Standardization of EFL Undergraduate Skill Exam Papers: A Case Study at Majmaah University

Majed Othman Abahussain
English Department
College of Education, Majmaah University, Saudi Arabia

Muhammad Iqbal
English Department
College of Education, Majmaah University, Saudi Arabia

Imran Khan
English Department
College of Education, Majmaah University, Saudi Arabia

Abstract
Designing standardized exams for EFL undergraduates is a matter of great concern in pedagogical practices that assess learners' academic achievements. This case study was conducted at Majmaah University, Saudi Arabia, to explore the significance of educational exam standardization and its impact on the process of implementation, leading to a balanced and equitable assessment of English language skills. During the study, 250 final English skill exam questions of two semesters were analyzed. A mixed-method approach using a questionnaire and Focused Group Discussion was adopted to collect relevant data. The quantitative data were analyzed using SPSS 16, and the qualitative data was processed by thematic analysis. The quantitative data results showed that questions were balanced, valid, and reliable in language skill exam papers designed for undergraduate English students. The qualitative data revealed four major themes – standardization of questions, the efficacy of Bloom's taxonomy, assessment/evaluation, and alignment of the question with learning outcomes. These emerging themes highlighted the effectivness of standardizing exam papers, identified problems, and suggested ways of improvement. The study indicates that exams must be scrutinized for the sake of standardization in terms of maximum course coverage; the variety in the question item; clear, concise, and precise instructions; validity and reliability of exam papers; and the alignment of questions with the course and program learning outcomes to ensure the implementation of the revised Bloom's taxonomy. The study suggests, to generalize the results, that further investigation should be done in another Saudi university for the sake of comparison.
Keywords: Standardization, EFL undergraduate students, Majmaah University, Bloom's taxonomy, Course Learning Outcomes

Introduction

Although designing questions for the final exam of ESL or EFL learners is a sensitive task, hardly any hard and fast method to ensure the standardization of question items exists (Inbar-Lourie, 2012). For informative evaluation, language teachers remain confined to classrooms to evaluate the learner's understanding (Black & William 2009). Hence, standardization of exams cannot be followed strictly as assessing communicative skills, as an informative evaluation is more an activity of listening and speaking, feedback, and oral observation. Conversely, the final achievement test as a tool of a summative assessmnet of listening, speaking, reading, and writing is more dependent on questions and test instructions (Taras, 2005). Questions prepared hastily and carelessly without professional judgment increase language learners' stress and badly influence their ultimate performance. Much depends upon the individual educator's expertise in how he designs the summative question papers for students, since assessment is essentially a process of professional judgment (Brookhart, 2011). If the assessor is competent enough, he will design the question items so that students find no impediment in comprehending what is required in the answers. However, in case of a flawed exam, even an excellent student may find himself perplexed when attempting to answer the questions. This inevitably undermines the validity of the exams and unjustly penalizes the students (Downing, 2005).

At the undergraduate level, cognitive skills assessment has to be associated with Blooms' taxonomy, which offers classification of learning processes (Adams, 2015). Hence, while designing question papers, the question items have to be aligned with course learning outcomes and program learning outcomes, as this is now a focal point in education (Crespo et al., 2010). The primary concern is that in an institutional language teaching milieu in which a series of programs and modules are imparted by multiple faculty members with varying learning objectives, the uniformity of evaluation tends to be a significant concern (Sivaraman & Krishna, 2015). Here, the different perceptions of the teachers about the relevance and efficacy of Bloom's taxonomy are implicitly at work in the course of designing question items for assessment of EFL skill courses. These teachers' perceptions of Bloom's taxonomy for assessing learners' the English language skills needs to be evaluated in quantitative terms for in-depth validity.

As mentioned above, the standardization of exams has always been a significant concern in the overall pedagogical process. Bloom's taxonomy offers a model for instruction and assessment. NCAAA (2019) in Saudi Arabia enjoined universities to follow Vision 2030, which aims to standardize the educational procedures, specifically standardization of their exams to facilitate the learners' final achievements. Hence, this research focuses on how questions are designed by various English teachers for their courses in terms of language, instructions, and variety and level of items. The results should lead to better validity and reliability of questions for fair and authentic summative assessment. Besides, the study can serve as a guideline for teachers to use questioning as a useful tool to stimulate students' cognition. In a congenial learning environment, effective question-designing occupies a cardinal place (Estes, Welch & Ressler, 2004). Another contribution of this study is in the use of Bloom's Taxonomy. The questions have to test different learning levels in knowledge, skills, and competence domains vis-à-vis the learning outcomes. The findings of the study will help teachers make Bloom's taxonomy more effective in the triangular process of instructions, objectives, and outcomes.

The objectives of this study are as follows:
1. To examine the accuracy and balance of the final exam question papers designed for assessing English language skills.
2. To check whether teachers adhered to Bloom's Taxonomy in designing question items.
3. To explore how far the Academic Board's performance is successful in streamlining final exam question papers and achieving the CLOs and POs.

The following are the research questions of this study:
1. How do teachers design final exam papers in terms of item patterns, variety, and balance?
2. To what extent do teachers adhered to Bloom's Taxonomy in designing questions?
3. How successful has the Exam Board been in achieving balance and standardization of final exam questions to achieve the outcomes?

Literature Review

   Revisiting the previous research on the subject reveals that very few studies have been conducted on designing and standardization of EFL or EFL final exam questions specifically. Most of this kind of research is confined to low-order and high-order questions based on Bloom's Taxonomy. Bloom's theory was first formulated in 1956 (Forehand, 2010), and over the last 63 years, it has undergone various changes to remain compatible with emerging trends in pedagogical practices and assessment criteria. The history of developing standardized test items dates back to 1956 when Benjamin S. Bloom, who worked as Associate Director of the Board of Examinations at the University of Chicago, emerged as a pioneer of this idea (Krathwohl, 2002). In the same article, Krathwohl provides details about the cognitive domain pyramid, classifying into six categories i.e., knowledge, comprehension, application, analysis, and evaluation, which later emerged in a new version in 2001 that changed the nouns to verbs: remembering, understanding, applying, analyzing, evaluating, and creating as shown in Figure one.

       The findings of a study by Köksal (2018) conducted in Turkey suggested that for a valid language assessment, there should be a balance in higher and lower order comprehension questions. A similar study was carried out in Jordan, analyzing the questions of secondary school exams exclusively based on Bloom's Taxonomy cognitive level, in the end suggesting that more high- level questions need to be incorporated in the exams (Alzu'bi, 2014). In his doctoral dissertation, Lach (2014) analyzed end-of-semester final exams conducted in Chicago high school science classrooms exploring the depth and breadth of content that students learn in science classrooms. Although this research deals with the final exams, there was no focus on the content of exams in terms of language accuracy, a variety of questions, and overall standardization. Many other studies focus on the assessment mechanism, validity, reliability, and wash-back effect. The questions are also analyzed based on the challenge involved in levels of learning for the examinees in various domains (Swart, 2009). Arshadi and Lubis (2017) make a distinction between higher-order and lower-order questions to assess the cognitive skills of ELT students and concluded that appropriate designing of questions plays a crucial role in developing and creative and analytical skills, which serve as an asset in their future academic and professional pursuits. As for writing multiple-choice questions (MCQs), the general view is that these types of questions assess a lower level of learning, but by adopting some appropriate strategies, higher learning levels can also be tested, as suggested by Jovanovska (2018).

Figure 1 Levels of learning and Bloom's Taxonomy (from Jovanovska, 2018)

While analyzing an English textbook, Qasrawi and BeniAbdelrahman (2020) highlight the role of Higher Order Thinking Skills (HOTS) and Low Order Thinking Skills (LOTS) in developing language skills. However, these skills have to be assessed through Higher Level Questions (HLQs) and Lower Level Questions (LLQs). Once created, the standardized questions can be used in question banks of a Learning Management System (LMS). While discussing the steps for designing questions for a question bank for a system such as Blackboard, Hamad and Alnuzaili (2020) remarked, "We can specify the number of questions, type of question and level of difficulty, and the system will choose the questions randomly from the folders in the pool with the desired criteria. Besides, that there is a great advantage that each student will have different questions, and if they have multiple attempts, they will have different questions each time."

However, assessments of speaking skills have always been very challenging. Ounis (2017) pointed out the issues relating to the designing of speaking exam rubrics. The issue of the imbalance of exam questions that affects assessment and evaluation was a crucial point in the findings of Serpil (2017). The same problem has been alluded to by Tsang (2020). Dunbar, Brooks, and Kubicka-Miller (2006) suggested pre-service training for teachers for assessing oral skills.

A close relationship exists among assessment (in which final exams play a crucial role), objectives (learning outcomes) ,and instruction (what is taught in the classroom). Continual alignment is needed among the three for learners to perform better (Airasian & Miranda, 2002). Anderson (2002) illustrates this relationship using the following triangle.
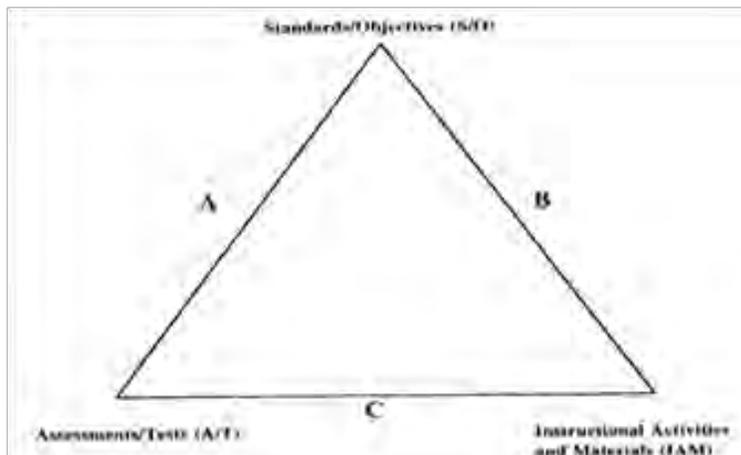
Figure 2. Triangular relationship of assessment, objectives ,and instructions (from Krathwohl, 2002)

Simultaneously, teachers' perceptions about Bloom's taxonomy and the whole process of standardization have been a focus of research because these do influence students' achievements in the long run. Malik (2019) focused on teachers' practices and perceptions to the extent of their conformity with the standards. Individual and group marking strategies evolved by the EFL/ESL teachers were examined, and subsequently, the focus shifted to students' involvement in the assessment process. The researchers confirmed the scarcity of studies that concentrate on both dimensions. English instructors responsible for designing substantial language tests may need strong test design skills, knowledge of educational measurement, and awareness of theory and concepts. The combination of these skills and this knowledge should help them bring about quality products (Giraldo, 2018). He and Shi (2008), while analyzing ESL students' views and understandings of standardized English writing tests pointed out in their concluding remarks that "Future study needs to find out how instructors who are teaching the test preparation program see their roles and the function of the test" (p. 143). ESL and EFL teachers and instructors differ a lot in their perceptions regarding the assessment process. According to Brown (2019), it is imperative for an English language evaluator must have an acute sense of the global and local standardization requirements in English proficiency tests. "The design and development of assessment criteria, procedures, and tasks should take full account of local practices and embrace a variety of assessment formats, activities, and reporting instruments that can help sample and reflect learner/user performance adequately. In other words, we are talking not just about 'assessing EFL' as such but about taking account of EFL use where appropriate in the conceptualization and design of English language assessment" (Jenkins & Leung, 2017, p. 13). Hamad (2019) emphasized the need for training for English language instructors in using Bloom's taxonomy and the specified verbs to acquaint themselves with the basics of writing test questions for achieving learning

objectives in a scientific pedagogical manner. Tajgozari and Alimorad (2019) suggested that using MCQs for assessment of writing skills is not appropriate because learners have to write sentences and paragraphs to demonstrate their writing ability. Hence, an overview of the literature on ESL/EFL teachers' perceptions regarding standardization of final tests about taxonomy and assessment rules admits further inquiry into the issue.

The standardization of a final exam paper in the English department, Majmaah University

According to Bloom's taxonomy and quality assurance departments, standardization of exams is a matter of great concern for pedagogical practices that assess learner's achievements. This study in the Department of English at the College of Education, Majmaah University, Saudi Arabia, endeavors to evaluate how far a balanced approach is realized in using Higher-Level Questions (HLQs) and Lower-Level Questions (LLQs) along with their alignment to Course Learning Outcomes (CLOs) and Program Outcomes (POs). Simultaneously, the implementation of Bloom's taxonomy and the standards set by the indigenous Academic Board for assuring the quality of exams has been measured in terms of perceptions and practice. As seen in the light of Blooms' taxonomy (Krathwohl & Anderson, 2009) and the quality documents of NCAAA, the Saudi accreditation authority, questions have to measure the learning outcomes. This consideration is supposed to be very much in the exam writer's mind while writing questions for the final exam. Generally, its implementation takes place through the Quality and Accreditation units that exist in every department. Recently, a requirement has been introduced for a cover page for each exam containing a table in which the Course Learning Outcomes (CLOs) and the Program Outcomes (POs) for each question are indicated.

Consistency with Program Outcomes

1. Knowledge			2. Cognitive Skills				3. IP Skills & Res.			4. Com., IT & Num.		5. Psychomotor	
1.1	1.2	1.3	2.1	2.2	2.3	2.4	3.1	3.2	3.3	4.1	4.2	5.1	5.2

Qs	CLO's	PO's	Total Marks	Marks Obtained	Verific a-tion
1	2.2	1.4	12		
2	2.4	2.3	5		
3	1.2	2.3	10		
4	1.1 & 2.1	1.3	10		
5	2.4	2.4	3		
6	3.1	3.1	20		
7					
			60		

Course Instructor
Iqbal
Reviewer
Mid –Terms Marks
40
Final Marks
60
Total Mark
100

Figure 3 Alignment of questions with CLOs and POs

This clipping from the cover page of Listening and Speaking's three final exams shows the the questions alignment with Course Learning Objectives and Program Objectives. Here, we also see a slight extension of Bloom's taxonomy. Since this is an advanced course, more questions fall

within the Cognitive Skills domain. Question six is about speaking that involves a higher level – Interpersonal Skills and Responsibility. As for weighting, 20 points each are allocated for knowledge, cognitive skills, and interpersonal skills. The same practice is followed in other skill exams – reading and writing with varying allocations of points, domain, and weights. At the end of the final exam, every teacher has to submit the Course Report for each course showing how the learning outcomes have been achieved by the exams and activities. Subsequently, a discussion occurs between the Course Coordinator and the course instructor, and actions for improvement to be implemented in the next exam are agreed upon. The practice just described is that which is explicitly followed by the Department of English, Majmaah University. However, for a better understanding of this practice, perceptions of the English teachers regarding Bloom's taxonomy and the Academic Board that sets national standards need to be surveyed in quantitative terms.

The Department of English at Majmaah University evolved a unique process for standardization of English language skill exams, taking into account the discrepancies observed by the internal reviewing committee when the exams had been revisited after the declaration of results and the analysis of students' feedback. To address the problem, an Academic Board was constituted to make arrangement to evaluate the exams before they were given next. This process has gone on for a couple of years but has not yet been evaluated or analyzed. However, as per the requirements of the National Commission for Academic Accreditation and Assessment (NCAAA, 2018), the questions in the exams must be standardized along with conformity with the learning outcomes that stem from the learning objectives covering the domains, such as "knowledge, cognitive skills, interpersonal skills and responsibility, communication, information technology, numerical, and psychomotor" as components of holistic assessment approach (Fathelrahman, 2019).

Hence, just before the commencement of final exams, the Head of Academic Board of the Department issued instructions for preparing examinations, which were taken as standards for this study.

Standard 1: Vary your questions (include a wide range of question types: essay, semi-essay, objective questions: multiple choice, filling in blanks, matching, true/false, etc.) whenever deemed appropriate. Remember that each type has its strengths and weaknesses.

Standard 2: Observe mark weight (distribution of marks). Keep it as balanced as possible.

Standard 3: Show good clarity (avoid ambiguous, tricky ,or vague questions). These questions will affect the validity of your tests (a test should test what it ought to be testing).

Standard 4: Proofread carefully whatever you write and keep doing this while writing exams. Remember that good writing is always rewriting, and that we very often don't see our own mistakes until shown to us by someone else.

Standard 5: Keep your test well-organized, well-presented, well-formatted, and preferably also neat-looking.

Standard 6: Write questions for all difficulty levels: advanced, complicated, moderate, easy, and very easy.

Standard 7: Use the Times New Roman font (12–14 point).

Last but not least, the evaluation of the Academic Board's functioning will also help in streamlining the process of designing and aligning final exam questions to course learning outcomes (CLOs) and program outcomes (POs). To summarize:

1) This research studies the perceptions and practices regarding the exam papers' standardization in Majmaah University.
2) It reviews the process of implementation of standardization in the university.
3) It evaluates the relevance and efficacy of Bloom's taxonomy for standardization of exams.
4) The study explores how Bloom's taxonomy is relevant to standardization of exams in particular and how it can provide a guideline for overall the standardization of exams throughout the Kingdom of Saudi Arabia under Vision 2030.
5) This will also serve as a guideline for the implementing standards and capacity building of English teachers in the rapidly progressing educational system in Saudi Arabia.
6) Standardization will lead to the elimination of fear of exams, lowered student anxiety levels, and improved student performance through an equitable exam system.

Following this, faculty members design questions and submit the question papers to the Academic Board for scrutiny for meeting the set standards. After necessary corrections, instructions, and alterations, the exam papers are finalized for printing. This practice needs to be examined for three reasons. First, how do the teachers design exams for assessing EFL students' English language skills? Second, how successful is it? Third, to what extent has it been beneficial for students to achieve the set objectives? The study aims to evaluate this internal mechanism of standardization of final EFL exams.

Methods

This case study employs a mixed method. The first segment deals with the analysis of the final exam questions pertaining to the assessment of English language skills—reading, writing, listening, and speaking—to see how this practice works. The second part of the study involves perceptions of English teachers as question designers of the role of Bloom's Taxonomy while writing the exam questions, and also their perception of the Academic Board that acts as a supervisory body to ensure that the questions are standardized, rationalized, valid, reliable, and error-free. This data was subjected to a qualitative method. The third part of the research needed to engage both the Academic Board members and the English teachers who design the question items for English language students for in-depth thematic analysis through qualitative methods of Focused Group Discussion (FGD). Hence, a sequential explanatory design of a mixed mode has been used to achieve the study aims.

Data Collection

The content of final examinations from two semesters (Spring and Fall, 2018) assessing English language skills proficiency at the Department of English, Majmaah University, Saudi Arabia, were analyzed. Then, the English teachers who design final exam papers for testing English language skills were engaged through a questionnaire. Finally, English language teachers and the Board members who set the standards participated together in Focused Group Discussion (FGD) for discussing issues arising during the course.

Sample and Sampling Techniques

As the research is a case study at Majmaah University, a non-probability convenient sampling technique was used. The first sample comprises 250 question items designed to assess English language skills during the two semesters. The second sample consists of 16 exam writers who are responsible for creating English language exams. The third sample consists of 12 English teacher

participants as exam writers and the English Academic Board members accountable for monitoring, editing, and proofreading the contents of EFL undergraduate exam papers.

Instruments
- Content of the final English exam questions to be analyzed.
- Questionnaire (Having two parts, one for the relevance of Bloom's Taxonomy in question writing and the other for evaluating the Academic Board's role in implementing the standards.)
- Focused Group Discussion (Taking participants from both English teachers who are also exam writers, and members of the English Academic Board for in-depth thematic analysis.)

The questionnaire to measure teachers' perceptions was adapted from Wozney and Abrami (2006). The number of items in the questionnaire is 20 and the responses to these items are based on a 5-point Likert scale ranging from Strongly Agree (five points) to Strongly Disagree (one point). The sample was requested to express their extent of agreement or disagreement with the items by selecting the answer from the five options of the Likert Scale. The questionnaire was vetted by veteran faculty members and was modified in view of valuable comments and feedback. A sample from 16 faculty members was subjected to Cronbach's alpha, and the internal consistency came to be 8.3, which amply validates the questionnaire reliability.

Focused Group Discussion (FGD) was conducted with members of the Academic Board and English teachers, six from each group, to see how effective the examination methods are, and how the practice of standardization of EFL final exams works.
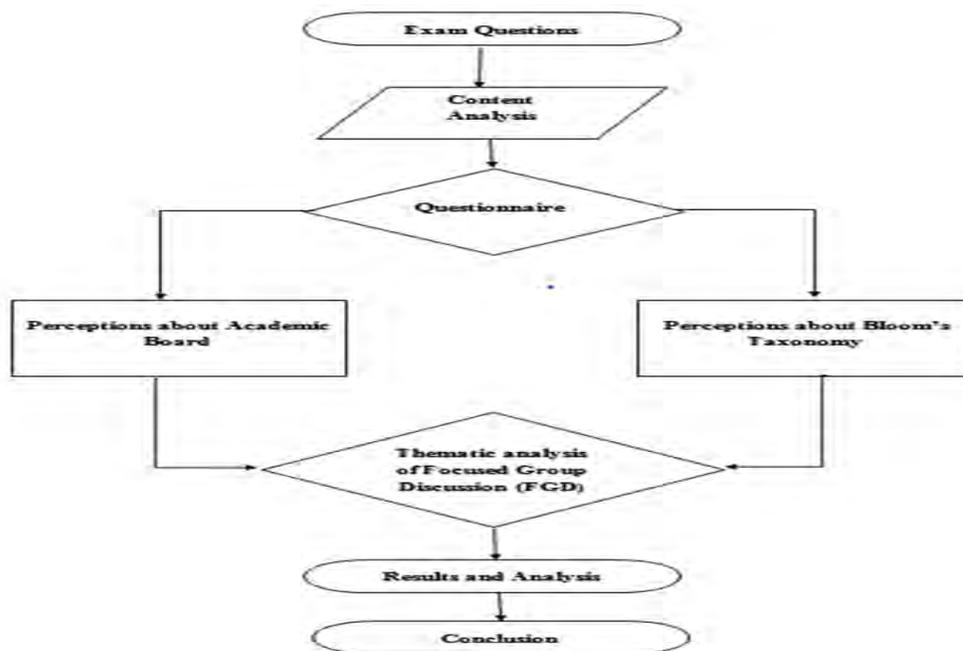The following is the theoretical framework of this study:

Figure.4 Theoretical framework of the study

Results
Content Analysis of Skill Exams
Question 1: How do the teachers design final exams in terms of item patterns, variety, and balance?

To examine how teachers distribute subjective and objective questions in terms of high and low levels, 250 exam questions based on four language skills—listening, speaking, reading, and writing—were analyzed using SPSS. The ratio of High Level Questions (HLQs) and Low Level Questions (LLQs) for multiple types of questions were calculated. Among listening skill questions, 25% were HLQs and 75% were LLQs. Speaking skill questions were 39% HLQs and 61% LLQs. Reading skill questions were 40% HLQs and 60% LLQs. Finally, among writing skill questions, 62% were HLQs and 38% LLQs. Table 1 presents detailed statistics for the questions sets in the exams.

Table 1: *Descriptive statistics of skill exam question papers*

Questions Type	Listening		Speaking		Reading		Writing		Total
	Subjective HLQs	Objective LLQs	Subjective HLQs	Objective LLQs	Subjective HLQs	Objective LLQs	Subjective HLQs	Objective LLQs	
MCQs	-	10	-	-	-	20	-	3	33
T/F or Y/N	-	10	-	15	-	5	-	5	35
FIB	8	13	-	-	4	9	1	1	36
WH Questions	-	5	12	6	15	5	10	10	63
Imperative Questions (Bloom's Verbs)	6	4	10	3	20	20	20	-	83
Total	14	42	22	24	39	59	31	19	250
%of HLQs and LLQs ratio	25%	75%	39%	61%	40%	60%	62%	38%	

Bloom's Taxonomy

Question 2: To what extent do teachers adhere to Bloom's Taxonomy in question designing?

Exam questions were classified based on Bloom's Taxonomy and the standards set by the Academic Board. The information gathered was tabulated for convenience. The questionnaire data was entered and analyzed using SPSS 22. Mean + S.D was given for quantitative variables. Frequencies and percentages were given for qualitative variables. Mean scores were calculated for all questionnaire Likert scale responses. The themes arising out of the discussion were analyzed in detail.

The data was collected using a questionnaire containing ten items that measured the English teachers' perceptions about the use of Bloom's Taxonomy by the faculty members' designing questions. Sixteen exam writers in the English Department at Majmaah University were requested to respond to the given items. The obtained data was fed into SPSS 22 for descriptive statistics. The range of the individual items' mean came to be M = 3.6–4.5 (SD = 0.51–0.71). Item #3 pertaining to striking a balance between high level and low level questions had the highest mean score (M = 4.5, SD = 0.51) with a high degree of satisfaction. The average mean score and standard deviation for the questionnaire sub-skill came to be (M=4.0, SD=0.49). The lowest mean score and standard deviation (M=3.6, SD=0.71) was received by item #8, which dealt with the relevance of Bloom's Taxonomy theory today. Table 2 illustrates the statistics of teachers' adherence to Bloom's Taxonomy.

Table 2: *Descriptive statistics of Bloom's Taxonomy*

No	Items	Mean	SD	Sum	Degree
1	I am fully conscious of Bloom's taxonomy while designing questions.	4.0000	1.03280	64.00	High
2	The instructional verbs are useful for writing questions.	4.1250	.61914	66.00	High
3	I try to strike a balance between high level questions and low-level questions.	4.5625	.51235	73.00	High
4	More questions are asked in the knowledge domain for beginners.	4.0625	.85391	65.00	High
5	More questions are asked in the skill domain for higher level students.	4.0000	.89443	64.00	High
6	I am always clear about the high level and low-level questions.	4.4375	.62915	71.00	High
7	The questions are designed to meet the learning outcomes.	4.1250	.50000	66.00	High
8	Bloom's taxonomy is still relevant for designing questions.	3.6250	.71880	58.00	High
9	Bloom's taxonomy is compatible with the course descriptions and course reports.	3.6875	.94648	59.00	High
10	Vision 2030 of the Kingdom of Saudi Arabia is followed while designing questions.	3.9375	1.12361	63.00	High
	Total	4.0	0.49	High	

Academic Board

Question 3: How successful is the Exam Board in achieving balance and standardization of final exam questions to achieve the desired outcomes?

The tool used for collecting data was a questionnaire comprising ten items regarding the faculty members' perceptions towards the Academic Board for editing and proofreading exams. The questionnaire was administered to 16 faculty members of the English Department at Majmaah University. The data was analyzed using SPSS 22. For descriptive statistics, the mean of the individual items in the questionnaire fell between the range of M=3.0–4.5(SD=0.6–1.09). The highest mean score and the standard deviation were received by item #10 pertaining to a variety of question types (M=4.5, SD=0.62), and the degree of satisfaction level was high. This is one of the significant aims of standardizing exams to offer the learner a variety of questions weighted in accordance with the time and coverage of the course items during pedagogical activities. This tends to make the exam more valid and reliable (Abuhattab & Yousef, 2017). The average mean score and standard deviation for this sub-skill of the questionnaire were (M=4.2, SD=0.45). Item #6 concerning encroachment by the Board on the exam writer's autonomy received the lowest mean score, and standard deviation (M=3.0, SD=1.09), and its level of satisfaction were moderate (2.5–3.4). This indicates a rift between the board members, who have to ensure balance in respect of numbers of items, and the allocation of grades, where the board tends to have its will and logic prevailed. Numbers of disputes also occur between the board and the exam writers, as is evident from item #7 (M=3.0, SD=1.1). This is because of a difference of opinion and the authority that lies with the board. Table 3 presents the results of the Academic Board in the implementation of standards.

Table.3 *Descriptive Statistics of Academic Board for implementing standards*

No	Items	Mean	SD	Sum	Degree
1	The Academic Board rationalizes the exams.	4.1875	.98107	67.00	High
2	The members of the board are competent enough for the job.	3.5000	.96609	56.00	High
3	The exam questions become valid after going through the board.	4.0000	.73030	64.00	High
4	Reliability of the questions is ensured.	3.9375	.57373	63.00	High
5	This practice should continue for standard exams.	4.0000	.89443	64.00	High
6	The members of the board encroach upon the autonomy of the exam writer	3.0000	1.09545	48.00	Moderate
7	This practice gives rise to disputes between the exam writers and the members of the board.	3.5625	.72744	57.00	High
8	Exam writers feel embarrassed when too many mistakes and errors are pointed out.	3.9375	.77190	63.00	High
9	The case of a course taught by more than one teacher, a uniformity of questions is achieved.	3.3125	1.01448	53.00	Moderate
10	A variety of question types with rational weightage is stricken.	4.5625	.62915	73.00	High
	Total	4.2	0.45		High

Focused Group Discussion

The contents of the questionnaire and the comments given by the respective faculty members gave way to specific themes and sub-themes for discussion. Hence, all six members were engaged in focused group discussion (FGD) – three from the exam writers and three from the Exam Academic Board for thematic analysis (Wilkinson, 2004). During the discussion, it came to the fore that initially, the Academic Board was established to rectify the typographical errors that created ambiguity and confusion for the examinees in the final exams. However, later the team responsible for corrections and post-examination feedback from the students and course coordinators needed to focus on consistency, validity, and reliability of exam items. "Sometimes, there were unintentional lapses by a few exam writers involving spelling mistakes, questions too easy or too difficult to be valid, inconsistent question items, and these needed to be corrected," remarked one of the Board members. Another Board member balked at giving comments by saying "As a member of the academic board, I am afraid my responses could not be very objective." Nevertheless, the exam writers admitted the utility of the members' tasks for streamlining the exams. "Even the frequency of dispute between the board and the paper setters lessened as the practice moved forward," reaffirmed the head of the board. It also came to the fore during the discussion that even in Finland's education system, which is considered the best in the world, standardization of exams is reported to have fairer results (Hendrickson, 2012).

The relevance of Bloom's Taxonomy to low level and high level questions also came under discussion. One of the participants in the FGD opined: "Although the taxonomy is now very old, it is amazing that it is still in many ways relevant and valid. Bloom was a legend. Appreciating him, should not, however, mean that his work shouldn't be questioned or criticized." Another pointed out that new theories show up all the time and old ones keep on adjusting their respective arguments according to the unique needs, along with the changing character of today's learners/readers of English. Also, on low level and high level questions, the FGD participants agreed that this technique helped grade the questions, starting from the knowledge and advancing to application and evaluation. Participants said that in listening, speaking, and reading skills exams, low level questions get more space, whereas writing skill exams are more topic-based rather than subjective, and these fall more in the domain of high level questions. The delineated themes and subthemes are shown in Table 4.

Table.4 *Thematic Analysis in FGD*

Sr. No.	Themes	Sub-themes
1.	Standardization of questions	a) Consistency of items b) Validity c) Reliability
2.	Efficacy of Bloom's Taxonomy	a) Relevance b) High level questions (HLQ) c) Low level questions (LLQ)
3.	Assessment/Evaluation	a) Uniformity of exam b) Allocation of grades c) Weighting

| 4. | Alignment of questions with Learning Outcomes | a) Course Learning Outcomes (CLOs) |
| | | b) Program Learning Outcomes (PLOs) |

Discussion

An analysis of the pattern of the questions in the final exam papers for four basic skills, i.e., listening, reading, and writing, has been done. The speaking exam does not involve any writing; the faculty member individually evaluates the speaking capabilities and awards the marks according to the prescribed standards in an assessment rubric to give it more objectivity as is mentioned by Tsang (2020). However, speaking exams are challenging and need more consideration for validity and reliability (Ounis, 2017). There are three graded listening in listening and speaking courses (Listening Speaking 1, Listening Speaking 2, and Listening Speaking 3). The first two courses have two groups in each semester, whereas the third one has only one group. Hence, in all ten questions, papers were scrutinized from the two previous semesters for each course. The similar distribution of groups was found in reading and writing as well. The number of questions in each paper was six to seven. In six questions, each question carried ten marks as the final exam was worth 60 marks. However, with seven questions, the division of marks varied slightly in a couple of questions.

The seven standards set by the Academic Board have been followed to a larger extent as is evident from Table 1. Standard one focuses on a variety of questions and standard two relates to the weight and balance of question items. In all, 40 question papers with around 250 questions came under examination. The ratio of objective questions is always on the high side for except writing skills, which is in line with the proposal of Tajgozari and Alimorad (2019). The same pattern exists in HLQs and LLQs, which leads to a rationalized mark weight for valid and reliable assessment. Similarly, the question items are also varied as MCQs, True/False, Yes/No, fill in the blanks (FIB), Wh-questions and imperative questions. Hence, the standards set in Bloom's taxonomy and the Academic Board's instructions cannot be followed blindly. Standard two relates to the weighting and balance. The distribution of marks varies according to the course item coverage; the Academic Board only ensures that the items in each question should not cross a particular limit. For instance, out of a total 60 marks, each question should not go beyond ten marks. This automatically keeps the weight in a state of balance. Clarity is ensured by Standard three. The practice is that the exams are submitted to the Academic Board, which scrutinizes them carefully. If any confusion or ambiguity exists, the paper is sent back to the exam writer, who follows the instructions and rephrases the questions or instructions wherever necessary. Hence Standard three is implemented without any problem. Standard four is just an instruction for being careful in writing exams. It is better if some peer review is made, which may help exam writers to be more vigilant. The format and organization of the exams were found be excellent, as Standard five is followed willingly by each exam writer. Standard six set by the Academic Board is an extension of Bloom's Taxonomy, which relates to the level of questions in terms of difficulty. Bloom's levels of questions are bifurcated between low level and high level. However, the Board extends it to excellent, above-average, average, below-average, and low. This standard tends to ensure that all the learners must get something to attempt; i.e., the questions should range from easy to difficult. At face value, the exams seem to follow this practice as well, particularly in terms of LLQs and HLQs. Standard seven deals with font size (12 or 14). All papers follow this standard for better readability. The practice mainly goes in line with the qualities of assessment enumerated

by Brown and Race (2013) as validity, reliability, transparency, inclusivity, authenticity and fairness.

The responses of English instructors as given in Table 2 reveal that the theory of Bloom's taxonomy is still considered relevant today. The principles as laid down in Bloom's Taxonomy offer a guideline for the faculty members responsible for designing question papers "but it is largely left to them to bring out a balanced paper as the final outcome" (Sivaraman & Krishna, 2015, p. 8). However, its level of satisfaction was still high (3.5 – 5). Although specifics may be argued, the matter is settled one way or another, even though in some cases either the exam writer or the board has to accede to the opposing view. Overall, the high numbers for all other item responses illustrate that the role of the Academic Board is viewed as highly positive in achieving a balance and standardization of the contents of question papers. Furthermore, the results shown in Table 3 pertaining to statistics about the performance of the Academic Board showed the satisfaction of the exam writers with the process. Nevertheless, some reservations are evident in terms of encroachment on the autonomy of the teachers in writing exams. Another issue is that if a course is taught by more than one instructor, the uniformity of questions is compromised, which sometimes leads to an imbalance of weight in assessment and evaluation, eventually affecting learners' achievements, as highlighted by Serpil (2017).

The FGD participants unanimously recommended 60:40 ratios for sessional and final assessments so as to have parity with other colleges offering the same courses. At the same time, the question of weighting was also on the anvil. "The issue of determining weights on the basis of topics and outcomes is often challenging. There is no clear cut mechanism or guideline for designing the weights for each topic and item on the exam," opined one of the members of the discussion group. The overall assessment and evaluation process were also discussed. As 40% of the total evaluation was based on formative assessment, standardization of all items or exam questions was impossible. However, for the final exam, the assessment is summative. This area where standardization of exams is most important for uniformity of questions, in case a course is taught by more than one teacher. Here again the matter of allocation of grades, marks, or points was debated.

Conclusion and Recommendations

The study looked into multiple dimensions of the process of exam writing for the final EFL exams by analyzing the practice and measuring perceptions. All of this does not occur in isolation. The language of the exams, the content of the items, the distributions of marks, and aligning the questions to the course learning outcomes (CLOs) and program outcomes (POs) are intimately interconnected. The content analysis revealed that more low level questions (LLQs) are incorporated in final exams than the high level questions (HLQs). However, despite all the grievances of the exam writers, the role of the supervising body—academic or exam board—is highly positive in making the exams flawless, uniform, balanced, and standardized (Holme, 2003). The very presence of this body rendered the faculty members more mindful of making use of the revised Bloom's taxonomy in designing questions to ensure validity and reliability. After editing and proofreading by the board, the questions on the exams were largely rationalized and standardized. In addition, the study reaffirms that the theory of Bloom's Taxonomy is still relevant though expounded long ago. It was felt that the application of Bloom's Taxonomy system had

enabled teachers to set examination papers that are well balanced, testing different cognitive skills without a tilt towards a tough or easy paper perception. The process also goes in line with the instructions of National Commission for Academic Accreditation and Assessment (NCAAA) and Vision 2030 of Saudi Arabia. Even the pedagogical implication of this study cannot be overlooked. Standardization of final exams not only streamlines the process of assessment but also has a direct effect on what is taught in the classroom. Finally, standardization of exam items as analyzed in this study goes to the benefit of the learners who get better and fairer assessment with less confusion and ambiguity in the Final Achievement Test.

It is, however, recommended that this practice needs to be replicated in other departments and institutions as well. The Quality departments should take into consideration the findings of this study to improve the quality of final exam papers. In view of the findings of this case study, it can be suggested that final exams prepared by an individual writer must be scrutinized for the sake of standardization in terms of maximum course coverage; the variety in the question item; uniformity if the course is taught by more than one teacher; clear, concise and precise instructions; validity and reliability of the exam papers for fair assessment; and the alignment of questions with CLOs and POs to ensure the implementation of revised Bloom's taxonomy. With awareness of the themes discussed in the FGD framework, it is suggested to hold seminars on assessment and quality issues and devise comprehensive guidelines for exam writers to follow.

Limitations
This study was confined to participants who are all male members of a single English Department. This was because the process of standardization of question papers was exclusively practiced here. Another limitation pertains to the role of the participants in the study. Some of the exam writers were also members of the Academic Board, so it was difficult for them to be entirely unbiased on both sides. However, this issue was partly resolved in FGD.

About the Authors:
Majed Othman Abahussain is an Assistant Professor at Majma'ah University, Saudi Arabia. He holds a Ph.D. in Applied Linguistics with more than 10 years of experience in ESL and EFL. His research mainly focuses on the use of communicative methods in ELT in Saudi Arabia.
ORCID ID: http://orcid.org/0000-0002-5424-9157

Muhammad Iqbal is a Lecturer at Majma'ah University, College of Education (Department of English). He has an M.A. in English language and literature with CELTA from Cambridge University. He has more than 20 years' teaching experience in ESL and English literature. Computer Assisted Language Learning (CALL), Communicative Methods in ELT and integration of technology in ESL and EFL classrooms, and electronic education are some of his major research interests.ORCID ID: http://orcid.org/0000-0003-2460-7504

Imran Khan is currently serving as a Lecturer in the Department of English, College of Education, Majma'ah University, Saudi Arabia. He has an M.A. in English. Presently, he is also a Ph.D. scholar at the University of Pahang, Malaysia. His research interests include technology-integrated teaching for EFL and ESL.   ORCID ID: http://orcid.org/0000-0002-4767-5051

References

Adams, N. E. (2015). Bloom's taxonomy of cognitive learning objectives. *Journal of the Medical Library Association: JMLA, 103*(3), 152-153. doi: 10.3163/1536-5050.103.3.010

Airasian, P. W., & Miranda, H. (2002). The role of assessment in the revised taxonomy. *Theory into practice*, *41*(4), 249-254. https://doi.org/10.1207/s15430421tip4104_8

Alfallaj, F. S. S., & Al-Ahdal, A. A. M. H. (2017). Authentic assessment: Evaluating the Saudi EFL tertiary examination system. *Theory and Practice in Language Studies*, *7*(8), 597-607. https://doi.org/10.17507/tpls.0708.01

Alzu'bi, M. A. (2014). The Extent of Adaptation Bloom's Taxonomy of Cognitive Domain in English Questions Included in General Secondary Exams. *Advances in Language and Literary Studies*, *5*(2), 67-72. https://doi.org/10.7575/aiac.alls.v.5n.2p.67

Anderson, L. W. (2002). Curricular alignment: A re-examination. *Theory into practice*, *41*(4), 255-260. https://doi.org/10.1207/s15430421tip4104_9

Ashadi, R. I., & Lubis, N. (2017). A Survey on the Levels of Questioning of ELT: A Case Study in an Indonesian Tertiary Education. *Advances in Language and Literary Studies*, *8*(3), 26-31.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice, 30*(1), 3-12. https://doi.org/10.1111/j.1745-3992.2010.00195.x

Brown, S. & Race, P. (2012). Using Effective Assessment to Promote Learning. In L. Hunt, & D. Chalmers, (Eds.), *University Teaching in Focus: a learning-centred approach* (pp. 74-91). Camberwell, VIC: ACER.

Brown, J. D. (2019). 5 Global Englishes and the International Standardized English Language Proficiency Tests. *Critical Perspectives on Global Englishes in Asia: Language Policy, Curriculum, Pedagogy and Assessment*, 64.

Black, P., & William, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education)*, *21*(1), 5-31.

Bates, C., & Hopkins, A. (2007). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education), 21,* 5-31. https://doi.org/10.1007/s11092-008-9068-5

Crespo, R. M. et al. (2010). *Aligning assessment with learning outcomes in outcome-based education*. In *IEEE EDUCON 2010 Conference* (pp. 1239-1246). IEEE. https://doi.org/10.1109/EDUCON.2010.5492385

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in health sciences education*, *10*(2), 133-143. https://doi.org/10.1007/s10459-004-4019-5

Dunbar, N. E., Brooks, C. F., & Kubicka-Miller, T. (2006). Oral communication skills in higher education: Using a performance-based evaluation rubric to assess communication skills. *Innovative Higher Education*, *31*(2), 115-128. doi: 10.1007/s10755-006-9012-x

Estes, A. C., Welch, R. W., & Ressler, S. J. (2004). Questioning: Bring your students along on the journey. *Journal of Professional Issues in Engineering Education and Practice, 130*(4), 237-242. https://doi.org/10.1061/(ASCE)1052-3928(2004)130:4(237)

Fathelrahman, A. (2019). *Measuring the Effectiveness of Holistic Approach to Teaching and Assessment used in Teaching a Decision Support Systems (DSS) Course.* https://doi.org/10.31124/advance.9771806.v1

Forehand, M. (2010). Bloom's taxonomy. *Emerging perspectives on learning, teaching, and technology*, *41*(4), 47-56. Retrieved from:
https://www.d41.org/cms/lib/IL01904672/Centricity/Domain/422/BloomsTaxonomy.pdf

Giraldo, F. (2018). Language assessment literacy: Implications for language teachers. *Profile Issues in TeachersProfessional Development*, *20*(1), 179-195.Hendrickson, K. A. (2012). Assessment in Finland: A Scholarly Reflection on One Country's Use of Formative, Summative, and Evaluative Practices. *Mid-Western Educational Researcher*, *25*(1-2), 33-43. Retrieved from.https://www.mwera.org/MWER/volumes/v25/issue1-2/v25n1-2-.pdf

Hamad, M. M., & Alnuzaili, E. S. (2020). Steps of Designing a Personal Questions Bank in a Pedagogical Way. *Theory and Practice in Language Studies*, *10*(6), 638-646.

He, L., & Shi, L. (2008). ESL students' perceptions and experiences of standardized English writing   tests. *Assessing   Writing*, *13*(2),   130-149.   https://doi-org.sdl.idm.oclc.org/10.1016/j.asw.2008.08.001

Holme, T. (2003). Assessment and quality control in chemistry education. *Journal of Chemical Education*, *80*(6), 594-596. https://doi.org/10.1021/ed080p594

Inbar-Lourie, O. (2012). Language assessment literacy. *The encyclopedia of applied linguistics*. https://doi.org/10.1002/9781405198431.wbeal0605

Jenkins, J., & Leung, C. (2017). Assessing English as a lingua franca. *Language testing and assessment*, 103-117.

Jovanovska, J. (2018). Designing effective multiple-choice questions for assessing learning outcomes. *Infotheca*, *18*(1), 25-42.

Karamustafaoğlu, S., Sevim, S., Karamustafaoğlu, O., & Çepni, S. (2003). Analysis of Turkish high-school chemistry-examination questions according to Bloom's taxonomy. *Chemistry Education Research and Practice*, *4*(1), 25-30. https://doi.org/10.1039/B2RP90034C

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into practice, 41*(4), 212-218. https://doi.org/10.1207/s15430421tip4104_2

Krathwohl, D. R., & Anderson, L. W. (2009). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage public

Köksal, D., & Ulum, Ö. G. (2018). Language assessment through Bloom's Taxonomy. *Journal of language and linguistic studies*, *14*(2), 76-88.

Kunen, S., Cohen, R., & Solman, R. (1981). A levels-of-processing analysis of Bloom's taxonomy. *Journal   of   Educational   Psychology*, *73*(2),   202-211. https://doi.org/10.1037/0022-0663.73.2.202

Lach, M. C. (2014). *Teacher-Generated Final Exams in High School Science: Content, Rigor, and Assessment Literacy,* (Unpublished Doctoral dissertation). Retrieved from http://www.indigo.lib.uic.edu:8080/bitstream/handle/10027/18824/Lach_Michael.pdf

Malik, S. I. (2019). Assessing the teaching and learning process of an introductory programming course with Bloom's taxonomy and Assurance of Learning (AOL). +*International Journal of Information and Communication Technology Education (IJICTE)*, *15*(2), 130-145.

National Center for Academic Accreditation and Assessment. (2018, October 10). Retrieved October 31, 2019, from https://ncaaa.org.sa/enportal/Pages/default.aspx.

NCAAA. (2019). Handbook for NCAAA Key Performance Indicators. *National Commission for Academic Accreditation & Assessment*, Riyadh, Kingdom of Saudi Arabia. Retrieved from https://ncaaa.org.sa/enportal/Pages/default.aspx

Nusche, D. (2008). *Assessment of Learning.* Paris: OECD publishing. doi:10.1787/244257272573

Ounis, A. (2017). The assessment of speaking skills at the tertiary level. *International Journal of English Linguistics*, *7*(4), p95.

Qasrawi, R., & BeniAbdelrahman, A. (2020). The Higher and Lower-Order Thinking Skills (HOTS and LOTS) in Unlock English Textbooks (1st and 2nd Editions) Based on Bloom's Taxonomy: An Analysis Study. *International Online Journal of Education and Teaching*, *7*(3), 744-758.

Serpil, Ö. Z., & Derin, A. T. A. Y. (2017). Turkish EFL instructors' in-class language assessment literacy: Perceptions and practices. *ELT Research Journal*, *6*(1), 25-44.

Sivaraman, S. I., & Krishna, D. (2015). Blooms taxonomy–application in exam papers assessment. *Chemical Engineering (VITU)*, *12*(12), 32. Retrieved from https://www.researchgate.net/profile/Ilango_Sivaraman/publication/285883763_Blooms_Taxonomy-_Application_in_Exam_Papers_Assessment/links/56641aaf08ae418a786d32d4/Blooms-Taxonomy-Application-in-Exam-Papers-Assessment.pdf

Staub, D., & Kirkgöz, Y. (2019). Standards Assessment in English Language Teacher Education. *Novitas-ROYAL (Research on Youth and Language)*, *13*(1), 47-61. Retrieved from https://files.eric.ed.gov/fulltext/EJ1214199.pdf

Swart, A. J. (2009). Evaluation of final examination papers in engineering: A case study using Bloom's Taxonomy. *IEEE Transactions on Education*, *53*(2), 257-264. https://doi.org/10.1109/TE.2009.2014221

Tajgozari, M., & Alimorad, Z. (2019). Iranian EFL students' perceptions of criteria for assessing students' written performance. *Global Journal of Foreign Language Teaching*, *9*(1), 2-9.

Taras, M. (2005). Assessment–summative and formative–some theoretical reflections. *British journal of educational studies, 53*(4), 466-478. https://doi.org/10.1111/j.1467-8527.2005.00307.x

Tsang, A. (2020). The relationship between tertiary-level students' self-perceived presentation delivery and public speaking anxiety: A mixed-methods study. *Assessment & Evaluation in Higher Education*, 1-13.

Vella Briffa, P. (2013). Explicating an English speaking examination: Challenges and considerations. *Symposia Melitensia, 9,* 193-208. Retrieved from https://www.um.edu.mt/library/oar/handle/123456789/8748

Wilkinson, S. (2004). Focus groups. In Breakwell, G. M. (ed.), *Doing social psychology research* (pp. 344-376). The British Psychological Society and Blackwell Publishing Ltd https://doi.org/10.1002/9780470776278.ch14

Wozney, L., Venkatesh, V., & Abrami, P. (2006). Implementing computer technologies: Teachers' perceptions and practices. *Journal of Technology and teacher education*, *14*(1), 173-207. Retrieved from https://www.learntechlib.org/p/5437/