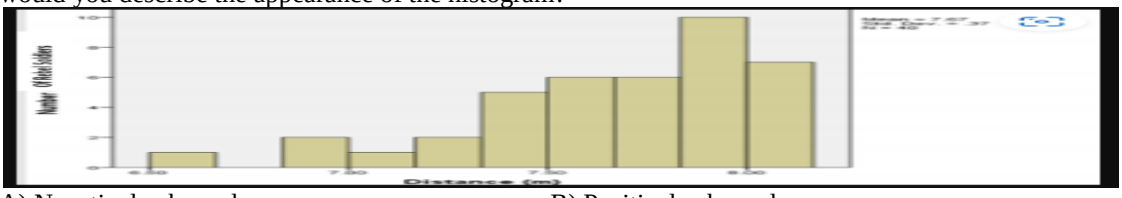
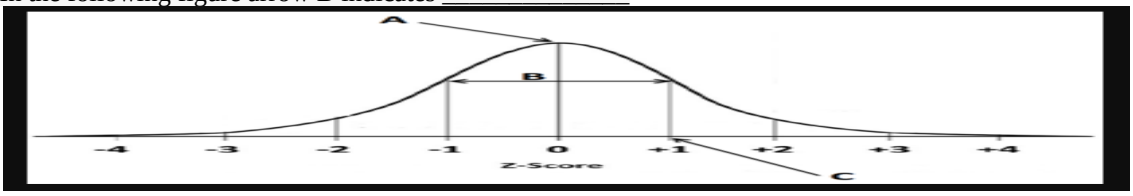
Part I

Note: Attempt all the MCQs. Each Item carries one (1) mark. Tick (√) the correct option (20)

1	A researcher conducts research on finding out which administrative style contributes more to institutional effectiveness? This will be an example of; A) Fundamental Research B) Basic Research C) Action Research D) Applied Research
2	Research to study the effect of certain policies, plans and programs is: A) Applied Research B) Descriptive Research C) Evaluation Research D) Casual Research
3	Which of the following best represents the stages of the experimental method of research? A) Research question, generate theory, hypothesis testing, data collection, data analysis, support/reject theory B) Generate theory, hypothesis testing, data collection, data analysis, support/reject theory C) Research question, generate theory, data collection, data analysis, support/reject theory D) Research question, hypothesis testing, data collection, data analysis, support/reject theory
4	Research intends to explore result of possible factors for the organization of effective mid-day meal interventions. Which research method will be most appropriate? A) Descriptive survey method B) Historical method C) Experimental method D) Ex-post facto method
5	A prediction regarding the outcomes of a study is a(n) _____ and an organized system of assumptions and principles that attempts to explain certain phenomena and how they are related is a(n)_____. A) Theory; Hypothesis B) Hypothesis; Theory C) Independent variable; Dependent variable D) Dependent variable; Independent variable
6	An educational psychologist was interested in the effects of a six-week mnemonics intervention on students' ability to remember areas of the brain such as the hippocampus. How might you define the mnemonics intervention? A) Independent variable B) Dependent variable C) Outcome variable D) Resultant variable
7	You are asked on a questionnaire to tick all the relevant boxes that apply to you on an item asking you which clubs or societies you belong to. Affiliation to a club or society would best be thought of what sort of variable? A) A hypothetical variable B) A categorical variable C) An absolute variable D) A continuous variable
8	Research design refers to the: A) Analysis of the data for the purpose of preparing the research report B) Steps necessary to define the research problem C) Suggestions made in the report about the research problem D) Plan that specify how data should be collected and analyzed for the purpose of research
9	Dr Martha Jones is interested in studying how indoor lighting can influence mood during the winter. She selects a sample of 110 households. Fifty of the homes were randomly assigned to the bright-light condition where Dr Jones replaced all the lights with 100-watt bulbs. In the other 50 houses, all the lights were changed to 60 watt

	bulbs. After two months Dr Martha Jones measured the level of depression for the people living in the houses. In this example, the level of depression is the _____ variable. A) Extraneous B) Dependent C) Independent D) Co relational
10	Which of the following is a preferred sampling method for the population with finite size? A) Systematic sampling B) Purposive sample C) Cluster sampling D) Area sampling
11	The figure below shows the type of sampling Population Sample  A) Systematic sampling B) Purposive sample C) Cluster sampling D) Stratified sampling
12	Sampling in qualitative research is similar to which type of sampling in quantitative research? A) Purposive sampling B) Quota sampling C) Probability sampling D) Stratified sampling
13	Which of the following is not true? A) Identity of respondents is known in case of schedule B) Identity of respondents is known in case of questionnaire C) A questionnaire is generally filled up by informants D) Schedule is costly than questionnaire
14	A researcher is interested in studying the prospective of a particular political party in an urban area. So, what tool should he/she prefer for the study? A) Schedule B) Interview C) Questionnaire D) Rating scale
15	When writing up a results section for most types of study it is good practice to use which ordering of detail? A) Descriptive statistics, effect size, figure/table/graph B) Figure/table/graph, descriptive statistics, effect size C) Descriptive statistics, figure/graph/table, effect size D) Effect size, figure/graph/table, descriptive statistics
16	In a population of N=6, five of the individuals all have scores that are exactly 1 point above the mean. From this information you can determine that the score for the sixth individual must be_____. A) Also, above the mean by one point B) Below the mean by one point C) Below the mean by five points D) We do not have enough information to describe the 6th score
17	Which of the following methods would you use to enter data on gender into SPSS so you can conduct statistical analyses? A) Enter data for males first and then for females B) Type male, female, or non-gender specific into the appropriate column of the data view C) Enter the data into different pages on SPSS depending on whether the answers are Male, Female or Non gender specific D) Numerically code the answers given with different numbers
18	The results of the Tauntaun long jump final at the Hoth Olympics a long time ago in a galaxy far away are given below. What is the mean distance jumped by the top 10 Rebel Soldiers (answers to two decimal places)?  A) 8.04 m B) 8.05 m C) 8.06 m D) 8.07 m

Position	1	2	3	4	5	6	7	8	9	10
Distance (m)	8. 31	8. 16	8. 12	8. 11	8. 10	8. 07	8. 01	7. 93	7.85	7. 80

19	Examine the histogram of the qualifying distance for the Tauntauns long jump final at the Hoth Olympics. How would you describe the appearance of the histogram?  A) Negatively skewed B) Positively skewed C) Positive kurtosis D) Negative kurtosis
20	In the following figure arrow B indicates _____  A) 50.00% of scores B) 34.13% of scores C) 68.26% of scores D) 95.44% of scores

Part II Marks:20

1- Create a hypothesis for the following research question:
What effects do instructional methods have on higher order thinking of graduate students? (2)

Hypothesis: …….

2- A researcher intended to determine the Effect of (3) Instructional Methods on Graduate Students' (3) Higher Order Thinking skills". Develop a conceptual framework for the study (5)

3- Suggest an appropriate research design for the study entitled "Effect of instructional methods on graduate students' higher order thinking in research" (3)

Research Design: …….
Justification:……

4- Draw the diagrams to differentiate between stratified and cluster sampling. (4)

Stratified Sampling	Cluster Sampling

5- A researcher wants to determine the use of Instructional Methods and activities for promoting the Higher Order Thinking of Graduate students. He/ She want to use a Questionnaire covering of six (6) constructs (Problem Solving Method, Inquiry Method, Group Discussion Method, Project/Assignment Method, HOT Activities, and Research Skills). **Develop at least three statements for anyone construct. (3)**

Responses: 1 Not at all, 2 Rarely, 3 Occasionally, 4 Frequently, 5 Always

Statements / Instructional Methods		**Level of Frequency**				
		1	**2**	**3**	**4**	**5**
1						
2						
3						
4						
5						

6- Critically interpret the results/ findings of the table given below. (3)

Table: I have the skill of selecting an appropriate research design for a specific study

No	Theme	Stat.	Responses					Total	SD	Mean
			SDA	DA	UD	A	SA			
CM7	Level of Agreement	F	9	17	17	129	112	284	0.98	4.12
		%	3	6	6	45	40	100		
	Need for Training	F	10	40	21	171	42	284	1.00	3.69

Interpretation:……
……
……
……