TESTING TEST VALIDITY AND RETENTION THROUGH THE WASHBACK EFFECT: A CASE STUDY

Igballe Miftari-Fetishi
International Balkan University, Skopje, North Macedonia, i.miftari@ibu.edu.mk

Abstract: The teaching process consists of many different components, from planning and design to teaching, assessment and feedback. The whole process goes hand in hand with the methodology and the sequence of activities that are presented and a huge part of the process lies in the effectiveness of the teaching i.e. students results at the end. This paper elaborates the issues of teaching, assessment and the ongoing process of learning. The crucial components of the process lie in the effectiveness of our teaching methodology, theory and of course, the average grade points that students achieve. The overall success achieved by students is the reflection of how well all was managed and of course, how much was in fact learned. In order to test the validity of the teaching methodology and assessment compared to the knowledge acquired, the washback effect was used. Namely, the process included 23 third year BA students of English language teaching at IBU. The case study tested test validity, student knowledge and the washback effect after an approximate 6 weeks of lectures, in the form of a Quiz. The Quiz contained 10 essay-form questions related to the topics which were previously discussed. The total time needed for the case study was 9 weeks, including 6 weeks of instructional teaching, the Quiz week, Retention week and the Washback effect week. Research questions include: "Do teachers teach to the test, and if so, are the obtained results better?" "Does score consistency prove retention of knowledge?""Can test validity be tested through the washback effect? The procedure in fact, proved that there was a discrepancy in the results, namely, some students did better, some gained less points and only a few remained the same in their average score i.e. half of the students received the same group of the Quiz during washback week, while half of the students received a different version of the Quiz (compared to the prior). The main aim was to identify the issue of retention and mechanical learning i.e. the extent to which information was remembered or the extent to which learners reflected on both Quiz versions after the actual procedure (they had one week of retention to do so).The overall results, discussion, limitations and future recommendations will be provided further in the paper.
Keywords: testing, validity, washback effect, assessment, retention etc.

1. INTRODUCTION

Testing and assessment are very broad terms and as so, not only are they difficult to establish but are difficult to monitor, distribute and evaluate. Both terms are used interchangeably however, as much as they have in common, they differ. Testing is referred to as the pen and pencil type of traditional grading, which is known and used by teachers for centuries now. Assessment, on the other hand, is the ongoing, systematic process of the gathering of information and is the alternative to testing. As so, both provide a grade or score at the end, however, with testing, the gained results refer to the material up until a given point and there is no second chance or improvement (then and there result). With assessment, the students have the opportunity to improve, show more skills and abilities and progress by the end (even though, there are cases when there is a decline in progress (this will be explained later). As stated by McNamara (2000), "…the very nature of testing has changed quite radically over the years to become less impositional, more humanistic, conceived not so much to catch people out on what they do not know, but as a more neutral assessment of what they do. Newer forms of language assessment may no longer involve the ordeal of a single test performance under time constraints. Learners may be required to build up a portfolio of written or recorded oral performances for assessment. They may be observed in their normal activities of communication in the language classroom on routine pedagogical tasks. They may be asked to carry out activities outside the classroom context and provide evidence of their performance…" (McNamara, 2000, pg.4)

Issues that occur with testing relate to dissatisfactory results at times or even failure. The pressure that is felt by the learners may have a negative impact on the overall scores, which differs with ongoing assessment, which is rather informal compared to the formality of testing. In both circumstances, teachers need to teach, prepare for and evaluate theirlearners competencies. The effect that testing or assessment might have on the learners is known as the washback effect. According to McKinley & Thompson (2018) "washback effect refers to the impact of testing on curriculum design, teaching practices, and learning behaviors. The influences of testing can be found in the choices of learners and teachers: teachers may teach directly for specific test preparation, or learners might focus on specific aspects of language learning found in assessments. (*McKinley & Thompson (Feb 2018). "Washback Effect in Teaching English as an International Language" (PDF).* According to Martel (2019), "washback, that is, the effect of tests on teaching and learning, has captured the interest of scholars in applied linguistics for almost 30 years.

However, most research has been conducted on large-scale tests, rather than classroom-based language assessments…" (2019, pg. 57)

In another study, Pitoyo et al. (2020) explored the washback effect of quizzes on students' learning. To collect the data, the researcher used a questionnaire, observation, and in-depth interview and analyzed qualitatively. The result of the study shows that students were motivated and wanted to learn more after doing integrated gamified tests and quizzes.

This paper hereby "tests" the issue of validity and retention after a period of time, with this, pointing out the positive and negative effects of only testing and not assessing learners' development. The procedure consists of 6 week lectures i.e. six chapters of course material and reviews after every chapter. The content is therefore, predicted, however, is not taught to the exam. The learners are providedwith extra materials as well as with additional explanations along the period. The main aim is to check learners' concentration, motivation and retention of the concepts and their implementation along the course. As stated by Abolfathi et. al (2022) "all summative assessments were treated very seriously by students who spent a considerable time on even small pieces of assessments such as weekly online quizzes with a very low weighting, however, they perceived completing harder and longer assessments contributed relatively less to their learning. This mismatch in the requirements (a lot versus a little) compared to the actual motivation on learning that students have is another issue- does this occur due to the idea that less requirement equals lack of seriousness? Or is too much material and information more difficult to retain? Another issue that should be mentioned here is the cultural issue, namely, the degree to which the stress is put on the teachers or vice versa and language barrier on the other hand. In a study by Lutfiano et al (2020) regarding cultural and language barriers, the results show that language barriers had an impact on the students" academic performance, while the cultural barriers on the other hand, did not have any impact on the students' academic progress.

2. INSTRUMENTS AND PARTICIPANTS

In order to test the validity of the teaching methodology and assessment compared to the knowledge acquired, the washback effect was used. Namely, the case study included 23 third year BA students of English language teaching at IBU. The case study tested test validity, student knowledge (retention) and the washback effect after an approximate 6 weeks of lectures, in the form of a Quiz. The Quiz contained 10 essay-form questions related to the topics which were previously discussed. The total time needed for the case study was 9 weeks, including 6 weeks of instructional teaching, the Quiz week, Retention week and the Washback effect week.

Research questions include:

- "Do teachers teach to the test, and if so, what are the overall obtained results?"
- "Is ongoing assessment better than testing?"
- "Can test validity be tested through the washback effect?

3. THE ISSUE OF TEST VALIDITY (THE PROCESS)

The process included the Quiz which referred to the 6 week course material and topics which were covered. The questions were essay form, open-ended questions, to which students were to reply to authentically i.e. expressing themselves and the knowledge they had stored. Prior to the Quiz, student participants were asked whether or not they believed that the Quiz was valid (A "valid" test is one which actually tests what it is designed to test. (Penny, Ur, A course in Language Teaching, 2024. New 3rdedition)

and whether the acquired knowledge was applicable i.e. whether the test had face-validity. ("Face validity" refers to the degree to which a test looks right, and appears to measure the knowledge or ability it claims to measure, based on the subjective judgment of the examinees who take it, the administrative personnel who decide on its use, and other psychometrically unsophisticated observers." (Mousavi, 2002, p. 244).

During this pre-task phase, students responded to with *yes* i.e. asserting do-ability and relevance as being part of the Quiz.

4. OBTAINED RESULTS AND DISCUSSION

The whole procedure of events included a 9 week process, by which the Quiz, retention week and the washback effect would test the validity issue and the effectiveness of the teaching/learning. Along the weeks, plenty of content drills, review and discussions were made, where students actually had the opportunity to discuss relevant issues and of course, retain information.

PHASE ONE: THE QUIZ WEEK

During the Quiz week, only 23 out of 26 students were present i.e**.** 3 students did not take the Quiz. The results were as follows:

5 students failed the Quiz (gained less than 50pts.)

7 students passed (gained 50-60 pts.)
11 students gained higher points than a passing grade (60-100) (this result is relatively good due to the fact that almost half of the group gained higher points than a passing grade. In case there is a consistency in results, we would automatically claim that the information is in fact retained and that students are continuously learning).

5. PHASE TWO: THE WASHBACK EFFECT

During the second phase, only 20 participants joined and were curious to see whether they had remembered and restored the information. This phase was conducted a week after the Quiz week, in order for the information to be stored accordingly in the brain. The hypothesis is that, when and if the information is remembered after the exam (or quiz), long term usage is applicable. This is tied to the general concept that states that under pressure, anxiety or even during exam week, the information stored is blurry, is mechanically available, concepts are misused etc. while after exam week has passed, the information stored is clear and understandable. (As for me, this was and still is valid as a concept). The results, however, were mixed, namely:
14 students had a *decline* in their previous scores;
6 students had a *balance* in results;
Overgeneralization: The fact that there was a discrepancy in the overall scores and results, with a decline between the first Quiz and the second, proved the following:
That most students did not retain the material for as long as 1 week after the Exam;
That the medium, standard students, in fact scored a balance (i.e. the medium average is more persistent than the higher scores, which may be based on luck or basic knowledge, or even cheating at times).
However, what is crucially important to be pointed out is the fact that I did not teach to the test, but rather required the logical pinpoints of the material, which, students had actually retained i.e. the basics of the concepts of the course were actually understood and this knowledge would help them in their near future as English teachers.
This form of assessment, however, also proved to be quite time-consuming and that there is in fact, a pattern that is followed in order to gain results at the end, just as in the alternative assessment(s). (Such as record-keeping, mixing up questions and concepts, balancing student progress etc.)

6. RESEARCH QUESTIONS OBTAINED

At the beginning of the paper, it was mentioned that the case study attempts to answer some research questions regarding the issue of test validity and the washback effect:

- *"Do teachers teach to the test, and if so, what are the overall obtained results?"* the procedure proved that I in fact did not teach to the test but rather provided the learners with overall concepts, definitions and hands on activities. On the contrary, the students would have gained maximum of points (due to memorization or mechanical learning). In cases where the teachers are taken accountable for the overall scores and results, such as standard exams, state exams, private courses etc. it is generally known (and practiced) by teachers to either teach to the test or provide the learners with the questions prior to the test.
- *"Is ongoing assessment better than testing?"* it is evident that both provide elements of grading and systematic evaluation. Testing provides a score based on a pen and pencil test, whereas assessment provides insights on the whole process of learning. In this case, both measurements were used in order to conclude a balance.
- *"Can test validity be tested through the washback effect?* In fact, it can, to a certain extent. The procedure proved that students remembered the basics of the course material and gained solid points at the end of the course. (ongoing, systematic learning)

7. IS THE WASHBACK EFFECT PRESENT? IS IT GOOD OR BAD?

According to Messick (1996), "Washback, a concept prominent in applied linguistics, refers to the extent to which the introduction and use of a test influences language teachers and learners to do things they would not otherwise do that promote or inhibit language learning. Some proponents have even maintained that a test's validity should be appraised by the degree to which it manifests positive or negative washback, a notion akin to the proposal of 'systemic validity' in the educational measurement literature. (1996, pg. 241)
The main reason behind the case study stood in the consistency and gradual obtaining of the issues related to the course, by which both test validity and the washback effect would be tested. The huge difference among the Quiz results versus the washback effect results brought up some issues of discussion and some reflecting on past experiences. Referring to the results, the issue of accountability and teaching to the test was examined. Due to stress, from other parties involved in the education system, teachers tend to teach to the test i.e, they provide learners with similar format questions, they provide them with different strategies of passing the exam or test, they skip chapters

and content materials…the results/scores achieved are better and all involved in the process are satisfied. This is the truth behind many washback effects that occur everywhere. Accordingly, as concluded by Rahman et. al (2021) "teachers usually engage students on drills of past examinations to familiarize them with the test format. During the interviews, comments of most teachers actually accepted the fact that in order to prevent stress and anxiety and in order to achieve high scores together with the students, they provided them with past tests and quizzes in the form of drills, thus, facilitating the issue of accountability on both sides.

8. REFLECTING ON PAST EXPERIENCES: THE TURTLE AND THE RABBIT COINCIDENCE IN LANGUAGELEARNING- IS IT THE TEST OR US?

Throughout my teaching experience, I have met the good learners and the not so good learners; the independent learners and the dependent ones; the shy learners and the active learners- they have all differed in many aspects. For starters, I have shown and taught them that, these differences actually make their uniqueness and that variety in the classroom is much appreciated. They were taught to accept mistakes and develop along the way and were shown how progress is made from day one. An ongoing issue that has followed me during my teaching career (apart from what is mentioned above) is language proficiency advancement, namely, the best students in the group have remained the same while the other lower leveled students have progressed and even at times, surpassed the grade A students. This example coincides with the story of *The rabbit and the turtle*, the moral of which is that even with slow steps and persistent struggle in the process, you may become a winner if you do not give up. The role of self-confidence in such cases has proven to be a negative influence (as much as we emphasize the positive role of self-confidence in language learning). The same issue occurred with a group of first year ELT students, namely, my praise and support (as assumed) affected their overall success negatively. The overall impression regarding this group of learners was quite good at the beginning. The students had much language knowledge and were fluent in their speech production; they showed great interest in partaking in the activities and were eager to voluntarily participate in any given task. This first *snapshot* of the group gave me an actual wrong impression. What actually happened was the total opposite. All the praise and self-confidence had a negative impact altogether. The group regressed- the majority started off as Rabbits, and somewhere along the way stopped. The remaining slower students proceeded as Turtles, pushing their way persistently. The rabbits, as in the story, failed. The turtles succeeded. Few rabbits did not fall asleep and remained the same-these few passed with great scores. The others did not even make it to the finish line. They, of course, will be repeating the course in the upcoming year. The same issue with the same students occurred in other courses as well. Our overall positive impression proved to be incorrect. What, in fact happened along the way? Where did we go wrong? Did we teach to the test or were the courses too difficult for them to comprehend? In most cases, there is lack of motivation and self-responsibility on behalf of the students' however, as teachers there is always a level of accountability that is in fact part of the process. The relevance of test validity and the washback effect would not be possible due to failure of accepting responsibility in learning and progress and due to the misconception tied to knowing a language. (this is already part of recent research which I have conducted) The additional circumstances and issues will be looked into in near future research.

9. LIMITATIONS AND FUTURE RECOMMENDATIONS

The case study, of course, had its limitations such as the methodological limitations i.e. small sample sizes, lack of control groups etc. Improvement of the research methods, including the use of mixed-methods approaches and longitudinal studies, can enhance the validity and reliability of the findings. Studies often rely on the teachers' perceptions of washback effects rather than directly observing the teaching practices, short term compared to long term practices etc. The procedure could have had a third phase of Quiz or testing (maybe positive washback effect would be more present).With all its' limitations, the findings, however proved that this form of assessment is more solid than only one form of a test and that a balance between both results in fact provides the real knowledge that students have gained. It is recommended that more research is conducted on the issue(s).

REFERENCES

Abolfathi, A., Tahir, N., & Ahmed, K. (2022). Different types of assessments and their effect on students' learning and workload in remote learning September 2022 Conference: The 8th International Symposium of Engineering Education.(https://www.researchgate.net/publication/366876462_Different_types_of_assessments_and_their_effect_on_students'_learning_and_workload_in_remote_learning)

Douglas Brown, H. (2018) Language assessment: Principles and classroom practice. Pearson Longman.

Lutfiana, L., Tono, S., & Mahmuda, A. (2020). Overseas students' language and culture barriers towards acquiring academic progress: A study of Thai undergraduate students. Int. J. Curr. Sci. Multidiscip. Res. 3, 107–114.

Martel, J. (2019). Washback of ACTFL's integrated performance assessment in an intensive summer language program at the tertiary level. Lang. Educ. Assess. 2, 57–69.

Messick, S. (1996). Validity and washback in language testing. Lang. Test. 13, 241–256. doi: 10.1177/026553229601300302

McKinley & Thompson (2018). "Washback Effect in Teaching English as an International Language" (PDF). TESOL Encyclopedia of English Language Teaching. **1**: 1–12. doi:10.1002/9781118784235.eelt0656. ISBN 9781118784228.)

Mousavi, Seyyed Abbas. (2002). An encyclopedic dictionary of language testing. Third Edition. Taiwan: Tung Hua Book Company

Penny, Ur, (2024) A course in Language Teaching, Third Edition. Cambridge University Press

Pitoyo, M.D., Sumardi, S., & Asib, A. (2020). Gamification- based assessment: The washback effect of quizzes on students learning in higher education. Int. J. Lang. Educ. 4, 1-10. Doi:10.26858/ijole.v4i2.8188

Rahman, K., Seraj, P. M. I., Hasan, M., Namaziandost, E., & Tilwani, S. A. (2021). Washback of assessment on English teaching-learning practice at secondary schools. Lang. Test. Asia 11, 1–23. doi: 10.1186/s40468-021-00129-2