University Teachers' Test Design Practices in Kurdistan: Compliance, Question Types, and Bloom's Taxonomy Levels

Soma Hassan [1*], and Fatima Al BAJALANI [2]

[1]English Language Teaching Department, Tishk International University, Erbil, Iraq
[2]College of Languages Salahaddin University-Erbil Erbil, Iraq
Correspondence: Soma Hassan, English Language teaching Department, Tishk International University, Erbil, Iraq
Email: soma.hassan@tiu.edu.iq

Doi: https://doi.org/10.23918/ijsses.v13i2p3

Abstract: This study aims to assess university teachers' performance in following testing design requirements. The study additionally checks the way the classification of the university whether public or private and the type of exams including final, midterm, and make up, affect teachers' compliance to testing design specifications. The data for this study was collected using a checklist for evaluating the exam papers that teachers designed for testing students. The sample consisted of 104 exam papers designed by university teachers in four universities in Erbil, in Kurdistan of Iraq. The data was analyzed by the researchers through the IBM SPSS V. 28 program. The findings regarding teachers following testing design requirements showed that teachers do not have major issues regarding following testing design requirements, and essay items are one of the most used testing techniques by the teachers. Moreover, the teachers mostly used lower-level thinking questions rather than higher-level thinking questions. Besides, the teachers' workplace (public or private), and the type of exams did not affect the teachers' performance in following testing design requirements.

Keywords: Test Design, Exam Papers, Performance, University Teachers

1. Introduction and Literature review

Brown (2004) defined assessment as a continuous procedure that covers a considerably greater scope. The teacher unconsciously assesses the student's performance each time he or she answers a question, makes a comment, or tries out a new term or structure (Brown, 2004). Bachman (1995) identified assessment as the "process of quantifying the characteristics of persons according to explicit procedures and rules". It is widely known that there are two types of assessment namely formative and summative. The main purpose of formative assessment is to "form" or change ongoing classroom procedures or activities. They share knowledge while it is still feasible to shape or "form" the routine activities that involve education (Russell & Airasian, 2012; Harlen, 2005). On the other hand, summative assessments are used to evaluate the outcomes of the instruction and take the form of tests, projects, term papers, and final exams (Russell & Airasian, 2012; French et al., 2023).

Received: 12.10.2025
Accepted: 21.06.2026
Published: 21.06.2026
Hassan, S., & Al BAJALANI, F. (2026). University Teachers' Test Design Practices in Kurdistan: Compliance, Question Types, and Bloom's Taxonomy Levels. *International Journal of Social Sciences & Educational Studies*, 13(2), 53-70.

Clay and Root (2001) defined some benefits for the students and for the teachers. Researchers have displayed the benefits of testing. Such benefits are: 1- Tests are a record of the students' grades. 2- The students and the teacher will form an idea of the level of understanding of each student. 3- Tests will motivate and provide a learning experience for students. 4- Each student will be motivated either by good or bad grades to do better. 5- Each test provides a base for the other tests in which students will gain an understanding of each teacher's testing strategies. 6- It shows the students' progress in achieving the course objectives. 7- Tests provide teachers feedback on what they have achieved and what needs to be focused upon (Clay & Root, 2001; Ali et al., 2023; Boyle et al., 2023). Mertler (2004) stated that by testing students and grading them, teachers would improve their instruction, enhance students' motivation, and increase their achievement levels.

1.2 Criteria for a Good Test Design

It is crucial for teachers to take into account practicality as well as authenticity, accessibility, reliability, validity, and washback when they design their exam papers. Researchers mention the importance of reliability in language testing and point out Some factors that can affect it, including unclear instructions, student-related problems, how the test is administered, and the subjective nature of some tests (McMillan, 2018; Brown & Abeywickrama, 2018; Douglas, 2014). Test designers should provide clear instructions, employ a sufficient number of items or activities, involve unbiased raters, provide explicit evaluation criteria, maintain transparent and unbiased scoring methods, and use shorter tests more frequently in order to achieve high reliability. Researchers also emphasize how crucial it is to assess students consistently and to compile a variety of their works in order to achieve reliability (McMillan, 2018; Brown & Abeywickrama, 2018; Douglas, 2014).

Testing and assessment revolve around the idea of validity, which refers to whether an evaluation captures what it is intended to capture and generates actionable findings (Douglas, 2014). According to Fitzner (2007) and Fulcher and Davidson (2007), there are various categories of validity, including criterion-oriented validity, content validity, and construct validity. Two requirements for testing design that are challenging to have at a high level simultaneously are validity and dependability. High reliability tests frequently have lower validity, and vice versa. The test designers can increase validity while retaining reliability by using a rubric that denotes degrees of performance (Fulcher & Davidson, 2007; Clark-Fookes, 2025).

The consequences of testing on teaching and learning are referred to as washback, also known as backwash or test impact. Washback has been defined by several writers as a psychological reaction to assessment, the effect of testing on instruction and classroom management, and the extent to which tests compel teachers and students to perform actions that either facilitate or obstruct language learning. Positive washback happens when testing encourages outstanding teaching techniques, whereas negative washback happens when test material or format is based on a limited grasp of language skills and restricts the instructional environment (Chicho & Hassan, 2022).

Another criterion of evaluation known as authenticity demands students to demonstrate skills that are similar to those employed outside of the classroom and simulates real-world experiences (Brown, 2004). Genuineness and authenticity are contrasted, with genuineness being a quality of spoken or written material created in an actual communicative setting and authenticity being tied to the appropriate response to communication (Douglas, 2007). Since authenticity is a subjective and relative concept, test designers must take into account all the other elements of an effective test design, including washback, reliability, validity, accessibility, and practicality (Zheng & Iseni, 2007). Natural language, contextualized things,

meaningful and relevant topics, thematic structure, and tasks that are an exact duplicate of real-world tasks are the main components to ensuring authenticity (Brown, 2004; Douglas, 2007).

Both students and teachers must be able to take part in language testing. Accessibility for students refers to their capacity to comprehend and participate in the test, whereas for teachers it refers to the degree to which each item captures the intended construct (Russell & Airasian, 2012). Since washback is less frequently taken into account than practicality, reliability, and validity, teachers' assessment literacy and adherence to assessment norms are crucial (Brown, 2004). The logistical and administrative elements required in designing, delivering, and scoring a test are referred to as practicality in testing (Cohen, Swerdlik & Sturman, 2004). A practical test should be affordable, simple to administer and score, and finished in a reasonable amount of time (Brown, 2004).

1.3 What Should Be Taken into Consideration in Designing Tests

Bloom's Taxonomy and Rubrics are two important points that should be taken into consideration in designing and scoring exams and tests. Bloom's taxonomy was developed to pinpoint educational goals and conduct which is essential to the learning process (Forehand, 2010). The taxonomy has three levels and is divided into three domains: cognitive, affective, and psychomotor. There are six levels in the cognitive domain, which receives the most focus in schooling. The levels were changed in the 2001 revision of Bloom's Taxonomy to better suit needs of the twenty-first century, and each level's associated verbs were defined. The rubric should be taken into account when creating a test paper. Teachers and students utilize rubrics, which are lists of specified requirements or standards, to concentrate on a topic or task (Russell & Airasian, 2012). Rubrics support the growth of common knowledge, improve the validity and reliability of exams, and support students in understanding how they will be evaluated. According to Russell and Airasian (2012), there are two different sorts of rubrics: holistic and analytical, which assess performance as a whole or in accordance with each criterion.

1.4 Previous Studies

The teacher-made test was examined in terms of testing procedures, Bloom's taxonomy, and item building errors in the paper titled "An Analysis of Teacher-Made Tests: Item Types, Cognitive Demands, and Item Construction Errors" by Marso and Pigge (1991). 175 teacher-made examinations from Ohio State in the United States were gathered as a result. Multiple-choice, Matching, and Short Response Items are the testing methods that are most frequently utilized, according to data analysis. The researchers realized that these testing procedures only gauge pupils' knowledge of Bloom's taxonomy. In terms of the perspective on construction errors, the researchers discovered that only matched items have the most construction faults and are unrelated to the experience of the teachers.

A study by Öz (2014) titled "Turkish Teachers Practices of Assessment for Learning in English as a Foreign Language Classroom" looked into the teachers' use of assessment strategies. The Assessment for Learning Questionnaire was utilized to gather the study's data. The results showed that rather than essay questions or short response questions, teachers favored employing supply items like True-False, Matching, and Multiple Choice.

Soyucok and Batur (2021) did another study in Turkey to find out the most popular testing methods used by teachers and the difficulty of the questions related to the updated Bloom's Taxonomy. A qualitative study was conducted by the researchers using document analysis of 100 exam papers from the academic years 2019 to 2020. The results demonstrated that open-ended and multiple-choice questions were more frequently employed by the teachers. The teachers employed lesser levels of the taxonomy on their

questions, like knowledge and understanding, for the new Bloom's Taxonomy levels of cognitive understanding.

The mentioned papers investigated the usage of testing design requirements by teachers in general. The studies did not examine all aspects like Bloom's taxonomy, linguistic and layout accuracy, and testing techniques in their investigation. Rather than looking at all aspects at the same time, they focused on only one aspect at a time. In this study, the researchers focused on all of those aspects. Moreover, they studied the effect of the university sector, and type of exams on teachers following testing design requirements.

1.5 Purpose of the Study

The purposes of this study are to find out the teachers' performance regarding following testing design requirements, and to discover the effect of the university sector (public or private), and the type of exam papers (final, midterm, and makeup) on teachers following testing design requirements.

The research tries to answer the following questions:

1. To what extent do the teachers follow the testing design requirements in their exam questions?
2. What are the differences in the testing design requirements followed between teachers at private and public universities?
3. Are there significant differences in the testing design requirements followed in terms of the type of exam papers (final, midterm, and makeup)?

2. Methods

2.1 Participants

The context of this study was across four universities in Erbil Kurdistan, Iraq. Two universities were public universities (University of Kurdistan Hawler, Salahaddin University- Erbil), while the other two were private (Knowledge University, Tishk International University). The selected universities offer a bachelor's program in the English language. The sample of this study was the exam papers designed by the university teachers of the mentioned universities. The number of exam questions that the universities provided for this study was as follows: UKH 48, TIU 31, Knowledge 9, and Salahaddin 16 exam papers. These papers were taken from the 2020–2021 and 2021-2022 academic years. Therefore, the most recent years have been selected to prevent the gap in experience to have an influence on the analysis. The exam papers consisted of three types: final, midterm, and makeup (retake, makeup). More information is provided in tables 1, 2and 3. The exam papers were obtained from all four universities through formal office requests submitted to their respective research and exam units, requesting access to the exam papers of English departments. Following a careful review of the research tools employed in this study, and after securing approval from the relevant universities, the exam papers were collected.

Table 1: Number of the Exam Papers Obtained from the Universities

Frequency			Percent	Valid Percent	Cumulative Percent
Valid	UKH	48	46.2	46.2	46.2
	TIU	31	29.8	29.8	76.0
	Knowledge	9	8.7	8.7	84.6

	Salahaddin	**16**	**15.4**	**15.4**	**100.0**
	Total	**104**	**100.0**	**100.0**	

Table 2: The Types of the Analyzed Exam Papers

Frequency			**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	**Final**	**76**	**73.1**	**73.1**	**73.1**
	Midterm	**15**	**14.4**	**14.4**	**87.5**
	Makeup (makeup)	**13**	**12.5**	**12.5**	**100.0**
	Total	**104**	**100.0**	**100.0**	

Table 3: The Years of the Conducted Exam Paper

Frequency			**Percent**	**Valid Percent**	**Cumulative Percent**
Valid	**2020-2021**	**69**	**66.3**	**66.3**	**66.3**
	2021-2022	**35**	**33.7**	**33.7**	**100.0**
	Total	**104**	**100.0**	**100.0**	

2.2 Data Collection and Analysis

The tool of this study was a checklist prepared based on the literature (Bapir, 2016; Clay & Root, 2001; Gronlund & Waugh, 2013; McMillan, 2016). The checklist consists of 67 items to evaluate teachers' exam questions. The checklist consisted of two parts. The first was demographic information. Those questions are set to obtain information about the exam papers: university, grade, year, department, question types, and courses. The second part was the checklist items that have three sections examining different aspects of the exam papers. Part (A) examined the question types (testing techniques). Those types were (multiple-choice items, true-false items, matching items, completion items, fill-in-the-blanks, reordering/rearranging/unscrambling items, classification, cloze test items, and essay-type items). Part (B) examined the accuracy of the exam questions that had two major points (linguistic accuracy and layout accuracy). The last part (C) examined Bloom's taxonomy categories (Remember, Understand, Apply, Analyze, Evaluate, Create). Each part was divided into multiple items containing the testing design criteria. Those exam papers were analyzed based on the criteria prepared by the researchers, by ticking a box of three choices (yes, no, and non-existing). The data analysis was carried out to answer all the study questions by IMB SPSS program. Nonparametric tests like Kruskal-Wallis and Mann-Whitney tests were used to analyze the data. Due to the non-normally distributed data, those two tests were used instead of the T-test and ANOVA test. More information is provided in table 4.

Table 4: Normality test for the Checklist

	Kolmogorov-Smirnova			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
all the multiple- choice items computed together	.429	104	<.001	.609	104	<.001
all the true false items computed together	.518	104	<.001	.408	104	<.001
all the matching items computed together	.494	104	<.001	.476	104	<.001
all the completion items and fill-in-the-	.438	104	<.001	.603	104	<.001
all the rearranged and reordered, and unscrambled items computed together	.539	104	<.001	.248	104	<.001
all the classification items computed together	.531	104	<.001	.341	104	<.001
all the cloze item tests computed together	.526	104	<.001	.327	104	<.001
all the essay item tests computed together	.418	104	<.001	.413	104	<.001
all the linguistic accuracy items computed together	.536	104	<.001	.119	104	<.001
all the layout accuracy items computed together	.523	104	<.001	.361	104	<.001
all the blooms taxonomy items computed together	.244	104	<.001	.880	104	<.001

2.3 The Scale

The Chronbach Alpha test shows the alpha coefficient result of the checklist. Internal consistency is "the degree to which the items that make up the scale 'hang together'" (Pallant, 2016, p. 9). The reliability score of the checklist is .929 which is one of the highest reliability scores for 65 items while two items were deleted to ensure higher reliability score (see table 5). These scores prove that the tool used is reliable. As a consequence, the results that are to be obtained from the analysis will not be affected by reliability.

Table 5: Reliability Statistics of the Checklist

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.929	.901	65

Brown defined validity as " the extent to which inferences made from assessment results are appropriate, meaningful, and useful in terms of the purpose of the assessment " (2004, p. 31). Fitzner classified validity into seven types that are “face validity, content validity, criterion validity, construct validity, concurrent validity, external validity, and internal validity” (2007, p. 776). For this study, face validity used. The face validity of the tool was ensured by giving them to 11 jury members that are specialized in Applied Linguistics and Assessment. These jury members were asked to evaluate the checklist and examine its suitability for the research aims.

3. Findings

To answer the first research question (*To what extent do the teachers follow the testing design requirements in their exam questions?*), the researchers chose the frequency test. The frequency test shows the percentage of usage of each item in the checklist (see Appendix 1). The items have been tested by classifying them in three ways (yes, no, and non-existence).

The frequency table (see Appendix 1) shows the percentage of using each item in the exam papers. The testing technique that was mostly used in all the 104 exam papers was essay items, with 96% (N = 100) usage by the teachers. The least used testing technique was reorder/rearrange/unscramble items by only 6% (N = 6) of the teachers in their exam papers.

The multiple-choice technique was one of the techniques that the teachers did not have major problems with. However, they had a slight problem with two items in the category: providing the options in a list form and naming them by uppercase letters rather than lowercase letters. The item (The options are provided in a format of a list) was used only in 33% (N = 34) of the papers, and 14% (N = 14) of that 33% did not provide the options in the format of a list.

Matching items was one of the other categories that was analyzed. Only 17% (N = 18) of the teachers used matching items in the exam papers. There was a problem with one item in the whole Matching items category. This item was (The premises are less than the responses). Only 4% (N = 4) of the teachers designed their matching questions in the correct way regarding this item.

As the most used testing technique, essay items were used well by the teachers. The only problem in designing this test technique was that it did not provide a word limit for the students in the instruction of the test items. Only 26% (N = 27) out of 96% (N = 100) managed to provide the word limit in their essay items from their exam papers.

The linguistic accuracy, layout accuracy and blooms taxonomy of all the exam papers were analyzed. Furthermore, all the exam papers were free of any layout, and linguistic accuracy problems. Bloom's taxonomy is important to consider when choosing testing techniques. The usage of each level of bloom's taxonomy from most used to least used was: Remember (88%, N = 91), Understand (83%, N = 86), Create (28%, N = 29), Apply (20%, N = 21), Evaluate (14%, N = 14), and Analyze (13%, N = 13)). The analysis showed that the lower level of thinking was mostly used to assess students. Another feature of a good test

is validity. All the exam papers analyzed were valid as they measured what they are supposed to measure. For example, a test paper of grammar includes only testing grammatical structurers, and a test of literature tests only literary issues.

The answer of second question (*What are the differences in the testing design requirements followed by teachers at private and public universities?*) was obtained by doing a Mann-Whitney test. According to the normality test, the data had a P-value less than 0.05 through all categories. Therefore, the Mann-Whitney test was the most suitable test for answering this research question.

Table 6: Reliability Statistics of the Checklist

Test Statistics											
	all the multiple-choice items computed	all the true false items computed together	all the matching items computed together	all the completion items and fill-in-	all the rearranged and reordered, and	all the classification items computed	all the cloze item tests computed	all the essay item tests computed	all the linguistic accuracy items computed	all the layout accuracy items commuted	all the blooms taxonomy items computed
Mann-Whitney U	1131.000	1044.500	920.000	1164.000	1088.000	1168.000	1061.500	900.500	1268.000	1185.500	1184.000
Wilcoxon W	1951.000	1864.500	1740.000	1984.000	1908.000	1988.000	1881.500	2980.500	3348.000	3265.500	3264.000
Z	-1.200	-2.655	-3.573	-.951	-3.176	-1.464	-2.857	-3.179	-.337	-1.185	-.680
Asymp. Sig. (2-tailed)	.230	.008	<.001	.342	.001	.143	.004	.001	.736	.236	.497

In the data analysis, there were differences in the scores of public and private teachers in using some testing techniques. The six categories (multiple-choice, fill in the blank, classification items, linguistic accuracy, layout accuracy, and Bloom's Taxonomy) that had retained the null hypothesis (there is no significant difference between the variables). All the significant scores are greater than 0.05; therefore, they all have an insignificant difference between public and private university teachers.

The other five categories showed a significant difference in the teachers' usage of testing design standards regarding the university they work at, either public or private. With the rejection of the null hypothesis (there is no significant difference between the variables), the five categories that had a significant difference between public and private university teachers were:

1. True/false items: the test revealed a significant difference between public (median = 36, N = 64) and private universities (median = 36, N = 40), U = 1044.500, z =-1.200, p=.008, r =.11. The p-value is less than .05, which means there is a significant difference with a very low effect size of r =0.11. The mean score for the public was higher (56.11) than for the private (46.51). Thus, public university teachers use true-false items better than private university teachers with a very small effect size.
2. Matching items: the results showed a significant difference between both groups regarding the use of testing design standards, public with (Median =24, N = 64) and private with (Median =24, N = 40), U = 920.000, z =-3.573, p =.001, r = 0.35. The p- value of matching items was 0.001, which is less than 0.05, meaning there is a significant difference between both groups with a very small effect size of 0.035. The mean scores show which group used the testing technique better. Public university teachers had a mean score of 58.13, and private teachers had a mean score of 43.50. Therefore, public university teachers are better at using matching items than private university teachers.
3. Rearrange, reorder, and unscramble items: as mentioned in the other two points, there has been a significant difference in this category as well. The test revealed a significant difference between public (median = 6, N = 64) and private universities (median = 6, N = 40), U = 1088.00, z = -3.176, p =.001, r =.31. The p-value is less than 0.05, which means there is a significant difference with a very low effect size of r =0.31. Compared to private (mean = 55.50), the average score for the public is greater (mean = 47.70). Consequently, with a relatively small effect size, public university teachers use rearrange, reorder, and unscramble items more effectively than private university teachers.
4. Cloze items: the considerable difference between public (median = 6, N = 64) and private (median = 6, N = 40), U =1061.500, z =-2.857, p = 0.004, r = 0.28, was revealed by the test. There is a significant difference with a relatively small effect size of r = 0.28 because the p-value is less than 0.05. Compared to private (mean = 55.91), the mean score for the public is higher (mean = 47.04). Ultimately, with small effect size, public university teachers use cloze items more successfully than private university teachers.
5. Essay items: the last test that had a significant difference is the essay type, with the public showing a median of 6 (N= 64) and private showing a median of 6 (N= 40), U=900.500, z=-3.179, p=-001, r= 0.31. The p-value is less than 0.05. There is a significant difference with a relatively small effect size of r = 0.31. The mean score for the public (mean = 46.57) is higher than the private (mean = 61.99). In the end, private university teachers use essay items more efficiently than public university teachers, with a modest impact size.

In summary, the test showed significant differences in some categories. The categories that had no difference in their usage by both sectors were multiple-choice, fill in the blank, classification items, the linguistic accuracy, the layout accuracy, and Bloom's Taxonomy. The categories that have a significant difference in usage between both sectors were true/false items, matching items, rearrange, reorder, and unscramble items, cloze items, and essay items. True-false items have the smallest effect size of 0.11 where public university teachers performed better than private. Cloze item was the second lowest effect size of 0.28 where public sector teachers performed better than private. Finally, the three highest effect size where matching 0.35, rearrange, reorder, and unscramble items 0.31, and essay items 0.31. In the first two, again teachers from public sector performed better than private, but the last one which is essay items the private sector teachers performed better than public. With only five categories that are significantly different from eleven categories, we cannot definitely say that public or private teachers are better or worse, we can only say that there is a difference which is not a very big difference. The sectors have more similarities than differences. As for validity, both sectors' test papers were proven to be valid.

According to the normality test, non-parametric test of ANOVA (Kruskal-Wallis H) was performed to answer the third research question (*Are there any significant differences in the testing design requirements regarding the type of exam papers (final, midterm, and makeup)?*

Table 7: Kruskal Wallis test between the type of exam questions and checklist items

Test Statistics[b]											
	all the multiple-choice items computed together	all the true false items computed together	all the matching items computed together	all the completion items and fill-in-the-blank items computed together	all the rearrange and reorder and unscramble items computed together	all the classification items computed together	all the cloze item tests computed together	all the essay item tests computed together	all the linguistic accuracy items computed together	all the layout accuracy items computed together	all the blooms taxonomy items computed together
Kruskal-Wallis H	3.352	2.693	1.911	3.748	2.322	.290	4.030	2.402	.744	.541	.785
df	2	2	2	2	2	2	2	2	2	2	2
Asymp. Sig.	.187	.260	.385	.154	.313	.865	.133	.301	.689	.763	.675

Table 7 shows the results of the Kruskal Wallis test and its significant scores. The test showed an insignificant difference in the mean scores of all-checklist items in the type of exam papers, either final, midterm, or makeup exams. The Asymp. Sig score should be less than 0.05 for a test to be considered significantly different. The test results showed that the significant levels of all the scores mentioned are higher than 0.05. This means that they are insignificant to the difference in the type of the questions, either final, midterm, or makeup.

4. Discussions and Conclusion

The first research question investigated the frequency of usage of each item in the checklist. It was found that the university teachers have the least number of problems regarding following testing design requirements. All test items are used correctly by the teachers. Not giving a word limit by the teachers while designing essay questions was one of the problems most of the teachers were facing. The analysis presents that the most used type of questions among the university teachers was essay-type questions; however, the least used question type was re-arrange, re-order, and unscramble questions. The following findings contradict the findings of Arhin and Quaigrain (2017), Öz (2014), Soyucok and Batur (2021), Maros and Pigge (1991), which mentioned that the essay items were least used by the teachers in the exam papers. Nevertheless, it aligns with their finding that matching items and multiple-choice questions are one of the common ways of designing exam papers (Arhin & Quaigrain, 2017; Öz, 2014; Batur &

Soyucok, 2021; Maros & Pigge, 1991). The findings suggested that the lower level of thinking was used mostly by teachers rather than a higher level of thinking regarding Bloom's taxonomy. This finding aligns with the findings of Maros and Pigge (1991) and Gall (1970) who had a collection of five studies that found the same usage of lower levels of thinking by the teachers than higher order thinking types. Although the teachers mostly used essay questions in their exam papers, those questions were not from the higher levels of Bloom's taxonomy. The teachers' use of essay type is because of the nature of the program they are teaching and because the mentality of the students and the subjects they study is eligible to have essay items mostly that needs higher level of thinking. The case is contradicting for the rearrange, reorder, and unscramble items because they are mostly used in lower-level thinking and are not suitable for such modules. Using lower-level thinking can be due to the reason that some of the modules studied at the undergraduate level are using books and the information inside the books does not have the potential for using higher-level thinking. On the other hand, the level of the students can be another problem facing the teachers to try to choose lower level-thinking so that their students be able to answer the questions in the exam papers. The reason why the teachers have the least number of problems regarding following testing design requirements is that in some universities the heads of the departments review the exam papers before giving the permission to be used. All the mistakes will be corrected, which also happens in some other universities in the form of peer review for the exam papers.

The effect of public and private sectors on the teachers' ability to follow testing design requirements was examined. The findings propose no difference between the level of following testing design requirements in both sectors. The findings coordinate with the findings of Öz (2014) and Özkan and Yılmazer (2017) which revealed no difference between the private and public universities regarding their question types and following such standards. This insignificant difference might be due to the reason that teachers who teach in private universities are graduates of public sector universities and vice versa. Those teachers who graduated from private universities might teach in both private and public universities, the same is true for public university graduates.

The type of exam paper and its influence on following the testing design requirements was examined. The findings indicated that the type of the exam whether final, midterm, or makeup exam paper does not make any difference regarding teachers' following testing design requirements. However, these findings were not connected to any previous studies because the researchers found no studies that identified or researched this issue. The results might be because the time between midterm and final or makeup is not enough to see a significant difference.

In conclusion, the major aim of this study was to investigate university teachers' performance in designing tests. It is concluded that the university teachers do not have major problems following the testing design requirements adopted by the researchers for analyzing the teachers' exam papers. Another conclusion is that neither the type of university (public or private) nor the type of exam papers (final, midterm, and makeup) made any difference in following the testing design standards in the exam papers. This study provides some practical implications to the field of assessment in the universities in Kurdistan Iraq and even around the world. First, the tool that has been used as the Checklist is a tool that can be used for further research. This tool’s validity and reliability have been proven by SPSS program. Second, the university teachers can benefit from the tools and the results of this study to be familiar with the testing design requirements and apply them in their future exams. Moreover, the information in this study can be turned into workshop, or training courses for university teachers in Kurdistan.

As any other study, this study suffered from some limitations in the process of conducting it. First of all, due to the novelty of the topic in the Kurdistan context and the difficulty of obtaining the necessary data this point was one of the limitations faced by the researchers for the paperwork and getting permission. Moreover, the second limitation was the limited data access. The exam papers were treated very securely

and secretly by the universities. Most universities did not have the authority to provide the exam papers for research purposes. Only four universities in Erbil provided the researchers with exam papers for this research. Regardless of these limitations, this paper's contributions are far greater than the limitations. This paper provided a tool for checking the performance of the teachers, and data as an eye opener to the reality of the university teachers teaching and raising the next generation.

4.1 Recommendation

1. Some of the key necessities for this situation that teachers encounter are courses, webinars, and training sessions. Those webinars and courses will provide a reflection or stage for learning that makes the teachers lifelong learners in assessment. All university teachers need to undergo a course or be informed about the requirements for designing each type of testing technique. Each testing technique is unique and has its strong and weak aspects; the teachers need to know about these aspects and how to sufficiently use them in their papers.
2. Developing the current assessment courses being studied in universities. Many students who are graduating to be teachers lack the needed skills to test students. Moreover, their attempts to learn it outside their academic circle (universities) might cause them to misunderstand the information. Thus, more assessment courses are needed on assessment, and how to relate assessment to learning.
3. Promoting the culture of peer review for the exam papers that the teachers design. This peer review will increase the quality of the exam papers and will reduce the number of small mistakes. In addition, this will provide a platform for the teachers to learn from each other.
4. The teachers need to be aware to use higher-level thinking in their exams, which are suitable for the age of their students who are mature enough to answer the high-level questions from Bloom's taxonomy.

4.2 Suggestion for Further Research

Further studies on this topic are crucial to view and understand teachers' situations clearly. Therefore, the following are some suggestions for further research on this topic:

1. A collection of a wider range of data that contains all the universities in Kurdistan, Iraq, to give an approximate finding about the performance of the teachers.
2. Research evaluating the assessment courses in public and private universities is needed.
3. Another research can be conducted by examining the oral exam that is being performed in the universities.

APPENDIX

Appendix 1

Items Category	Checklist Items	Scale	Frequencies	Percentages
Multiple-choice Items	It has only one correct answer	Yes	33	31.7
		No	1	1.0
		Non-existence	70	67.3
	It contains three or four options	Yes	27	26.0
		No	7	6.7
		Non-existence	70	67.3
	It examines only one point at a time	Yes	34	32.7
		No	0	0
		Non-existence	70	67.3
		Yes	34	32.7

	The options are grammatically suitable for the stem	No	0	0
		Non-existence	70	67.3
	The distracters are closely related to the correct choice	Yes	34	32.7
		No	0	0
		Non-existence	70	67.3
	There are no negative stems	Yes	33	31.7
		No	1	1.0
		Non-existence	70	67.3
	The options are relevant to each other regarding word class	Yes	34	32.7
		No	0	0
		Non-existence	70	67.3
	The options are provided in a format of a list.	Yes	20	19.2
		No	14	13.5
		Non-existence	70	67.3
	There is a random distribution of the correct answers in the options.	Yes	33	31.7
		No	1	1.0
		Non-existence	70	67.3
	Excessive wordings are eliminated from the stem	Yes	34	32.7
		No	0	0
		Non-existence	70	67.3
	Irrelevant words are eliminated from the stem.	Yes	34	32.7
		No	0	0
		Non-existence	70	67.3
	The options are titled in upper case letters (A B C D) rather than lower case letters (a b c d) to avoid confusion.	Yes	8	7.7
		No	26	25.0
		Non-existence	70	67.3
True false Items	Items can be classified unambiguous as either true or false.	Yes	14	13.5
		No	0	0
		Non-existence	90	86.5
	Each item includes a single major point.	Yes	14	13.5
		No	0	0
		Non-existence	90	86.5
	Negative statements are NOT used.	Yes	12	11.5
		No	2	1.9
		Non-existence	90	86.5
	Words like (always, all, or never), which tend to make the statement false; words like (usually, often, or many) usually make the statement true are NOT used in the statement.	Yes	14	13.5
		No	0	0
		Non-existence	90	86.5
	Statements should be clear and direct.	Yes	14	13.5
		No	0	0
		Non-existence	90	86.5
	The items are equal, or false items are slightly more than true items.	Yes	12	11.5
		No	2	1.9
		Non-existence	90	86.5
	The items are equal, or false items are slightly more than true items.	Yes	12	11.5
		No	2	1.9
		Non-existence	90	86.5
	True and false items are randomized	Yes	12	11.5
		No	2	1.9
		Non-existence	90	86.5
	The statements are briefly stated	Yes	13	12.5
		No	1	1.0
		Non-existence	90	86.5

	Enough information is given so that the students do not depend on memorization.	Yes	14	13.5
		No	0	0
		Non-existence	90	86.5
	The statements do not include common knowledge so that the students who do not know the material can predict the answer.	Yes	14	13.5
		No	0	0
		Non-existence	90	86.5
	The items are ONLY answered by True or False, not by other possible responses like Yes or No.	Yes	13	12.5
		No	1	1.0
		Non-existence	90	86.5
	A distinguishable pattern of answers is avoided.	Yes	13	12.5
		No	1	1.0
		Non-existence	90	86.5
Matching Items	The premises and the responses to be matched are homogenous	Yes	19	18.3
		No	0	0
		Non-existence	85	81.7
	The premises are in a numbered column at the left, and the responses are in a lettered column at the right.	Yes	11	10.6
		No	7	6.7
		Non-existence	86	82.7
	There is only one correct response for each premise (unless mentioned otherwise in the instructions of the question).	Yes	18	17.3
		No	0	0
		Non-existence	86	82.7
	The items to be matched are on the same page.	Yes	16	15.4
		No	2	1.9
		Non-existence	86	82.7
	The premises are less than the responses	Yes	4	3.8
		No	14	13.5
		Non-existence	86	82.7
	Enough directions are provided to the students so they will not get confused about which part is the premises and which one is the response.	Yes	18	17.3
		No	0	0
		Non-existence	86	82.7
	The premises are longer than the responses	Yes	18	17.3
		No	0	0
		Non-existence	86	82.7
	No clues are provided on the premises that declare the answer in the responses	Yes	17	16.3
		No	1	1.0
		Non-existence	86	82.7
Completion and fill-in blanks Items	Only significant keywords are omitted from the statement.	Yes	32	30.8
		No	0	0
		Non-existence	72	69.2
	The omitted keywords should not make the statement lose its meaning.	Yes	32	30.8
		No	0	0
		Non-existence	72	69.2
	The items have only one correct answer.	Yes	32	30.8
		No	0	0
		Non-existence	72	69.2
	Blanks of the same length are used throughout the test.	Yes	21	20.2
		No	11	10.6
		Non-existence	72	69.2
	The blank in an item is in the middle or at the end of the statement.	Yes	25	24.0
		No	7	6.7
		Non-existence	72	69.2
		Yes	32	30.8

	Grammatical clues to the correct answer are not found in the statement.	No	0	0
		Non-existence	72	69.2
	A list of acceptable responses is developed (optional).	Yes	21	20.2
		No	11	10.6
		Non-existence	72	69.2
	The items test vocabulary and grammar.	Yes	32	30.8
		No	0	0
		Non-existence	72	69.2
Re-ordering/ Re-arranging/ Unscramble items	The items test factual information	Yes	5	4.8
		No	1	1.0
		Non-existence	98	94.2
	The items test grammar points and coherence.	Yes	6	5.8
		No	0	0
		Non-existence	98	94.2
Classification Items	There is a category column and a list of words in the question.	Yes	6	5.8
		No	4	3.8
		Non-existence	94	90.4
	The items measure simple learning outcomes (lower-order thinking).	Yes	10	9.6
		No		
		Non-existence	94	90.4
	A word fits only one category; otherwise, it should be stated in the instruction.	Yes	10	9.6
		No	0	0
		Non-existence	94	90.4
	There is more than one category column in the question.	Yes	9	8.7
		No	1	1.0
		Non-existence	94	90.4
Cloze Items	Vocabulary knowledge is tested.	Yes	9	8.7
		No	1	1.0
		Non-existence	94	90.4
	The deletion rate is mechanically/ systematically set.	Yes	8	7.7
		No	0	0
		Non-existence	96	92.3
Essay Items	The instructions for the questions are stated in a way that clearly mentions what is required of the students to do.	Yes	99	95.2
		No	1	1.0
		Non-existence	4	3.8
	A larger number of questions that require shorter answers are provided instead of a smaller number of questions with longer answers.	Yes	90	86.5
		No	10	9.6
		Non-existence	4	3.8
	Optional questions are avoided.	Yes	89	85.6
		No	12	11.5
		Non-existence	3	2.9
	Time considerations are thought through.	Yes	97	93.3
		No	4	3.8
		Non-existence	3	2.9
	The Word count limit for the answers is provided.	Yes	27	26.0
		No	73	70.2
		Non-existence	4	3.8
Linguistic Accuracy Items	The test is free from any grammatical mistakes.	Yes	102	98.1
		No	2	1.9
		Non-existence	0	0
	The test is free from any spelling mistakes.	Yes	104	100.0
		No	0	0
		Non-existence	0	0
		Yes	104	100.0

	The test is free from any punctuation mistakes.	**No**	**0**	**0**
		Non-existence	**0**	**0**
. Layout Accuracy Items	**Each test is preceded by clear instructions.**	**Yes**	**103**	**99.0**
		No	**1**	**1.0**
		Non-existence	**0**	**0**
	Questions are separated from one another.	**Yes**	**103**	**99.0**
		No	**1**	**1.0**
		Non-existence	**0**	**0**
	The duration of the test has been thought through for the number of questions asked	**Yes**	**102**	**98.1**
		No	**2**	**1.9**
		Non-existence	**0**	**0**
	Each question is provided with its grade.	**Yes**	**101**	**97.1**
		No	**3**	**2.9**
		Non-existence	**0**	**0**
Blooms Taxonomy Items	**Can the student recall or remember the information? (Remember)**	**Yes**	**91**	**87.5**
		No	**13**	**12.5**
		Non-existence	**0**	**0**
	Can the student explain ideas or concepts? (Understand)	**Yes**	**86**	**82.7**
		No	**18**	**17.3**
		Non-existence	**0**	**0**
	Can the students use the information in a new way? (Apply)	**Yes**	**21**	**20.2**
		No	**83**	**79.8**
		Non-existence	**0**	**0**
	Can the student distinguish between the different parts? (Analyze)	**Yes**	**13**	**12.5**
		No	**91**	**87.5**
		Non-existence	**0**	**0**
	Can the student justify a stand or decision? (Evaluate)	**Yes**	**14**	**13.5**
		No	**90**	**86.5**
		Non-existence	**0**	**0**
	Can the student create a new product or point of view? (Create)	**Yes**	**29**	**27.9**
		No	**75**	**72.1**
		Non-existence	**0**	**0**

References

Ali, H., Mjenda, M. (2024). Teachers' understanding of classroom assessment: Insights from English language teachers in Dodoma municipality, Tanzania. *Cogent Education, 11*. https://doi.org/10.1080/2331186x.2024.2380627.

Bachman, L. F. (1995). *Fundamental consideration in language testing* (3rd ed.). Oxford: Oxford University Press.

Boyle, C., & Hashemi, N. (2023). From Struggles to Success: Investigating the Impact of Early Learning Assessments on Students Performance and Motivation. *Education Sciences*. https://doi.org/10.3390/educsci13030225.

Brown, D. H. & Abeywickrama, P. (2018). *Language assessment: Principles and classroom practices* (3rd ed). London: Pearson Education.

Brown, D. H. (2004). *Language assessment: Principles and classroom practices* (2nd ed). London: Longman.

Chicho, K.Z.H., & Hussein, S.H. (2022). The Washback of Midterm Examination on First-Year Students' Perception Regarding the Final Exam. *International Journal of Social Sciences and Educational Studies, 9*(2), 267-277. https://doi.org/10.23918/ijsses.v9i2p267

Clark-Fookes, T. (2025). Arts education's wicked problem: tensions between reliability and validity in arts assessment. *Arts Education Policy Review*. https://doi.org/10.1080/10632913.2025.2515044

Clay, B., Root, E. (2001). *Is This a Trick Question? Kansas State. Kansas Curriculum Center*.

Douglas, D. (2014). *Understanding language testing*. London: Routledge.

Fitzner, K. (2007). Reliability and validity: a quick review. *The Diabetes Educator, 33*(5), 775-780. https://doi.org/10.1177/0145721707308172

Forehand, M. (2010). Bloom's taxonomy. *Emerging perspectives on learning, teaching, and technology, 41*(4), 47-56.

French, S., Dickerson, A., & Mulder, R. (2023). A review of the benefits and drawbacks of high-stakes final examinations in higher education. *Higher Education*, 1-26. https://doi.org/10.1007/s10734-023-01148-z.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. London: Routledge.

Gall, M. D. (1970). The use of questions in teaching. *Review of educational research, 40*(5), 707-721. https://doi.org/10.3102/00346543040005707

Harlen, W. (2005). Teachers' summative practices and assessment for learning – tensions and synergies. *The Curriculum Journal, 16*, 207 - 223. https://doi.org/10.1080/09585170500136093.

Marso, R. N., & Pigge, F. L. (1991). An analysis of teacher-made tests: Item types, cognitive demands, and item construction errors. *Contemporary Educational Psychology, 16*(3), 279-286. https://doi.org/10.1016/0361-476X(91)90027-I

McMillan, J. H. (2018). *Classroom assessment: Principles and practice that enhance student learning and motivation*. New York: Pearson.

Mertler, C. A. (2004). Secondary teachers' assessment literacy: Does classroom experience make a difference? *American secondary education, 33*(1), 49-64. https://www.jstor.org/stable/41064623

Öz, H. (2014). Turkish teachers' practices of assessment for learning in the English as a foreign language classroom. *Journal of Language Teaching and Research, 5*(4), 775- 785. http://dx.doi.org/10.4304/jltr.5.4.775-785

Özdemir- Yılmazer. M. & Özkan, Y. (2017). Classroom assessment practices of English language instructors. *Journal of Language and Linguistic Studies, 13*(2), 324-345. https://search.informit.org/doi/10.3316/informit.337212271866366

Pallant, J. (2016). *SPSS survival manual: A step-by-step guide to data analysis using IBM SPSS*. London: Routledge.

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education, 4(*1), 1-11. http://dx.doi.org/10.1080/2331186X.2017.1301013

Russell, M. K.s &, Airasian, P. W. (2012). *Classroom assessment: Concepts and applications*. New York: McGraw-Hill.

Swerdlik, M.E., Cohen, R.J. & Sturman, E. D. (2009). *Psychological Testing and Assessment: An Introduction to Tests and Measurement* (7th ed.). New York: McGraw−Hill.

Zheng, Y., & Iseni, A. (2017). Authenticity in language testing. *Journal of the Association-Institute for English Language and American Studies, 6(*8), 9-14.