Republic of the Philippines
Department of Education
REGION IV-A CALABARZON
CITY SCHOOLS DIVISION OF BIÑAN CITY

OPTIMIZING MULTIPLE CHOICE ITEMS OF QUARTERLY TESTS IN GENERAL CHEMISTRY 1 AND 2: TOWARDS THE DEVELOPMENT OF A TEST BANK

RUBEN P. GARCIA
Teacher II
Biñan Integrated National High School

FREDDIE JOHN V. CALUMNO, Ed.D.
Teacher II
Biñan Integrated National High School

ABSTRACT

A staple of achievement testing is multiple-choice questions. Numerous high-quality items are required to sufficiently cover all learning domains and demonstrate achieving mastery by giving meaningful, exact results. This study determined the quality of quarterly tests in General Chemistry using item properties such as reliability, difficulty index, discrimination index, and distractor efficiency. It sought to use the findings of the study by optimizing item properties to develop a test bank for future General Chemistry tests. To address the research questions, item analysis of questions and answers from three quarterly tests was performed. The findings revealed that the item properties in General Chemistry 2 was more reliable than the test in General Chemistry 1. Furthermore, a trend exposed that the quality of test items improved across three quarters, which may entail improving test construction as well. Based on the findings, recommendations and implications on test construction and development of a test bank in General Chemistry are provided.

Keywords: item analysis, difficulty index, discrimination index, distractor efficiency

INTRODUCTION

The process of teaching and learning is fundamentally centered around assessment. It is characterized as the methodical gathering and evaluation of data to enhance student education The test is a part of student assessment and should be an objective and standardized measure of a sample of behavior (Rezigalla, as cited in Firstenberg and Stawicki, 2022). Any educational test should gauge how well students have mastered the subject matter. Additionally, it results in a general evaluation of learners' development to determine their academic standing.

Factual memory, levels of comprehension, and learning application can

all be evaluated using multiple-choice exam items. A great pre-assessment indicator of student understanding may be found in multiple choice tests, and they can also serve as a starting point for a post-test assessment. Multiple choice questions are somehow challenging to create and can occasionally be complicated. It can be an effective and efficient means to assess learning outcomes.

The theoretical foundations of the study are based on the Classical Test Theory (CTT)and Item Response Theory (IRT). The CTT states that a learner's observed score can be decomposed into her/his true score and a random error component. A true score is the score obtained if the test were measuring the ability of interest perfectly (i.e., with no measurement error). Reliability and validity are crucial ideas in both the CTT and IRT theories. A measure that consistently assesses a construct across time, people, and contexts is considered dependable. This suggests that if it were performed repeatedly, it would probably result in the same outcomes each time. A measure is considered valid if it captures the required information.

A reliability coefficient estimates the level of concordance between observed and true scores of a learner (De Champlain, as cited in Ali, Carr, & Ruit, 2016). Cronbach's alpha is one of the widely used statistical tool for evaluating internal consistency (Downing, as cited in Ali, Carr, & Ruit, 2016).

Item analysis provides the avenue for establishing the validity and reliability of tests or exams. This analysis examines the properties of test items. A test's validity is determined by how well it samples the range of knowledge, skills, and abilities that learners were supposed to acquire in the period covered by the test. In item analysis, validity is established in terms of difficulty index (DFI), discrimination index (DI), and distractor efficiency (DE). DFI is relevant for determining whether students have learned the concept being tested (Bai and Ola, 2017). DI refers to the ability of an item to differentiate among learners based on how well they know the material being tested (Thompson, 2021). Shin, Guo & Gierl (2019) defines DE as the ability of incorrect answers to distract learners. If a distractor is not chosen as the correct answer by any of the low ability learners, this means it is not an efficient distractor.

The tedious tasks of preparing and designing multiple choice questions for a formative or summative assessment can take a toll on the teacher. Analyzing the item properties of MCQ items from previous tests or exams can provide the teacher with valid and reliable items that can really measure the learning of the students. Reliability of the MCQ items, based on internal consistency, is ensured through item analysis of the test or exam.

Item analysis should be performed on test items to create a test bank. A test bank is a ready-made electronic testing resource that can be used and customized by teachers for their assessment of students' learning. It is a compilation of a teacher's test questions from past semesters that is stored for future use (Corwin, 2019). A test bank attempts to provide a variety of questions in different objective formats (recall, recognition, compare, contrast, critical, analytical thinking) as students vary in how they can best demonstrate that they have learned. Item properties such as distractor efficiency and difficulty index are essential in aiding the teacher to develop a test bank of questions that may be used or modified repeatedly that will suit a batch of learners every semester.

This study analyzed and determined the item properties of quarterly tests in General Chemistry 1 and 2 that can be used for the development of a test bank. In terms of action planning, the findings may serve as basis for the development of test banks in other subjects in senior high school that may be provided by the researchers through learning action cells or seminar workshops.

METHODOLOGY

This study utilized a descriptive research design employing an item analysis procedure. Quantitative methods involved descriptive analysis that provided the answers to the research problems.

The population of the study are senior high school students enrolled at Biñan Integrated National High School, SY 2022-2023. The inclusion characteristics for the participants of the study specified that the student is a Grade 12 STEM student enrolled for the current academic year, have taken General Chemistry 1 and 2. The sampling technique used was total enumeration sampling which resulted in 331 test results for each quarterly test in General Chemistry.

The main data gathering instruments of the study are the quarterly tests in General Chemistry. The test items are aligned with learning domains specified in the table of specifications. After validation by a subject coordinator and approval by the school head, the test is administered.

Test results of the students from the quarterly tests in General Chemistry are scanned and tabulated using an online test scanning app. After downloading the scores, it was transferred to the item analysis template developed by @Carlo Excels (2022). The template analyzes item properties of test questions designed in multiple-choice format along with other formats. A summary sheet provides the validity and reliability of the test items. The validity of test items was measured in terms of the item difficulty index, item discrimination, and distractor efficiency of the responses. Reliability was measured using Cronbach's alpha.

RESULTS

The study investigated the item properties of the quarterly tests in General Chemistry, SY 2022-2023 and established a pool of test items leading to the development of a test bank.

The 4th quarter test indicated a better reliability (α=.835) than the other two quarterly tests in Quarter 1 (α=.651) and Quarter 3 (α=.704). Overall, the quarterly tests in General Chemistry 2 are more reliable than the 1st Quarter test in General Chemistry 1. The findings also suggest increasing reliability of the tests after each quarter (Table 1).

Table 1
Reliability of Quarterly Tests in General Chemistry

Quarterly Test	Cronbach's Alpha	Internal Consistency	Reliability
GenChem1-Q1	0.651	Low	Questionable
GenChem2-Q3	0.704	Moderate	Acceptable
GenChem2-Q4	0.835	High	Good

The mean difficulty level of the quarterly tests has indicated an average level of difficulty (*.30 ≤ mean ≤ .70*). The standard deviations indicate some extreme scores in each test but suggest increasing performance across the three quarterly tests at similar levels of difficulty (Table 2). The difficulty level of the items also shows that there are more items with an average difficulty level and less items with an easy difficulty level compared to the number of items specified in the table of specification of each quarterly test.

Table 2
Difficulty Index of Items in Quarterly Tests in General Chemistry (N=50 Items/Test)

		Chem1-Q1		Chem2-Q3		Chem2-Q4	
DFI	**Level**	**Number of Items**	**%**	**Number of Items**	**%**	**Number of Items**	**%**
<0.30	Difficult	6	12.0%	6	12.0%	5	10.0%
0.30 - 0.70	Average	31	62.0%	33	66.0%	32	64.0%
>0.70	Easy	13	26.0%	11	22.0%	13	26.0%
Mean, SD		0.52 ± 0.21		0.54 ± 0.20		0.56 ± 0.19	

DFI=Difficulty Index

The 1st and 3rd quarter test have been found to be discriminating and the 4th quarter test to be very discriminating (Table 3). The standard deviations indicated some extreme discrimination between the upper and lower groups in the 1st and 3rd quarter tests but better discriminating capability in the 4th quarter test.

Table 3
Summary of Discrimination Index of Quarterly Tests in General Chemistry (N=50 items)

		Chem1-Q1		Chem2-Q3		Chem2-Q4	
DI	**Level**	**Number of Items**	**%**	**Number of Items**	**%**	**Number of Items**	**%**
>0.34	Very Discriminating	16	32%	17	34%	32	64%
0.25-0.34	Discriminating	11	22%	16	32%	8	16%
0.20-0.24	Average	4	8%	5	10%		
<0.20	Not Discriminating	19	38%	12	24%	10	20%
Mean, SD		0.26 ± 0.16		0.28 ± 0.14		0.37 ± 0.22	

DI=Discrimination Index

The 3rd quarter test indicated highest test quality in terms of distractor efficiency as it has the highest number of functional distractors and no item with less than 66.7% DE (Table 4). The 1st quarter test indicated the lowest test quality in terms of distractor efficiency as it has the lowest number of functional distractors and 4 items with 33.3% DE.

Table 4
Distractor Efficiency of Items in Quarterly Tests in General Chemistry (N=50 Items/ Test)

	Chem1-Q1		Chem2-Q3		Chem2-Q4	
	f	%	f	%	f	%
Functional Distractors (FD)	128	85%	139	93%	135	90%
Non-functional Distractors (NFD)	22	15%	11	7%	15	10%
Number of Items with 100% DE	30	60%	39	78%	35	70%
Number of Items with 66.7% DE	16	32%	11	22%	13	26%
Number of Items with 33.3% DE	4	8%			2	4%

DE=Distractor Efficiency; Number of Items=50; Total Distracters=150

The 4th quarter exam indicated the greatest number of retained items for the test bank after item analysis and more than half the total number of test items in the table of specifications characterized as easy, average, and difficult (Table 5).

Table 5
Summary of Retained Items of Quarterly Tests in General Chemistry (N=50 Items/ Test)

		Chem1-Q1		Chem2-Q3		Chem2-Q4	
Domain	**Total Items**	**Number of Items**	**%**	**Number of Items**	**%**	**Number of Items**	**%**
Remembering	30	8	16%	9	18%	14	28%
Understanding		9	18%	13	26%	13	26%
Applying	15	3	6%	1	2%	4	8%
Analyzing		6	12%	8	16%	6	12%
Evaluating	5	1	2%	2	4%	3	6%
Creating							
Total Items		27		33		40	

Table 6 shows the recommended actions on items of the quarterly tests not retained after item analysis. All items to be rephrased will be used again in the respective tests that will be given again for school year 2023-2024 to verify if their item properties will improve.

Table 6
Summary of Recommended Actions on Items not Retained in General Chemistry

	Chem1-Q1			Chem2-Q3			Chem2-Q4		
Domain	**Revise Item**	**Revise NFD**	**Discard**	**Revise Item**	**Revise NFD**	**Discard**	**Revise Item**	**Revise NFD**	**Discard**
Remembering	8	4		6		1	1	1	
Understanding	3	2	2	1			1	1	1
Applying	2		1	3	1				1
Analyzing	2	1	1	2	2	1	1	1	3
Evaluating	2	1	2	2	1	1	2	1	
Total Items	**17**	**8**	**6**	**14**	**4**	**3**	**5**	**4**	**5**

DISCUSSION

Scores of students in tests or any other form of assessment should not be the lone indicator for the quality of the test, test construction, mastery of competencies, and proficiency of teachers. As a common practice in public schools, only scores are analyzed and items with the highest score serve as indicator that the student has mastered a particular competency. In fact, one required report a teacher submits after each quarterly test is an item analysis report which does not even remotely resemble the item analysis process. The item analysis procedure performed in this study has shown the score obtained for a particular item is merely scratching the surface of the quality of each test item. The basis for a quality test item is not just reflected by the number of students who answered the item correctly.

The item properties of the quarterly tests in General Chemistry in this study have shown some weak and strong points of test construction. The weak points included the questionable reliability of the 1st quarter test and its lesser quality of item properties compared to the 3rd and 4th quarter tests. These provide basis and motivation to further improve test construction for General Chemistry 1 test items.

The strong point is reflected in the improving quality of test construction from the 1st quarter to the fourth quarter. There is an increasing trend in the item properties after each quarterly exam. This increasing trend across quarters is a welcome sign of improvement and hopefully can be maintained or further enhanced with respect to test construction.

The findings of this study have shown that to evaluate the quality of test items in a test, four properties should be considered, namely reliability, difficulty index, discrimination index, and distractor efficiency. Evaluating these four item properties for every test administered will eventually lead to better test construction. Distractor efficiency is crucial to test construction since most of the time, evaluation of the item quality is focused solely on how the item was stated and how many answered it correctly, and not on the kind of choices given.

Test banking would provide teachers with welcome relief from trying to come up with items every time a test is to be administered. It can provide quality test items that can be reused. If item analysis is performed on reused items, the quality of the test items may be maintained or further enhanced. Most importantly, time to construct a test will be lessened considerably if there is a sizeable number of banked test items.

ACKNOWLEDGEMENTS

The proponents would like to express their gratitude to the following:

Edward R. Manuel, SEPS for Planning and Research, for his outpouring support for teacher researchers and providing us with another opportunity to enhance the teaching-learning process through this study.

Oliver P. Caliwag, Principal of BINHS, for his constant efforts in trying to imbibe and provide a culture of research for our institution.

STEM 12 Faculty, who in one way or another have supported and encouraged the proponents of this undertaking to breach the limits of establishing quality assessments.

Our families, for being a perpetual source of inspiration and motivation through tumultuous times of this study.

Almighty Father, whose constant presence and guidance has made it possible for the proponents to accomplish this undertaking.

REFERENCES

Ali, S., Carr, P., and Ruit, K. (2016). Validity and Reliability of Scores Obtained on Multiple-Choice Questions: Why Functioning Distractors Matter. *Journal of the Scholarship of Teaching and Learning*, 16(1), pp.1-14. https://doi.org/10.14434/josotl.v16i1.19106

Bai, X. and Ola, A. (2017). A tool for performing item analysis to enhance teaching and learning experiences. *Issues in Information Systems*, 18(1), pp.128-136. https://doi.org/10.48009/1_iis_2017_128-136

Carlo Excels (2022, April 4). *All-in-One Item Analysis Template* [video]. Excel for Teachers.YouTube. https://www.youtube.com/watch?v=GpO31yImCNo

Corwin, H. (2019, March 27). Professors shouldn't consider using test banks cheating. *The Daily Texan*. https://thedailytexan.com/2019/03/27/professors-shouldnt-consider-using-test-banks-cheating/

Firstenberg, M. and Stawicki, S. (2022). Item Analysis: Concept and Application. *Medical Education for the 21st Century*. https://doi.org/10.5772/intechopen.95701

Shin, J., Guo, Q., and Gierl, M. (2019). Multiple-Choice Item Distractor Development Using Topic Modeling Approaches. *Frontiers in Psychology*,v10 https://doi.org/10.3389/fpsyg.2019.00825

Thompson, N. (2021) *Item analysis and statistics*. Assessment Systems Corporation (ASC). Retrieved from https://assess.com/what-is-item-analysis/