

Research Article

# Test Item Writing Competence Among Oman College of Health Sciences Nurse Faculty

Mohammed Khalfan Ambusaidi\* 

Nursing Program, Oman College of Health Science, ALDahiliya Governate, Nizwa Wilayate, Oman

## Abstract

**Background:** The demand for nursing faculty to deliver high-quality teaching and assessment has surged, emphasizing the need for accurate learning assessment through effective testing and outcome measurement. Yet, the literature reveals that many nursing faculty are underprepared in measurement and evaluation, lacking essential competencies in test development and often failing to follow established guidelines. While both teacher-made tests (TMT) and standardized test scores are used to inform nursing faculty what the students have learned, the validity and reliability of TMT remain a significant concern. This study, guided by Social Justice Theory, assessed nursing faculty competencies in TMT development within Oman's College of Health Sciences, evaluating the fairness and effectiveness of these assessments based on content coverage, difficulty level, and test validity and reliability. **Methodology:** Descriptive statistics were used to portray the sample; Pearson's correlation was used to determine the relationship between TMT and committee-designed standardized end-of-semester final examinations (CDESFE). **Results:** Results showed a strong positive correlation between student scores on teacher-made tests (TMT) ( $M = 23.95$ ,  $SD = 4.74$ ) and committee-designed standardized exams (CDESFE) ( $M = 31.99$ ,  $SD = 6.22$ ),  $r(1672) = .613$ ,  $p < .001$ . MANOVA analysis indicated no significant differences between TMT and CDESFE regarding best-practice guidelines and item analysis (Wilk's  $\Lambda = .93$ ,  $F(2, 21) = .78$ ,  $p = .47$ , partial  $\eta^2 = .69$ ). Multiple regression analysis further demonstrated that both TMT and CDESFE scores significantly predict students' overall academic achievement ( $F(2, 1671) = 2241$ ,  $p < .001$ ,  $R^2 = .73$ ), underscoring the predictive value of both testing methods for student success. **Conclusion:** With a gap in how nurse faculty implement TMT, there are potential negative consequences on students' progress towards licensure examination. This study contributes to nursing science and education through objective, efficient, fair, and equitable assessment measures in the classroom setting. These findings may also transfer to the clinical setting, where nursing students, staff, and faculty are assessed during and after educational sessions and workshops.

## Keywords

Teacher-Made Test, End-of-Semester Final Examination, Students' Test Scores, Academic Achievement, Best Practice Guidelines, Item Analysis

## 1. Introduction

Test item writing is challenging and time-consuming and imposes more faculty accountability to practice accurate as-

essment and evaluation procedures. Likewise, faculty are responsible for establishing test criterion-related, construct,

\*Corresponding author: [adonis8482@gmail.com](mailto:adonis8482@gmail.com) (Mohammed Khalfan Ambusaidi)

**Received:** 31 October 2024; **Accepted:** 12 November 2024; **Published:** 29 November 2024



Copyright: © The Author(s), 2024. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

and content validity. As a result, faculty members must know the necessary methods to gather validity evidence related to testing, as emphasized by the *Standards for Educational and Psychological Testing* and the National League for Nursing [1].

Guided by the Theory of Social Justice [2], this study shed light on test validity and test fairness practices among nursing faculty as demonstrated through faculty competence. Thus, the study aimed to evaluate nurse faculty competency in developing TMT. The main concepts and variables of the study were defined. The National League for Nursing [1] stressed the necessity of integrating test fairness by nurse faculty, which is an ethical obligation during testing. Likewise, the literature suggests the availability of best practice guidelines to guide nurse faculty in developing TMT. The literature search did not yield empirical data related to TMT validity and fair testing in the nursing discipline. Testing is the most commonly used evaluation method, especially in undergraduate education [3]. The study findings help nursing faculty ensure students are standardized using identical assessment methods, content, administration, scoring, and interpretation processes to avoid test bias in the evaluation process.

Testing is essential in teaching, and test development is challenging and time-consuming. Academic decisions are made based on students' academic achievement in tests. However, the lack of faculty competency in developing test items affects the credibility of the evaluation process [4]. The literature suggests conflicting views of faculty competency in test development. In some countries, the challenge of test development starts during the pre-service period of college life. Asim et al. [5] found that pre-service teachers are challenged in developing the best answer type multiple-choice questions, which requires higher-level reasoning. For this reason, Asim et al. [5] recommended additional training for pre-service teachers before the actual engagement in test development.

Literature indicates limited research studies with conflicting views on test item development and evaluation. However, the testing process may depend on the institution's administration and the view of the importance and fairness of the testing process. Educators must implement best practice guidelines in every institution responsible for developing and administering tests. Educators must also establish evidence of the validity and reliability of tests via appropriate evaluation strategies, mainly when test results are used for decision-making [1, 6, 7].

Most reviewed studies were conducted in the United States and Nigeria, with varying research quality and sample size. Although the studies were rigorous, most were descriptive designs, using either interviews or survey sources. There are some limitations regarding the generalizability of the study results, such as the number of study participant sources, single-site study sources, participants with limited years of experience, and assessment of the test item qualities of pre-service teachers.

There is a general consistency among the cited research studies that faculty are not well prepared and require regular training and development in test item writing, administration, and evaluation [3, 5, 8-10]. However, there is a need to assess faculty competency in writing, administering, and evaluating test items. In contrast, Oermann et al. [11] found that most nursing faculty had completed a formal graduate course, attended continuing education, or received guidance and mentoring in assessment, evaluation, and grading. Unlike other nursing studies, Oermann et al. [11] found that nursing faculty set standards for best practices of test item writing and were committed to attending formal evaluation training and implementing test review processes. However, Nedeau-Cayo et al. [9] asserted that whether faculty referred to objectives or not, it had no impact on the cognitive levels of the test items ( $p = .087$ ) nor the presence of item writing flaws IWF ( $p = .3270$ ).

This study intends to address the literature gap and contribute additional evidence to the domain of assessment and evaluation in nursing education. Specifically, this research examines nursing faculty competency in developing test items in Oman. This area has not been previously studied, thus enriching the nursing literature and promoting an in-depth understanding of faculty competency in test item development.

## 2. Methodology

### 2.1. Study Area

The study evaluated nurse faculty competency in developing teacher-made tests (TMT) at the Oman College of Health Sciences, particularly analyzing the extent to which multiple-choice questions (MCQs) adhered to best practice guidelines. It explored assessment practices across nursing branches and aimed to determine the fairness and standardization of evaluations among students. The target population included nursing faculty teaching specific courses and nursing students enrolled in the general baccalaureate nursing program.

### 2.2. Study Design

This research employed a descriptive, correlational, and comparative design, utilizing existing student test scores from nine nursing branches. The study assessed differences in adherence to best practice guidelines between TMT and committee-designed standardized end-of-semester final examinations (CDESFE). It analyzed various test item attributes, including reliability, difficulty levels, and potential bias. A checklist developed by Haladyna & Rodrigues [7] was used for evaluation, while demographic data was collected from nursing faculty to inform the study's context.

## 2.3. Sample Size Determination

The sample size for the study was calculated using the GPower calculator Version 3.1. For Pearson's correlation analysis with a medium effect size (0.3), an alpha level of .05, and a power of .95, a minimum of 138 student test scores was required. For multiple regression analysis with a medium effect size (0.15) and two predictors, a minimum of 107 scores was needed. For MANOVA with a medium effect size (.0625), two predictors, and the same alpha and power levels, a minimum of 151 scores was necessary. Consequently, the study aimed for a sample size of 151 to minimize Type I and Type II errors.

## 2.4. Sampling Technique

The study was conducted across nine nursing branches at OCHS, located in various governorates of Oman, governed by the Ministry of Health. The faculty and student distribution was documented, with approximately thirty faculty teaching relevant nursing courses during the 2019-2020 academic year.

## 2.5. Ethical Approval

Ethical approval was obtained for the study, ensuring adherence to ethical standards in research involving human participants. This included obtaining informed consent from all participants and ensuring confidentiality throughout the study.

## 2.6. Data Collection Instrument

Data were collected using a combination of instruments, including a demographic data questionnaire, teacher-made tests (TMT), and end-of-semester final examinations (CDESFE). The reliability of the tests was assessed using expert reviews and a ten-item best practice guidelines checklist. Additionally, item analysis was performed to evaluate the quality and effectiveness of test items, focusing on statistics such as reliability coefficients, item difficulty, and discrimination indices.

## 2.7. Method of Data Collection

Data collection involved a multi-step procedure initiated by sending requests to associate deans across nine nursing branches. These requests included letters to course chairs and invitations to faculty teaching adult-health nursing, pharmacology, and pathophysiology courses. Faculty were invited to complete an anonymous 15-item online demographic questionnaire. The researcher requested copies of teacher-made tests (TMT) and their item analyses from faculty, as well as standardized end-of-semester final examinations and corresponding analyses. Student scores from TMT and standardized examinations were also collected,

ensuring anonymity by removing identifying information. Institutional Review Board (IRB) approval was secured before data collection, and an information letter was provided to faculty to clarify voluntary participation and confidentiality. Data were stored securely, and participants had the option to enter a drawing for gift cards as an incentive for completing the questionnaire.

## 2.8. Data Analysis

Statistical analysis was conducted using SPSS Version 27. Descriptive statistics were used to calculate the samples' frequencies, percentages, means, standard deviations, ranges, and midpoints. MANOVA was employed to assess relationships between students' final grades and item difficulty levels. Multiple linear regression analysis was used to explore relationships between dependent variables (best practice guidelines, item analysis, and academic achievement) and independent variables (teacher-made tests and end-of-semester final examinations). Research questions were addressed through various statistical methods, including Pearson's correlation for assessing relationships between test scores and MANOVA for comparing adherence to best practice guidelines between the two types of tests. Ancillary analyses included comparisons of demographic data related to faculty academic preparation and experience.

## 3. Results

The research variables included in this study were best practice guidelines, item analysis, TMT, CDESFE, students' scores in TMT and the CDESFE, and academic achievement. Best practice guidelines were measured at the ratio level by calculating the mean scores on the multiple-choice item writing best practice guidelines violation checklist. The item analysis was calculated at the interval level by calculating the mean item difficulty levels of teacher-made and CDESFE. Finally, the students' scores and academic achievement were calculated at the interval level. Descriptive findings for the study variables are presented in Table 1 and Table 2.

**Table 1.** Descriptive Findings of Teacher-Made Tests (TMT) and Committee-Designed Standardized End-of-Semester Final Examinations (CDESFE) (N = 24).

Variables	Mean	SD	Range	Midpoint
TMT				
Best Practice Guidelines	9.46	0.37	0-10	7
Item Analysis	0.59	0.13	0.30 – 0.80	0.47
CDESFE				

Variables	Mean	SD	Range	Midpoint
Best Practice Guidelines	9.73	0.21	0-10	7
Item Analysis	0.63	0.03	0.30 – 0.80	0.63

Note. TMT = Teacher-made test, CDESFE = Committee-designed end-of-semester final examination.

**Table 2.** Descriptive Findings of Students' Test Scores ( $N = 1672$ ).

Variables	Mean	SD	Range	Midpoint
TMT Students' Score	23.98	4.64	0-40	23.29
CDESFE Students' Score	32.04	6.13	0-50	29.99
Academic Achievement	54.94	10.69	60.41	52.09

Note. TMT = Teacher-made test, CDESFE = Committee-designed end-of-semester final examination.

The results indicated that teacher-made tests and committee-designed standardized end-of-semester final examinations adhere to best practice guidelines for test item writing ( $M = 9.46$ ,  $SD = 0.37$ ;  $M = 9.73$ ,  $SD = 0.21$ ) and item analysis ( $M = 0.59$ ,  $SD = 0.13$ ;  $M = 0.63$ ,  $SD = 0.03$ ) respectively.

Pearson's correlation coefficient tested the hypothesis to

find the relationship between students' test scores on teacher-made tests and committee-designed standardized end-of-semester final examinations. A total of 1672 students' test scores were analyzed. The results revealed a positive correlation between students' test scores in teacher-made ( $M = 23.95$ ,  $SD = 4.74$ ) and students' test scores on committee-designed standardized end-of-semester final examinations ( $M = 31.99$ ,  $SD = 6.22$ ),  $r(1672) = .613$ ,  $p < .001$ . Therefore, the hypothesis was accepted.

Multivariate analysis (MANOVA) was used to determine whether there is a statistical difference in best practice guidelines and item analysis between teacher-made tests and committee-designed standardized end-of-semester final examinations. The results revealed no significant difference between teacher-made tests and committee-designed standardized end-of-semester final examinations on the variables item analysis and best practice guidelines, Wilk's  $\Lambda = .93$ ,  $F(2, 21) = .78$ ,  $p = .47$ , partial  $\eta^2 = .69$ .

ANOVA results for each dependent variable, at an alpha level of .005, indicated that there was no significant difference between TMT and CDESFE on item analysis  $F(1, 22) = .40$ ,  $p = .53$ , partial  $\eta^2 = .02$ , or with teacher-made tests ( $M = 0.59$ ) committee-designed standardized end-of-semester final examinations ( $M = 0.63$ ). Additionally, there was no significant difference between TMT and CDESFE on best practice guidelines  $F(1, 22) = 1.54$ ,  $p = .22$ , partial  $\eta^2 = .07$ , or with teacher-made tests ( $M = 9.46$ ) committee-designed standardized end-of-semester final examinations ( $M = 9.73$ ). Therefore, the hypothesis was rejected. Tables 3 and 4 present multivariate analysis results.

**Table 3.** MANOVA – Differences between Best Practice Guidelines and Item Analysis ( $N = 24$ ).

Variable		Mean	SD	Value	F	df	p	Partial Eta Squared
Best Practice Guidelines	TMT	9.46	0.12	0.93	0.78	1 - 22	0.471	0.07
	CDESFE	9.73	0.37					
Item Analysis	TMT	0.59	0.13					
	CDESFE	0.63	0.03					

\*Statistically significant differences  $p < 0.05$

Note. TMT = Teacher-made test, CDESFE = Committee-designed end-of-semester final examination.

**Table 4.** MANOVA – Differences between Best Practice Guidelines and Item Analysis (Tests of between-subject effects) ( $N = 24$ ).

Source	Dependent variable	df	Mean Square	F	p
Corrected Model	Best Practice Guidelines	1	.193	1.544	.066
	Item Analysis	1	.006	.402	.018

Regression analysis determined the relationship between the dependent variable (academic achievement) and independent variables (teacher-made tests and end-of-semester final examinations). The results revealed that students' test scores from teacher-made tests and scores from committee-designed standardized end-of-semester final examinations predict students' overall academic achievement,  $F(2, 1671) = 2241$ ,  $p < .001$ ,  $R^2 = .73$ . Therefore, the hypothesis was accepted.

Students' test scores from teacher-made tests and scores from committee-designed standardized end-of-semester final examinations predict students' overall academic achievement,

$\beta = .30$ ,  $t(1672) = 18.45$ ,  $p = .00$ ,  $R^2 = .73$ . and  $\beta = .64$ ,  $t(1672) = 39.55$ ,  $p = .00$ ,  $R^2 = .73$  respectively. Table 5 presents multiple regression analysis results.

A separate correlation was calculated to find the best predictor for academic achievement. The results indicated that committee-designed standardized end-of-semester final examination scores ( $M = 31.99$ ,  $SD = 6.22$ ),  $r(1672) = .822$ , had a higher correlation than teacher-made test scores ( $M = 23.95$ ,  $SD = 4.74$ ),  $r(1672) = .689$ ,  $p < .001$ , which indicates that the CDESFE scores were a better predictor for students' academic achievement.

**Table 5.** Regression Coefficient of Students' Test Scores ( $N = 1672$ ).

Variables	<i>B</i>	<i>SE</i>	<i>t</i>	<i>p</i>	95.0% CI
(Constant)	3.76	.79	4.78	.000	[2.217, 5.300]
TMT Students' Score	.67	.04	18.45	.000	[-.600, .743]
CDESFE Students' Score	1.10	.03	39.55	.000	[1.042, 1.151]

Dependent Variable: Academic Achievement

Note. TMT = Teacher-made test, CDESFE = Committee-designed end-of-semester final examination.

#### *Ancillary Analyses*

To compare the scores of the students taught by faculty leading the curriculum committee in a specific course with the scores of other students in branches, a t-test was performed. There was a significant difference in the mean scores between students' test scores in teacher-made tests at the college branches ( $M = 23.58$ ) and the teacher-made test by the faculty leading the course ( $M = 20.91$ ),  $p < .001$  for the adult-health

nursing. Likewise, there was a significant difference in the mean scores between students' test scores in committee-designed end-of-semester final examinations (CDESFE) at the college branches ( $M = 33.84$ ) and the scores of the students taught by the faculty leading the course ( $M = 32.09$ ),  $p < .001$  for adult-health nursing. The results of the t-test are presented in Table 6.

**Table 6.** Independent T-Test Comparing Students' Scores in Teacher-Made Tests and Committee-Designed End-of-Semester Final Examinations for Adult-Health Nursing Course ( $N = 165$ ).

Variables	Mean	<i>SD</i>	<i>t</i>	<i>df</i>	<i>p</i>
Adult-Health Nursing TMT scores			-5.92	332	.000
Faculty Leading course	20.91	4.19			
Branches	23.58	4.01			
Adult-Health Nursing CDESFE scores			-3.103	332	.002
Faculty Leading course	32.09	5.58			
Branches	33.84	4.66			

Note. TMT = Teacher-made test, CDESFE = Committee-designed end-of-semester final examination.

For the pathophysiology course, the results indicated a significant difference in the mean scores between students'

test scores in teacher-made tests at the college branches ( $M = 24.62$ ) and the teacher-made test by the faculty leading the



course ( $M = 30.54$ ),  $p < .001$ . However, there was no significant difference in the mean scores between students' test scores in committee-designed end-of-semester final examinations (CDESFE) at the college branches ( $M = 31.29$ ) and

the scores of the students taught by the faculty leading the course ( $M = 33.22$ ),  $p = .159$ . The results of the t-test are presented in Table 7.

**Table 7.** Independent T-Test Comparing Students' Scores in Teacher-Made Tests and Committee-Designed End-of-Semester Final Examinations for Pathophysiology Course.

Variables	<i>n</i>	Mean	<i>SD</i>	<i>t</i>	<i>df</i>	<i>p</i>
Pathophysiology TMT				9.63	54.92	.000
Faculty Leading course	31	30.54	2.92			
Branches	249	24.62	5.00			
Pathophysiology CDESFE				1.41	278	.159
Faculty Leading course	31	33.22	5.50			
Branches	249	31.29	7.35			

Note. TMT = Teacher-made test, CDESFE = Committee-designed end-of-semester final examination.

For the pharmacology course, the results indicated a significant difference in the mean scores between students' test scores in teacher-made tests at the college branches ( $M = 11.64$ ) and the teacher-made test by the faculty leading the course ( $M = 24.92$ ),  $p < .001$ . Likewise, there was a significant difference in

the mean scores between students' test scores in committee-designed end-of-semester final examinations (CDESFE) at the college branches ( $M = 31.37$ ) and the scores of the students taught by the faculty leading the course ( $M = 34.45$ ),  $p = .005$ . The results of the t-test are presented in Table 8.

**Table 8.** Independent T-Test Comparing Students' Scores in Teacher-Made Tests and Committee-Designed End-of-Semester Final Examinations for Pharmacology Course.

Variables	<i>n</i>	Mean	<i>SD</i>	<i>t</i>	<i>df</i>	<i>p</i>
Pharmacology TMT scores				13.14	291	.000
Faculty Leading course	33	24.92	5.72			
Branches	260	11.64	2.75			
Pharmacology CDESFE scores				2.80	291	.005
Faculty Leading course	33	34.45	6.12			
Branches	260	31.37	5.92			

Note. TMT = Teacher-made test, CDESFE = Committee-designed end-of-semester final examination.

#### Item Analysis

An individual option analysis was done for teacher-made tests with five or more tagged multiple-choice questions identified as problematic questions. The options analysis of the problematic questions indicated that correct options were not chosen, the lower group chose correct options, the upper and lower group equally selected correct options, the higher group chose distractors which the lower group did not choose, distractors were not chosen at all, an equal number of the high and low group had chosen distractors.

#### Item Difficulty Index

The mean item difficulty index represents the proportion of the examinees who answered the item correctly, indicating the total test's difficulty. A high percentage suggests an easy test, and a low percentage indicates a difficult test. An acceptable item difficulty ranges from 30 % to 80% [12]. Extreme low or high item difficulty does not contribute to the discriminating power of a test. For this reason, the item difficulty index is the most frequently reported in item analysis statistics. The item analysis summary of teacher-made tests indicated that teach-

er-made tests were easy (57%), moderately difficult (33.3%), and difficult (9.5%).

On the other hand, the summary of committee-designed end-of-semester final examinations (CDESFE) item analysis indicated that the committee-designed end-of-semester final examinations (CDESFE) for the three courses were moderately difficult (100%).

#### *Item Discrimination Index*

The item discrimination index indicates the ability of the item to differentiate between those who scored at the high end and those who scored at the low end of the test. The discrimination index is based on the correlation of the item score to the total score. A mean discrimination index from -1.0 to +1.0. A positive mean discrimination index indicates that the students

who achieved high scores on the test chose the correct answer for the item more frequently than those who had low scores. A negative mean discrimination index indicates that the students who achieved low scores chose the correct answer on the test more frequently than those who achieved high scores [12].

The summary of teacher-made tests item analysis indicated that teacher-made tests were not discriminating (9.5%), moderately discriminating (42.9%), discriminating (28.6%), and very discriminating (19%). In contrast, the summary of committee-designed end-of-semester final examinations (CDESFE) item analysis indicated that the committee-designed end-of-semester final examinations (CDESFE) were moderately discriminating (66.7%) and discriminating (33.3%).

**Table 9.** Summary of Teacher-Made Tests Item Analysis.

Test	Number of Items	Number of Examinees	Mean Scores	SD	Mean Diff. Index	Mean Disc. Index	Reliability KR20 (Alpha)
1	20	31	13.00	2.300	0.650	0.289	0.287
2	20	31	11.323	2.291	0.566	0.273	0.276
3	15	31	10.484	1.899	0.699	0.227	0.421
4	15	31	9.839	1.588	0.656	0.211	0.028
5	15	31	8.097	2.401	0.540	0.408	0.518
6	14	32	7.750	2.437	0.554	0.407	0.569
7	14	32	8.344	2.532	0.596	0.398	0.615
8	14	33	9.364	2.101	0.669	0.367	0.533
9	14	33	8.030	1.766	0.574	0.267	0.304
10	20	35	12.743	2.589	0.637	0.291	0.505
11	20	36	10.778	2.562	0.539	0.315	0.364
12	20	34	4.471	1.242	0.224	0.097	-0.156
13	20	34	13.382	2.413	0.669	0.246	0.462
14	21	35	13.800	2.594	0.657	0.295	0.475
15	20	36	4.861	1.512	0.243	0.150	-0.158
16	15	33	9.000	2.309	0.600	0.335	0.525
17	15	31	10.032	2.456	0.669	0.385	0.609
18	12	29	6.724	2.196	0.560	0.397	0.578
19	12	29	7.448	1.849	0.621	0.294	0.403
20	12	34	8.676	2.152	0.723	0.409	0.614
21	12	32	7.813	2.338	0.651	0.651	0.636

**Table 10.** Summary of Committee-Designed End-of-Semester Final Examinations Item Analysis.

Test	Number of Items	Number of Examinees	Mean Scores	SD	Mean Diff. Index	Mean Disc. Index	Reliability KR20 (Alpha)
1	35	131	20.893	4.059	0.597	0.289	0.592
2	23	121	15.240	3.471	0.663	0.355	0.654
3	30	132	19.242	3.583	0.641	0.269	0.574

#### *Coefficient Alpha (KR-20)*

The coefficient alpha measures the internal consistency reliability of the test items represented as alpha or KR-20 in item analysis. Miller et al. [13] suggested that teacher-made tests should have reliability coefficients between 0.60 and 0.85. However, the summary of teacher-made tests indicated that only 19% (n = 4) of teacher-made tests have a coefficient alpha greater than 0.60. For most teacher-made tests, 81% (n = 17) have a coefficient alpha less than 0.60. On the contrary, the summary of committee-designed end-of-semester final examinations (CDESFE) item analysis indicated that two of the committee-designed end-of-semester final examinations (CDESFE) have coefficient alpha below 0.60 (66.7%) and one has a coefficient alpha of 0.65 (33.3%) which is at the acceptable level.

## 4. Discussion

The first research question was, “To what extent do teacher-made tests and committee-designed standardized end-of-semester final examinations adhere to best practice guidelines for test item writing and item analysis?” The results indicated no difference in the mean score of teacher-made tests and committee-designed standardized end-of-semester final examinations. Likewise, nurse faculty adherence to best practice guidelines was indicated by high mean scores in TMT and CDESFE for best practice guidelines and item analysis. These results were in line with the findings of Oermann et al. [11], who noted that most nursing faculty were prepared in assessment and evaluation and grading practice via continuing education, graduate-level courses, guidance from mentors, or self-study. In the same study, faculty followed standards, completed formal evaluation training, and used standard tests, and multiple test reviewers.

On the other hand, most of the literature reported that faculty do not follow best practice guidelines when developing TMT [3, 5, 7, 9, 12-16].

In this study, the mean scores of adherence to best practice guidelines and item analysis for TMT were higher than the midpoint range. Similarly, the mean scores of the adherence to best practice guidelines and item analysis for the CDESFE were higher than the midpoint range. Additionally, the demographic online questionnaire result indicated that most

participants had either academic coursework, an online course, or attended workshops on test construction. These results suggest that the nursing faculty and the central committees responsible for developing final examinations follow best practice guidelines in test item development. The findings of this study were consistent with the findings of Oermann et al. [11], in which nursing faculty were prepared for assessment and evaluation via continuing education or graduate-level courses. The results also support the work by Bristol et al. [14], who found that most nursing faculty had a graduate-level course in assessment and evaluation and reported confidence in identifying different levels of test items and analyzing examination items.

In contrast, Bristol et al. [14] found that clinical and laboratory educators were less confident in assessment and evaluation and were likely to use or modify existing test items from previous exams, test banks, or textbooks. Likewise, Hijji [15] reported a lack of faculty adherence to best practices in test item development and that most TMT items contained one or more flaws related to item writing, content, or structure. The researcher attributed the item writing flaws to linguistic reasons as the test items were written in English while the faculty were non-English native speakers.

The findings of this study were in opposition to research findings in other non-nursing disciplines, which indicated that faculty were not well prepared in assessment and evaluation and did not follow best practices guidelines. For example, Wright et al. [3] found that most of the biology faculty did not depend on the learning objectives for test item development, and despite that the students were taught the same course, TMTs were of varying degrees of difficulty, resulting in variation in students' examination scores. Likewise, Asim et al. [5] found that pre-service teachers of different disciplines were challenged in constructing best-answer type questions, and thus faculty lacked competency in developing TMTs. Zhang and Burry-Stock [17] supported these findings by stating that trained faculty were more skilled in measurement and evaluation. The researchers suggested a lack of faculty preparation in test development and recommended faculty training in assessment and evaluation in both studies.

A possible explanation for the findings of this study may be that the Oman College of Health Sciences conducted multiple continuing education workshops to enhance the quality of TMT development. Therefore, continuing education might



have improved faculty competence and adherence to test development. Another possible explanation is that nursing education has taken a step forward in preparing nursing faculty for different aspects of education, including assessment and evaluation. Additionally, nurse faculty are provided with mentorship to enhance faculty competency in providing quality education and ensure a fair assessment. Another explanation for this finding is that the course leaders and the curricular committee members are recent, within the last few years, doctorally prepared nursing faculty. Finally, as international organizations have emphasized the importance of fair assessment and evaluation of students, the OCHS and other educational institutions should ensure that the students are fairly assessed and evaluated.

#### *Research Question 2 and Hypothesis 1*

The second research question was, “What is the relationship between students’ test scores on teacher-made tests and students’ test scores on committee-designed standardized end-of-semester final examinations?” The hypothesis stated that there will be a significant relationship between students’ test scores on teacher-made tests and committee-designed standardized end-of-semester final examinations. Using Pearson's correlation coefficient, the present study's data analysis revealed a positive correlation between students’ test scores in TMT and CDESFE, indicating that TMT and CDESFE were constructed based on standards of test construction without ambiguity. The findings of this study are congruent with the findings of Oermann et al. [11] that TMTs were the most frequently used and had the highest weight toward the course grade. However, Hijji [15] found that violating item writing flaws (IWF) reduced the reliability of test scores.

In a secondary data analysis, Brown and Abdulnabi [18] evaluated students learning in an introductory educational psychology course in New Zealand. The study assessed the quality of 100 teacher-made MCQs used in an undergraduate midterm test (50 four-options MCQ) and final exam (50 four-options MCQ), making up 50% of the course grade responses of 372 students enrolled in an introductory educational psychology course. Item difficulty, discrimination, and chance properties were determined using classical test theory (CTT) and item response theory (IRT) statistical item analysis models. The two-parameter logistic (2PL) model consistently fit the data best. In addition, the analysis was conducted to check dimensionality with confirmatory factor analysis of a single factor with 50 items using the weighted least square estimator with robust standard errors and mean- and variance-adjusted  $\chi^2$  test statistic in Mplus version 7.4 to account for the dichotomous nature of the items.

The misfitting items and distractors were removed, and test statistics were obtained. The internal consistency of both the midterm and the final exam was acceptable. However, the researchers found that the grade results changed for nearly two-thirds of the 372 participants [18]. The researchers concluded that the MCQs used in the course were problematic

and that the midterm test had more problematic MCQs than the final examination. Additionally, the findings suggest that problematic MCQs result in a negative direct impact on students’ course grades.

In line with the findings of Brown and Abdulnabi [18], One explanation for the positive correlation between students’ test scores in TMT and CDESFE in this study is that the source of questions for both examinations was the teachers, given that the committees request sample questions from college branches to support the development of the final examination. However, this does not mean the exact questions are used because the committees screen, review, and modify the test items. Additionally, the test form must be sent to reviewers who provide feedback, including the clarity of test items.

Another possible explanation for the findings is that the committees comprise teaching faculty at the college branches with the same qualities as other nurse faculty who develop and administer TMTs. It could also be that test items were derived from each course's required and recommended references. Therefore, no test items are accepted outside the course references, limiting the number of test items developed. As a result, the students may see the same concept/question more than once but were asked in a different test form.

#### *Research Question 3 and Hypothesis 2*

The third research question was, “What are the differences in best practice guidelines incorporation and item analysis between teacher-made tests versus committee-designed standardized end-of-semester final examinations?” The hypothesis stated there will be statistical differences in best practice guidelines incorporation and item analysis between teacher-made tests and committee-designed standardized end-of-semester final examinations.

The multivariate analysis (MANOVA) revealed no significant difference between TMT and CDESFE on the variables item analysis and best practice guidelines. Additionally, the ANOVA results for each dependent variable indicated no significant difference between TMT and CDESFE on item analysis and best practice guidelines. As a result, the hypothesis was rejected.

In line with research question 2, TMT and CDESFE followed best practice guidelines. The item analysis results indicated that TMT and CDESFE were within the acceptable discrimination index and item difficulty. However, the coefficient alpha of most TMT was below the recommended range. These results indicated that the TMT may not meet the intended measure. In contrast, the coefficient alphas of the three CDESFE were at the recommended range.

Despite the adherence to best practice guidelines and the impressing results of the item analysis, teacher-made tests require intense individual item and options analysis. Additionally, test experts may provide other opinions regarding the quality of teacher-made test items.

The results of this study are partially in line with the findings of D'Sa and Visbal-Dionaldo [19], who conducted a descriptive study in the College of Nursing of a Governmental

University in Saudi Arabia. The researchers analyzed 48 single-answer multiple-choice questions from an anatomy examination administered in 2015 to second-year Baccalaureate Nursing program students. In addition, the researchers retrieved 56 answer scripts with students' responses. The researchers found that fifty percent of the items had average difficulty. The researchers reported that the reason was that the anatomy of the test items measured factual content.

Most items exhibited an excellent or good discrimination index [19]. However, the researchers found a negative Pearson's correlation between difficulty and discrimination indices. Additionally, three-fourths of the distractors were functional, while one-fourth were dysfunctional. Therefore, the researchers recommended assessing items for ambiguity, clues, and wrong keys and recommended intense analysis to identify dysfunctional distractors. Moreover, due to sampling bias and a small sample size, the researchers warned about using this study results. They also recommended careful decision-making, especially when educational assessment was considered. In addition, the researchers recommended faculty training in test item writing to enhance test item quality.

Amelia et al. [20] recently conducted a descriptive study to explore the quality of teacher-made chemistry tests such as item fit and person fit, item difficulty, and test reliability. The sample consisted of 356 senior students from senior high schools in Yogyakarta, who were selected using a cluster random sampling technique. In addition, the researchers analyzed a final national test try-out, which consisted of 40 multiple-choice questions to measure students' cognitive chemical abilities.

The results indicated that all the teacher-made chemical test instrument items were proven to fit the Rasch model [20]. This means no items have a fit statistic that was too high or too low in the analyzed instrument. According to the researchers, this result reflects the students' actual chemical abilities. The results also revealed that only two items were classified as difficult and low. The remaining items were classified as medium item difficulty. These results indicated that the chemistry teacher-made test met test requirements and guidelines. Additionally, the item statistics indicated excellent internal consistency, as evidenced by Pearson's coefficient.

The current study results were partially in line with the results of Amelia et al. [20]. While the evaluation of best practice guidelines suggested that TMT and CDESFE met the best practice guidelines, they lacked internal consistency. The current study results indicate a need to conduct regular item analysis and test reviews after administration to students.

Using Ex post facto design, Ramadhan et al. [21] analyzed the content validity (quantitative and qualitative), empirical validity, reliability, and standard error of measurement of semester teacher-constructed final exam test instruments on Physics Grade XII Senior High School academic year 2017/2018. The sample consisted of five test forms with 135 multiple-choice questions administered to five high schools in Bima Regency in Indonesia. In addition, the researchers col-

lected 555 answer sheets from the high school after obtaining permission from school principals.

Content validation was determined through the agreement of three experts in physicists, educational research, and evaluation. The experts used validity analysis on the question sheet developed by the Republic of Indonesia's Minister of National Education Regulation Number 20 of 2007 concerning Educational Assessment Standards (Indonesia, 2015). The evaluation tool was designed to evaluate test items' substance, construction, and language. The researchers added a fourth evaluation area, the level of thinking based on Bloom's Taxonomy. For the reliability coefficient, Iteman 4.3 was used to calculate Cronbach's alpha.

The expert reviewers identified approximately 15% of items as invalid based on the four areas of evaluation [21]. Substance analysis indicated that three-fourths of the items matched the requirements, and construction analysis indicated that most items were clearly stated, not confusing, and had good functioning distractors. Language analysis revealed that most items used appropriate communicative language. However, the thinking level analysis indicated that more than half of the items did not meet the expectations. This signified that the teachers were not skilled in making test items with high thinking levels.

The reliability analysis indicated that three of the test forms were unreliable, with the reliability estimate below 0.70, and only two forms had reliability above 0.70. Item characteristics were evaluated based on Classical Test theory. The results indicated that only half of the items had good difficulty and discrimination indices. The researchers expressed that the item difficulty of the final semester examination was inconsistent with the participants' ability, and the item discrimination and the distractor did not function properly [21].

The researchers concluded that TMT possesses good quality based on substance, construction, and language, but teachers still had difficulty constructing high-level thinking test items. As a result, teachers' ability to prepare for the final semester exam test items was limited.

The findings of this study are congruent with the findings of Ramadhan et al. [21] in that the expert reviewers determined that TMT and CDESFE were well constructed without linguistic difficulties. Likewise, the item analysis results indicated that teacher-made tests maintained acceptable reliability and good item difficulty and discrimination. However, the CDESFEs were mostly difficult but discriminating.

#### *Research Question 4 and Hypothesis 3*

The fourth research question was, "To what extent do students' test scores from teacher-made tests versus committee-designed standardized end-of-semester final examinations predict students' overall academic achievement?" The hypothesis stated that "students' test scores from teacher-made tests and from committee-designed standardized end-of-semester final examinations predict students' overall academic achievement."

The results revealed that students' test scores from TMT

and scores from CDESFE predicted students' overall academic achievement, for which the hypothesis was accepted. Additionally, students' test scores from TMT and scores from CDESFE predicted students' overall academic achievement. A separate correlation was calculated to identify the best predictor for academic achievement. The results showed that students' scores in CDESFE were better than teacher-made test scores. Thus, CDESFE scores better predicted students' academic achievement.

Moore et al. [22] conducted a retrospective, correlational study to determine predictors of BSN student success on the HESI Exit Exam in a BSN program within a liberal arts university in the southeastern United States. Via nonrandom sampling, the researchers collected records of 486 BSN students enrolled in eight cohorts within the program from August 2014 to May 2018. In addition, the researchers examined factors such as test anxiety, academic motivation, and academic demographics associated with students' success in the HESI Exit Exam.

The results indicated that most students were white females with no previous nursing course failure and passed the NCLEX-RN on the first attempt [22]. A high percentage of the students completed the program within three years. Additionally, for most students, test anxiety scores were low, academic motivation scores were high, and they had earned a letter grade of A or B in five pre-nursing science courses and courses with clinical components.

The results also indicated that HESI examination scores and the exit exam scores were within the recommended performance level except for community health. To predict the HESI Exit Exam pass rate, the researchers performed an independent-sample t-test. The results indicated that students with low test anxiety scored significantly higher on the HESI Exit Exam than the students with high test anxiety. Likewise, students with high academic motivation scores had higher HESI Exit Exam scores than students who reported low academic motivation scores [22].

Furthermore, the researchers performed correlational analysis to predict the NCLEX-RN pass rate on the first attempt. The results indicated that students who scored higher on the HESI Exit Exam were significantly more likely to pass the NCLEX-RN examination on the first attempt. Additionally, the researchers conducted multiple linear regression to test independent variables, academic barriers of SAT/ACT score, final GPA, TEAS score, and scores on previous Specialty HESI examinations and nonacademic barriers of age at graduation, test anxiety, and academic motivation [22].

The results indicated that medical-surgical and obstetrics nursing topics significantly correlated to higher HESI Exit Exam scores; however, there were no significant correlations in fundamentals, pediatrics, and psychiatric nursing topics [21]. Additionally, students' final GPA was a significant predictor of NCLEX-RN success.

The researchers reported some limitations, such as a homogenous student population [22]. Additionally, the

NCLEX-RN pass rate was high, limiting the generalizability of the results. The researchers attributed the high pass rate despite high anxiety levels to the students' high academic motivation. Additionally, the researchers noted that the students in this study may have had less stress, fewer demands, and availability of assistance when required.

The findings of this study correspond with the findings of Moore et al. [22] in that students' test scores in TMT and CDESFE predict students' overall academic achievement. While Moore et al. [22] found that students' final GPA significantly predicted NCLEX-RN success, the current study's findings indicated that students' scores in the CDESFE were better predictors of overall academic achievement.

Gillespie and Nadeau [23] conducted a retrospective record study to explore the relationships between Kaplan integrated exam scores and the HESI exit exam score to determine early indicators of success on the exit exam. The study involved the assessment of online student scores for standardized testing across five semesters of a nursing program in a private, faith-based university. The researchers analyzed the standardized exam scores of 131 students who had completed the BSN program between May 2011 and December 2012. In addition, the researchers accessed students' scores from Kaplan Resource and Evolve Elsevier websites. Kaplan assessment modules include Critical Thinking questions, Fundamentals, Pediatrics, Comprehensive Medical-Surgical, Obstetrics, Mental Health, and Pharmacology. Additionally, students were administered the HESI exam with 180 questions during the last semester of the program.

Data analysis included a simple bivariate correlation of continuous data between each Kaplan exam and each student's first attempt on the HESI exit exam [23]. The results indicated significant positive relationships between the score on the HESI exit exam and the Kaplan exam modules. There was a significant correlation between Kaplan exam modules administered each semester with performance on the first attempt of the HESI exit exam. As a result, the researchers suggested that Kaplan exam scores for individual students were highly sensitive to the HESI exit exam performance. Additionally, the results suggested that Kaplan diagnostic exam was highly correlated with the initial NCLEX RN attempt.

Gillespie and Nadeau [23] concluded that the correlation of standardized exam scores in the study may not mirror exams administered at the national level. However, the Kaplan module exams administered throughout the nursing program in the current study were highly sensitive to students' success on the HESI exit exam. Moreover, the researchers warned that if standardized testing programs are implemented, individual test results should be analyzed continuously and used to alert students as early as possible and enforce remedial plans.

The current study findings were consistent with the findings of Gillespie and Nadeau [23], who found a significant correlation between Kaplan exam modules administered each semester and the HESI exit exam. In addition, the current

study indicated a correlation between students' test scores in TMT and academic achievement. Likewise, Gillespie and Nadeau [23] found that Kaplan diagnostic exam was highly correlated with initial NCLEX-RN attempts. Finally, the current study indicated that CDESFE scores best predict students' overall academic achievement.

Spurlock and Hunt [24] conducted a retrospective, descriptive, correlational study to explore the differences between the actual NCLEX-RN pass rate and the expected pass rate for the nursing program. The researchers collected data from students' records from a large, single-purpose nursing college in a large midwestern town after completing the HESI Exit Exam and the NCLEX-RN. Records of 179 nursing graduates were collected. The data were extracted from student records for those students graduating from January 2004 to July 2005. In addition, the data were collected from the online licensure verification system, which provides more timely results than those in the quarterly board of nursing reports.

The researchers provided descriptive statistics in which the first attempt and final attempt of Exit Exam scores were calculated [23]. Additionally, one-way ANOVA was used to assess differences in the first and final scores of the HESI Exit Exam between students who passed and those who failed the NCLEX-RN. Furthermore, the researchers calculated point-biserial correlation coefficients to assess the relationship between students' first Exit Exam scores, final Exit Exam scores, and NCLEX-RN outcomes. The researchers found a significant relationship between first-exit exam scores and NCLEX-RN outcomes. However, there was no significant relationship between final Exit Exam scores and NCLEX-RN outcomes. According to the researchers, allowing the students multiple attempts to pass the Exit Exam introduced errors in the form of spurious Exit Exam scores into the relationship, which caused the relationship to decrease in strength.

Moreover, a binary logistic regression was calculated to predict NCLEX-RN outcomes based on HESI Exit Exam scores. The results indicated a good model fit for the first Exit Exam scores, which meant that the first Exit Exam scores distinguished between who would pass and who would fail the NCLEX-RN. To further examine how well first Exit Exam scores predicted NCLEX-RN outcomes, the researchers calculated the Receiver Operating Characteristics (ROC) curve, which shows the tradeoff between sensitivity and specificity. The researchers found the first Exit Exam scores as "fair" predictors of NCLEX-RN outcomes. However, the researchers found that final Exit Exam scores were insignificant and poor predictors of NCLEX-RN outcomes [24].

While students scoring 850 or higher are predicted to pass, and students scoring less than 850 are predicted to fail, Spurlock and Hunt [24] set the cutoff score at 900, and each of the test qualities of sensitivity, specificity, positive predictive value, negative predictive value, and odds ratios. Further, the cutoff scores were tested in 25-point increments to a minimum score of 550. Students were classified as predicted to fail

if they scored below 900, down to 550. The researchers found that the best cutoff score for the students in this research was 650. Thus, A HESI Exit Exam score of 650 on the first attempt yields the best classification of students' positive predictive value, and the Odd Ratio increases, which means that students who score less than 650 were predicted to fail compared to those scoring above 650.

Spurlock and Hunt [24] further tested each cutoff score against actual student data and tested prediction accuracy. Finally, the researchers categorized the HESI Exit Exam into eight categories with a predicted probability of failing NCLEX-RN. Then, the researchers confirmed the prediction by calculating predicted probabilities of failures derived from the logistic regression model. The results indicated that students scoring very highly on the first Exit Exam have little chance of failing the NCLEX-RN. In contrast, students scoring low on the first Exit Exam have a high chance of failing the NCLEX-RN.

The researchers [23] warned that schools may prevent students' graduation due to low Exit Exam scores, which, means that such students could be better prepared to pass NCLEX-RN. Additionally, the researchers recommended using lower cutoff scores than the scores recommended by HESI to increase the predictive accuracy of the test. An essential lesson to be learned from the result of Spurlock & Hunt [24] is that, while the predictive probability of success or failure is essential for educators, especially in the NCLEX-RN pass rate and the progression policy, educators must be cautious in analyzing and interpreting the predictive probability of students' success and academic achievement.

The current study findings were partially consistent with the findings of Spurlock and Hunt [24], whose findings indicated a significant relationship between first-exit exam scores and NCLEX-RN outcomes. In contrast, the final Exit Exam scores were insignificant and poor predictors of NCLEX-RN outcomes. Additionally, the researchers found that the first Exit Exam scores were "fair" predictors of NCLEX-RN outcomes. In contrast, the current study findings indicated that committee-designed end-of-semester final exam scores were the best predictor of students' overall academic achievement. In addition, contrary to the findings of Spurlock and Hunt [24], the findings of this study indicated a significant relationship between teacher-made tests and overall academic achievement.

The predictive probability of students' success and academic progression has been given adequate attention, mainly to the progression policies, HESI Exit Examination, and NCLEX-RN success in nursing schools. For example, Moore et al. [22] found a direct relationship between test anxiety and academic motivation, HESI examination scores, and exit examination scores. A significant finding is that students with high HESI Exit Exam were predicted to pass NCLEX-RN on the first attempt. Likewise, students' final GPA was used to predict NCLEX-RN. Similarly, Gillespie and Nadeau [23] found a positive relationship between the HESI Exit Exam



and Kaplan exam modules to predict NCLEX-RN success.

The findings of these two studies are consistent with the findings of Spurlock and Hunt [24]. Conducted a decade earlier, they found a significant relationship between first-exit exam scores and NCLES-RN outcomes. However, the repeated attempts of Exit Exams impacted NCLEX-RN success prediction. As a result, the researchers suggested that the first Exit Exam scores were “fair” predictors of NCLEX-RN outcomes. However, the final exit exam scores were insignificant and were poor predictors of NCLEX-RN success in repeated attempts.

## 5. Conclusions

This descriptive, correlational, comparative design evaluated nurse faculty competency in developing teacher-made tests. Using the Theory of Social Justice [2], this researcher explored teacher-made test validity and test fairness practices among nursing faculty at Oman College of Health Sciences as demonstrated through faculty competence. Some of this study's results were congruent with available published literature, while other results conflicted with the literature. Additionally, the results of this study provided empirical evidence of the validity and reliability of best practice checklists, which can be used to evaluate teacher-made tests.

The study underscores a critical need to enhance nursing faculty competencies in developing teacher-made tests (TMT) and to establish consistent assessment standards across branches. We recommend that academic institutions provide structured, ongoing development programs focused on the essentials of test item writing, administration, and evaluation, equipping faculty with the skills necessary for creating reliable, valid assessments. Such programs should emphasize best practices in test construction, including item analysis and question formulation, which are vital for high-quality testing. Equally important, academic institutions should adopt standardized testing policies and guidelines to ensure uniformity in test item quantity and duration, thereby promoting equity in student assessment across different academic institutions. Additionally, we advocate for further research into using best-practice checklists in TMT development to verify their validity and reliability and to support faculty in consistently applying these guidelines. Future studies should also investigate nursing faculty's perceptions and self-assessed competencies in test development, as understanding their viewpoints can inform and refine training programs to meet specific needs. Integrating these competencies into clinical assessments is essential to ensure a seamless approach to evaluation from the classroom to practical settings, ultimately benefiting nursing students and preparing them comprehensively for clinical excellence. Finally, we advocate that Liberal Arts and Science Colleges concerned with faculty preparation should offer intense courses, not only programs, to prepare future faculty for assessment and evaluation.

The current study had limitations. It focused on only one

semester, offering a limited view of nursing faculty competency in test item writing. While many teacher-made tests were reviewed, only half of the test forms were included, which could impact the results. Due to convenience sampling, findings may not be generalizable beyond the sample and geographic area. Manual data entry for student responses introduces potential human error, and some test forms were missing from certain college branches. Online testing is recommended for electronic data collection to improve accuracy. Although nursing faculty experts reviewed the test items, their input doesn't replace that of dedicated test experts. Despite limited evidence and challenges in educational research, this study offers a foundation for future exploration.

## Abbreviations

MCQs	Multiple-Choice Questions
TMT	Teacher-made Tests
CDESFE	Committee-Designed Standardized End-of-Semester Final Examinations
IWF	Item Writing Flaws

## Acknowledgments

I would like to express my gratitude and thanks to the people who supported me during this journey. A special thanks to the dissertation committee, without whom this dissertation would not have been possible. Dr. Barbara Patterson was an asset and was always available to guide me through this dissertation. Dr. Patterson was more than a committee chair. She always shared suggestions and ideas and encouraged students to step up and catch learning opportunities. I would also like to sincerely thank Dr. Donna Callaghan and Dr. Kim Nobel for their valuable comments, guidance, and thoughtful critique throughout my dissertation. I also would like to thank the readers, Dr. Mary Baumberger-Henry and Dr. Rose Rossi, for their comments and feedback. A special thanks to Dr. Darrell Spurlock, who was the committee chair and who helped me fine-tune and narrow the research ideas. Dr. Spurlock also helped me build a foundation of knowledge in assessment and evaluation early in the program. I always hoped we could go through this project and celebrate this achievement.

## Author Contributions

Mohammed Khalfan Ambusaidi is the sole author. The author read and approved the final manuscript.

## Conflicts of Interest

The authors declare no conflicts of interest.



## References

- [1] National League for Nursing. (2020, November). The fair testing imperative in nursing education: A living document from the National League for Nursing. Retrieved from <http://www.nln.org/docs/default-source/advocacy-public-policy/nln-fair-testing-vision-series.pdf?sfvrsn=2>
- [2] Rawls, J. (1971). A theory of justice. The Belknap Press of Harvard University Press.
- [3] Wright, C. D., Huang, A. L., Cooper, K. M., & Brownell, S. E. (2018). Exploring differences in decisions about exams among instructors of the same introductory biology course. *International Journal for the Scholarship of Teaching and Learning*, 12(2). <https://doi.org/10.20429/ijstl.2018.120214>
- [4] Simsek, A. (2016). A comparative analysis of common mistakes in achievement tests prepared by school teachers and corporate trainers. *European Journal of Science and Mathematics Education*, 4(4), 477–489.
- [5] Asim, A. E., Ekuri, E. E., & Eni, E. I. (2013). A diagnostic study of pre-service teachers' competency in multiple-choice item development. *Research in Education*, 89(1), 13–22. <https://doi.org/10.7227/RIE.89.1.2>
- [6] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. AERA.
- [7] Haladyna, T. & Rodriguez, M. (2013). Developing and validating test items (1st ed.). Routledge.
- [8] Ing, L., Musah, M. B., Al-Hudawi, S., Tahir, L. M., & Kamil, M. N. (2015). Validity of teacher-made assessment: A table of specification approach. *Asian Social Science*, 11(5), 193–200. <https://doi.org/10.5539/ass.v11n5p193>
- [9] Nedeau-Cayo, R., Laughlin, D., Rus, L., & Hall, J. (2013). Assessment of item-writing flaws in multiple-choice questions. *Journal of Nurses Professional Development*, 29(2), 52–57. <https://doi.org/10.1097/nnd.0b013e318286c2f1>
- [10] Ugodulunwa, C. A., & Wakjissa, S. G. (2016). What teachers know about validity of classroom tests: Evidence from a University of Nigeria. *Journal of Research & Method in Education*, 6(3), 14–19.
- [11] Oermann, M. H., Saewert, K. J., Charasika, M., & Yarbrough, S. S. (2009). Assessment and grading practices in schools of nursing: National survey findings part I. *Nursing Education Perspectives*, 30(6), 352–357.
- [12] McDonald, M. (2014). *The nurse educator's guide to assessing learning outcomes* (3<sup>rd</sup> ed.). Jones & Bartlett Learning.
- [13] Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching* (10<sup>th</sup> ed.). Pearson Education.
- [14] Bristol, T. J., Nelson, J. W., Sherrill, K. J., & Wangerin, V. S. (2018). Current state of test development, administration, and analysis: A study of faculty practices. *Nurse Educator*, 43(2), 68–72. <https://doi.org/10.1097/nne.0000000000000425>
- [15] Hijji, B. (2017). Flaws of multiple choice questions in teacher-constructed nursing examinations: A pilot descriptive study. *Journal of Nursing Education*, 56(8), 490–495. <https://doi.org/10.3928/01484834-20170712-08>
- [16] Kinyua, K., & Okunya, L. O. (2014). Validity and reliability of teacher-made tests: Case study of year 11 physics in Nyahururu District of Kenya. *African Educational Research Journal*, 2(2), 61–72. Retrieved from <http://www.netjournals.org/pdf/AERJ/2014/2/14-015.pdf>
- [17] Zhang, Z., & Burry - Stock, J. A. (2003). Classroom assessment practices and teachers' self - perceived assessment skills. *Applied Measurement in Education*, 16(4), 323–342. [https://doi.org/10.1207/S15324818AME1604\\_4](https://doi.org/10.1207/S15324818AME1604_4)
- [18] Brown G. & Abdulnabi H. (2017) Evaluating the Quality of Higher Education Instructor-Constructed Multiple-Choice Tests: Impact on Student Grades. *Front. Educ.* 2: 24. <https://doi.org/10.3389/feduc.2017.00024>
- [19] D'Sa, J. L., & Visbal- Dionaldo, M. L. (2017). Analysis of multiple-choice questions: item difficulty, discrimination index and distractor efficiency. *International Journal of Nursing Education*, (9)3, 109-114. <https://doi.org/10.5958/0974-9357.2017.00097.4>
- [20] Amelia, R., Sari, A., & Astuti, S. (2021). Chemistry learning outcomes assessment: how is the quality of the tests made by the teacher?. *Journal of Educational Chemistry (JEC)*, 3(1), 11-22. <https://doi.org/10.21580/jec.2021.3.1.6582>
- [21] Ramadhan, S., Sumiharsono, R., Mardapi, D., & Prasetyo, Z. K. (2020). The quality of test instruments constructed by teachers in bima regency, Indonesia: document analysis. *International Journal of Instruction*, 13(2), 507-518. <https://doi.org/10.29333/iji.2020.13235>
- [22] Moore, L. C., Goldsberry, J., Fowler, C., & Handwerker, S. (2021). Academic and nonacademic predictors of BSN student success on the HESI Exit Exam. *Computers, Informatics, Nursing: CIN*, 39(10), 570–577. <https://doi.org/10.1097/CIN.0000000000000741>
- [23] Gillespie, M. D., & Nadeau, J. W. (2019). Predicting HESI® Exit Exam success: a retrospective study. *Nursing Education Perspectives*, 40(4), 238–240. <https://doi.org/10.1097/01.NEP.0000000000000410>
- [24] Spurlock, D. R., Jr, & Hunt, L. A. (2008). A study of the usefulness of the HESI Exit Exam in predicting NCLEX-RN failure. *The Journal of Nursing Education*, 47(4), 157–166. <https://doi.org/10.3928/01484834-20080401-07>