*International Journal of Instruction* *January 2018 ● Vol.11, No.1*
*e-ISSN: 1308-1470 ● www.e-iji.net* *p-ISSN: 1694-609X*
*pp. 195-210*

Received: 21/06/2017
Revision: 11/09/2017
Accepted: 16/09/2017

# Test Specifications and Blueprints: Reality and Expectations

**Ibrahim S. AlFallay**
Prof., Faculty of the Department of English Language and Literature, College of Arts, King Saud University, Riyadh, Saudi Arabia, *ifallay@ksu.edu.sa*

This study investigates to what extend do teachers of English as a school subject (ESS) in Saudi schools follow recommendations and guidelines suggested by language testing specialists in developing tables of specifications and preparing blueprints to their formative and summative language tests. To answer the study questions, a thirteen-statement Likert-scale questionnaire was developed and validated. The questionnaire was completed by 199 female and male ESS teachers in Saudi schools with different years of experience who teach ESS in public and private schools to intermediate and high school level students. The results indicated that the study participants rarely follow the recommended guidelines in preparing their test specifications and blueprints. It was also found that the participants usually prepare their tests without prior planning. They do not specify in advance language skills and elements they are going to include in their tests or the scoring methods they are going to follow. Language elements that lend themselves to be tested have the priority. Significant differences among the participants according to gender, years of experience, school types the participants work for and the level they teach were found and reported. The study is concluded with recommendations that might be helpful to in-service ESS teachers.

Keywords: tables of specifications, blueprints, test constructions, formative language tests, summative language test

## INTRODUCTION

Forty years ago, Spolsky in his description of the pre-scientific period in the history of language testing wrote: "During this period, and in this approach, language tests are clear the business of language teachers, or, in more formal situations, of language teachers promoted or specially appointed as examiner. No special expertise is required: if a person knows how to teach, it is to be assumed that he can judge the proficiency of his students" (1978:5). Spolsky was describing an era in the history of language testing that preceded the time of his writing. He might be referring to the 1940s and 1950s of the last century. Unfortunately, forty years after Spolsky's writing, the same misconception is prevalent. It is a common belief that a good classroom teacher is a good test developer. Besides, the lack of rigorous analyses of test items and students'

**Citation:** AlFallay, I. S. (2018). Test Specifications and Blueprints: Reality and Expectations. *International Journal of Instruction*, 11(1), 195-210. https://doi.org/10.12973/iji.2018.11114a

scores on tests written by teachers and/or in-house developed assessment procedures has deepened this misconception. Students may pass or fail the test without consideration of test psychometrics and/or test item analyses. With the wide spread of teacher made tests in almost all schools and universities and with the common utilization of these tests in all educational stages and for all purposes, there is a need to reconsider the statement that anyone who could teach could also accurately assess her/his students' achievement and proficiency. It is rare that test items are considered by teachers after they report their students' scores. Bad items or malfunctioning alternatives and options are overlooked.

With the importance given to test results developed by classroom teachers, investigating the ways teachers usually follow to construct their tests becomes necessary. We could claim that any testing situation consists of at least three related phases: the pre-testing phase, the testing phase and the post-testing phase. The pre-testing phase consists of three ordered stages. In the first stage, test developers prepare a list of items they will include in their tests. These test items are usually derived from course objectives, teaching syllabi and/or instruction goals. This stage is referred to as writing test specification tables blueprint preparations. The stage that follows is the stage of actual test writing; then the moderation stage comes as the last stage in this phase. The scope of this paper is limited to the first stage of the pre-testing phase: the stage of writing and preparing test specification tables and blueprints. The aim of this paper is to investigate how classroom teachers of English as a school subject (ESS) in Saudi schools, who are also test developers, deal with this stage of test preparation. Their practices will be compared to the guidelines and recommendations suggested by language testing specialists. At the end, it is hoped that the actual practices of classroom teachers would reveal how teacher-made tests are written. The practices that agree with or contradict the recommendations and suggestions of language testing specialists would help in drawing the attention to the behaviours that would be reinforced and those that would be modified. In other words, it is hoped that the real practices of ESS teachers while constructing their language tests (reality) meet the standards and guidelines of language testing specialists (expectations). It is an investigation of reality and expectations.

## REVIEW OF SELECTED LITERATURE

Preparing a table or tables of test specifications is sometimes referred to as test blueprints (Downing, 2006; Stuart-Hamilton, 2007; Seo & Jong, 2015; Ali, 2016, to name just a few). A subtle difference between the two terms might be claimed as follows: "Test specifications provide guidelines for item writers...on what content may be tested and how items must be written. These specifications lead to test blueprints that outline test design and the number of questions to be tested in each score reporting category" (Oregon Department of Education, 2016, p. 1). It seems that writing test specifications is a prerequisite to preparing blueprints. First, a test developer decides on language elements s/he wants to include in the test; then the actual writing of test items, their numbers, the scoring procedures that are going to be used and test divisions are issues related to blueprinting. A blueprint is a first draft that test developers are going to consider and, if necessary, modify to reach the final version of the test. In this paper, the use of one term entails the other, unless otherwise stated.

Stuart-Hamilton (2007:266) defined test specifications as "[the] collection of factors which a test is intended to measure". Noveanu (2015:84) contended that "...specifications refer to the design of a plan that is used to develop the assessment indicating the main features to be covered". Champagne (2015) provided a similar definition. She stated, "[test] specifications include the relative emphasis the different components of science knowledge and understanding will receive, the kinds of items (selected and constructed response items, hands-on) that will be used, and the content of the background material that will be surveyed" Champagne (2015:88). Alderson, et al. (1995) provided a more accessible definition. They claimed that "A test's specifications provide the official statement about what the test tests and how it tests it" (Alderson, Clapham, and Wall, 1995:9). If these definitions, and many more, are considered, test specifications might be operationally defined as a list, table of content and/or a prior plan that should include, but not limited to, details on language elements/components to be included in the language test, in addition to the total points allocated, the total number of items to be included, test duration, the proposed testing techniques to be used and the scoring methods. The factors when deciding on particular vocabulary and specifying language skills/elements and/or certain grammatical rules to be included in a language test are the backbone of test specifications.

At this stage of test development, a test developer does not need to specify the actual test items. For example, a classroom teacher decides to assess her students' comprehension of the present progressive tense. At the stage of writing test specifications, the teacher clearly outlines the number of items she is going to write to assess this grammatical element (for instance, three items), the points she will allocate to each testing item (2 points) and the test format she is going to use (multiple choice format).

A blueprint, according to Bachman & 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." Downing (2006:9) provided a clearer definition where he stated that " A *test blueprint* defines and precisely outlines the number (or proportion) of test questions to be allocated to each major and minor content area and how many (what proportion) of these questions will be designed to assess specific cognitive knowledge levels" [emphasis in the original]. Similar views on blueprints were also stated in the literature (for example, Van Dyk & Weideman, 2004; East, 2015; Foote, 2015; Hinenoya & Lyster, 2015; Nejad & Mahmoodi-Shahrebabaki, 2015; Van Dyk, 2015; Yaraki et al., 2015; Beaulieu-Jones & Proctor, 2016; Hiver & Al-Hoorie, 2016; Saadatnia et al., 2016; Sims & Kunnan, 2016; Freeman, 2017) to name just a few.

Davidson and Lynch's model (2002) does not help much; and similar suggestions to writing tables of specifications were given by Alderson et al. (1995), Zandi et al. (2014) and Ali (2016), to mention a few. There is no single table of specifications that fits all needs. In short, there should be a table of specification for each test section (Matlock & Turner, 2016). A more comprehensive and detailed table of specifications may also

include test setting and testees' score analyses. A blueprint might be operationally defined as a detailed plan that includes lists of actual language elements to be included in the test, in addition to specifying the number of test parts and their arrangement, the points allocated to each part/item and a consideration of test setting and directions. It is the first draft a test developer would develop and then revise as needed.

Tables of specifications are needed in the development of tests for all purposes, levels of education and various disciplines. Fives & DiDonato-Barnes (2013) provided guidelines which classroom teachers may find useful in developing in-class summative tests. Tables of specifications are also needed in developing formative tests, quizzes and continuous assessment. Besides, standardized tests that are administered to hundreds of thousands of students are usually constructed according to predetermined tables of specifications (Bay-Borellim et al., 2010; College Board, 2015; Schoenfeld, 2015; Embretson, 2016; In'nami et al., 2016; Scholtz, 2017).

The use of test specifications produces tests of equal difficulty and discrimination. Besides, their use would also increase test reliability, validity and practicality (Chase, 2007; Fives & DiDonato-Barnes, 2013; CoPo, 2015; Patil et al., 2015), test taking strategies (Kashkouli et al., 2015) and grading consistency and strictness (Bonner, 2016). This does not only add to the importance of preparing tables of specification before test construction, but also shows that the way a table of specification is developed may alter students' scores. Hence, the step of preparing tables of consideration is crucial since invalid and unreliable results may be obtained.

## METHOD

The study design is a quantitative, quasi-experimental design where a 13-statement Likert scale questionnaire was developed according to the suggestions and recommendations of language test specialists pertinent to the preparation and development of test specifications and blueprints (Taylor et al., 2016:202). The participants' responses to the questionnaire statements were statistically analyzed and the obtained results were reported.

### Aims

The paper aimed to investigate how ESS teachers in Saudi schools develop their formative and summative tests, with reference to the preparation of test specifications and blueprints. Their practices in preparing their tables of specifications and blueprints prior to the actual stage of test writing were explored and compared to the guidelines and recommendations of language testing specialists reported in the literature. By knowing their actual practices, it is hoped that positive practices will be reinforced and negative ones will be modified.

### Participants

199 ESS intermediate and high schools teachers in Saudi schools participated in the study. They were 87 female teachers and 112 male teachers. 82 ESS teachers work in public schools and 117 teachers are ESS teachers in private schools. The ESS teachers in private schools composed 58.80% of the study sample. The participants were 60 male

intermediate ESS teachers, 41 female intermediate ESS teachers, 52 male ESS high school teachers and 46 female ESS high schools teachers. The participants were 76 ESS teachers with less than 5 years of experience, 60 with 5 to 10 years of experience and 63 with more than 10 years of experience. 101 participants were ESS teachers in intermediate schools; 98 participants were high school teachers of ESS. All participants have a BA in English. Their ages ranged from 25 years to 53 years.

The questionnaire was distributed in April 2017. The study participants belonged to the five educational offices of Riyadh Educational Zone. School selection was random. All public and private intermediate and high schools in the five educational offices were assigned numbers. Then 23 public schools and 23 private schools, a total of 46 schools with 24 intermediate and 22 high schools, were drawn from the study population. Although school selection was random, the participant selection follows convenience or cluster sampling (Fraenkel & Wallen, 2009:98). It was hard, if not impossible, to randomly select the participants. 300 questionnaires were distributed; 211 of them were returned with a return percentage of 70.33%. 12 questionnaires were incomplete; hence, they were discarded, leaving only 199 questionnaire for further analyses.

**Data Collection Instruments**

The study instrument was part of a large project to investigate the actual practices of ESS in Saudi schools concerning the three phases of language test construction and administration, namely, the pre-testing phase, the testing phase and the post-testing phase. The participants were asked to complete the 13-statement Likert scale questionnaire. The participants would respond to each statement by choosing whether they “never”, “rarely”, “often”, “usually” or “always" practice the statement suggestion. If a participant admitted that s/he has never practiced the statement suggestion, a value of (1) was given to his/her response; 2 points were given to the response "rarely", 3 points to "often", 4 points to "usually" and 5 points to "always". A minimum score an ESS teacher would have is 13 points and the maximum score is 65 points. This means that a 65 point would indicate a total conformity with the suggestions of language test specialists while developing and preparing test specifications and blueprints. A score of 13 may imply that the ESS teacher does not take into consideration necessary steps in preparing test specifications and blueprints. The questionnaire was written in English; and since the study participants are ESS teachers, there was no need to hand a translated version (an Arabic version) of the questionnaire. There was no reverse item in the questionnaire.

The validation of the study instrument went through two phases. First, the reliability of the participants' responses was assessed using Cronbach's α statistic. Cronbach's α for the specification and blueprint total was .886. When the reliability index for each questionnaire statement was calculated, the reliability indices ranged from .664, the reliability index of the first statement, to .862, the reliability index of the eighth statement. The obtained reliability indices are considered acceptable (Gliem & Gliem, 2003:87; Lance et al., 2006:205). Three points are worth mentioning here. First, as Spiliotopoulou (2009:150) claimed, "Low size of the coefficient alpha might not always indicate problems with the construction of the tool; whereas large sizes do not always

suggest adequate reliability." Second, Cronbach's α statistic is considered a conservative reliability index (Cortina, 1993; Sijtsma, 2009); hence, it was used to get results that are more reliable. Third as Cronbach's α is sensitive to sample size, the calculation of the reliability indices for each group in the study may lead to false interpretation.

To ensure the questionnaire face and content validity, a panel of four professors in applied linguistics were consulted. The study aims and the questionnaire were reviewed; their comments and suggestions were taken into consideration. To assess the construct validity of the study instrument, Pearson product-moment correlations, corrected for item/total or part-whole overlap, were utilized. The obtained correlation indices ranged from .577, the corrected correlation between the third statement and its total, to .855, the corrected correlation between the first statement and its total. The statistical significance of the corrected correlation indices ranged from moderate statistical significance to strong statistical significance (Cox, 2014, p. 175).

**Questions of the Study**

The study sought to answer the following two question:

1- Do ESS teachers in Saudi schools prepare and develop test specifications and blueprints in accordance with specialists' suggestions and recommendations before they write or construct their formative and summative?

2- Would there by significant differences among the participants if the study four independent variables, namely, gender, years of experience, school type and school level, are considered?

**FINDINGS**

The analyses of the study will be divided into two parts. First, the analyses of the questionnaire statements without consideration of the study four independent variables will be undertaken. The results of these analyses would lead to the answer of the first study question. Then in-depth analyses of each independent variable: the participants' gender, years of experience, school type and school level, will be conducted with the aim of answering the second question of the study.

**Analyses of the total questionnaire scores**

The means (*M*), standard deviations (*SD*), mode and standard error of the means (*SEM*) of the participants' scores on the questionnaire as a whole regardless of the study variables are displayed in Table 1.

Table 1
Means, standard deviations, mode and standard error of means of the participants' scores on the questionnaire

| Statement | *M* | *SD* | Mode | *SEM* |
|---|---|---|---|---|
| 1 | 2.111 | 1.336 | 1.000 | 0.095 |
| 2 | 4.352 | 0.988 | 5.000 | 0.070 |
| 3 | 2.553 | 1.469 | 2.000 | 0.104 |
| 4 | 2.804 | 1.388 | 2.000 | 0.098 |

| | | | | |
|---|---|---|---|---|
| 5 | 2.286 | 1.390 | 2.000 | 0.099 |
| 6 | 3.678 | 1.305 | 4.000 | 0.093 |
| 7 | 1.985 | 0.670 | 2.000 | 0.048 |
| 8 | 1.899 | 0.932 | 2.000 | 0.066 |
| 9 | 3.211 | 1.444 | 4.000 | 0.102 |
| 10 | 3.151 | 1.675 | 5.000 | 0.119 |
| 11 | 2.663 | 1.471 | 2.000 | 0.104 |
| 12 | 2.397 | 1.340 | 2.000 | 0.095 |
| 13 | 2.764 | 1.463 | 1.000 | 0.104 |
| Mean | 2.758 | 0.305 | 2.846 | 0.022 |
| Total | 35.854 | 3.969 | 37.000 | 0.281 |

Table 1 shows that the participants' total mean was 35.854. This is a very low mean since the highest score a participant may get on the questionnaire is 60 points. This represents 59.76% of the highest possible score. The table also shows that the participants' means on the second statement which asked them whether they "decide in advance on the total points of ... [the] tests (Is it out of 20 points, 30 points or 40 points etc.?)" was the highest. This is understandable since the Ministry of Education determines the points allocated for each school subject activities. The lowest mean was that of the eighth statement which inquired whether the participants "come up in advance with a detailed list of language elements from which ... [they] choose to include in ... [their] test (example, Grammar: declarative statement in the simple past tense)". The mode of this statement (2 points) shows that the participants "rarely" do that. In fact, 68 of the participants (34.17%) chose the option "never" and 106 of them (53.27%) chose the option "rarely". 174 participants (87.44%) confessed that they "never" or "rarely" prepare their tables of specifications. It seems that the participants' common practice is to open their students' textbooks and include items that lend themselves to testing. The thirteen statement asked the participants whether they "ensure the clarity of test instructions [before test administration]". Although the mean of this statement was not the lowest among all, its mode was the lowest which indicates a variation in the participants' responses. 52 participants (26.13%) claimed that they "never" do that, 49 participants (24.62%) "rarely ensure the clarity of their tests before administration, 28 of them (14.07%) "often" do this pretest checking. 34 participants (17.09%) admitted that they "usually" check their tests before administration; 36 of them (18.09%) claimed that they "always" ensure their test clarity. In general, the descriptive statistics indices displayed in Table 1 are discouraging and disappointing. It is obvious that the behaviors of the majority of the study participants are at odd with the recommendations and suggestions of language testing specialists concerning the preparation and development of test tables of specifications and blueprints. The mean and mode of the participants' totals on the questionnaire confirm this remark.

To check whether there are statistically significant differences among the participants' responses to the questionnaire statements, MANOVA (Multivariate Analysis of Variance) was calculated. Table 2 displays MANOVA summary table.

Table 2
MANOVA summary tables for the participants' responses

| Effect | | Value | F | *df* | Error *df* | Sig. |
|---|---|---|---|---|---|---|
| Intercept | Pillai's Trace | .881 | 92.514 | 13 | 163 | .000 |
| | Wilks' Lambda | .119 | 92.514 | 13 | 163 | .000 |
| | Hotelling's Trace | 7.378 | 92.514 | 13 | 163 | .000 |
| | Roy's Largest Root | 7.378 | 92.514 | 13 | 163 | .000 |
| All Factors | Pillai's Trace | 1.947 | 1.622 | 247 | 2275 | .000 |
| | Wilks' Lambda | .072 | 2.008 | 247 | 1790 | .000 |
| | Hotelling's Trace | 4.133 | 2.696 | 247 | 2095 | .000 |
| | Roy's Largest Root | 2.655 | 24.457 | 19 | 175 | .000 |

A significant MANOVA *F* values led to the calculation of between group effects. This was done by calculating ANOVA (analysis of variance) for each questionnaire statement. The differences among the participants' means on six of the questionnaire statements were not significant. There seems to be no differences among the participants' responses to the statements that inquired whether they decide in advance on the language components they wish to test (the first statement; $F(23, 175) = .994$, $p = .475$), total point allocation (the second statement; $F(23, 175) = 1.342$, $p = .147$), the number of test items in each component (the fifth statement; $F(23, 175) = 1.558$, $p = .058$), test duration (the sixth statement; $F(23, 175) = 1.557$, $p = .058$), the arrangement of the parts in the actual test (the eleventh statement; $F(23, 175) = 1.278$, $p = .188$) and test setting (the twelfth statement; $F(23, 175) = 1.206$, $p = .245$). It seems that the participants, and before they construct their classroom tests, do not have a clear idea of what to include in their tests or the number of items they are going to devote for each language skill or element. The differences among the participants' responses to the remaining seven statements and the total were statistically significant as follows: the third statement ($F(23, 175) = 12.030$, $p = .000$), the fourth statement ($F(23, 175) = 7.329$, $p = .000$), the seventh statement ($F(23, 175) = 3.219$, $p = .014$), the eighth statement ($F(23, 175) = 1.852$, $p = .014$), the ninth statement ($F(23, 175) = 12.813$, $p = .000$), the tenth statement ($F(23, 175) = 1.629$, $p = .042$), the thirteenth statement ($F(23, 175) = 1.816$, $p = .017$) and the total ($F(23, 175) = 8.253$, $p = .000$).

Since the emphasis in this part is on the statements rather than on the study four independent variables, ANOVA with between group effects for all factors was calculated. When the four independent variables were taken as a whole, a different picture emerged. Only four significant differences among the participants' scores on the questionnaire were found. The significant differences were in the participants' responses to the third ($F(19, 175) = 6.441$, $p = .000$), seventh ($F(19, 175) = 2.525$, $p = .001$) and ninth ($F(19, 175) = 13.513$, $p = .000$) statements, in addition to the total scores ($F(19, 175) = 2.585$, $p = .001$) on the whole questionnaire. There are significant differences among the participants in their perception of the degree to which their students are aware of what to be achieved, their decisions on the testing techniques they are going to use and their prior determination of the scoring method they are going to use. Besides, there were significant differences among the participants in their responses to the questionnaire when its items are taken as a whole.

**Analyses of the questionnaire scores according to the study variables**

To answer the second study question, the participants' responses to the questionnaire statements were analysed according to the four study independent variables: gender, years of experience, school type and school level.

*Gender*

Since the participants' gender was an independent variable in this study, Table 3 below displays the means (*M*), standard deviations (*SD*) and standard error of the means (*SEM*) of the participants' scores on the questionnaire according to their gender.

Table 3
Means, standard deviations and standard error the means of the participants' scores on the questionnaire according to gender

| Statement | Male teachers | | | Female teachers | | |
|---|---|---|---|---|---|---|
| | *M* | *SD* | *SEM* | *M* | *SD* | *SEM* |
| 1 | 2.009 | 1.312 | 0.124 | 2.241 | 1.364 | 0.146 |
| 2 | 4.420 | 0.992 | 0.094 | 4.264 | 0.982 | 0.105 |
| 3 | 3.071 | 1.587 | 0.150 | 1.885 | 0.958 | 0.103 |
| 4 | 2.491 | 1.470 | 0.139 | 3.207 | 1.163 | 0.125 |
| 5 | 2.080 | 1.274 | 0.120 | 2.552 | 1.492 | 0.160 |
| 6 | 3.705 | 1.271 | 0.120 | 3.644 | 1.355 | 0.145 |
| 7 | 1.920 | 0.673 | 0.064 | 2.069 | 0.661 | 0.071 |
| 8 | 1.741 | 0.803 | 0.076 | 2.103 | 1.046 | 0.112 |
| 9 | 3.482 | 1.266 | 0.120 | 2.862 | 1.586 | 0.170 |
| 10 | 3.455 | 1.610 | 0.152 | 2.759 | 1.684 | 0.181 |
| 11 | 2.420 | 1.399 | 0.132 | 2.977 | 1.509 | 0.162 |
| 12 | 2.277 | 1.254 | 0.118 | 2.552 | 1.437 | 0.154 |
| 13 | 2.696 | 1.488 | 0.141 | 2.851 | 1.435 | 0.154 |
| Total | 35.768 | 3.651 | 0.345 | 35.966 | 4.363 | 0.468 |

To investigate whether the observed differences between the male and female participants' means on the questionnaire statements were statistically significant, independent-samples t-tests were calculated. The observed mean differences between the male and female ESS teachers were not statistically significant in their responses to the first (t (197) = 1.219, p < . 224), second (t (197) = 1.100, p < . 273), sixth (t (197) = .330, p < . 742), seventh (t (197) = 1.565, p < . 199), twelfth (t (197) = 1.439, p < .152) and thirteen (t (197) = 0.736, p < .462) statements, in addition to their totals on the questionnaire (t (197) = 0.348, p < . 728).

The ESS male teachers' mean was higher than that of their female counterparts on the third statement *t* (197) = 6.156, *p* < . 000) which asked the participants whether their "students have a clear picture of what they have to achieve and to what degree of success". By the same token, the male instructors' means on the ninth statement *t* (197) = 3.067, *p* < . 002) ("I decide in advance on the scoring method") and on the tenth statement *t* (197) = 2.968, *p* < . 003) "I decide in advance on the number of parts I will include in my test" were statistically higher than the female instructors' means. The data

also shows that the ESS female teachers were more apt to follow the guidelines and recommendations when it comes to the allocation of "points to each component" (the fourth statement $t$ (197) = 3.725, $p$ < . 000)), the decision on "the number of items for each component" (the fifth statement $t$ (197) = 2.401, $p$ < . 017), a prior planning on "language elements ... to include in ... [the] test" (the eighth statement $t$ (197) = 2.766, $p$ < . 006) and a prior decision on "the arrangement of parts in the actual test if it consists of more than one part" (the eleventh statement $t$ (197) = 2.693, $p$ < . 008). The analyses also show that there was no statistically significant difference between the means of the male and female participants on the questionnaire as a whole. Their low overall means on the questionnaire, reported in Table 3 above, show that the two groups, regardless of their gender, "rarely" develop tables of specification; they also "rarely" prepare blueprints for their tests.

*Years of experience*

The participants' years of experience in teaching English at intermediate and high schools was a variable in this study. Table 4 shows the participants' means, standard deviations and standard error of means calculated according to their years of experience.

Table 4
Means, standard deviations and standard error the means of the participants' scores on the questionnaire according to years of experience

| Statement | Less than 5 years | | | 5-10 years | | | More than 10 years | | |
|---|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *SEM* | *M* | *SD* | *SEM* | *M* | *SD* | *SEM* |
| 1 | 2.316 | 1.426 | 0.164 | 2.217 | 1.367 | 0.176 | 1.762 | 1.132 | 0.143 |
| 2 | 4.553 | 0.839 | 0.096 | 4.200 | 1.086 | 0.140 | 4.254 | 1.031 | 0.130 |
| 3 | 2.711 | 1.477 | 0.169 | 2.617 | 1.530 | 0.198 | 2.302 | 1.387 | 0.175 |
| 4 | 2.895 | 1.382 | 0.158 | 2.833 | 1.368 | 0.177 | 2.667 | 1.426 | 0.180 |
| 5 | 2.605 | 1.541 | 0.177 | 2.267 | 1.339 | 0.173 | 1.921 | 1.154 | 0.145 |
| 6 | 4.026 | 1.233 | 0.141 | 3.683 | 1.228 | 0.159 | 3.254 | 1.356 | 0.171 |
| 7 | 1.895 | 0.723 | 0.083 | 2.200 | 0.632 | 0.082 | 1.889 | 0.599 | 0.075 |
| 8 | 2.118 | 1.032 | 0.118 | 1.883 | 0.885 | 0.114 | 1.651 | 0.786 | 0.099 |
| 9 | 3.487 | 1.371 | 0.157 | 3.150 | 1.436 | 0.185 | 2.937 | 1.501 | 0.189 |
| 10 | 3.250 | 1.706 | 0.196 | 3.033 | 1.697 | 0.219 | 3.143 | 1.635 | 0.206 |
| 11 | 3.039 | 1.536 | 0.176 | 2.500 | 1.420 | 0.183 | 2.365 | 1.360 | 0.171 |
| 12 | 2.592 | 1.406 | 0.161 | 2.350 | 1.287 | 0.166 | 2.206 | 1.297 | 0.163 |
| 13 | 3.263 | 1.464 | 0.168 | 2.367 | 1.327 | 0.171 | 2.540 | 1.435 | 0.181 |
| Total | 38.750 | 3.476 | 0.399 | 35.300 | 2.670 | 0.345 | 32.889 | 3.064 | 0.386 |

In general, the participants' means are low. A quick glance at the data shows that the means of the participants with less than 5 years of experience was the highest, followed by the means of the participants with 5 to 10 years of experience. ESS teachers with more than 10 years of experience had the lowest means. To investigate whether the observed differences among the participants' means were statistically significant, ANOVA was used.

There were eight significant differences among the participants' means. Scheffe's post hoc comparison was calculated. Although the mean differences among the three groups

in their responses to the first statement $F(2, 196) = 3.305$, $p = .039$) was statistically significant at $p \leq .039$, Scheffe's was not able to capture the source of the significant differences. This is justifiable since Scheffe's test is considered a very conservative statistics. However, when Tukey's HSD (honest significant difference) was used, a statistically significant difference between the means of the first and third groups was given. This indicates that the ESS teachers in Saudi schools with less than 5 years of experience are keener on preparing their tables of specifications with the aim of deciding on language elements and skills they want to test. It seems that they have a clear idea in advance of what to test, grammar, reading comprehension, vocabulary etc. With reference to the observed significant differences among the means on the fifth statement $F(2, 196) = 4.326$, $p = .015$), the only statistically significant difference was between the means of the first and third groups, in favor of the first group. Again, the ESS teachers with less years of experience seem to follow the guidelines more than their colleagues with more years of experience. The differences among the participants' means on the sixth statement was also significant $F(2, 196) = 6.356$, $p = .002$). The mean of the first group was significantly higher than the mean of the third group. However, when it comes to deciding on the testing techniques that are going to be used in the test, the seventh statement, the participants with 5 to 10 years of experience had the highest mean. Their mean was significantly higher than the means of the two other groups $F(2, 196) = 4.584$, $p = .011$). The means of the ESS teachers with less than 5 years of experience on the eighth $F(2, 196) = 4.503$, $p = .012$) and eleventh $F(2, 196) = 4.287$, $p = .015$) statements were significantly higher than the means of the third group on the same statements. Besides, the means of the first group on the thirteenth statement $F(2, 196) = 7.888$, $p = .001$) and overall questionnaire $F(2, 196) = 62.099$, $p = .000$) were significantly higher than the means of the two other groups. When the results are taken as a whole, the participants with few years of experience are the ones who would prepare their test specifications and blueprints. The lack of many years of experience may be the reason behind this observation; they have to cautiously deal with test preparation.

*School type*

The participants' school type was one of the study variables. Private schools are usually better equipped with language labs and libraries. Hence, ESS learners who study in private schools are usually more competent English language learners. Table 5 displays the means of the ESS teachers in Saudi schools according to the school type they work for.

Table 5
Means, standard deviations and standard error the means of the participants' scores on the questionnaire according to school type

| Statement | Public schools | | | Private schools | | |
|---|---|---|---|---|---|---|
| | *M* | *SD* | *SEM* | *M* | *SD* | *SEM* |
| 1 | 2.21 | 1.349 | .149 | 2.04 | 1.329 | .123 |
| 2 | 4.29 | 1.071 | .118 | 4.39 | .928 | .086 |
| 3 | 2.45 | 1.380 | .152 | 2.62 | 1.530 | .141 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 4 | 2.83 | 1.386 | .153 | 2.79 | 1.395 | .129 |
| 5 | 2.51 | 1.484 | .164 | 2.13 | 1.303 | .121 |
| 6 | 3.45 | 1.433 | .158 | 3.84 | 1.189 | .110 |
| 7 | 1.95 | .683 | .075 | 2.01 | .663 | .061 |
| 8 | 1.94 | .837 | .092 | 1.87 | .996 | .092 |
| 9 | 3.22 | 1.414 | .156 | 3.21 | 1.471 | .136 |
| 10 | 3.05 | 1.728 | .191 | 3.22 | 1.641 | .152 |
| 11 | 2.57 | 1.423 | .157 | 2.73 | 1.506 | .139 |
| 12 | 2.32 | 1.413 | .156 | 2.45 | 1.290 | .119 |
| 13 | 2.90 | 1.445 | .160 | 2.67 | 1.474 | .136 |
| Total | 35.70 | 3.918 | .433 | 35.97 | 4.017 | .371 |

Again, the participants' means are low regardless of their school type. To investigate whether there are statistically significant differences between the means of the two groups, independent sample t-test was used; the results indicated that there were no statistically significant differences between the means of the two groups, except on their responses to the sixth statement $t$ (179) = 2.072, $p$ < . 040). The ESS teachers in private schools are more aware of the importance of limiting the time of their tests. This might be because they do not usually use the whole class time for assessment.

*School level*

When the schools level, intermediate versus high school teachers, was taken as a factor, the means of the teachers in high schools are higher than those of the intermediate school teachers. A summary of the descriptive statistics and t-test values of the participants' means on the questionnaire statements is given in Table 6.

Table 6
A summary of the descriptive statistics and t-test values of the participants' means on the questionnaire statements

| Statement | Intermediate school teachers | | | High school teachers | | | *df* | *t* | *p* |
|---|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *SEM* | *M* | *SD* | *SEM* | 197 | 2.380 | 0.018 |
| 1 | 1.891 | 1.207 | 0.120 | 2.337 | 1.428 | 0.144 | 197 | 3.304 | 0.001 |
| 2 | 4.574 | 0.726 | 0.072 | 4.122 | 1.160 | 0.117 | 197 | 6.749 | 0.000 |
| 3 | 3.178 | 1.609 | 0.160 | 1.908 | 0.953 | 0.096 | 197 | 10.845 | 0.000 |
| 4 | 1.970 | 0.877 | 0.087 | 3.663 | 1.292 | 0.130 | 197 | 1.735 | 0.084 |
| 5 | 2.119 | 1.314 | 0.131 | 2.459 | 1.451 | 0.147 | 197 | 1.249 | 0.213 |
| 6 | 3.792 | 1.275 | 0.127 | 3.561 | 1.332 | 0.135 | 197 | 4.309 | 0.000 |
| 7 | 2.178 | 0.669 | 0.067 | 1.786 | 0.613 | 0.062 | 197 | 2.441 | 0.016 |
| 8 | 1.743 | 0.594 | 0.059 | 2.061 | 1.165 | 0.118 | 197 | 1.645 | 0.102 |
| 9 | 3.376 | 1.593 | 0.158 | 3.041 | 1.259 | 0.127 | 197 | 2.913 | 0.004 |
| 10 | 3.485 | 1.629 | 0.162 | 2.806 | 1.660 | 0.168 | 197 | 1.157 | 0.249 |
| 11 | 2.545 | 1.432 | 0.142 | 2.786 | 1.508 | 0.152 | 197 | 0.412 | 0.681 |
| 12 | 2.436 | 1.374 | 0.137 | 2.357 | 1.310 | 0.132 | 197 | 1.669 | 0.097 |
| 13 | 2.594 | 1.250 | 0.124 | 2.939 | 1.642 | 0.166 | 197 | 0.097 | 0.923 |
| Total | 35.881 | 4.129 | 0.411 | 35.827 | 3.818 | 0.386 | 197 | 2.380 | 0.018 |

t-test revealed that half of the comparisons were statically significant. The means of the intermediate school teachers on the second, third, seventh and tenth statements were significantly higher than the means of the high school teachers. The high school teachers outperformed their intermediate school counterparts on the first, fourth and eighth statements. There were no statistically significant differences between the two groups in their responses to the twelfth statement which inquired whether they check test settings before test administration. The participants who teach high school level students usually know in advance the components they are going to include in their tests, the points allocated to each test part and the specific elements they are going to write as testing items.

## CONCLUSION AND RECOMMENDATION

The results of the study are not encouraging. The participants' means on the questionnaire statements revealed that the participants are either not aware of the importance of preparing test tables of specifications and blueprints or not motivated to do so. The results showed that the majority of the participants began the task of writing their formative/summative in-class tests without a clear idea of what to include in their tests. In other words, they do not plan on language elements and components they are doing to include in their tests. Their common practice seems to be to include the items that lend themselves to be tested. Having a multi-components test that cover language productive and receptive skills, in addition to grammar and knowledge of vocabulary, does not seem to be a planned task. Besides, ESS teachers in Saudi schools overlook the importance of introducing their students to and informing them about the goals and objectives of the course they study. In addition, they do not set what should be achieved and to what degree of success. The study results also showed that ESS teachers do not determine in advance the points they will allocate to each component. For example, they do not decide on the allocation of 20 points to assess their students' knowledge of grammar neither do they decide in advance on how many points they will allocate to assess their students' reading comprehension. The study participants stated that they "rarely" decide on the number of testing items in each test components. They also do not set their test duration in advance. A testing session may either last for the whole teaching period or it may last for few minutes. This depends on the time their students usually need to finish the task.

The results of the study reveal that there is much work has to be done. The craft of test writing and construction should be taken seriously. ESS teachers, regardless of their gender or level of school they teach, should begin test construction with clear tables of specifications. These tables of specifications will guide them during the task of test construction. They will save teachers' efforts and time. They will also guide them to construct representative, reliable and valid tests. University departments responsible for teacher preparation programs should include in their study plans courses pertinent to language testing. As a student teacher graduates from college, she would be able to construct her tests according to norms and guidelines of how language tests ought to be constructed.

**REFERENCES**

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. Retrieved 20 April, 2017 from (www.journals.aiac.org.au/index.php/IJALEL/article/view/2599). http://dx.doi.org/10.7575/aiac.ijalel.v.5n.6p.64

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

Bay-Borellim M., Rozunick, C., Way, W., & Weisman, E. (2010). *Considerations for developing test specifications for common core assessments*. Upper Saddle River, New Jersey: Pearson.

Beaulieu-Jones, L., & Proctor, C. (2016). A Blueprint for implementing small-group collaborative discussions. *The Reading Teacher, 69*(6), 677 - 682.

Bonner, M. (2016). Grading rigor in counselor education: A specifications grading framework. *Educational Research Quarterly, 39*(4), 21-42.

Champagne, A. (2015). Assessment: An overview. In R. Gunstone (Ed.), *Encyclopedia of Science Education* (pp. 87-89). New York: Springer Science+Business Media Dordrecht.

Chase, C. (2007). *Measurement for educational evaluation*. Boston, MA: Addison-Wesley Publishing Company.

College Board. (2015). *Test specifications for the redesigned SAT*. New York: College Board.

CoPo, A. (2015). Students' initial knowledge state and test design: Towards a valid and reliable test instrument. *Journal of College Teaching & Learning, 12*(4), 189-194.

Cortina, J. (1993). What is coefficient Alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98-104.

Cox, M. (2014). Conundrums in benchmarking government capabilities? Perspectives on evaluating European usage and transparency. *Electronic Journal of e-Government, 12*(2), 170-178. Retrieved 25 April, 2017 from www.ejeg.com/issue/download.html?idArticle=351.

Davidson, F., & Lynch, B. (2002). *A teacher's guide to writing and using language test specifications.* New Haven, CT: Yale University Press.

Downing, S. (2006). Twelve steps for effective test development. In S. Downing and T. Haladyna (Eds.), *Handbook of test development* (pp. 3-25). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

East, M. (2015). Coming to terms with innovative high-stakes assessment practice: Teachers' viewpoints on assessment reform. *Language Testing, 32*(1), 101-120.

Embretson, S. (2016). Understanding examinees' responses to items: implications for measurement. Educational Measurement: *Issues and Practice, 35*(3), 6–22.

Fives, H., & DiDonato-Barnes, N. (2013). Classroom test construction: The power of a table of specifications. *Practical Assessment, Research & Evaluation*, *18*(3), 1-7.

Foote, R. (2015). The production of gender agreement in native and L2 Spanish: The role of morphophonological form. *Second Language Research*, *31*(3), 343 – 373.

Fraenkel, J., & Wallen, N. (2009). *How to design and evaluate research in education* (7th ed.). New York, NY: McGraw-Hill.

Freeman, D. (2017). The case for teachers' classroom English proficiency. *RELC Journal*, *48*(1), 31 – 52.

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 20 April, 2017 from www.ssnpstudents.com/wp/wp-content/uploads/2015/02/Gliem-Gliem.pdf.

Hinenoya, K., & Lyster, R. (2015) Identifiability and accessibility in learning definite article usages: A quasi-experimental study with Japanese learners of English. *Language Teaching*, *19*(4), 397 – 415.

Hiver, P., & Al-Hoorie, A. (2016). A Dynamic ensemble for second language research: Putting complexity theory into practice. *The Modern Language Journal*, *100*(4), 741-756.

In'nami, Y., Koizumi, R., & Nakamura, K. (2016). Factor structure of the Test of English for Academic Purposes (TEAP®) test in relation to the TOEFL iBT® test. *Language Testing in Asia*, 6(3), 1-23.

Kashkouli, Z., Barati, H., & Nejad Ansari, D. (2015). Test-taking strategies and item specifications: Focusing on a high-stake test. *International Journal of Research Studies in Education*, 4(2), 43-56.

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.

Matlock, K., & Turner, R. (2016). Unidimensional IRT item parameter estimates across equivalent test forms with confounding specifications within dimensions. *Educational and Psychological Measurement*, 76(2), 258-279.

Nejad, B., & Mahmoodi-Shahrebabaki, M. (2015). Effects of metacognitive strategy instruction on the reading comprehension of English language learners through cognitive academic language learning approach (CALLA). *International Journal of Languages' Education, 3*(2), 133-164.

Noveanu, G. (2015). Assessment specifications. In R. Gunstone (Ed.), *Encyclopedia of Science Education* (pp. 84-85). New York: Springer Science+Business Media Dordrecht.

Oregon Department of Education. (2016). *Test specifications and blueprints*. Salem, Oregon: Oregon Department of Education.

Patil, S., Gosavi, M., Bannur, H., & Ratnakar, A. (2015). Blueprinting in assessment: A tool to increase the validity of undergraduate written examinations in pathology. *International Journal of Applied and Basic Medical Research*, 5, 76-9.

Saadatnia, M., Ketabi, S., & Tavakoli, M. (2016). EFL learners' levels of comprehension across text structures: A comparison of literal and inferential comprehension of descriptive and enumerative expository texts. *Journal of Psycholinguistic Research, 45*(6), 1499–1513.

Schoenfeld, A. (2015). Summative and formative assessments in mathematics supporting the goals of the common core standards. *Theory into Practice, 54*(3), 183-194.

Scholtz, D. (2017). The appropriateness of standardised tests in academic literacy for diploma programmes of study. *Language Matters, 48*(1), 27-47.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psycometrika, 74*(1), 107–120.

Spiliotopoulou, G. (20029). Reliability reconsidered: Cronbach's alpha and paediatric assessment in occupational therapy. *Australian Occupational Therapy Journal, 56*(3), 150–155.

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.

Seo, D., & Jong, G. (2015). Comparability of online- and paper-based tests in a statewide assessment program: Using propensity score matching. *Journal of Educational Computing Research, 52*(1), 88–113.

Sims, J., & Kunnan, A. (2016). Developing evidence for a validity argument for an English placement exam from multi-year test performance data. *Language Testing in Asia, 6*(1), 1-14.

Stuart-Hamilton, I. (20017). *Dictionary of psychological testing, assessment and treatment second edition*. London, UK: Jessica Kingsley Publishers.

Taylor, S., Bogdan, R., & DeVault, M. (2016). *Introduction to qualitative research methods*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Van Dyk, T., & Weideman, A. (2004). Switching constructs: On the selection of an appropriate blueprint for academic literacy assessment. *Journal for language teaching, 38*(1), 1-13.

Van Dyk, T. (2015). Tried and tested: Academic literacy tests as predictors of academic success. *Tijdschrift voor Taalbeheersing, 37*(2), 159-186.

Yaraki, H., Jahandar, S., & Khodabandehlou, M. (2015). A study on the effectiveness of thesaurus Iranian upper intermediate EFL learners listening comprehension ability. *Modern Journal of Language Teaching Methods (MJLTM), 5*(4), 397-410.

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