UCP Journal of Languages & Literature
Vol. 1, Issue 2 (July - December 2023)
Journal website: http://ojs.ucp.edu.pk/index.php/ucpjll

An Estimation of Test-Retest Reliability of the Verbal Section of Entrance Test for M.Phil at the University of Azad Jammu & Kashmir, Muzaffarabad

Nouman Hamid[1*], Sehrish Shafi[2]

Abstract

This study examines the consistency of test scores as a gauge of test takers' performance on the verbal section of two editions (administrations) of the MPhil entrance exam at the University of Azad Jammu & Kashmir. The objective was to find out the extent to which test takers performed similarly in the two administrations of the same test. For this purpose, the verbal section of University Graduate Aptitude Test (UGAT) was administered to thirty students twice with an interval of eighteen days. The results indicate that students' overall performance on the test was constant between the two administrations. However, on each of the five activities in the verbal portion of the test, students' performance was less consistent as compared to their performance in the overall test. The results have important implications for future research and pedagogy.

Keywords: Entrance exam, test-retest reliability, university graduate aptitude test, verbal section

Introduction

Reliability is one of the most important characteristics of test usefulness (Bachman & Palmer, 1996) because a) it contributes in establishing the validity of test scores and b) it indicates the extent to which various kinds of generalizations are justifiable (Miller, Linn, & Gronlund, 2009). Reliability or *Reproducibility* (Fulcher & Davidson, 2007) refers to the degree to which test scores or other assessment results are consistent from one measurement to another (Bachman & Palmer, 1996; Fulcher, 2010; Mackey & Gass, 2005; Miller, Linn, & Gronlund, 2009). In

[1] Assistant Professor, Department of English, University of Azad Jammu and Kashmir, Muzaffarabad
[2] Assistant Professor, Department of English, Mirpur University of Science and Technology, Mirpur
[*]Corresponding author's E-mail: noumanhamid@gmail.com
Available online: 25-03-2024

assessment, reliability refers to the consistency of scores, not of the instrument itself as the same instrument may produce different reliability indices depending on the test takers and the situation in which it is used. Understanding the variation in retesting is essential for a variety of reasons, including evaluating the reliability of measures, expectations of validity and prediction, and providing bounds and reasonable expectations within which test preparation may have an impact, among other situations in which tests are used for hiring decisions or for more in-depth research (Lee & Wai, 2023). Moreover, reliability is an essential but not sufficient condition for validity. It is possible that consistent results may not be measuring what they claim to measure but for a test to provide valid information about the performance that it measures, it is necessary that the test yields consistent scores (Miller, Linn, & Gronlund, 2009). Furthermore, reliability is assessed through correlation coefficient which ranges from -1 to +1. A correlation coefficient of +1 means a strong positive relationship between two sets of scores whereas -1 means strong negative relationship and 0 means no relationship (Dörnyei, 2007; Fulcher, 2010).

Reliability or consistency is of different types, including test-retest (consistency across time), equivalent forms (consistency across different forms of the same test), interrater and intra-rater (across raters), and internal consistency (across tasks that measure the same thing) (Mackey & Gass, 2005). Of these, the present study is restricted to the test-retest reliability. To estimate test-retest reliability, the same test is administered to the same group of individuals at two points in time with a considerable time interval between the two administrations. The scores are correlated and the resulting correlation coefficient provides a measure of stability which means how consistent the scores are over the given period of time. A high correlation coefficient means that the individuals with high scores on test 1 will tend to score high on test 2 and vice versa. An important consideration in establishing and interpreting test-retest reliability is the time interval between the two administrations. If time interval is short, the scores will exhibit higher consistency. On the other hand, the consistency will be lower in case of longer interval of time (Mackey & Gass, 2005; Miller, Linn, & Gronlund, 2009).

Although test-retest reliability provides important information about the stability of assessment scores across time, it is surprising to find that most standardized tests do not provide test-retest reliability coefficients. Brooks (2005) provides the test-retest reliability of *Bilingual Syntax Measure* (BSM) for a sample of test takers who took the English and

Spanish versions of the test two times with an interval of approximately two weeks. A correlation coefficient of 0.62 was found for the two administrations of the English version of the test, while a coefficient of 0.64 was reported for that of the Spanish version. Both correlation coefficients show less stability of scores across time.

Association of Classroom Teacher Testers (USA) publishes a test called *Combined English Language Skills Assessment in a Reading Context* (CELSA). It has two versions, that is, a paper-based version and a computer-based version. Test-retest reliability estimate was 0.93 for the paper-based version and 0.96 for computer-based version (Strong-Krause, 2005). However, the time interval between the two administrations of both versions of the test is not mentioned in the study.

Educational Testing Service (ETS) provides an estimate of test-retest reliability with equivalent forms for *Test of English as a Foreign* Language (TOEFL) based on the analysis of candidates repeating the test within 30 days from January 2007 to August 2007. Reliability estimate was 0.91 for the whole test, 0.78 for reading, 0.77 for listening, 0.84 for speaking, and 0.77 for writing (http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_iBT_Reliability.pdf).

All the above studies report test-retest reliability estimates for language tests. No published research on test-retest reliability could be located for aptitude tests like *Graduate Record Examination* (GRE General) published by ETS or *Graduate Assessment Test* (GAT General) published by *National Testing Service, Pakistan* (NTS): A national testing agency mandated by *Higher Education Commission, Pakistan* (HEC) to conduct aptitude tests for admissions to post-graduate degree programs from 2005 to September 2014. After Lahore High Court's verdict that the universities are not under any obligation to be bound by the results of NTS in offering admissions at post-graduate levels, one of the alternatives given to the universities by the HEC was that universities could conduct their own tests. The University of Azad Jammu and Kashmir, Muzaffarabad (UAJK) conducted its first entrance test in September 2014. The test contained three parts, namely verbal section, quantitative section, and analytical reasoning. The researchers of the present study were part of the team that developed the verbal section. The purpose of this experimental study is to find out the extent to which the verbal section of the test produced consistent scores when administered twice to the same group of individuals with an interval of 18 days. This paper attempts to answer the following research questions.

a) How stable is test-takers' performance on the two administrations of the verbal section of the entrance test?

b) How stable is test-takers' performance on various tasks in the verbal section of the entrance test across time?

Methodology

Participants

Participants included 30 students who were in the final semester of MA English at the University of AJ&K. They were appropriate for this study as they were on the verge of graduating and many of them would opt for a test like this in order to get admission in M.Phil. Although the ideal sample for this study would have included participants from different disciplines, the participants with only English major were selected due to the reason that it was not possible to gather students from various departments at one place twice and conduct the test since academic departments in UAJK are located in different parts of District Muzaffarabad. There were 48 students in the final semester of MA English but only 30 volunteered to participate in the study. Out of these 30, there were 10 males and 20 females.

Development of the Test

A five-member committee was constituted by the competent authority to construct items for the verbal section. The committee in its first meeting decided on the number of items, types of tasks, and number of items to be included in each task. The committee mandated the researchers (us) to construct a pool of items for the verbal section. In the next meeting, the researchers presented the constructed test items before the committee. After a thorough discussion, the committee finalized the items that were to be included in the test. Due to the lack of time, the test was not piloted for item analysis.

Description of the Test

The verbal section of the entrance test consisted of 40 multiple choice items. Test-takers were given 50 minutes to respond to these 40 items. This section contained five sub-tasks. In what follows is the description of these tasks (See Appendix for the test.).

There were 11 items in the first task. Each item comprised a sentence with a blank space. Below each sentence, four options were given. For

each item, test-takers were required to choose the most appropriate option to fill in the blank. Task two consisted of 8 items. Each item consisted of a word followed by four synonyms. Students were to select the most appropriate synonym for the given word. For task three, test-takers were provided with 6 items. In each item, there was a pair of words followed by four pairs of words. For this task, students had to choose the pair that exhibited the closest terms of relationship as that of the given pair. Task four contained 7 items. Each item consisted of a word followed by four antonyms. Students were to select the most appropriate antonym for the given word. The final task included 8 items. In this task, students were given a reading passage followed by 8 multiple-choice questions/items based on that passage. Students were supposed to answer the questions based on their comprehension of the passage.

Procedures

The first administration of the test was held at room 5, block B, City Campus, UAJK, Muzaffarabad on Wednesday, December 03, 2014. The test was administered by one of our colleagues. Test-takers were requested to respond to the test items with the best of their knowledge and effort. They were also informed that their identities will be kept confidential and their scores will not be shared with anyone other than the researchers. They were told that they were required to take another test (they were not told that it will be the same test) two weeks later (i.e., December, 17). However, due to some unavoidable circumstances, the test was not conducted before Monday, December 22, 2014. The interval of eighteen days (originally fifteen days) between the two administrations was deemed appropriate by the researchers as it followed the suggestion of Miller, Linn, and Gronlund (2009) that the interval between the two administrations should not be too short or too long. The second test was administered at the same place by the same colleague. Both tests were scored by the researchers themselves with the help of an answer key. Each correct item was awarded one point and 0 was given to each incorrect response. Then, cumulative scores were given for each task by adding up the points given to each item belonging to that task. In the end, Scores on all tasks were added to give an overall score for the whole test (out of forty). The data were keyed in SPSS for analysis.

Results

Results of the two administrations of the verbal section of entrance test show that test-takers performed slightly better in test 2 as compared to

test 1 as the overall mean score for test 2 (21.93) is slightly higher than the mean score for test 1 (21.10). The same trend can be observed for four out of five tasks, that is, Fill in the blanks (FITB), word synonyms (WS), word antonyms (WA), and reading comprehension (RC). Table 1 presents the descriptive statistics for two test scores.

Table 1 *Descriptive Statistics*

	N	Range	Minimum	Maximum	Mean	SD	R	SEM
T_T1	30	21	10	31	21.10	4.84	0.63	2.94
T_T2	30	22	10	32	21.93	5.67	0.70	3.11
FITB_T1	30	8	3	11	6.70	1.84		
FITB_T2	30	8	3	11	6.93	2.03		
WS_T1	30	5	1	6	3.40	1.25		
WS_T2	30	6	0	6	3.87	1.33		
Anl_T1	30	6	0	6	3.10	1.32		
Anl_T2	30	5	1	6	2.83	1.34		
WA_T1	30	7	0	7	4.30	1.53		
WA_T2	30	7	0	7	4.37	1.97		
RC_T1	30	7	1	8	3.60	1.63		
RC_T2	30	5	2	7	3.93	1.57		

Table 1 clearly shows an increase in mean score from 6.70 to 6.93 for FIIB, from 3.40 to 3.87 for WS, from 4.30 to 4.37 for WA, and from 3.60 to 3.93 for RC. On the other hand, Analogy (Anl) is the only task on which test-takers' performance decreased from test 1 to test 2 (3.10 to 2.83). Similarly, there is also an increase in the standard deviation of scores for test 2 (5.67) as compared to the standard deviation of scores for test 1 (4.84), which means that test 2 scores show more variability as compared to test 1 scores. Furthermore, scores on four out of five tasks also exhibit this trend with WA showing the largest variability for test 2 scores. However, RC displays more variability for test 1 scores (1.63) as compared to test 2 scores (1.57). Internal consistency reliability coefficient (Chronbach's alpha) was found to be 0.63 for test 1 and 0.70 for test 2. As the Standard Error of Measurement (SEM) for test 1 is 2.94, it means that test-takers' true score range within one unit of SEM for test 1 would be 18.16-24.04. With an SEM of 3.11, the true score range within one unit of SEM for test 2 would be 18.82-25.04.

In order to establish the degree to which scores were consistent from one administration to the other, Pearson product-moment correlation coefficients were computed for the overall test scores as well as for the scores on five tasks. Table 2 presents a summary of the findings.

Table 2 *Test-retest Reliability*

	Test 1 + Test 2	Sig.
Total	0.80	.000
FITB	0.52	.003
WS	0.34	.063
Anl	0.59	.001
WA	0.54	.002
RC	0.67	.001

Table 2 displays a correlation coefficient of r (28) = 0.80, p = .000 for the overall test scores. For the sub-parts, the table shows significant correlation for four out of five tasks. For FITB, the correlation coefficient is r (28) = 0.52, p = .003, for Anl, it is r (28) = 0.59, p = .001, for WA, it is r (28) = 0.54, p = .002, and for RC, it is r (28) = 0.67, p = .001. For WS, the correlation, r (28) = 0.34, p = .063, is not significant as the significance value exceeds .05.

Discussion and Conclusion

The present study aimed at finding out the extent to which the verbal section of the entrance test for M.Phil at the UAJK produced consistent scores when administered twice to the same group of individuals with an interval of 18 days. For this, the study set out to find answers to the following research questions.

a) How stable is test-takers' performance on the two administrations of the verbal section of the entrance test?

b) How stable is test-takers' performance on various tasks in the verbal section of the entrance test across time?

For the first question, this study shows that students' performance was highly stable on the two administrations of the test as indicated by the correlation coefficient of 0.80 for the overall test scores. Measures of stability in the 0.80 range are commonly reported for standardized tests of aptitude and achievement over occasions within the same year (Miller, Linn, & Gronlund, 2009, p, 110). A glance at the mean scores for overall tests also reveals that there is not much difference in the performance of test-takers on the two tests. Moreover, as shown in table 1, both tests exhibit moderate internal consistency reliability coefficients (alpha = 0.63 for test 1 and alpha = 0.7 for test 2), which show that the scores on sub-parts of the test are also reliable as an alpha coefficient between 0.6

and 0.8 is considered acceptable for non-standardized tests (Morgan, Leech, Gloeckner, & Barrett, 2007; Subkoviak, 1988).

Although test-retest reliability coefficient for this experimental study is higher than that reported for BSM by Brooks (2005), it is much lower than that for TOEFL and CELSA. The reason may be that these standardized tests are constructed, piloted, edited, and revised by assessment experts, whereas the test for this study did not undergo these stages and also because this was the very first test of its kind that was developed at UAJK. Considering these facts, it can be safely concluded that the test-takers' performance was highly stable across the two administrations.

As far as the second research question is concerned, the data show that students' performance was not highly stable on any of the five tasks. Table 2 reveals that students' performance was moderately stable on only one task (i.e., RC). For FITB, Anl, and WA, their performance was somewhat stable but their performance on the WS part lacked stability. There could be many possible reasons for this lack of stability, including increased learning during the interval between the tests, anxiety, lack of motivation, mood, and the possibility of guessing --- to name just a few.

The results of the present study have important implications. For assessment practitioners, the study shows that the true performance of test-takers cannot be judged by the single administration of a test. It is very important to establish the extent to which learners' performance is consistent or improves over time. They should not make decisions regarding students' performance on the basis of a single test but rather on the basis of multiple administration of a test. Future studies can take participants from various disciplines as well as conduct detailed analysis of the factors that contribute in consistency or inconsistency in test-takers' performance across two administrations of the same test. Moreover, to better understand students' experiences and enhance the test, longitudinal studies, rewriting test components, and qualitative research may all be included in future research.

Reference

Bachman, L., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests.* New York: Oxford University Press.

Brooks, L. (2005). Bilingual syntax measure (BSM) I and II. In S. Stoynoff & C. Chapelle (Eds.), *ESOL tests and testing* (pp. 38-42). Alexandria, VA: TESOL.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford: Oxford University Press.

Educational Testing Service. (2008). Reliability and comparability of TOEFL iBT scores. Retrieved from http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_IBT_Reliability.pdf.

Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource Book.* New York: Routledge.

Mackey, A., & Gass, S. (2005). *Second language research: Methodology and design*. New York: Routledge.

Miller, M., Linn, R., & Gronlund, N. (2009). *Measurement and evaluation in teaching.* Upper Saddle River, NJ: Merrill, Prentice Hall.

Morgan, G., Leech, N., Gloeckner, G., & Barrett, K. (2007). *SPSS for introductory statistics:Use and interpretation.* London: Lawrence Erlbaum Associates Publishers.

Strong-Krause, D. (2005). Combined English language skills assessment in a reading context (CELSA). In S. Stoynoff & C. Chapelle (Eds.), *ESOL tests and testing* (pp. 57-58). Alexandria, VA: TESOL.

Subkoviak, M. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25,* 47-55.

Appendix
The University of Azad Jammu & Kashmir

Time Allowed: 50 minutes

Note: Cutting and/or over-writing will result in the loss of points.

Directions for questions 1 –11: In each sentence given below, there is a blank space. Below each sentence four options, marked (a), (b), (c) and (d) are given. Out of these, choose the most appropriate option to fill in the blank.

1 - Alexander Fleming, _____, received the Nobel Prize in 1945.

a) who discovered Penicillin
b) which discovered Penicillin
c) he discovered Penicillin
d) that discovered Penicillin

2 - In the ear, just above the cochlea, there are three small semi-circular canals _____ as an organ of balance.

a) that function together
b) function together
b) are functioning together
d) they function together

3 - _____ riding a bicycle is good leg exercise; it does not use up a lot of calories.

a) as
b) although
c) because
d) so that

4 - She turned off the tape recorder _____ she could study.

a) now that
b) even if
c) so that
d) in case

5 - Pearl Bock, who was the first American woman to win a Nobel Prize for Literature, _____ for her novel Good Health.

a) who is best known
b) which is best known
c) is best known
d) best known

6 - It was the impact of railroad _____ agriculture in the west.

a) it expanded
b) expanded
c) that it had expanded
d) that expanded

7 - _____ to the issuance of stamps, letters were marked ‘paid’ by pen & ink or hand stamps.

a) in accordance b) before c) due
d) prior

8 - Fredrick J Turner, _____, argued that the Frontier shaped a distinctive way of life.

a) a famous American historian who b) a famous American historian
c) despite a famous American historian d) he is a famous American historian

9 - Solar Beat Penetrates more deeply into water than _____.

a) it is penetrating into the soil b) it does into soil
c) does it into soil d) that it does into soil

10 - The more distant a star happens to be, _____ to us.

a) the dimmest it seems b) the dimmer it seems
c) it seems dimmer d) it seems dimmest

11 - The research conducted by the students was so _____ that every detail was included in their report.

a) definite b) complete c) comprehensive
d) specific

Directions for questions 12 – 19: Select the word which is similar in meaning to the given word.

12 - NEPOTISM

a) Favoritism b) Classicism
c) Nationalism d) Socialism

13- PLAGIARISM

a) Theft of funds b) Theft of ideas
c) Belief in God d) Ethical theory

14- MENTOR

a) Guide b) Genius
c) Philosopher d) Stylist

15- FRANCHISE

a) Discount b) License
c) Reason d) Fashion

16- TENTATIVE

a) Prevalent b) Experimental
c) Mocking d) Provisional

17- ENSUE

a) Come afterward b) Precede
c) Introduce d) Praise

18- TURMOIL

a) Peace b) Destruction
c) Order d) Disturbance

19- MESMERIZE

a) Remember b) Delay
c) Hypnotize d) Analyze

Directions for question 20– 25: Select the pair of words that show the closest terms of relationship as that of the given pair.

20 – SECOND: MINUTE

a) Year: Month b) Centimeter: Inch
c) Mile: Hour d) Day: Week

21 – AUTHOR: WRITE

a) Mother: Annoy b) Janitor: Clean
c) Banker: Preach d) Firefighter: Study

22 – BROKEN: REPAIRED

a) Foreign: Strange b) Unusual: Weird
c) Selfish: Generous d) Nasty: Mean

23 – ZEBRA: ANIMAL

a) Green: Colour b) Shape: Cube
c) Country: Brazil d) Food: Dessert

24 – EVIL: MORALS

a) Special: Manners b) Famous: Friends
c) Blind: Sight d) Beautiful: Opinions

25 – RACIST: PREJUDICED

a) Clerk: Hungry b) Nurse: Sick
c) Worker: Tired d) Saint: Holy

Direction for questions 26 – 32: Select the word which is opposite in meaning to the given word.

26 - MAGNIFICENT

a) Big b) Small
c) Unimpressive d) Gigantic

27 - ARROGANT

a) Snooty b) Stupid
c) Cunning d) Humble

28 - FILTHY

a) Clean b) Nice
c) Dirty d) Pretty

29 - GLOOMY

a) Dark b) Cheerful
c) Impressive d) Ugly

30 - OPTIMIST

a) Naturalist b) Realist
c) Pessimist d) Activist

31 – ACCIDENTAL

a) Surprising b) Sudden
c) Incidental d) Deliberate

32 – PROFESSIONAL

a) National b) Amateur
c) Mature d) Traditional

Directions for question 33-40: Read the following passage carefully and answer the questions by referring to it.

PASSAGE

It is bad to have food stuck between your teeth for long periods of time. This is because food attracts germs, germs produce acid, and acid hurts your teeth and gums. Flossing helps to remove the food that gets stuck between your teeth. This explains why flossing helps to keep your mouth healthy, but some doctors say that flossing can be also good for your heart.

It may seem strange that something you do for your teeth can have any effect on your heart. Doctors have come up with a few ideas about how flossing works to keep your heart healthy. One idea is that the germs that hurt your teeth can leave the mouth and travel into your blood. Germs that get into the blood can then attack your heart. Another idea is based on the fact that when there are too many germs in your mouth, the body tries to fight against these germs. For some reason, the way the body fights these mouth germs may end up weakening the heart over time.

Not every doctor agrees about these ideas. Some doctors think that the link between good flossing habits and good heart health is only a coincidence. A coincidence is the occurrence of two or more events at one time apparently by mere chance. The incidence of these events is completely random, as they do not admit of any reliable cause and effect relationship between them. For example, every time I wash my car, it rains. This does not mean that when I wash my car, I somehow change the weather. This is only a coincidence. Likewise, some doctors think that people who have bad flossing habits just happen to also have heart problems, and people who have good flossing habits just happen to have healthy hearts.

The theory that flossing your teeth helps to keep your heart healthy might not be true. But every doctor agrees that flossing is a great way to keep your teeth healthy. So even if flossing does not help your heart, it is sure to help your teeth. This is enough of a reason for everyone to floss their teeth every day.

33 - Which of the following would be the best title for this passage?
a) Why Doctors Disagree about Flossing
b) How to Keep Your Teeth Healthy
c) Flossing Your Way to a Healthy Heart
d) Flossing by Coincidence

34 - Flossing effectively helps to keep your mouth healthy by preventing
a) germs from producing acid
b) food from entering your body
c) germs from entering into your blood
d) acid from contacting your teeth and gums

35 - In paragraph 2, the author introduces ideas about how flossing works to keep your heart health y. Exactly how many of these ideas does the author put forth in this paragraph?
a) 1 b) 2 c) 3 d) 4

36 - In paragraph 2, the author explains how having too many germs in your mouth can "end up weakening the heart." Using the passage as a guide, it can be understood that with respect to the actual way in which this occurs, doctors are
a) reluctant to hypothesize
b) confident in their estimations
c) extremely knowledgeable
d) uncertain but speculative

37 - In paragraph 3 the author writes, "Not every doctor agrees about these ideas." The author's purpose in writing this sentence is to
a) provide an example
b) introduce a new topic
c) change a previous statement
d) clarify an earlier assertion
38 - Using information in paragraph 3 as a guide, which of the following is the best example of a **coincidence**?

a) Jim wakes up with a sore throat. He eats a piece of bacon for breakfast. By noon, he feels much better. Jim decides that the bacon has cured his sore throat.
b) Laura remembers to brush her teeth every day, but she only remembers to floss once a week. She writes a note to herself, reminding herself to floss and sticks it to her bathroom mirror.
c) Mario is not very good at baseball. He practices playing every day. After a several months of practice, he is a much better baseball player.
d) Jai has a bad heart. Her doctor tells her to eat more vegetables and less junk food. After nearly a year of doing this, the doctor tells Jai that her heart is doing much better.

39 - Based on its use in the final paragraph, it can be inferred that **theory** belongs to which of the following word groups?
a) query, question, interrogation
b) assertion, declaration, affirmation
c) hypothesis, supposition, belief
d) idea, thought, notion

40 - Which of the following best states the main idea of the final paragraph?
a) Because doctors do not agree that flossing will help your heart, it is useless to floss.
b) It is a fact that flossing can help your heart as well as your teeth.
c) Even if flossing is only good for your teeth, you should still do it every day.
d) There is no good reason to believe that flossing will help your heart, but it is still a good idea to do it every day.