Test Development Strategies Used by Saudi In-Service Teachers of English as a School Subject

Ibrahim S. Al-fallay

Department of English Language and Literature, King Saud University, Riyadh, Kingdom of Saudi Arabia

E-mail: ifallay@ksu.edu.sa

Received: June 8, 2017 Accepted: June 18, 2017 Published: June 25, 2017

doi:10.5296/ijl.v9i3.11368 URL: https://doi.org/10.5296/ijl.v9i3.11368

Abstract

This study investigates how in-service teachers of English as a school subject (ESS) in Saudi schools deal with test construction, administration, and score analyses. To answer the study's questions, a 60-item questionnaire was constructed according to the suggestions and recommendations of language testing specialists reported in the literature. The questionnaire was built according to seven dimensions: specifications/blueprints, test writing, moderation, test administration, scoring, analyses of students' scores, and item banking. 199 female and male intermediate and high ESS teachers in Saudi schools, with different years of experience in teaching English, completed the questionnaire. The findings indicated that ESS teachers in Saudi schools do not usually follow suggested recommendations pertinent to the above-mentioned dimensions. Their tests are written without planned specifications or clear blueprints. Besides, ESS teachers- regardless of their gender, years of experience, or the level they teach- rarely moderate their tests. They rarely analyze their student test scores or store good testing items in an item bank for future use. The study provides suggestions and recommendations to improve the current situation. Workshops, seminars, and on-the-job training should be conducted. Academic departments in Saudi universities responsible for English teacher preparation programs should introduce courses related to language testing if they do not have any or increase their numbers if they already have some. The study concludes with suggestions for future research.

Keywords: Test specifications, Blueprints, Test construction, Moderation, Item bank, Test development

1. Introduction

Amid great challenges and high competition, having reliable, valid, and more importantly, accurate assessment tools and procedures in the Saudi educational system is necessary. The great challenges are manifested in the current situation of all Saudi high school graduates who have to obtain high scores on standardized achievement tests, administered nationwide by the National Center for Assessment in Higher education, an autonomous assessment agency. The high competition among Saudi youths to win a seat in Saudi public universities is not an easy task to accomplish. The frantic race among Saudi public university to win national and international accreditation has annually decreased the number of newly admitted undergraduates to these universities.

In the year 2000, the Saudi Ministry of Education began to implement a new assessment system based on the concepts of criterion-referenced testing, called the continuous assessment program (AlDawood, 2004). The application began in the classrooms of first-year elementary school children. Six years later, this system was the only official assessment procedure used in Saudi elementary schools. Elementary school teachers are asked to assess learners' performance based on their achievement of certain objectives. The used techniques usually include formative tests, short quizzes, observations, and assignments (AlDawood, 2004). This adds to the importance of having trained teachers who could accurately assess students' performance.

The importance of reliable, valid and practical assessment is needless of extra emphasis. The unchallenged fact is that a well-written test not only protects students’ rights but also makes the pedagogical process more meaningful and beneficial backwash more possible. It has always been contended that a person who is capable of teaching a subject is also able to accurately assess the performance of her/his students. Spolsky (1978) drew the attention to a commonly advocated misconception that a person who knows how to teach, knows also how to test. In fact, such misconception has always been there even before the time of Spolsky’s writing; it is unfortunately a common belief nowadays. The chief factors behind such misconception might be teachers' unfamiliarity with adequate procedures for writing good test, the lack of sufficient information related to test constructions in the used textbooks/teacher manuals, the dearth of teachers' in-service training programs and the disregard for language testing essences in teacher preparation programs (Grinnell, 1991; Haladyna, 1994; McNanara, 2000; Conderman, 2002). Hence, the present study aims at exploring the common practices of ESS teachers in Saudi schools while constructing, administering and analyzing their language tests.

2. Review of Selected Literature

Downing (2010:159) defined test development as "the science and art of planning, preparing, administering, scoring, statistically analyzing, and reporting results of tests." The process of assessing students’ achievement and/or proficiency passes through at least three distinct phases. In the first phase, or the pre-administration phase, three are three steps: specifying test specifications and blueprints, the actual process of test writing and test moderation. In the second phase or during testing phase, the test is usually administered to test-takers; and this is usually followed by the third stage, the post-testing phase. In the last phase of testing, a teacher or test developer usually scores the tests, analyzes the performance of testees to investigate test item psychometrics, in addition to exploring test item statistics such as item-facility and item discrimination. Finally, good testing items are continuously added to an item bank.

Although language testing specialists tackle the problem of any testing situation differently, there is an unanimous agreement on the importance of composing test specifications and specifying blueprint contents as the first major step in test construction. Stuart-Hamilton (2007:266) defined test specifications/blueprint as "[the] collection of factors which a test is intended to measure". The importance of determining in advance what to include in a test has been underscored by many language-testing specialists (Harrison, 1983; Heaton, 1988; Weir, 1990; Davies, 1990; Popham, 1992; Alderson et al., 1995; Bachman & Palmer, 1996; Alderson, 2000; Hughes, 2003; Zandi, 2014, to name just a few). Heaton (1988:13) for example stated that

> Before starting to write any test items, the test constructor should draw up a detailed table of specifications showing aspects of the skills being tested and giving a comprehensive coverage of the specific language elements to be included.

Test specifications should be devised according to some logical and sequential steps. Briefly stated, the teacher may first begin by considering language skills and elements s/he wishes to include in the test; then the time allocated to the test should be considered. The two logical steps that may follow are deciding on the testing techniques (multiple-choice format, essay questions, true/false statements, etc.) to be utilized and the number of items to be written according to each technique. A list of the major language elements should be clearly stated. In addition, the scoring processes should also be descried in details. Such test specifications may save the teacher’s time and effort during the actual process of test writing. They also serve as a guide in the selection of language elements in that a representative sample of those elements could be included instead of including only those elements that lend themselves to assessment.

A blueprint, according to Bachman and Palmer (1996:90)

> Consists of characteristics pertaining to the structure, or over all generalization, of the test, along with test task specifications for each task type to be included. … A blueprint … describes how actual test tasks are to be constructed, and how these tasks are to be arranged to form the test.

It is obvious that what Bachman and Palmer have in mind is test structure. According to them, the blueprint should include the number of parts the teacher wishes to include in his/her test. It should also include the arrangement of the actual test parts if the test consists of more than one part. The blueprint should also state the number of items in each part, in addition to some relevant information about test setting and instructions. When the two concepts, i.e. test specifications and blueprints, are taken into consideration by the teacher before devising the test, it is more likely that the constructed test will adequately serve his/her needs.

Jafarpur (2003) sought to investigate the role test-developers play in constructing multiple-choice items geared to assess reading comprehension when those test-developers work without prior constructed specifications. More precisely, he wanted to “explore how reading comprehension items constructed by different test-developers on the same text with limited moderation but without resource to specifications compare with one another” (Jafarpur, 2003:61). To achieve this goal, 6 Ph.D. candidates, who were attending a program in teaching English as a foreign language and shared the same mother tongue and cultural background, were asked to develop three sets of multiple-choice items for six different reading passages. Each set was administered to a group of students. There were no specifications guiding test-developers. Jafarpur found that there were statistically significant differences among the three groups, although the testing items were constructed for the same reading passage. He concluded that test-takers’ performance on the multiple-choice items varied considerably according to who was the test-developer. In addition, he noted that some multiple-choice items were more suitable for high-ability students whereas others suited low-ability testees. More seriously, Jafarpur reports that his “results suggest that the items may be tapping into different underlying constructs of reading comprehension. … The items constructed by different item writers without resource to specifications and blueprints exhibit variation in the skill(s) they intend to measure” (Jafarpur, 2003:69). It is clear that multiple-choice items constructed by different individuals for the same passage differ in their construct validity, although the test-developers share the same first language, cultural background and academic standards.

Although Downing (2010) proposed a 12-step model for test development, the two steps of determining test specifications and producing its blueprint are of paramount importance. He contended that "blueprinting of content ensures systematic, comprehensive, and representative sampling of the domain" to be tested (Downing, 2010:159). It follows that a fair, representative and accurately developed test would increase its reliability and validity. Coniam (2009:232) supported this claim by stating that "in terms of construct validity, they were made aware of the need for detailed specifications that were required not only for each test but also for each subtest – and the testing points that they would consciously adopt to meet these criteria." Hence, specifications and blueprints are not only essential for the whole test, but for each subsection of that test as well. Besides, the effects of test specifications on test content and construct validity were echoed by Ali (2016:64) who claimed that "the key recommendations to increase the content and construct validity of these tests include developing test specifications...".

Test moderation is another key concept in test development. It refers to the process through which a teacher surveys his/her colleagues' opinions, who usually share the same specialty, concerning his/her test items, grading and scoring. Coniam (2009) examined the quality of objective tests produced by 31 Chinese teachers of English as a foreign language. He cited one of his participants who attributed the success in test development to "careful test design and moderation procedures"(2009:265). Although his participants realized the importance of test moderation, Coniam concluded, "little moderation [actually] takes place" (2009:268).

Test moderation has also been emphasized by Daugherty (2010) and James (2010), to name just a few. Daugherty (2010:269) believed that test quality control and its quality assurance "depend on moderation procedures designed to produce consistency across assessors in qualitative judgments of student performance".

Appropriate test administration is a key factor in achieving accurate assessment and beneficial backwash. Heaton (1988) stated five considerations to be considered during the test-administration stage. The first consideration is test practicality. This is reflected in an easy and straightforward test administration. The time allocated for testing may be unexpectedly short; and hence test administrators usually find it difficult to allocate sufficient time for test administration, answer sheet collection and instruction reading. Another consideration is the availability of stationery and the clarity of the answering sheets. If test takers are asked to answer the test items on separate sheets, instructions to this effect should be clearly stated and supplemented with examples. Moreover, test administrators should be experienced with the test administration process. They should be trained to efficiently respond to emergencies. The availability of necessary equipment is also considered a major consideration in test administration. This is very clear in the case of oral/aural examinations (speaking and listening tests). Finally, test paper presentation is also important. Whether printed or typewritten, a test should be neat and legible. Heaton (1988) added that test instructions should be clear. Teachers may use their students' first language if necessary, especially in the early stages of language learning.

Hughes believed that "The best test may give unreliable and invalid results if it is not well administered" (1989:152). He divided the administration process into two parts: preparation and administration. During the preparation stage, attention should be paid to materials and equipment, examiners, invigilators, candidates and rooms. Hughes (1989) provided easy to follow instructions to ensure best results of test administration. Among the most relevant ones are detailed instructions that should be prepared for examiners and invigilators, checking all equipments and the choice of quiet and large rooms to comfortably accommodate the testees. For an ideal administration of language tests, Hughes (1989) gave, yet, more instructions, such as the early arrival of testees in testing rooms, the way test materials are distributed to test takers and the precise timing of the test.

Calderon and Gonzales (1993) pointed out that although standardized tests and teacher-made tests are similar in function, teacher-made tests are hurriedly constructed and they are not usually subject to investigation as to their reliability and validity. This statement is very serious since it is well known that the scores students obtain on teacher-made achievement tests often determine their future. Based on their scores, students may either be promoted to following levels or asked to repeat the same level. Hence, the decisions made based on those scores are very serious; post analyses of testing items become necessary. If teacher-made tests lack reliability and/or validity, how can we be confident that the scores we report reflect the actual academic standard of our students?

Without item analyses, tests and students' scores become meaningless. Bad items could be used frequently. Jafarpur (2003) reported that he found it necessary in his study to ask his

Ph.D. candidates to revise the testing items they wrote because of their low quality. However, views such as those of Calderon and Gonzales (1993) were rejected by Fitt et al. (1999) who claimed that standardized tests and teacher-made tests would give different results even when both are constructed around the same content. They also claimed that standardized tests provided by textbook publishers are low in their validity and poor in their reliability. They also contended that teacher-made tests are more difficult than standardized ones. According to them, teacher-made tests require more critical thinking. Their conclusions questioned the validity of standardized tests; they also claimed that students may perform poorly if the testing format is novel or unfamiliar. Such remarks, when added to those of Witt et al., (1994) and Stiggins (1997), demonstrate that the investigation of teacher-made tests' validity and reliability should be taken into consideration since they might be better than standardized ones. Their results could be more valid, reliable and practical.

Using a simile from financial institutions, an item bank refers to a database where testing items with acceptable indices of item facility/discrimination, reliability and validity are stored for future use (Beeston 2000; Heydari, 2015). The notion of item bank has been credited to Rasch model of item response theory (IRS) (Anzalduam 2002; Crocker & Aigina, 2008). Although Rasch model did not deal with item banking directly, the notion of item psychometric analyses and the obtained indices paved the way to the rise of item bank in language testing situations (Yuji (2010). The idea of having an item bank based on Rasch model is not new. Choppin (1976:216) wrote: "Full exploitation of the advantages inherent in the item bank concept depends on the adoption of an explicit model of test - taking behaviour, such as that proposed by Rasch.". Bergstrom & Gershon (1995) claimed four advantages to adopting IRT in item bank development. These advantages are "[1] Easy preparation of parallel test forms, [2] Comparison of individual candidate performance over time (for candidates who repeat the test), [3] Comparison of group performance over time (to evaluate overall candidate proficiency or proficiency by school, program, or specific content area) [4 and] Usage of the item bank for computerized adaptive testing" (Bergstrom & Gershon (1995:200).

It is worth mentioning that test items that are usually stored in an item bank are items of objective tests. The most popular among all are multiple-choice test items. The question stem, the correct response, the distractors and the item statistics are recorded in the item bank. Item banking has developed from a system where " Test questions (items) were frequently written, or perhaps typed, on index cards ... item statistics... were frequently written on the backs of the cards, identified by test form and date" (Weiss, 2013, p. 185) to a more sophisticated and customized computer software. The advantages of test banks could be summarized as follows: the ability to develop parallel tests of balanced difficulty whose items enjoy acceptable levels of test psychometrics, the availability of good test items that might be added to the test if needed, time and effort saving and the possibility of reviewing and assessing curricula goals (Rudner, 1998; Anzaldua, 2002; Heydari, 2015). The major drawback of item banks is that database is limited to items of objective testing.

3. Method

3.1 Aims

The paper aimed at exploring the actual process followed by in-service ESS teachers in Saudi schools during the three stages of language test preparation and administration, namely the pre-administration, test administration, and post-administration stages. It is hoped that this paper will draw the attention of teachers to their positive practices which might lead to the enforcement the positive sides; it may also enable teachers to discover the drawbacks in their practice during testing. Suggestions and recommendations that may improve the current state of affairs will also be provided.

3.2 Participants

199 intermediate and high school teachers participated in this study. They were 87 females and 112 males. 82 of the participants were teachers at public schools; 117 participants were ESS instructors at private schools. Each participant holds a BA degree in English. They varied in their years of experience as ESS teachers. Table 1 summarizes the number of participants, their gender, and their years of experience in teaching ESS.

Table 1. The Study participants according to the study independent variables

Gender	Experience	Group	Intermediate school	Group	High school	**Total**
Male	Less than 5 years	1	23	7	18	41
	5-10 years	2	19	8	19	38
	More than 10 years	3	18	9	15	33
Subtotal			**60**		**52**	**112**
Female	Less than 5 years	4	17	10	18	35
	5-10 years	5	10	11	12	22
	More than 10 years	6	14	12	16	30
Subtotal			**41**		**46**	**87**
Total			**101**		**98**	**199**

3.3 Materials

To achieve the study goals, a sixty-item Likert-type scale questionnaire was developed. The questionnaire items were built according to the following seven dimensions: test specifications and blueprints, test writing, moderation, test administration, analyses of students' scores and item bank. The questionnaire was written in English since the participants are ESS teachers. No translated version of the questionnaire was required. The questionnaire items were mostly derived from the reviewed literature pertinent to the recommended practices in the field of language testing. The questionnaire is divided into three parts. The cover page of the questionnaire asked the participants demographic questions about their gender, years of teaching ESS, the type of school they work for (public versus private) and the school level they teach. The first part of the questionnaire contained 27 items (items 1-27); it was designed to explore the actual practices of the participants during the pre-administration stage. The second part consisted of 14 items (items 28-42, with the exception of item 37). It was designed to investigate the teachers' practices during the testing stage. For the post- administration stage, 19 items (item 37 and items 43-60) were used to examine the teachers' practices after test administration. The questionnaire was distributed in April 2017. The seven dimensions of the study and their corresponding items are given in Table 2.

Table 2. Dimensions of the study and their corresponding items

Dimensions of the study	**Corresponding questionnaire items**
Specifications and blueprints	1-6, 13-15, 17-20
Test writing	7-12, 16, 23-27
Moderation	21, 22
Test administration	28-36, 38-42
Scoring	54-56, 60
Analyses of students' scores	43-49, 57-58
Item Banking	37, 50-53, 59

The participants were asked to respond to the questionnaire items by choosing from among five alternatives: "never", "rarely", "often", "usually" and "always". Five points were allocated for a response with "always", four points for "usually", three points for "often", two points for "rarely", and one point for "never". Two reverse items were used (items 8-9). The possible score on the questionnaire ranged from 300 points, which indicate that the teacher completely abides by the instructions and recommendations of language testing specialists, to 60 points, which means that the teacher does not follow the recommended guidelines while dealing with language-testing situations.

3.4 Procedure

The questionnaire was distributed in Riyadh five educational offices, Saudi Arabia. School selection was random in that all intermediate and high school names were assigned numbers and decoded into a computer. Then random numbers corresponding to various schools were generated. 46 schools were contacted, 23 public and 23 private. The participants were selected from 24 intermediate schools and 22 high school. 300 questionnaires were distributed; only 211 questionnaires were returned. The returning percentage was 70.33%. Of these 211 returned questionnaires, 12 questionnaires were discarded due to incomplete responses. 199 questionnaires were statistically analyzed.

3.5 Questions of the Study

The study sought to answer the following two questions:

1- How do in-service ESS teachers in Saudi schools deal with the three stages of language test preparation and administration?

2- Are there any differences among the participants' practices based on their gender, years of experience and/or the school level they teach?

4. Results and Discussions

The first step in our analyses was to ensure the reliability and validity of the study questionnaire. Reliability or internal consistency was calculated using Cronbach's α statistic. Cronbach's α coefficient for all responses was (α = .832), for male intermediate school teachers (α = .852), for female intermediate school teachers (α = .894), for male high school teachers (α = .851) and for female high school teachers (α = .7594). The obtained Cronbach α coefficients were of acceptable reliability level (Gliem & Gliem, 2003:87; Lance et al., 2006:205).

The content and face validity of the study instrument was assessed by asking four professors of applied linguistics, two full professors and two associated professors, to referee the questionnaire. Their comments and suggestions were taken into consideration. Besides, the construct validity of the questionnaire was measured by investigating the inter-item correlations between items of a specific dimension and the total of that dimension. The inter-item correlations between dimension items and the dimension total ranged from .923, the correlation coefficient between the total scoring dimension and its items, to .776, the correlation coefficient between the total analyses of students' scores dimension and its total. The remaining correlation coefficients were as follows: .886 for the specification and blueprint dimension, .887 for the test writing dimension, .819 for the moderation dimension, .835 for the test administration dimension and .918 for item banking dimension. All correlations were statistically significant at N = 199, $p \leqslant$ .0001. The statistical significance was either strong or very strong (Cox, 2014, p. 175). This means that the items of a specific dimension are more related to that dimension than to the other ones. This adds to the construct validity of the questionnaire.

The analyses of the study data are divided into two parts. The first part takes into consideration the responses of the participants to the questionnaire as a whole based on the three independent variables of the study: gender, school level and years of experience. The second part deals with the analyses of individual questionnaire items.

4.1 Analyses of the Total Scores on the Questionnaire

The means and the standard deviations of all participants' responses to all questionnaire items were calculated. For clarity and ease of presentation, only mean totals are reported.

Table 3. Means and standard deviations of the participants' totals on the questionnaire

Gender	Experience	Intermediate school		High school	
		Mean	Standard deviation	Mean	Standard deviation
Male	Less than 5 years	152.652	5.936	178.556	3.347
	5-10 years	148.421	4.694	173.421	2.567
	More than 10 years	140.889	5.758	169.800	4.570
Female	Less than 5 years	158.764	4.521	162.722	5.529
	5-10 years	151.700	0.823	160.333	3.229

	More than 10 years	144.500	3.674	152.125	7.482

All means of the participants are low. The highest mean was that of the male high school teachers with less than 5 years of experience (178.556 points). This mean represents only 59.52% of the possible height score on the questionnaire (300 points). The male intermediate teachers with more than 10 years of experience had the lowest means (140.889), which means that they were the group of teachers whose practices were not in accordance with the instructions of language testing specialists. As a whole, this may indicate that the study participants do not usually follow instructions and suggestions pertinent to the recommended ways of test construction, administration, and score analyses. However, there are obvious differences in the magnitude of the participants' abidance by these guidelines and recommendations. Table 3 also shows that the male high school teachers outperformed their female counterparts. However, this was not true in the case of intermediate school teachers. The means of the female intermediate school teachers were higher than those of their male counterparts.

To investigate whether the observed differences among the means of the participants on the questionnaire totals were statistically significant, A One-Way Analysis of Variance (ANOVA) was utilized. The dependent variable was the participants' totals on the seven dimensions of the questionnaire; the independent variable was the participants' factors combined (gender, school level and years of experience). The observed differences among the groups' means were statistically significant [*F* (6 , 1386) = 71.660 , *p* = .000].

With a significant *F* value, Scheffe's post hoc comparison statistic was used to compute the significance level of the mean differences among the participants' means. The mean difference (MD) and the *p* values are given in Table 4.

Table 4. Mead differences and *p* values of Scheffe *F*-test for the differences among the participants’ means on the questionnaire

Groups		1	2	3	4	5	6	7	8	9	10	11	12
1	MD	0.000											
	p ≤	1.000											
2	MD	4.231	0.000										
	p ≤	.706	1.000										
3	MD	11.76*	7.532*	0.000									
	p ≤	.000	.025	1.000									
4	MD	6.113	10.34*	17.88*	0.000								
	p ≤	.159	.000	.000	1.000								
5	MD	.952	3.279	10.81*	7.065	0.000							
	p ≤	1.000	.989	.001	.265	1.000							
6	MD	8.152*	3.921	3.611	14.27*	7.200	0.000						
	p ≤	.013	.909	.953	.000	.296	1.000						
7	MD	25.90*	30.14*	37.67*	19.79*	26.86*	34.06*	0.000					

	$p \leq$	.000	.000	.000	.000	.000	.000	1.000					
8	MD	20.77*	25.00*	32.53*	14.66*	21.72*	28.92*	5.135	0.000				
	$p \leq$	.000	.000	.000	.000	.000	.000	.485	1.000				
9	MD	17.15*	21.38*	28.91*	11.04*	18.10*	25.30*	8.756*	3.621	0.000			
	$p \leq$	.000	.000	.000	.000	.000	.000	.007	.940	1.000			
10	MD	10.07*	14.30*	21.83*	3.958	11.02*	18.22*	15.83*	10.70*	7.078	0.000		
	$p \leq$	.000	.000	.000	.875	.001	.000	.000	.000	.098	1.000		
11	MD	7.681	11.91*	19.44*	1.569	8.633	15.83*	18.22*	13.09*	9.467*	2.389	0.000	
	$p \leq$	.051	.000	.000	1.000	.102	.000	.000	.000	.010	.999	1.000	
12	MD	.527	3.704	11.24*	6.640	.425	7.625	26.43*	21.30*	17.68*	10.60*	8.208	0.000
	$p \leq$	1.000	.921	.000	.162	1.000	.074	.000	.000	.000	.000	.053	1.000

As the mean of the male high school teachers with less than 5 years of experience (Group 7) was the highest, the mean differences between their mean and other participants' means were statistically significant with one exception. The exception was the mean difference between the mean of the male high school teachers with less than 5 years of experience and the mean of the male high school teachers with 5 to 10 years of experience (Group 8). Similarly, the mean of the male high school teachers with 5 to 10 years of experience was the second highest mean. The mean differences between that group's mean and all other means were statistically significant. The third highest mean was that of the male high school teachers with more than 10 years of experience (Group 10). The mean differences between the mean of Group 10 and all means were statistically significant. It seems that the male high school teachers are more inclined to follow the recommended guidelines. The lowest mean was that of the male intermediate-school teachers with more than 10 years of experience (Group 3). In fact, their mean was significantly lower than all means. The second lowest mean was that of the female intermediate school teachers with more than 10 years of experience.

Although not all comparisons displayed in Table 4 were statistically significant, general remarks might be made. With the exception of the female high school teachers with more than 10 years of experience, it is obvious that high school teachers, regardless of their years of experience, were more aware than intermediate school teachers of the optimal procedures in classroom testing. This might be because their students are more advanced with reference to English proficiency than intermediate school students. High schools are considered the last stage of formal instructions. Besides, it seems that as the teachers' years of experience increase their abidance by the suggested guidelines and recommendations decreases, regardless of the school level they teach. In each group, the means of the experienced teachers were the lowest. Although this might contradict research on the role of experience in effective teaching career (Dial, 2008; Ladd, 2008; Xie, 2014), this remark seems to be supported by the claim that the early years of experience are the most important ones (Rice, 2010).

4.2 Analyses of the Individual Items of the Questionnaire

The study participants showed different responses to individual items. When the participants' responses to individual questionnaire items were considered, the second item of the specifications/blueprints dimension had the highest mean (*M* = 4.352, *SD* = 0.988). This item asked the participants whether they "decide in advance on the total points of [their] tests (Is it out of 20 points, 30 points or 40 points etc.?)".

The frequencies of this item showed that 8 participants (4.020%) chose “never”, 7 participants (3.518%) “rarely”, 5 (2.513%) participants “often”, 66 participants (33.166%) “usually” and 113 participants (56.784) “always”. This seems natural since the points allocated to short quizzes, midterms and finals are set by the Ministry of Education. However, when the participants were asked whether they "decide in advance on language components [they] ... wish to test (for example, speaking, reading, grammar)." their mean was very low (*M* = 2.111, *SD* = 1.336). 152 participants stated that they either never (43.216%) or rarely (33.166%) do that. Only 41 participants believed in the importance of specifying in advance the language elements that they will include in their tests. 21 participants (10.553%) stated that they "usually" and 20 participants (10.050%) claimed that they “always” write their tests specifications. The second highest mean (*M* = 3.940, *SD* = 0.068) of an individual questionnaire item was the mean of the seventh item of the test writing dimension. This item inquired whether the participants "include language elements that lend themselves to be assessed". The majority of the participants include items that lend themselves to be assessed. 90 participants (45.226%) agreed with such practice. 33 participants (16.583%) chose “often”, 34 (17.086%) selected “usually” and 23 (11.556%) answered with “always”.

The third lowest mean in the questionnaire (*M* = 1.950, *SD* = 0.074) was that of the first item of item banking dimension. This item investigated whether the participants have an item bank where items with good psychometric indices are stored. The lowest mean of all questionnaire items (*M* = 1.769, *SD* = 0.068) was that of the sixth item of the item banking dimension. This item asked the participants who used item bank if they would "change the order of the items and/or the options within each item". The picture we get is that the majority of the participants do not have an item bank; those who do would just copy an item from the bank without any modifications. It seems that the study participants write new test items every time they have/want to assess the abilities of their students. Such practice wastes their time, which could be wisely utilized in devising and writing exercises and supplementary materials.

5. Conclusion

The findings of the present study are not encouraging. They showed that the behavior of the study participants was in odd with the recommendations and suggestions of language testing specialists. The participants seem to write their tests with no clear and/or prior specifications or blueprints. They haphazardly include items in their tests and they usually start to prepare their tests as they flip through their students’ textbooks. Language components to be included in their tests, utilization of various testing techniques and the arrangement of those components in the testing sheets do not seem to go through careful planning. The number of items to be included in the test and the test setting are rarely considered before test administration. Test Means and standard deviations, along with item facility/discrimination statistics and reliability/validity indices, are not usually calculated. In fact, teachers do not usually perform even simple calculations such as frequencies. Hence, weak and strong students are rarely identified and misfunctioning or malfunctioning items are seldom considered or revised. With the absence of item bank where good items are stored, teachers have to write new test items every time they need/want to assess their students' abilities and achievement.

It has always been assumed that more experienced instructors are more effective teachers and testers than novice ones, the data obtained in this study lead to a different conclusion. It was found that teachers with moderate experience (teachers with less than 5 years of experience) were keener than more experienced teachers in following the suggested recommendations. Male and female teachers also differ. It was found that male teachers were more apt than female teachers to moderate their tests, analyze their test' scores, and store good items in item bank for future use. The female teachers were found to be keener than male teachers in following recommendations and suggestions when it comes to test administration and scoring. The study also showed that the teachers' practices differ according to the level they teach. High school teachers were found to utilize item bank and to analyze their test scores more than intermediate school teachers do.

What should be done to make up for such deficiencies? How could the current state of affair be improved? First, prospective teachers should be introduced to issues related to language testing during the various stages of English teacher preparation. Departments in Saudi colleges should pay more attention to language testing by introducing courses pertinent to language testing and/or increasing the number of language testing courses presently available. Language testing specialists at Saudi universities and experienced English language supervisors should hold seminars and workshops to English language teachers. This would expose teachers to new techniques and strategies in the field. Their questions and remarks could also be discussed during these seminars and workshops. On-the-job-training may be a valuable asset that could improve teachers' performance.

Future research could explore whether there are differences in the practice of public vs. private schools. The nationalities and age of English language teachers could also be taken as variables. This study was limited to teachers with BA in English; future studies may consider MA or diploma holders or teachers who have graduated from teachers' colleges. The impact of seminars and workshops on teachers' performance could also be examined. The effect of on-the-job-training may also be investigated. This study was limited to teachers' practice concerning language testing. Future studies could investigate the actual practice of teachers related to language learning and teaching and whether their practices are in consonance with theories of second/foreign language learning and teaching and recommendations and suggestions of specialists in the field. More dimensions related to language testing such as practicality may also be considered in the future.

References

AlDawood, H. (2004). *The Reality of Continuous Assessment of Mathematics in Early Grades of Girls' Elementary Schools*. Unpublished MA Thesis. King Saud University, Riyadh, Saudi Arabia.

Alderson, J. (2000). *Assessing Reading*. Cambridge: Cambridge University Press.

Alderson, J., Clapham, C., & Wall, D. (1995). *Language Test construction and Evaluation*. Cambridge: Cambridge University Press.

Ali, M. (2016). A Study of the Validity of English Language Testing at the Higher Secondary Level in Bangladesh. *International Journal of Applied Linguistics & English Literature*, *5*(6), 64-75. http://dx.doi.org/10.7575/aiac.ijalel.v.5n.6p.64

Anzaldua, R. (2002). *Item Banks: What, Where, Why and How*. Paper presented at the Annual Meeting of the Southwest Educational Research Association, Austin, TX February 2016.

Bachman, L., & Palmer, A. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Bergstrom, B., & Gershon, R. (1995). Item Banking. In James C. Impara (Ed.), *Licensure Testing: Purposes, Procedures, and Practices* (pp. 187-204). Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska-Lincoln.

Conderman, G., & Koroghlanian, C. (2002). Writing Test Questions Like a Pro. *Intervention in School and Clinic, 37*(2), 83-87.

Coniam, D. (2009). Investigating the quality of teacher-produced tests for EFL students and the effects of training in test development principles and practices on improving test quality. *System*, *37*(2), 226–242.

Cox, M. (2014). Conundrums in Benchmarking eGovernment Capabilities? Perspectives on Evaluating European Usage and Transparency. *Electronic Journal of e-Government*, *12*(2), 170-178. Retrieved from www.ejeg.com/issue/download.html?idArticle=351

Crocker, L., & Aigina, J. (2008). *Introduction to Classical and Modern Test Theory*. Mason, Ohio Cengage Learning.

Daugherty, R. (2010). Summative Assessment by Teachers. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (3rd ed., Vol. 3, pp. 384-391). Oxford, UK: Academic Press an imprint of Elsevier.

Davies, A. (1990). *Principles of Language Testing*. Oxford: Blackwell.

Dial, J. (2008). *The Effect of Teacher Experience and Teacher Degree Levels on student Achievement in Mathematics and Communication Arts*. (Unpublished doctoral dissertation). Baker University, Baldwin City, Kansas, United States. Retrieved from https://www.bakeru.edu/images/pdf/SOE/EdD_Theses/Dial_Jaime.pdf

Downing, S. (2010). Test Development. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education*, (3rd ed., Vol. 4, pp. 159-169). Oxford, UK: Academic Press an imprint of Elsevier.

Gliem, J., & Gliem, R. (2003). *Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales*. Paper presented at the Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education, Columbus, OH: The Ohio State University, October 2003. Retrieved from www.ssnpstudents.com/wp/wp-content/uploads/2015/02/Gliem-Gliem.pdf

Grinn, P. (1991). *Developing Effective Classroom Tests*. West Lafayette, Indiana: Kappa Delta Pi.

Haladyna, T. (1994*). Developing and Validating Multiple-choice Test Items*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Harrison, A. (1983). *A Language Testing Handbook*. Bethlehem, PA: ELTS.

Heaton, J. (1988). *Writing English Language Tests*. London: Longman.

Heydari, P. (2015). Item Response Theory (IRT): State of the Art. *Modern Journal of Language Teaching Methods*, *5*(1), 134-144. Retrieved from http://mjltm.org/files/cd_papers/r_17_170607182144.zip

Hughes, A. (2003). *Testing for Language Teachers* (2nd ed.). Cambridge University Press, Cambridge.

Jafarpur, A. (2003). Is the Test Constructor a Facet? *Language Testing, 20*(1), 57-87.

James, M. (2010). An Overview of Educational Assessment. In P. Peterson, E. Baker, and B. McGaw (Eds.), *International Encyclopedia of Education* (3rd ed., Vol. 3, pp. 161-171). Oxford, UK: Academic Press an imprint of Elsevier.

Ladd, H. (2008). "*Value-Added Modeling of Teacher Credentials: Policy Implications*." Paper presented at the second annual CALDER research conference, "The Ins and Outs of Value-Added Measures in Education: What Research Says," Washington, D.C., November 21, 2008.

Lance, C., Butts, M., & Michels, L. (2006). The Sources of Four Commonly Reported Cutoff Criteria What Did They Really Say? *Organizational Research Methods*, *9*(2), 202-220. https://doi.org/10.1177/1094428105284919

McNamara, T. (2000). *Language Testing*. Oxford: Oxford University Press.

Popham, W. (1992). A Tale of Two Test-Specification Strategies. *Educational Measurement: Issues and Practice*, *11*(2), 16-17.

Rudner, L. (1998). *Item banking* (Report No. EDO-TM-98-05). Washington, DC : ERIC Clearinghouse on Assessment and Evaluation. Retrieved from ERIC database. (ED423310)(https://eric.ed.gov/?id=ED423310).

Spolsky, B. (1978). Introduction: Linguists and Language Testers. In B. Spolsky (Ed.), *Advances in Language Testing Series: 2: Approaches to Language Testing*. Arlington, VA: Center for Applied Linguistics, pp. v-x.

Stuart-Hamilton, I. (20017). *Dictionary of Psychological Testing, Assessment and Treatment* Second Edition. London, UK: Jessica Kingsley Publishers.

Weir, C. (1990). *Communicative Language Testing*. New York: Prentice Hall.

Weiss, D. (2013). Item banking, test development, and test delivery. In K. Geisinger, B. Bracken, J. Carlson, J. Hansen, N. Kuncel, S. Reise, & M. Rodriguez (Eds.), *APA Handbook of Testing and Assessment in Psychology, Vol. 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology* (pp. 185-200). Washington, DC, US: American Psychological Association.

Xie, M. (2014). The Relationship between Teachers' Knowledge, Attitude and Belief with the Implementation of Inquiry-Based Learning in Zhengzhou, China. *International Journal of Learning, Teaching and Educational Research*, *8*(1), 149-161. Retrieved from http://www.ijlter.org/index.php/ijlter/article/view/169/70

Yuji, N. (2010). Application of Rasch measurement to item banking in language testing. *Educational Studies*, *52*, 167-177.

Zandi, H. (2014). The effect of test specifications review on improving the quality of a test. *Iranian Journal of Language Teaching Research*, *2*(1), 1-14.

Copyright Disclaimer