International Council of Association for Science Education

ORIGINAL ARTICLE

How Does the Washback of Two Different Formats of Assessment Impact Chemistry Postgraduate Students

Saadia Y. Qureshi*

Institute of Education and Research, University of Punjab, Punjab, Pakistan

***Corresponding Author:** saadiayousouf@gmail.com

ABSTRACT

Washback has been defined as an effect of assessment on teaching and learning which may be negative or positive. This study investigated the washback effect of multiple choice question (MCQ) format of assessment on learning of concepts in physical sciences (chemistry) as compared to constructed response tests (CRTs). This study collected perceptions of students through open-ended questionnaires about these two different formats of assessment in the subject of chemistry at the postgraduate level. Perceptions were validated through diagnostic analysis of midterm assessment consisting of CRT and MCQ format revealing their comparative washback. Post-test data were used to compare the performance of students for two sets of comparable chapters. This study revealed that students chose MCQ format to avoid narration and organization of responses, ultimately avoiding creativity, which lead to the proposal of a washback model. This study refutes the perception that MCQ format results in higher marking, is quicker, and is a more objective way of assessment. MCQs produced an equal level of comprehension of concepts as that produced by CRTs as washback applying paired sample t-test. MCQs did elicit higher order thinking but should be used along with other formats to design a comprehensive assessment.

KEY WORDS: washback; multiple choice questions; perceptions; formats of assessment; diagnostic analysis

INTRODUCTION

Teaching and learning in the classroom are affected and often driven by the assessment held at the end of term. This is referred to as "washback of assessment" (Alderson and Wall, 1993). As a result, teachers often align their curriculum according to the assessment (Shohamy, 1992). This phenomenon becomes more prominent when the assessment is high stakes and has been reported to result in both test-driven schooling and impacting on areas such as curriculum breadth, pedagogy, staff morale, student's educational experiences, and their well-being (Dulfer et al., 2012). As students are one of the main stakeholders, they perceive the effect of assessment more acutely and often respond accordingly to the ramifications of this in their classroom practices. The whole process of test taking creates a belief system about the context and format of the particular assessment, which is manifested in the performances and responses of the test takers. Student assessment performances are evaluated and scrutinized by other stakeholders such as policymakers, employers, higher education administrators, and test authors. As a result, tests are amended accordingly, which spreads the effect of washback to a wider audience (Wall, 1997).

LITERATURE REVIEW

As Cheng (1997) suggested, washback is a complex phenomenon that is an outcome of the interaction of a variety of intervening variables such as tests, test-related teaching, learning, and perspectives of stakeholders. Taking into account that complexity, washback studies often involve "naturalistic," "observational," and "descriptive" elements. Alderson and Wall (1993) proposed washback hypotheses based on who or what might be affected such as:

1. Teaching
2. Learning
3. Content
4. Rate of learning
5. Sequence of teaching/learning
6. Degree/depth of curriculum coverage
7. Attitudes of teachers/learners and others.

It has been argued in a range of contexts that a test has a powerful influence on the learner who is preparing to take the test and on the teacher who tries to help him/her prepare. Pearson (1998) pointed out that public examinations do influence the attitudes and behaviors of the examinees. Glaser and Bassok (1989) and Glaser and Silver (1994) reported that the beliefs about testing tended to follow the beliefs about teaching and learning. Cheng (1997) studied intended change produced through public examination. She found that there was no doubt that the examination could be used as a vehicle for bringing changes in curriculum in the domain of teaching and learning, but there are also side effects or unintended outcomes. These side effects or unintended outcomes resulted because there is no single track to trace washback rather it is a result of an extensive and interwoven web of causes and effects.

Washback acts on a continuum stretching from harmful at one end through neutral to beneficial at the other end (Figure 1).

Washback may be termed as negative if it causes:
1. Restriction of content leading or narrowing of the curriculum
2. Too much time spent practicing for the test.

And positive if it causes:
1. Clear objectives and outcomes
2. Increase in the motivation of learners
3. Effective accountability of teachers

Swain (1984) stated the prevailing opinion "it has frequently been noted that teachers will teach to a test: That is, if they know the content of a test and/or the format of a test, they will teach their students accordingly" (p. 43). This effect becomes much prominent when the assessment or measurement is career-defining like in high-stake examinations. A high-stakes test could affect teachers and learners directly and negatively by focusing more on teaching test-taking skills. As high-stakes tests may directly affect promotions or graduation decisions and therefore affect their and their family's life, these decisions may become crucial or controversial (Kadriye and Bekir, 2013). A test result is not usually controversial in and of itself until high-stakes decisions are based on a single test and questions on whether the measurement was valid or not.

Qi (2004) noted that stakeholders such as students, parents, teachers, and administrators perceive high-stake examinations as the foremost basis for the decisions of their lives, directly or indirectly affecting them. On the other hand, Davis (1995) argued for positive washback through creative and innovative testing. Davis also highlighted that the greater the outcome of the test, the more likely it would be to have an effect on teaching.

Washback effect being a complex phenomenon has been investigated by different researchers. Though first defined by Alderson and Wall (1993), washback encompasses many aspects of assessment (Djuric et al., 2008; Melor and Hadi, 2011; Tsagri, 2007; 2009). Messick (1996) reported that interpretation of scores of a particular assessment required specifying the context of that assessment. This complex context was simplified by Hughes (1994) presenting the concept of 3Ps: Participants (every stakeholder), processes (every action taken in this regard), and products (learned and taught).The interplay of the 3Ps produced what Watanabe (2004) termed as intra-washback. Baily (1996) suggested that inter-washback resulted ultimately into intra-washback. Inter-washback is the interaction of contextual variables (participants, process, and products) in the environment, while intra-washback is after effect produced by such variables which make a belief system varying individual to individual. Watanabe (2004) proposed varied intensity of such interplay between patterns of assessments, context of interaction, and participants, process, and products.

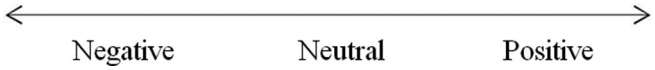

Figure 1: Washback continuum

Researchers like Baker (1989) and Shepard (1993) reported that multiple choice format of standardized tests has limited relevance and meaningfulness resulting in the narrowness of teaching content and neglect higher order thinking skills. Rather, this format produced more receptive skills than productive skills. A washback study of university entrance examination in Iran on teacher perceptions revealed through survey analysis that teaching was focused on the abilities tested through particular examination format (multiple-choice questions [MCQs]) such as reading, comprehension skill, vocabulary learning, and basic grammar knowledge and abilities such as communicative activities and practicing productive skills in the classroom which were ignored (Salehi and Yunus, 2012). Akpinar and Cakildere (2013) reported that most of the washback studies were based on classroom observation, surveys, interviews, or a combination of these. The washback effect is a characteristic of each test that drives the teaching and learning.

WORKING MECHANISM OF WASHBACK

Baily (1996) raised the concern that relatively little empirical research has been done to document the exact nature or mechanism of how washback works. To explain how this phenomenon works, Messick (1996) advised considering the contextual aspects along with the interpretation of scores of that particular assessment. As stated, Hughes (1994) divided this complex contextual aspect in three broad divisions: Participants (every stakeholder), processes (every action taken in this regard), and products (learned and taught). Extending the concept of Hughes, Baily (1996) proposed a more thorough model of washback based on the 3Ps (Figure 2).

Baily explained interactions of all stakeholders labeled as participants with the test, and going through the process of interaction, where the products produced were learning, teaching, research results, new material, and new curricula. All products again affect the test practices. According to this model, process, participants, and products are directly affected from the test introduced, which in turn causes new material and new curricula ultimately causing innovations in teaching and learning methodologies. Watanabe (2004) explained this as the evolution of washback theory moving from the traditional through the 1990's black box model to cognitive model.

Watanabe (2004) termed washback validity, coined by Morrow (1986), as a single and uniform reaction to the quality of the test rather than from the teachers, hence suggesting a consistency of response. He elaborated that this model does not explain the findings of observational and evidence-based research. The data collected presented an individual's responses to a variety of tests and revealed patterns in behavior ultimately giving way to curriculum innovations. The cognition model of washback

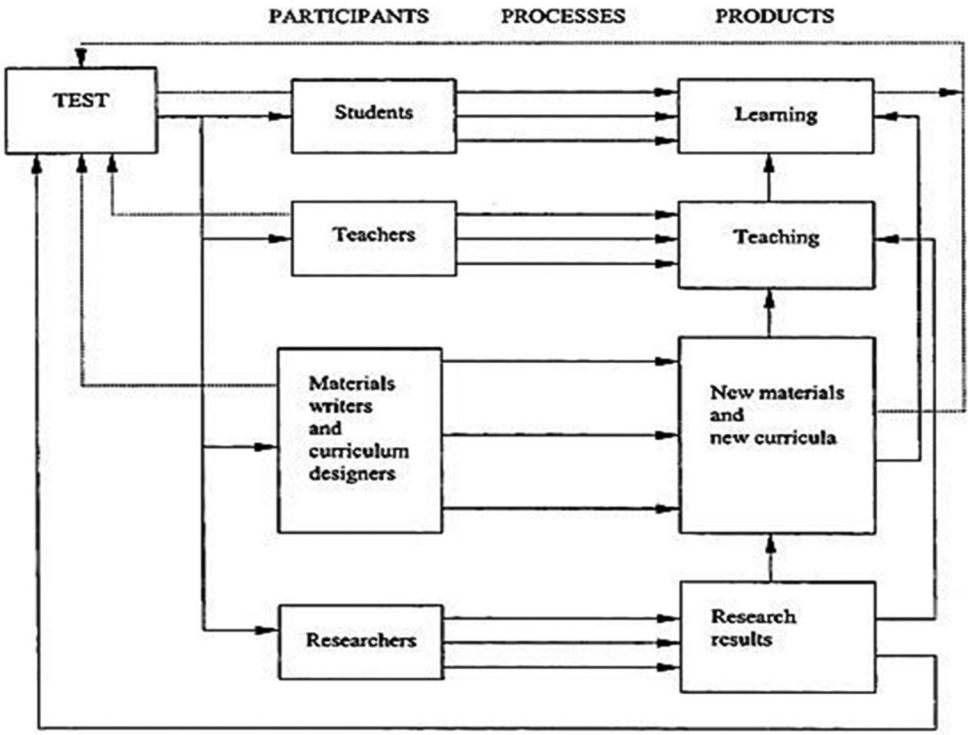

Figure 2: Model of washback (reprinted from Baily, 1996. p. 264)

showed itself in terms of different beliefs, assumptions, and knowledge making washback mechanism more complex.

Green (2007) examined participant and process variables such as learner background, motivation, class activities, and learning strategy used through pre and post-tests. His findings produced two implications: First, a more advantageous way to improve students' scores would be to integrate material covered on the test with regular teaching rather than test-driven instructions; and second, intentions for taking the test should be clear to both students and teachers to foster learning. Results provided a comprehensive list of factors, which were considered to influence learning outcomes (Figure 3). His list of factors provided working lines for researchers interested in exploring washback effects. Green stated that, if teaching the targeted skills boosts scores, this will have profound implications for the relationship between teaching and testing.

Washback being a complex phenomenon, it is not easy to establish whether it is primarily due to a test-prep course itself or other factors such as learning experience, motivation, and age. Burrows (2004) studied classroom-based assessment in implementation of the certificate in Spoken and Written English in an Adult Migration English Program in Australia. This assessment was competency based, came with assessment guidelines that were aligned with the curriculum tasks. It affected the courses significantly allied to this assessment as teaching objectives were assessment outcomes and it was teaching to the test. Patterns revealed through observations established causality of change which led to a new model of washback for curriculum innovation which was based on the findings that teachers perceive and interpret a new test to shape up their beliefs, assumptions, and knowledge that is reflected in their teachings.

Ana and Marta (2010) did an investigative study in classroom context in the Universidad EAFIT in Columbia. They did acknowledged that the washback effect in the classroom would not be on the same scale as in a high stakes examination. They used quantitative and qualitative methodology in a comparative study between an experimental and comparative group. Positive washback of an oral assessment was reported in some of the areas of teaching and learning.

On the other hand, many researchers have reported negative aspects allied with standardized tests such as the narrowness of teaching content, neglect of higher order thinking skills, and the limited relevance and meaningfulness of their multiple choice formats (Baker,1989; Shepard, 1993). Studies working on receptive and productive skills revealed that students developed more receptive skills than productive skills with the overuse of MCQs. In the Iranian context, washback of university entrance examination was studied on teacher perceptions through survey analysis (Salehi and Yunus, 2012). The format of the assessment entrance examination was MCQs. It was concluded that teaching was focused on the abilities tested through the particular examination format (MCQs) such as reading, comprehension skill, vocabulary learning, and basic grammar knowledge; abilities such as communicative activities and practicing productive skills in the classroom were ignored. As seen, the MCQ format of assessment has been studied for its washback effect, but the findings could not declare whether the format of assessment was responsible for the produced washback as other factors were not considered

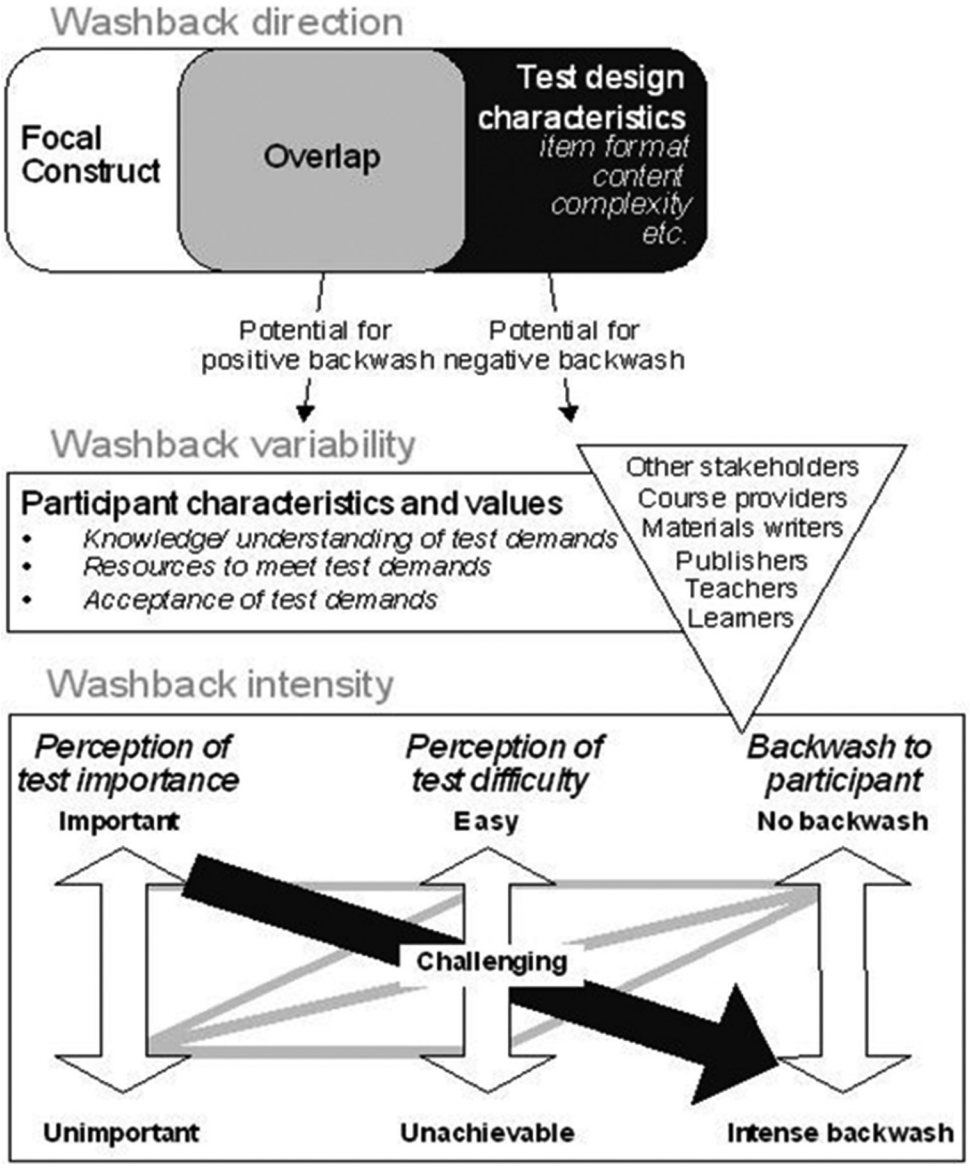

Figure 3: Washback model (reprinted from Green, 2007. p. 194)

or controlled. To address this, this study was endeavored to present the washback of classroom-based assessment in action. Specifically, this study was proposed to compare two formats of assessment keeping other contextual conditions as constant as possible.

METHOD

This study used a mixed method of data collection and data analysis. This study included three phases: Studying comparative washback of two formats of assessments quantitatively using a post-test design, studying perceptions qualitatively, and a diagnostic analysis of mid-term assessments. A sample of 32 students studying Chemistry in a Master of Science Education at the Institute of Education and Research, University of Punjab, was selected. Course content was divided into two compatible sets based on scores of a pre-test consisting of 60 comprehension level items that covered all the contents of course. The opinion of experts was obtained to ensure the validity of pre-test while the Cronbach's Alpha reliability of the instrument was 0.892. This pre-test required only one word or one sentence to answer. The test took 90 min and was conducted before the semester started, and two compatible sets of content revealed through scores were subjected to different formats of assessment, MCQs, and constructed response tests (CRTs) throughout the semester of 3 months (Table 1). Other factors such as teaching methodologies, environment of classroom, and context were kept as constant as possible during the semester.

Participants were selected following single-subject design in natural setting to allow for the diagnostic analysis of their performance. Perceptions of students about the two formats of assessment were collected before and after the mid-term assessment of the semester through open-ended questionnaires that were analyzed qualitatively using thematic

Table 1: Equivalent sets of scores of pre-test

Serial no	Set 1 (MCQs)	Sum of scores	Set 2 (CRTs)	Sum of scores
1	Electrochemistry	64	Chemistry of P block elements	64
2	Nuclear chemistry	57	Solutions	60
3	D block elements	69	Acids and bases	67

Scores of each chapter were calculated by adding scores of items from that chapter for each case revealing two equivalent sets of chapters. MCQs: Multiple choice questions, CRTs: Constructed response tests

analysis. Emerging themes were quantified as percentages corresponding to the two formats being studied. Mid-term assessment was conducted choosing two chapters from each equivalent portion of content and was assessed through MCQ and CRT format. Achievement data were quantitatively analyzed as lower, middle, and higher level of performance.

At the end of semester, a post-test was conducted consisting of 60 comprehension level items, which covered all of the course content. This 90-min test was conducted like the pre-test, and achievement scores for two portions were tested for a statistical difference of means applying paired sample t-test using Statistical Package for the Social Sciences (SPSS) software.

STUDYING PERCEPTIONS OF STUDENTS

Student responses were collected through open-ended questionnaires. Individual responses were analyzed qualitatively through thematic analysis. Emerging themes were rated as percentages to reveal the distinctive preferences of students for both of the formats addressed. As stated, a questionnaire was administered before the mid-term assessment and again at the end of semester. Student perceptions were matched with the performance by diagnosing mid-semester assessments. This study investigated the perceptions about the washback effect of MCQ and the CRT format of assessment. Results of qualitative analysis of questionnaires collecting perceptions were established through the statistical significance of post-test results and diagnostic analysis of midterm assessment.

Instrument Collecting Perceptions of Students

Instrument contained open-ended questions collecting individual responses about perceptions of students about the format of the assessments. Questions directly addressed the issues to be investigated in the study. Instrument asked about their preferred assessment format and their reason for choosing it. Responses were collected to investigate if either of the formats required more comprehension. Respondents were asked to elaborate on their answers about if they believed the format of assessments as fair or not. Reliability of instrument found after statistical analysis of individual responses and categorizing the data was 0.7.

Collection of Data

Perceptions about the preferred format of assessment were collected before and after the mid-term assessment. Students were given enough time to narrate their opinions on the questionnaire. Questions addressed their general perceptions about the format of assessment and particularly about the assessments held during semester. Oral reminders were provided about assessments of particular chapters of coursework that was required to answer corresponding questions.

Perceptions of students were treated as individual responses and analyzed qualitatively using thematic analysis, and then, emerging themes were quantified as percentages to produce a distinctive clarity of the belief system of students for two formats of assessment.

DIAGNOSTIC STUDY

Mid-semester assessment was designed to include comprehension level as well as application level questions comprising of MCQ and CRT type assessments. A mid-semester assessment is a usual component of these participating students' normal assessment during the semester and is worth 35% of total assessment.

Mid-semester Assessment Instrument

This was an achievement test that covered two chapters assessed through MCQs and two chapters assessed through CRT type written test. These tests were analyzed diagnostically to establish the levels of achievements attempting two types of formats (MCQs and CRTs).

Collection of Data

This assessment was a 90-minute written test. The results were analyzed to determine the comprehension and application level of performances for both MCQs and CRTs. Data analysis highlighted the perceptions of these students about the two formats of assessmen. Data were analyzed to validate the perceptions of students about two the formats of assessment through diagnostic analysis of mid-term assessment and that was to be depicted in post-test performances accordingly to see if there was any difference in washback effect produced by two formats of assessments.

RESULTS

Perceptions of Students before Mid-semester Assessment

Perceptions of students about the two different assessment formats were collected through an open-ended questionnaire. Any ambiguity in their written answers was resolved through discussions. Analysis of perceptions provided the basis of evaluation of the mid-semester assessment.

Q1. What is your preferred format of assessment?

71% of students preferred MCQs as their preferred format of assessment, and 21% declared CRTs as their preferred format, while 8% students stated a mixed type of assessment that included both MCQs and CRTs would be preferable.

Q2. Is your preferred format of assessment based on the assumption that its questions and required responses would be covered in the semester's content?

Almost all of the respondents declared that their preferences (either MCQ or CRT) would work for them if the content required to address each question was provided during the course.

Q3. Which of the two assessment formats, do you think requires a deeper level of comprehension to respond?

Students who preferred MCQs stated that, to respond to this type of question, you needed a more critical comprehension of the content; however, this type also meant that you did not have to memorize the course content. Students who preferred CRTs justified this by stating their belief that this type of test needs students to have a detailed comprehension of concepts and be able to organize their responses. Those students (8%) choosing the mixed type of assessment felt that CRTs needed a detailed understanding, but MCQs require a critical understanding of concepts.

PERCEPTIONS OF STUDENTS AT THE END OF SEMESTER

At the end of semester, perceptions were again collected from students about their preferred assessment format and reasons for that preference. This questionnaire consisted of sixteen items.

Q1. Did You like the Chemistry 1 Course being Assessed in the Same Manner as your other MSEd Courses (i.e., 20% Academic Assignments + 35% Mid-Term Written Tests + 45% Final Written Test)?

80% of the participating students liked that the Chemistry 1 course was assessed in the same manner as their other course, while 20% stated that they did not want it to be assessed at all.

Q2. Do you think you are able to demonstrate your abilities to your maximum level through assessments held during semester?

All of them responded that the two types of assessments provided them with the opportunities to demonstrate their abilities as the assessments consisted of both MCQs and CRTs.

Q3. How do you rate the assessment criteria: Bad, fair, good, or excellent?

70% of students rated the assessment criteria as good, 20% of students as fair, and 10% declared the assessment criteria to be excellent.

Q4. What do you think are the positives and negatives of your assessments during the semester?

40% of students rated the assessments as good because the course's teaching methodology suits their learning style. They reported that their assessments were supported by the course's teaching methods of inquiry, discussion, feedback, and participation. 60% of these students noted that they were required to study the content thoroughly for their assessment, especially, as they had to cover lengthy portions of the content for the CRT assessment. As a result, these students highlighted that they did some selective study for preparing for CRT assessment, but they reviewed the whole range of content when they prepared for the MCQ portion.

Q5. You are given some content explaining a scientific discovery. This content has an introduction and a part explaining its links between the facts. You are asked to provide a conclusion and predict its future applications. To do this, do you memorize the facts or try to understand the facts?

90% of these students stated that, if they were given the content explaining a scientific discovery and then were required to provide a future application of this scientific discovery, they would have to comprehend the facts.

Q6. How would you prepare if you knew you were going to be assessed through MCQ format

All of the students gave the opinion that, if the content was assessed through MCQs, they would study each and every detail of content thoroughly.

Q7. Chemistry 1 covered seven chapters, half assessed by MCQs and half by CRTs, which set did you find the most interesting and why?

50% of the students reported that the set assessed through MCQs contained the content they were most interested in while 50% reported the same for CRTs. Further exploration revealed that the type of assessment was not the determining factor but the students own personal interest.

Q8. Which part of the chapter did you find more relevant to your assessment, the part stating scientific facts or the mathematical parts?

40% of these students said that that mathematical part was relevant to their assessment, while 60% said that factual part was relevant to their assessment. Further investigations revealed that those students who chose their most interesting chapter from those assessed through CRTs found the mathematical part to be relevant while those who chose their most interesting chapter from MCQs set reported the factual part to be relevant to the assessment.

Q9. Which format of assessment MCQs or CRTs suited you best?

85% of students stated that the MCQ format of assessment suited them best.

Q10. Did your academic background influence which format you think is most suitable for you?

70% of students declared that their BSc (Bachelor of Science) background helped them being assessed through the MCQ format of assessment. 30% of these students chose the CRT format due to the fact that in their BSc they were largely assessed through narrative assessment similar to this study's CRT format where students were required to provide a written response.

Q12. How did these two assessments affect your learning?

Q13. What difference did you find while preparing for these two formats?

80% of students declared that, while preparing for the MCQ format of assessment, they had to comprehend the content more deeply and they had to review all the content. 20% of these students said that their preparation for the CRT assessment produced a deeper understanding of the concepts and provided them with a basis for this content's future application in their degree. It should be noted that 50% of the students who choose the MCQ format also reported that they believed that the MCQs were easier to complete as the CRTs required not only written responses to the question but also memorization of concepts, which if they could this was to be avoided by the students.

Q14. Which format of assessment do you think leads to developing a deeper understanding of the subject?

While answering this question, 60% of these students concluded the MCQ format of assessment would lead them to develop a deeper understanding of subject, while 40% reported the CRT format of assessment.

Q15. Which chapter did you find difficult to prepare for?

Q16. Was this difficulty due to the format of assessment for that chapter?

85% of students indicated that their most difficult chapter came from the MCQ format set. These students then reported that this difficulty was not due to the format of assessment but due to their lack of interest in that chapter. The rest of the students (15%) noted their most difficult chapter came from the CRT set and reported similar reasons. Most of these students argued that it was good that the chemistry 1 course was assessed as it gave them the opportunity to demonstrate their abilities. They noted that the criteria assessment was suitable and also highlighted that the course's teaching methodologies were responsible for making the assessment criteria compatible with their preparation. There were some negative issues reported by these students such as they felt they had to memorize lengthy and difficult portions of content, while students.

These results indicate that the teaching methodologies and other contextual aspects affected learning for these participating students. Students declared that, to conclude a scientific concept and to predict its future application, this concept was necessary to be understood. Rote memorization was not the best way to learn in the physical sciences. They also clearly stated that, to prepare for the MCQ assessments, concepts had to be understood thoroughly. Students chose an equal number of interesting chapters from both sets meant for MCQs and CRTs. Students who chose chapters from the set assessed through CRT assessment found the mathematical part to be relevant while students who chose from the MCQ set found the factual portion to be relevant. It would seem that, in preparing for CRTs, students prepare for detailed mathematical applications of the concepts, while in preparing for MCQs, they consider the factual part.

Overall, 70% of these students chose the MCQ format of assessment as more suitable for them. They did think preparing for the MCQs required deeper critical thinking, but most of them also admitted that assessments through the MCQ format saved them from rote memorization and having to write responses to test questions. For this study's students, at the end of semester, most of these students favored the MCQs as their preferable assessment even when the more difficult content was going to be assessed by MCQs.

DIAGNOSTIC ANALYSIS OF MID-TERM ASSESSMENT

In the MCQ mid-semester assessment part, student performances <50% correct response at the comprehension level were rated as low and marked with a "0," while performance having 50% correct responses as a "1" and those having more than 50% were marked with a "2." For the CRT portion, the students' constructed responses consisting of a correct narrative part at the comprehension level were labeled as a "1," while a correct narrative along with a mathematical derivation was labelled with a "2." The application level responses that were partially correct were labelled with a "1" and correct responses labelled with a "2". Adding individual scores for both the MCQ and CRT portions separately for comprehension and application level gave the performance levels for the mid-term assessment and were recorded as high, middle, and low (Table 2).

Questions at the comprehension level for the MCQs were prepared to produce answers only after a critical thought process by the students as these answers were not explicitly discussed in the classroom. For the CRTs, the comprehension

Table 2: Performance levels in mid-term assessment revealed after diagnostic analysis

Performance level	Comprehension level (%)		Application level (%)	
	MCQs	CRTs	MCQs	CRTs
High	29	51	45	51
Middle	29	25	29	25
Low	42	24	25	24

Content of course was divided into two matched groups of content to be assessed through MCQs and CRTs and assessments were assessed for percentage level of performance as high, middle or low for the two sets. MCQs: Multiple choice questions, CRTs: Constructed response tests

level answers could be produced directly from the content provided in the course. At the comprehension level, it was evident that higher-level performances (29%) for the MCQ portion were lower when compared to the CRT portion (51%). Similarly, the percentage of lower performance was higher (42%) for the MCQs than for the CRT portion. At the application level for the MCQ portion, performances were higher at 45% than for the comprehension level. The application level of student performance for the CRT portion was consistent to the comprehension level. These results revealed that although a smaller number of students performed at the higher level, these students reported the MCQs required a more critical comprehension but as noted by these students the MCQs required a more critical comprehension seen in the mid-term assessment results. The CRT assessment produced the same level of performance at both comprehension and application level for all three levels of high, middle, and low.

Post-test Data

Two students did not take test leaving behind 30 cases. Set 1 was assessed by the MCQ format while set 2 underwent CRT assessment. The test applied was paired sample t-test. Software used was SPSS. Statistics revealed are presented in Tables 3-5.

t-test

Statistics showed the mean of 12.06 for CRT portion (set 1) was slightly higher than the mean for MCQ portion (set 2) which was 11.50. CRT portion had a standard deviation SD 6.169 which showed a higher variation in performances than found in the MCQ portion SD 4.946.

Analysis gave the t-value to be 0.851 with 29 df at 0.05 significance level as confidence interval was set to be 95%. These results showed the negligible difference or no difference between two portions for the comprehension of concepts of chemistry. Analysis showed that washback of both the formats of assessment in terms of approach toward concepts of subject at comprehension level turned out to be the same.

Paired sample correlation was found to be of higher value as 0.807, which showed strong relationship for both. This indicated tha the performances for both the sets have strong and direct relationship with each other showing higher similarity for producing washback effect regarding comprehension of concepts.

Table 3: Paired samples statistics

Pair 1	Mean	n	Standard deviation	Standard error mean
Sumvar1	12.0667	30	6.16963	1.12642
Sumvar2	11.5000	30	4.94626	0.90306

Table 4: Paired samples correlations

Pair 1	n	Correlation	Significant
Sumvar1 and sumvar2	30	0.807	0.000

DISCUSSION

In collecting perceptions about the preferable format of assessment and the reason for that preference, the majority of these students preferred the MCQ format. This was mainly because of its objectivity in scoring, increased chances of obtaining higher scores, and a quicker response to answering the questions. They reported being more at ease if the questions' content was provided in the course's content. They believed that the teaching practices in the course prepared them better for MCQs.

This study's data showed that these students performed well at the comprehension level for CRTs but low for MCQs. It is also reported by these students that questions on the CRT portion could be answered from the content studied during semester, while answers for the MCQ portion could not be produced by just remembering the content. It would seem that these students performed well when the content was first covered in the course and then assessed by questions about the content. These students did not do as well when they were required to evaluate content beyond the comprehension level. On the other hand, these students claimed that, to answer the MCQs successfully, they needed to comprehend the concepts and to review the whole of the course's content. They noted that they were able to do selective study to prepare for the CRT assessment, but they were required to operate at a deeper level of comprehension.

MCQ format does allow for some guessing, but this study results indicate that it is an objective way of assessing the comprehension level of students. MCQs required higher order thinking which cannot be answered simply by remembering facts. It appears that the students' claim "To try MCQs one has to comprehend facts effectively" is accurate.

Research has stated that the MCQ format is believed to result in students securing a high level of marking, but this study refutes that idea. Specifically, if MCQ questions are carefully planned, this format does not result in artificially higher scores. In this study, the MCQs were worded carefully to require student comprehension rather than only recalling facts,

Table 5: Paired sample test

Pair 1	Paired differences					t	df	Significant (two-tailed)
	Mean	Standard deviation	Standard error mean	95% Confidence interval of the difference				
				Lower	Upper			
Sumvar1–Sumvar2	0.56667	3.64534	0.66555	−0.79453	1.93786	0.851	29	0.402

which resulted in lower scores for the MCQ portion. While referring to CRTs, the high, middle, and low equivalence of performance at comprehension and application level indicates that students do perform equally well for comprehension and application level along with the written evidence produced at the application level.

This students' performance at the application level of understanding in the MCQ portion was much improved over their comprehension level, shifting from 29% to 45%, while it remained the same for CRT portion at 51% for both comprehension and application level. It would seem that, although students scored low at the comprehension level for MCQs, it required them to study these chapters more critically. It has been reported that MCQs do allow for guessing; however, this study results showed that these students performed better at a higher order of thinking. This raises the question was the difference in performance between MCQs and CRTs because the MCQ assessment included content that was not directly provided in the course? This question was addressed by designing the instrument for the post-test of concepts consisting of two parts. One part of course studied was assessed through MCQ format and the other part of post-test for the part of course assessed through CRT format, but both the formats comprised of questions which were not plainly discussed in the class. Paired sample statistics showed no significant difference between the scores of both the parts.

These students when compared to their CRT assessments refute the fact that the MCQ format has been criticized for producing higher marks. For this study students, the CRT assessment privileged mathematical learning. In addition, this was one of the advantages of the MCQ formats for some students as it allowed them to avoid this mathematical learning.

To address this study question, this study was endeavored to present the washback of classroom-based assessment in action. Specifically, this study was proposed to compare two formats of assessment keeping other contextual conditions as constant as possible. If the negatives of these formats are washed out, for example, the bluffing side of CRTs and the guessing side of MCQs, then what is the comparison of comprehension of concepts for these formats? The answer was provided. This study showed equal capability of both formats to produce comprehension of concepts. The MCQ format is similar to the CRT format in producing positive washback at the comprehension level. Both the formats produced equal levels of washback up to the comprehension level, though both the formats reached this by different routes. The MCQs were by critical appreciation of individual ideas and the CRTs were through the holistic view of ideas. It could be concluded that perceptions gave way to modification of making and evaluating the tests. If the concerns of learners were addressed in making valid tests, their performances do modify accordingly but within the design of the assessment format. The results of this research do not support the idea that MCQ format produces lower comprehension or results in higher scoring when

Table 6: Pathway of washback working within classroom for different formats of assessment

Teaching methods	Perceptions	Test formats	Validity factors	Measuring content	Pathway of washback	washback at the end
Lecture method	Preferred format	Constructed response test	Content validity	Taught skills	Holistic perception of concepts	Performance at comprehension level
Participation method	Reasons for preference		Construct validity	Covered directly from content		(Constructed response test scores)
Activity method	Level of comprehension involved		Expert's opinion	Holistic measurement		
		Multiple choice questions		Taught skills	Critical appreciation of concepts	Performance at comprehension level
				Covered indirectly from content		(Constructed response test scores)
				Individual concepts measurement		
		Other formats?		?	?	Application level. (?)

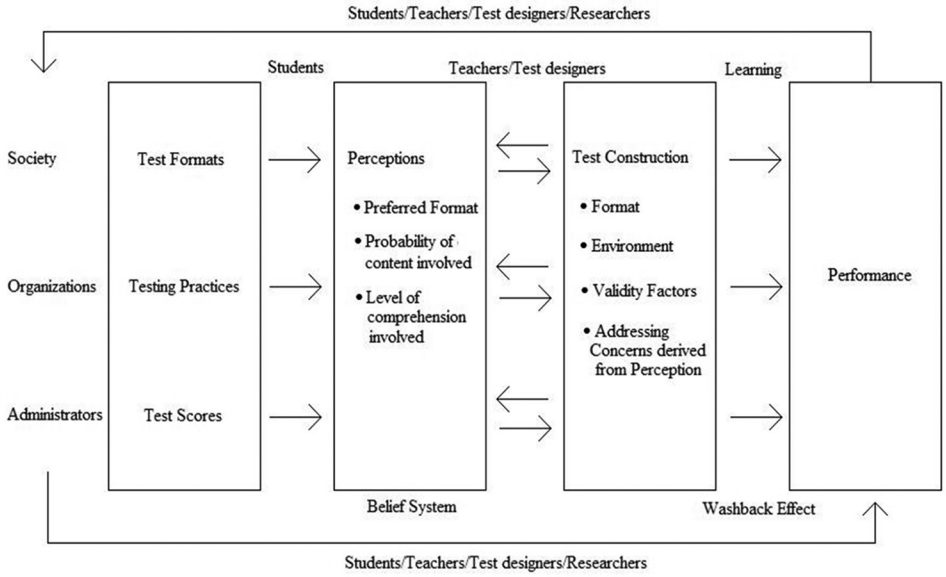

Figure 4: Feedback mechanism and proposed model of washback within classroom context

compared to CRT format, when the tests are valid. As a result, this study proposes a model for the mechanism of the washback effect within classroom context (Table 6).

This model suggests that washback works more or less in a feedback fashion. Perceptions are derived from general testing practices. Teachers or test designers modify the tests accordingly.

It acknowledges that the test does affect the learning, which then modifies the performance on the next tests accordingly. Test scores modify the perceptions and testing practices at large.

Washback is the product of tests or assessment directly or indirectly, but the tests are never alone in producing this phenomenon. Washback is inclusive and is aided by other educational practices besides the testing practices. However, comparative part played by test format in question can be isolated if formats of tests are varied in particular educational setting besides other factors such a teaching methodologies and environment of learning, which are kept constant but of course through valid tests. The washback effect can be designed as feedback mechanism working within the classroom context and society at large see Figure 4.

CONCLUSION

Critical comprehension was produced while going through MCQ testing which improved these students' application level, which was not evident from CRT testing. This requires test not only to be valid but also to be written to provoke critical thought. Carefully planned and well-constructed MCQs can test higher order reasoning. These tests rely on providing planned stem or statement of question to serve as stimulus materials to think about assessing higher order thinking abilities in MCQ format (Coderre et al., 2004). Such kind of tests has been developed by the Australian Council for Educational Research like the Scholastic Aptitude Test which is different from regular achievement test as it tests the abilities which would be developed from effective teaching and interested learning.

This study would argue that the MCQ format is in no way inferior to CRT testing if designed to produce positive washback effect on comprehension of concepts of subjects in physical sciences. This study revealed that MCQs are best at testing comprehension level and objective way of scoring and maintaining transparency required in high-stakes examinations. At the same time, however, the MCQ format is becoming the preferred choice of students. Students believe that this format requires less work to organize ideas and therefore a shortcut to better grades. Students reported that the mathematical part was a relevant portion of the CRT format, which is a clear advantage of this assessment over assessment through MCQs, which required a more critical comprehension of the facts presented. At the same time, these students reported that a negative of the CRT assessment was it encompassed application of ideas through mathematical calculations, and the MCQ allows students to avoid mathematical learning. In this study, the MCQ format produced effective comprehension of concepts, but further study would be warranted to investigate the long-term effects of students avoiding mathematical learning in the physical sciences.

REFERENCES

Akpinar, D.K., & Cakildere, B. (2013). Washback effects of high-stakes language tests of Turkey (KPDS and UDS) on productive and receptive skills of academic personnel. *Journal of Language and Linguistic Studies*, 9(2), 81-94.

Alderson, J.C., & Wall, D. (1993). Examining washback: The Sri Lankan impact study. *Language Testing,* 10(1), 41-69.

Ana, P.M., & Marta, E.A. (2010). Washback of an oral assessment system in EFL classroom. *Language Testing,* 27(1), 33-49.

Baily, K.M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing*, 13, 257.

Baker, J.B. (1989). *Can We Fairly Measure the Quality of Education.* Los Angeles, CA: Centre for Study of Evaluation (CSE).

Burrows, C. (2004). Washback in classroom-based assessment: A study of the washback effect in the Australian adult migrant English program. In: Cheng, L., Watanabe, Y., & Curtis, W., (Eds.), *Washback in Language Testing*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc. pp. 113-118.

Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and Education,* 11(1), 38-54.

Coderre, S.P., Harasym, P., Mandin, H., & Fick, G. (2004). The Impact of Two Multiple-Choice Question Formats on the Problem-Solving Strategies used by Novices and Experts. BMC Medical Education. p1. Available from: https://www.doi.org/10.1186/1472-6920-4-23. [Last retrieved 28 Sep 2018].

Davis, K.A. (1995). Qualitative theory and methods in applied linguistics. *TESOL Quarterly,* 29, 427-453.

Djuric, M. (2008). Dealing with situations of positive and negative washback. *M Djuric/Scripta Manent,* 4(1), 14-27.

Dulfer, N., Polesel, J., & Rice, S. (2012). *The Experience of Education: The Impact of High Stakes Testing on School Students and Their Families. An Educator's Perspective*. Sydney, Australia: The Whitlam Institute within the University of Western Sydney.

ESP in the Classroom: Practice and Evaluation. pp. 89-107. Available from: https://www.academic.oup.com/eltj/article-abstract/45/3/273/3113725. [Last Retrieved 2018 Sep 10].

Glaser, R., & Bassok, M. (1989). Learning theory and the study of instruction. *Annual Review of Psychology,* 40, 631-666.

Glaser, R., & Silver, E. (1994). *Assessment, Testing, and Instruction: Retrospect and Prospect (CSE Technical Report 379).* University of Pittsburgh: CRESST/Learning Research and Development Centre.

Green, A. (2007). Washback to learning outcomes: A comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education,* 14(1), 75-97.

Hughes, A. (1994). *Backwash and TOEFL 2000*. Reading, UK: University of Reading, Educational Testing Service (ETS).

Kadriye, D.A., & Bekir, C. (2013). Washback effects of high-stakes language tests of Turkey (KPDS and UDS) on productive and receptive skills of academic personnel. *Journal of Language and linguistic Studies*, 9(4), 81-94.

Melor, M.Y., & Hadi, S. (2011). *The Washback Effect of Entrance Exam of the Universities on the Iranian Pre-University Student's English Learning.* 18th International Conference on Learning. Mauritius: University of Mauritius. pp. 18-22.

Messick, S. (1996). Validity and washback in language testing. *Language Testing,* 13, 241.

Morrow, K. (1986). The Evaluation of Tests of Communicative Performance. In: Portal, M., (Ed.), *Innovations in Language Testing: Proceedings of the IUS/NFER Conference*. London: NFER/Nelson. pp. 1-13.

Pearson, I. (1998). Test as levers for change. In: Chamberlain, D., & Baumgardner, R.J., (Eds.),

Qi, L. (2004). Has a high-stakes test produced the intended changes? In: Cheng, L., Watanabe, Y., & Curtis, A., (Eds.), *Washback in Language Testing: Research Contexts and Methods*. Mahwah: NJ: Lawrence Erlbaum Associates. pp. 56-71.

Salehi, H., & Yunus, M.M. (2012). The washback effect of the Iranian universities entrance exam: Teachers' insights. *GEMA Online™ Journal of Language Studies,* 12(2), 609-628.

Shepard, L. (1993). The place of testing reform in educational reform: A reply to Cizek. *Educational Researcher,* 2(4), 10-14.

Shohamy, E. (1992). Beyond proficiency testing: A diagnostic feedback testing model for assessing foreign language learning. *Modern Language Journal,* 76, 513-521.

Swain, M. (1984). Large-scale communicative language testing: A case study. In: Savignon, S.J., & Berns, M., (Eds.), *Initiatives in Communicative Language Teaching*. Reading, MA: Addison-Wesley. pp. 185-201.

Tsagri, D. (2007). Review of Washback in Language Testing: How has Been Done? What More Needs Doing? Lancaster, UK. Available from: https://www.files.eric.ed.gov/fulltext/ED497709.pdf. [Last retrieved 2018 Sep 10].

Tsagri, D. (2009). *The Complexity of Test Washback: An Empirical Study*. Frankfurt, Germany: Centre for Technical Studies.

Wall, D. (1997). Impact and Washback in Language Testing. In: Clapham, C., & Corson, D., (Eds.), *Encyclopaedia of Language and Education. Vol. 7. Language Testing and Assessment*. Dordrecht: Kluwer Academic. pp. 291-302.

Watanabe, Y. (2004). Teacher factors mediating washback. In: Cheng, L., Watanabe, Y., & Curtis, A., (Eds.), *Washback in Language Testing: Research Contexts and Methods*. Mahawa: NJ: Lawrence Erlbaum Associates. pp. 129-146.