Sprin Journal of Arts, Humanities and Social Sciences

ISSN: 2583-2387 (Online)

Vol. 4(11), Jan 2026, pp, 24-32

Journal homepage: https://sprinpub.com/sjahss

Research Article

Analysis on Teacher Made Tests in JRMSU: Basis for Departmental Tests

Jayson A. Dapiton

School Principal I, Department of Education, Dipolog City Division, Philippines

ARTICLE INFO

Keywords:

Teacher-Made Tests, Assessment Quality, Bloom's Taxonomy, Test Item Analysis, Departmental Testing

Article History:

Received: 20-09-2025

Revised: 25-12-2025

Accepted: 23-01-2026

Published: 31-01-2026

ABSTRACT

Aimed at evaluating the quality of teacher made tests within the Jose Rizal Memorial State University (JRMSU) system, guided by Bloom's taxonomy and Gronlund's typology of test formats, informed by psychometric perspectives from Classical Test Theory and Item Response Theory, this study documents how locally constructed assessments distribute cognitive demand, adhere to language in use standards (grammar and mechanics), and align with institutional learning outcomes. Using documentary analysis of test papers and Tables of Specifications across five JRMSU campuses, the research identifies over reliance on lower order thinking items, occasional misalignment between targeted and actual cognitive levels, and sporadic violations of item writing conventions. The discussion argues for faculty development, peer review of test items, and the institutionalization of departmental testing to stabilize validity and fairness. The sample concludes with actionable recommendations for assessment literacy, including calibration routines, item banks with annotated rationales, and alignment audits that link curriculum, instruction, and testing.

Cite this article:

Dapiton, J. (2026). Analysis on Teacher Made Tests in JRMSU: Basis for Departmental Tests. Sprin Journal of Arts, Humanities and Social Sciences, 4(11), 24–32. https://doi.org/10.55559/sjahss.v4i11.575

1. Introduction and Rationale

Assessment is integral to teaching and learning. It not only measures achievement but also shapes instructional focus and learner behavior. In higher education contexts where standardized external examinations are rare, teacher-made tests often function as the primary assessment instrument. Their design, therefore, has disproportionate influence on what and how students' study. Ensuring that such tests are valid, reliable, and cognitively balanced is essential to educational quality and equity.

Within JRMSU, instructors and professors frequently construct their own tests across general education courses, including English, Mathematics, and Natural Sciences. While such autonomy enables contextualized assessment, it also requires consistent assessment literacy. This work examines the cognitive profile and language quality of teacher-made tests across five campuses of JRMSU located in Dapitan City (main campus), Dipolog City, Katipunan, Tampilisan, and Siocon campuses in the province of Zamboanga del Norte, Philippines. It used institutional Tables of Specifications (TOS) and a standardized checklist derived from Bloom's (1956) taxonomy as revised by Anderson and Krathwohl (2001) and Gronlund's (1998) typology to analyze alignment.

The rationale for the investigation is two-fold. First, faculty and program leaders require diagnostic evidence to guide professional development in assessment construction. Second, the institution must ensure that assessments support higher-order learning outcomes associated with communicative competence, problem solving, and critical reasoning-outcomes that transcend rote recall and grammatical minutiae.

1.1 *Theoretical and Conceptual Framework*

This study is anchored on Item Response Theory (IRT) as cited by Embretson and Reise (2013), which states that the probability of answering an item correctly and the ability of a test-taker can be modeled in different ways depending on the nature of the test. It is common to assume unidimensionality on the items in a test to measure one single latent ability.

According to IRT, test-takers with high ability should have a high probability of answering an item correctly. Another assumption is that it does not matter which items are used in order to estimate the test-takers' ability. This assumption makes it possible to compare test-takers' result despite the fact that they have taken different versions of a test. The computer programs can now perform the complicated calculations that IRT requires (Ven der Linden & Glas, 2000).

This is backed up by another theory which existed prior to IRT- the Classical Test Theory (CTT), cited by Sharkness and DeAngelo (2011). It has dominated the area of testing and is based on the assumption that a student has an observed score and a true score. His observed score is usually seen as an estimate of his true scores plus/minus some unobservable measurement errors. An advantage with CTT is that it relies on weak assumptions and is relatively easy to interpret.

However, CTT can be criticized since the true score is not an absolute characteristic of a student since it depends on the content of the test items. If there are students with different ability levels, a simple or more difficult test would result in different scores.

These theoretical lenses are embedded in the Schema of the Study, presented below as Figure 1, which serves as the conceptual framework guiding the analysis of teacher-made tests. It integrates

Corresponding Author:

jayson.dapiton@deped.gov.ph (*J. Dapiton*)

three principal dimensions- Language-in-Use, Bloom's Questioning Levels, and Gronlund's Types of Tests- to provide a comprehensive evaluation of both the linguistic quality and the cognitive validity of classroom assessments. At the center of the schema are the teacher-made tests, which form the primary unit of analysis. Each of the three dimensions interacts with these tests to address distinct but interconnected research concerns.

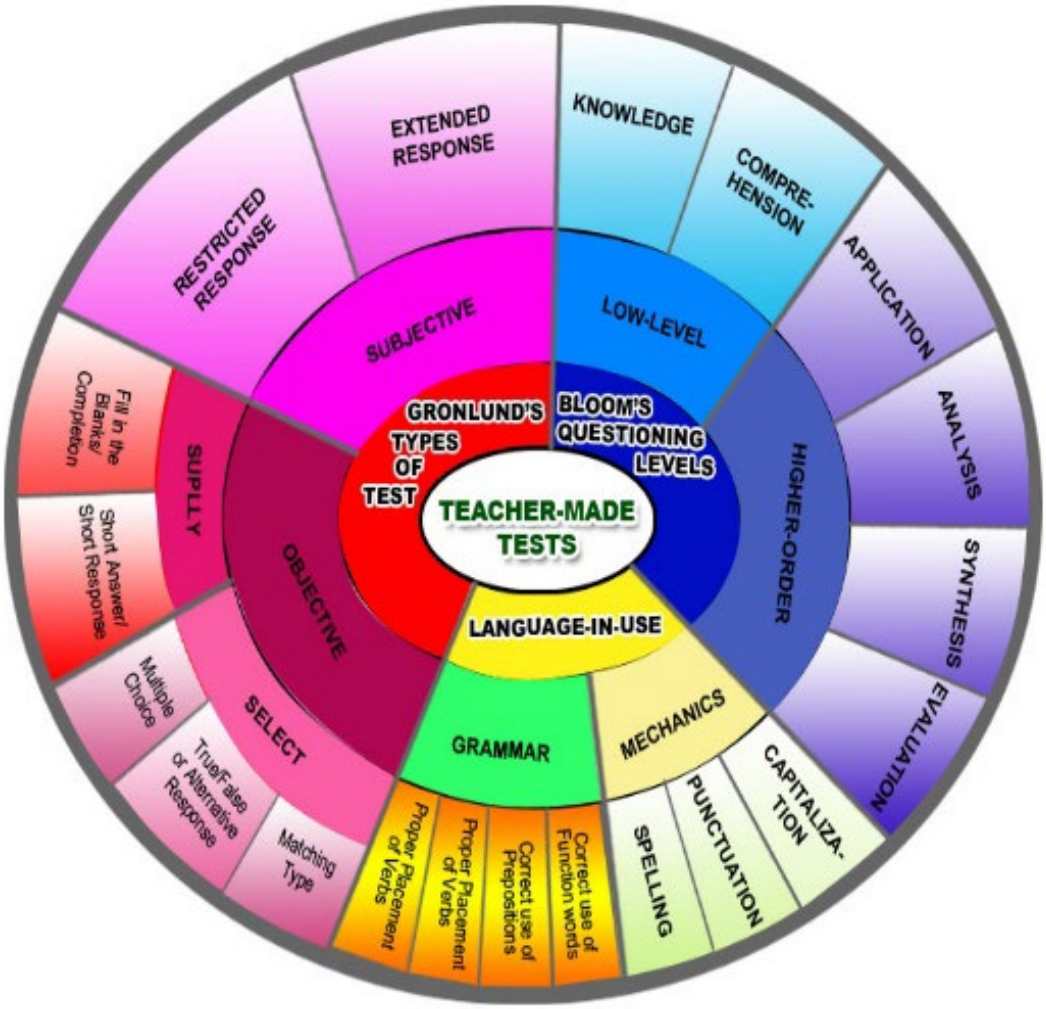

Figure 1: Schema of the Study

Language-in-Use Dimension- the first dimension focuses on the linguistic accuracy of the tests as instructional tools. This component examines both mechanics and grammar- including capitalization, punctuation, spelling, function word use, prepositions, and proper placement of verbs. By identifying mechanical precision and grammatical lapses, this dimension determines the clarity, correctness, and communicative integrity of teacher-made assessments. It thus reflects the teachers' command of written English and their ability to construct grammatically sound test items that accurately convey intended meanings.

Bloom's Questioning Levels- the second dimension is based on Bloom's Taxonomy of Questioning Levels, which classifies test items according to the cognitive skills they elicit. These levels range from lower-order thinking skills (LOTS)- Knowledge and Comprehension- to higher-order thinking skills (HOTS)- Application, Analysis, Synthesis, and Evaluation. This component measures the depth of cognition that the teacher-made tests demand from students. By categorizing each question according to Bloom's framework, the study evaluates whether the tests promote mere recall of information or foster critical and creative thinking aligned with higher cognitive objectives.

Gronlund's Types of Tests- the third dimension draws from Gronlund's (1998) Quality Types of Tests, which distinguishes between Objective and Subjective assessments. Objective tests include select-type items- such as True/False (Alternative Response), Multiple Choice, and Matching Type and supply-type items, such as Completion (Fill-in-the-Blank) and Short Answer. Subjective tests encompass Restricted Response and Extended Response formats. This dimension examines not only the form and structure of the test items but also their alignment with learning outcomes and validity as measures of performance. It reflects how teachers balance efficiency and reliability (typical of objective tests) with depth and expressiveness (typical of subjective tests).

At the core of the schema lies the Teacher-Made Test, where all three dimensions converge. The linguistic evaluation (Language-in-Use) ensures the test's textual quality and clarity; Bloom's taxonomy provides insight into its cognitive rigor; and Gronlund's typology assesses the methodological soundness and appropriateness of test types. Together, these frameworks produce a holistic appraisal of teacher-made assessments- linking language accuracy, cognitive demand, and test design quality.

In this way, the schema encapsulates the study's guiding principle: that effective classroom assessment is both a linguistic and cognitive act. Teachers are not only evaluators of learning outcomes but also language users whose written tests reflect their pedagogical competence, linguistic precision, and commitment to promoting higher-order thinking.

1.2 *Statement of the Problem*

The study aimed to determine and analyze the instructors'/professors' language-in-use, their use on Bloom's Standard Questioning Levels and Gronlund's Quality Types of Tests in English 11 (Communication Arts/Skills 1), Math 11 (College Algebra), and Natural Sciences 11 (Physical Science) in their teacher-made tests in Jose Rizal Memorial State University (JRMSU) System of the first semester of academic year 2014-2015. It sought to answer the following specific questions:

1.2.1 What is the researcher's analysis on the instructors'/professors' language-in-use in their teacher-made tests?

1.2.2 What is the researcher's analysis on the instructors'/professors' teacher-made tests based on their Table of Specifications (TOS) and the Standardized Qualitative Checklist when grouped as English 11, Math 11, and Natural Sciences 11?

1.2.3. What is the researcher's analysis on these tests based on Gronlund's Quality Types of Tests when grouped as above? and

1.2.4. What is the researcher's overall collective analysis on these teacher-made tests?

2. Literature Review

Stiggins and his colleagues (2004) pointed out that assessment, or the process of gathering and synthesizing information through measurement or quantification of students' learning via tests, involves evaluation of personality in which an individual meets and solves a variety of lifelike problems. Even problem-solving is a major learning task, holistic appraisal of a learner, his environment and accomplishments is the principal objective of educational assessment.

Brown (2010), however, asserted that assessment needs to be fit-for-purpose, that it should enable evaluation of the extent to which learners have learned and the extent to which they can demonstrate that learning. Teachers then need to consider not just what they are assessing and how they are doing it, particularly which methods and approaches, but also why, which is the rationale for assessing on any particular occasion and in any context.

Most classroom assessment involves tests that teachers have constructed themselves. It is estimated that 54 teacher-made tests are used in a typical American classroom per year and worldwide, millions of unique assessments, perhaps billions, are produced yearly. Regardless of the exact frequency, teachers regularly use tests they have constructed themselves.

Many teachers believed that they need strong measurement skills and reported that they were confident in their ability to produce valid and reliable tests. Other teachers, however, reported a level of discomfort with the quality of their own tests or believe that their training was inadequate. This was because the formal assessment training teachers received often focused on large-scale test administration and standardized test score interpretation, rather than on the test construction strategies or item-writing rules that teachers need (Stiggins et al., 2004).

A quality teacher-made test should follow valid item-writing rules, but as many researchers point out, empirical studies establishing the validity of item-writing rules are in short supply and often inconclusive, and item writing-rules are based primarily on common sense and the conventional wisdom of test experts.

Even after decades of psychometric theory and research, some bemoaned the almost complete lack of scholarly attention paid to achievement test items. Haladyna et al. (2002) reasserted this claim, stating that the body of knowledge about multiple-choice item writing was still quite limited and added recently that "item writing is still largely a creative act".

The current empirical research literature for item-writing rules-of-thumb is most often of two kinds: studies which look at the relationship between a given item format and either test performance or the psychometric properties of the test; and studies which have evaluated the quality of teacher-made tests by applying some set of item-writing standards or criteria. Reviewing these studies for an agreed upon list of classroom assessment rules, however, is not overly fruitful, as few rules present themselves.

Guidelines for multiple-choice, matching and alternate-choice/true-false items with at least some evidence of validity by examining textbook endorsement and empirical studies were catalogued. Though the authors did find empirical support for general advice such as "avoid trick items" and many studies testing particular rules, only four specific rules on their final revised inventory were supported without contradiction across studies and two of those were supported by the existence of only one study.

Though there has been greater research emphasis on the importance and value of other types of assessments in the classroom such as performance-based, authentic, formative, and informal, the majority of tests that teachers construct themselves continue to follow a paper-and-pencil, objectively scored format (Earl, 2012).

Along this concern, Williams (1991) constructed rules and guidelines for each type of teacher-made tests which elaborately specified what the test items should contain and possess in each type of test. The first was subjective test, where the student is required to answer questions through his own constructed words in sentence form; and objective test, where there is only one correct answer with no judgment entering into the correctness of the answer.

The essay is the only form of subjective test. However, the latter has two subcategories: the objective select test, which requires recognition of material and where the student selects the correct answer from among given alternatives; and the objective supply test, which requires recall of information in a cognitive task and where the student has to supply the correct answer. Objective select types of tests are multiple choice, true-false or alternative response, and matching type. Objective supply types of tests are fill in the blanks or completion and short answer or short response. The signal words of short answers /short responses are "name, list, identify, give, enumerate, and mention," which requires the students to list the information requested; "state," which requires the student to describe, define, or point out the requested information; and "give the principle of," which requires the student to provide the law, rule, or principle.

All these types of tests require different scoring procedures. However, Reise (2012) said that there is a need to have those different scores combined as a composite and find out how mixtures of such scores may be efficaciously combined. In addition, exploring the implications for test construction of some typical findings is a need. There are many modes of possible assessment, each with strengths and weaknesses. He further elaborated that constructed-response items, for example, are more difficult to score objectively and reliably, but they provide a task that may have more systemic validity. A review of studies in four domains -writing, word knowledge, reading, and quantitative- was more equivocal about the value of constructed response but concluded that if differences do exist for any domain, they are very likely to be small. Another is portfolio assessment, which requires smaller leaps of faith to specialize to a specific situation, but it loses ground in the areas of objectivity, equity, generality, and standardization.

Along this view, Bloom's six cognitive processes/domains test both the memory skills and HOTS of the learners and these are also used by teachers in their test plan or table of specifications (Tejero & Catchillar, 2004). These are Knowledge, which requires the students to recall or recognize information as it was exactly learned and rely on memory and/or senses to provide the answer; Comprehension, which requires them to go beyond simple requires them to interpret or to use previously learned information by putting it in his/her own words and rephrasing it; Application, which requires them to apply previously learned information to solve or answer a problem where they use a rule, a definition, a classification system, directions, or the like in solving a problem with a specific correct answer; Analysis, which requires them to identify causes, reasons or motives (when these have not been provided to the student previously), analyze information to reach a generalization/conclusion, find evidence to support a specific occurrence, event or situation, and establish interrelationship between and among things ; Synthesis, where they are required to create something new based on a set of criteria or use original and creative thinking in developing original communication, making predictions, and solving problem for which there is no single answer; and Evaluation, which requires the students to judge the merits or value of an aesthetic work, an idea, or a solution to a problem based on one's own criteria or the well- understood criteria of another.

A revision, primarily by name, of Bloom's taxonomy was designed by Anderson and Krathwohl (2001) to help teachers understand and implement standards-based curricula, to refocus the attention of educators on the original Bloom's Taxonomy as a document not only historical in nature but valid in context of today's standards, and, secondly, to incorporate new knowledge and thought into Bloom's framework. The revised Bloom's Taxonomy incorporated a framework that is no longer simply linear but a grid.

The original six components were renamed so that they still relate directly to the original taxonomy but in terms that are both more relevant to today and simplified. "Knowledge" becomes "remember", "comprehension" becomes "understand", "application" is simplified to "apply", "analysis" to "analyze", and "synthesis" becomes synthesis becomes somewhat confusingly "evaluate" as "evaluation" changes to the more descriptive "create". Cognitive psychologists, curriculum specialists, teacher educators, and researchers had developed a two-dimensional framework, focusing on knowledge and cognitive processes. In combination, these two define what students are expected to learn in school. Curricula target to explore the three unique perspectives-cognitive psychologists (learning emphasis), curriculum specialists and teacher educators (C&I emphasis), and measurement and assessment experts (assessment emphasis).

The six questioning levels by Bloom as revised by Anderson and Krathwohl (2001) and the types of tests by Gronlund (1998) along with the guidelines in test construction of each type and subtype of tests by Williams (1991) are used by the researcher in this study for the analysis of teacher-made tests in JRMSU.

2.1 *Related Studies*

The study of Walstad and Becker (1994) revealed that there was little difference in the knowledge, skills, or abilities measured by multiple-choice and essay or constructed-response tests. The result of the study practically negate the study of Advanced Placement (AP) tests in seven college subjects: calculus, computer science, chemistry, biology, history, French, and music, which concluded that whatever is being measured by the constructed-response section is measured better by the multiple-choice section.

Another study was that of Frey and his colleagues (2005) on "Item Writing Rules: Collective Wisdom". It revealed that multiple choice, matching, and short response item types were most frequently used, and the essay items were infrequently used. The test items functioned primarily at the knowledge cognitive level, that matching exercises contained the most types of construction errors, and that neither the frequency nor the nature of the construction errors varied with teachers' experience. These construction errors can be attributed on some teachers who placed much emphasis on non-test assessment and evaluation strategies.

In a study by Gareis and Grant (2015) on "Teacher-Made Assessments: How to Connect Curriculum, Instruction, and Student Learning", they revealed that the average teacher did not perceive college measurement courses to be pertinent to his/her classroom testing needs and that most teachers learned how to test their students through their on-the-job experiences. Teachers in higher education were perceived to have a limited understanding of the nature of assessment practices in K-12 classrooms. From the perspective of the classroom teacher, the implication was a need for the reorientation of college instruction, with respect to measurement issues and concepts.

Unlike these types of tests which did not involve higher-order thinking skills of the students, Williams (1991) revealed in her "Writing Quality Teacher-Made Tests: A Handbook for Teachers" that the use of Objective Type of Test assesses students' learning better and promotes HOTS while the use of Subjective Test assesses students' creative thinking and organization skills which is applicable towards effective communication. She further claimed that the multiple-choice test questions tap the students' higher-order thinking skills namely the application, analysis, synthesis, and evaluation.

This is supported by a study of Walstad and Becker (1994) on the achievement differences on multiple choice and essay type tests where they also claimed that a multiple-choice or fixed-response format allows for a wider sampling of the content because more questions can be given in a testing period. Multiple-choice tests also offer greater efficiency and reliability in scoring than an essay. However, they added that the major disadvantage of a multiple-choice item is that the fixed responses tend to emphasize recall and encourage guessing. In an essay or subjective/constructed-response test, students generate responses that have the potential to show originality and a greater depth of understanding of the topic. The essay also provides a written record for assessing the thought processes of the student.

Wainer and Thissen (1993) in their study on "Combining Multiple-Choice and Constructed-Response Test Scores: Toward a Marxist Theory of Test Construction" said that they had never found any test that is composed of an objectively and a subjectively scored section for which this is not true.

A study of Sharkness and DeAngelo (2011) on "Measuring Student Involvement: A Comparison of CTT and IRT in the Construction of Scales from Student Surveys" indicated a finding that although both CCT and IRT can be used to obtain the same information about the extent to which the scale items tap into the latent trait being measured, the two measurement theories provide very different pictures of scale precision. On the whole, IRT provides much richer information about measurement precision as well as a clearer roadmap for scale improvement. The findings support the use of IRT for scale construction and survey development in higher education.

Teachers also have indicated a need for more training in constructing tests. One situation to this was that of Corder's (1982) in his study entitled "Error Analysis and Interlanguage", who found out that teacher-made tests indicated grammar errors

and for that he said that the information arrived at through error analysis could be useful to teachers, textbook writers, and learners. Thus, analysis of grammatical errors offered significant insights into the nature of difficulties in writing and test construction.

The need for error analysis and test construction training is indeed an important consideration for two reasons: studies have shown that teachers construct over half of the tests used in their classes during the school year; and analyses of teacher-made tests have found that multiple choice items are often used and, indeed, item construction errors occur frequently.

The analysis of many teacher-made examinations and patterns of evaluation by Earl (2012) on his "Assessment as Learning: Using Classroom Assessment to Maximize Student Learning" revealed that the majority of the teacher's questions were of the 'low-level' type, requiring students to recall facts. Moreover, very few questions were used to cause students to analyze statements, to put ideas together, to formulate hypotheses and plans, or simple to employ high cognitive levels.

To add up, according to Mertler (1999) on his study "Assessing Student Performance: A Descriptive Study of the Classroom Assessment Practices of Ohio Teachers", teachers construct test items that at some point and at various causes misalign their targeted objective. What is more important, however, is that teachers are post aware of their mistakes and that they further engage themselves in test construction enhancement activities.

On the other hand, Haladyna et al. (2002) through her study "A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment", whom after analyzing a number of studies conducted during the last fifty years, concluded that in a half century, there has been no essential change in the types of questions which teachers emphasize in the classroom, that is the select type of objective test. She added also that about 60 % of teachers' questions require students to recall facts, about 20% require students to think, and the remaining 20% are procedural. That analysis was made many years ago, so the researcher would like to determine, through his study, if changes on teacher-made tests are already made.

Indeed, the above review substantiates the heavy use of teacher-made tests, their reliance on the tests for measurement and assessment, and their self-identified need for greater skill in test item construction. Such literature and studies provided information and insight to the researcher on the subject understudy. Basically, the literature and studies dealt with profound similarity to the present investigation along the subject matter, that is teacher-made tests and the subject area of the test made. However, the difference lies on the locale and year the studies were conducted. This investigation is unique as it focuses on the analysis of teacher-made examinations among instructors/professors of a state university located in Zamboanga del Norte, Mindanao, Philippines.

3. Methodology

3.1 Design. The study used documentary analysis of teacher-made paper-and-pencil tests and their corresponding Tables of Specifications (TOS) collected across five JRMSU campuses namely Siocon, Tampilisan, Katipunan, Dipolog and Dapitan (Main) Campuses during the first semester of academic year 2014-2015. The analysis triangulated quantitative tabulations of item types and cognitive levels with qualitative judgments of language-in-use and item quality.

3.2 Corpus and Participants. The dataset comprised examinations from English 11 (Communication Arts/Skills 1), Math 11 (College Algebra), and Natural Sciences 11 (Physical Science). Instructors and professors who authored these tests were considered indirect participants for the purposes of artifact analysis.

3.3 Instruments. Four checklists were employed: (1) a standardized Bloom's taxonomy checklist for coding cognitive level; (2) a Gronlund-based typology checklist for objective vs. subjective formats and subtypes (multiple choice, true/false, matching, completion, short answer, restricted/extended response); (3) a language-in-use error analysis checklist for grammar and mechanics; and (4) an overall test-quality checklist integrating clarity of directions, blueprint alignment, and scoring transparency.

3.4 Procedures. Each TOS was first coded for its intended cognitive distribution. Each item in the corresponding test was then independently coded for its actual cognitive demand. Divergences were flagged as misalignment cases. Item formats were classified per Gronlund. Grammar and mechanics were reviewed for correctness and clarity in stems, options, and directions. Frequency counts and percentages were computed, supplemented by descriptive comparisons across campuses and subjects.

4. Results and Discussions

4.1 Problem 1. Analysis of Instructors' and Professors' Language-in-Use in Teacher-Made Tests

The analysis of the instructors' and professors' teacher-made tests revealed distinct patterns of grammatical usage and accuracy. Although the 126 examined tests showed no errors in mechanics, several grammatical lapses were observed, most notably in the use of function words, prepositions, verb placement, and modifiers. These findings underscore that, while teachers generally maintain mechanical accuracy, subtle linguistic inconsistencies persist in areas that require grammatical precision and syntactic awareness. The most frequent grammatical issue involved the use of function words, particularly nouns. Seven of the forty-two respondents used the term "direction" instead of the plural "directions" in their test instructions. The singular form denotes a linear course or path, whereas the plural form properly conveys instructional guidance for test takers.

Similarly, three instructors wrote "each of the item/statement" instead of "each of the items/statements," while two others committed the reverse error with "each sentence" in place of "each sentence." These patterns suggest an underlying inconsistency in teachers' grammatical intuition regarding number agreement and the semantics of quantification. Errors in prepositional usage were also prevalent. Eight respondents interchanged the prepositions "in" and "on"- writing, for example, "in the space provided," "in the answer sheet," and "in the paper," where "on" would have been contextually appropriate. Conversely, three respondents used "on" in phrases such as "on the box" and "on the column," instead of the more correct "in." This misuse indicates a lack of consistent awareness of spatial and functional prepositional distinctions, which are fundamental to precise instructional language. Five errors were likewise recorded in the placement of linking verbs, particularly "is" and "are." A representative case is the directive "Identify what part of speech are the underlined words," in which the verb "are" was incorrectly positioned before the determiner "the." The correct phrasing should be "Identify what part of speech the underlined words are." This reflects a syntactic transfer from speech to writing, where teachers may unconsciously prioritize oral rhythm over grammatical order. Additionally, three instances of misplaced modifiers were observed in directions such as "Read each statement carefully," "Write the letter only on the space provided," and "Choose the verb from the parenthesis which agrees with the subject." The modifiers in these sentences were placed too far from the words they modify. Clearer and

grammatically accurate constructions would be "Read carefully each statement," "Write only the letter on the space provided," and "From the parenthesis, choose the verb which agrees with the subject."

These findings align with Corder's (1982) assertion that grammatical errors in teacher-made assessments can provide valuable insights into teachers' linguistic competence and instructional practice. Error analysis, as Corder notes, offers a diagnostic lens for identifying patterns of difficulty that may hinder effective communication in testing contexts.

The present study reinforces this view, highlighting the critical need for ongoing training in test construction and language precision. Since research shows that teachers design over half of the tests used in classrooms each academic year, the quality of these assessments depends significantly on teachers' mastery of both subject content and linguistic accuracy. Systematic support for teachers in test-writing and grammatical awareness is therefore essential for improving the validity and reliability of teacher-made assessments across disciplines.

4.2 Problem 2. Analysis of Instructors' and Professors' Teacher-Made Tests Based on Their Table of Specifications (TOS) and Standardized Qualitative Checklist

The analysis of instructors' and professors' teacher-made tests, as reflected in their respective Tables of Specifications (TOS) and corroborated by a standardized qualitative checklist grounded in Bloom's taxonomy, revealed both commendable strengths and notable inconsistencies in cognitive-level alignment. Across subjects, the findings demonstrate that most teachers favored test items that measure application and analytical thinking, though issues of misclassification and conceptual misunderstanding of questioning levels were also evident.

In English 11, the 3,465 test items revealed that roughly 26 percent targeted the Application level, while Synthesis comprised the smallest proportion. Upon validation through the standardized checklist, however, the least-represented cognitive level was Evaluation rather than Synthesis, although Application still remained dominant. This suggests that English instructors prioritized tasks requiring the practical use of learned knowledge-problem solving through definitions, rules, and classifications- while giving limited attention to evaluative or creative reasoning. The discrepancy between the TOS and the checklist indicates that some test items were misaligned with the intended cognitive level. For instance, several questions originally tagged under Evaluation were more appropriately classified as Synthesis, while others categorized under Knowledge in fact tested Application or Comprehension. Such inconsistencies point to the instructors' occasional uncertainty in framing questions that accurately reflect the cognitive domains prescribed by Bloom.

In Mathematics 11 (College Algebra), both the TOS and the checklist agreed that Application dominated, whereas Synthesis was entirely absent. Substantiation revealed that many questions initially coded as Analysis were, in fact, application-type items, and nearly one-fourth of those labeled as Knowledge tested evaluative reasoning. This pattern suggests that mathematics instructors emphasize procedural and rule-based problem solving- consistent with the nature of mathematical learning- but seldom frame items that invite open-ended reasoning, original problem creation, or multiple possible solutions.

For Natural Sciences 11 (Physical Science), further divergence emerged between the TOS and checklist results. The TOS identified Knowledge as the most frequently targeted level, yet checklist validation revealed that analysis actually predominated. Many items initially labeled as Knowledge or Comprehension required students to perform analytical reasoning, such as identifying causes, interpreting relationships, and drawing conclusions rather than simple recall. About fifteen percent of Knowledge-level items were found to assess Analysis, and nearly nineteen percent fell under Evaluation. This reclassification demonstrates a general upward shift toward higher-order questioning, suggesting that science instructors naturally incorporate analytical reasoning even when intending to test factual recall.

Nonetheless, Synthesis remained the least emphasized domain across the sciences, with only a marginal increase in representation after validation. Few instructors asked students to formulate predictions, design hypotheses, or produce original explanations- skills associated with creative and synthetic thinking. The prevalence of analytical tasks over synthetic and evaluative ones reflects a disciplinary bias toward empirical reasoning rather than generative thinking.

When viewed collectively, the data show that across English, Mathematics, and Science, instructors predominantly assessed higher-order thinking skills (HOTS)- Application, Analysis, Synthesis, and Evaluation- more frequently than lower-order skills such as Knowledge and Comprehension. This overall trend toward complex cognition is noteworthy, even if some misclassifications blur categorical distinctions. The results imply that faculty members, whether consciously or intuitively, tend to value the cognitive engagement of their students beyond rote learning. Students were often required to analyze statements, relate ideas, make predictions, or judge the merit of arguments or solutions- activities that promote deeper learning and intellectual autonomy.

These findings diverge from that of Earl (2012), whose analyses of teacher-made tests in other contexts found that most questions targeted only lower-level cognitive skills. In contrast, the JRMSU data reflect a more progressive orientation toward fostering critical and analytical reasoning. However, the identified discrepancies between intended and actual questioning levels underscore the continuing need for faculty training in assessment literacy, particularly in the accurate application of Bloom's taxonomy to classroom evaluation. Refining teachers' understanding of cognitive domains can enhance the validity of departmental examinations and ensure a closer alignment between instructional objectives and the assessments that measure them.

4.3 Problem 3. Analysis of Teacher-Made Tests Based on Gronlund's Quality Types of Tests

An examination of the teacher-made tests used by JRMSU instructors across English 11, Mathematics 11, and Natural Sciences 11, grounded in Gronlund's (1998) Quality Types of Tests, revealed distinct disciplinary tendencies in test design and construction. The data underscore a strong institutional preference for objective-type assessments, particularly of the select and supply forms, and a corresponding underutilization of subjective test types that require extended written responses. In English 11, the data show a pronounced dominance of objective-type items, constituting approximately 98.7 percent of all 3,465 test items. Of these, more than half (56.17 percent) were select-type and the remainder supply-type. Within the select-type items, the Alternative Response (AR) format overwhelmingly prevailed, comprising 86.83 percent of all select-type tests. Interestingly, these AR items frequently expanded beyond the conventional true/false dichotomy, offering multiple options and often requiring students to write their chosen responses rather than simply encircle them—an innovative adaptation that merges objectivity with student agency. The analysis further revealed that the AR items demonstrated commendable adherence to established principles of test construction. Each statement

contained a single clear idea, avoided ambiguity or trick questions, and was devoid of grammatical or syntactic errors. Teachers refrained from using double negatives, overly complex phrasing, or trivial content, ensuring that the items genuinely reflected intended learning outcomes. The absence of opinion-based or verbatim textbook statements also attests to the originality and pedagogical intent of the instructors. Collectively, these results suggest that English 11 teachers designed tests that moved beyond rote recall to measure comprehension and application, embodying a fair degree of construct validity. However, not all test forms reflected the same rigor. Matching-type (MT) tests, which represented the smallest proportion (3.9 percent), frequently omitted key structural elements such as explicit directions for matching premises and responses or clear titling of columns.

Although most sets displayed internal consistency and conceptual homogeneity, the lack of uniform response distribution sometimes led to multiple plausible matches—an issue that can reduce reliability. Despite these lapses, the MT items were generally organized logically, presented on a single page, and corresponded with the learning outcomes in the Table of Specifications (TOS). The English tests also included Restricted Response items but no Extended Response essays. The restricted form required students to recall, organize, and integrate ideas within a controlled scope, encouraging higher-order thinking without demanding extensive elaboration. While the directions were grammatically sound and the tasks aligned with learning outcomes, they lacked essential specifications such as time limits, point values, and scoring rubrics. This omission limits the reliability and transparency of scoring. Nonetheless, the dominance of the AR and restricted-response types reflects English instructors' preference for assessments that balance objectivity with conceptual engagement. In contrast, Mathematics 11 and Physical Science 11 displayed parallel qualitative patterns. Both subjects were characterized by a high prevalence of objective supply-type tests comprising 64.98 percent of Mathematics and 58.06 percent of Science objective test items. Within these, short-answer questions predominated, making up 98.55 percent in Mathematics and 63.33 percent in Science. These questions were typically phrased in clear and concise language, directly tied to the TOS-specified outcomes, and often required computation, definition, or brief explanatory responses. Multiple-choice (MC) questions also figured prominently in both disciplines. In Mathematics, MC items accounted for 64.34 percent of the select-type tests, while in Science, they comprised 75.38 percent. The construction of these items generally followed recognized best practices: stems were concise and complete, alternatives were plausible, and distractors were logically related to the content. Teachers avoided syntactic clues, unequal option lengths, or obvious patterns that could cue correct answers. Items were typically written as direct questions rather than incomplete statements, thereby improving clarity and fairness. Importantly, all keyed responses were unique, and each question stood independently- traits conducive to reliable scoring and standardized testing preparation.

Nevertheless, some lapses were noted. In several tests, alternatives were not presented in ascending order, and distractors occasionally included problematic options such as "all of the above," "none of the above," or compound responses like "both A and B." While such issues are common in teacher-made assessments, their presence highlights the need for more systematic training in item-writing conventions.

Despite these imperfections, multiple-choice questions remained among the least error-prone formats and the most frequently used, affirming Frey et al.'s (2005) finding that teachers favor MC, matching, and short-response items for their efficiency and objectivity. A notable distinction among disciplines lies in the absence of subjective-type items in Mathematics, and their marginal presence in Science and English. This pattern suggests a prevailing pedagogical orientation toward structured, quantifiable assessment rather than open-ended evaluation. Frey (2005) similarly observed that essay items are rarely employed by teachers, often due to time constraints and perceived grading subjectivity. Overall, the aggregate data reveal that objective select-type tests held the highest mean proportion across all subjects (0.41), while subjective-type tests registered the lowest (0.02). This confirms that teachers across JRMSU rely predominantly on structured, objective measures of student performance.

The findings substantiate Haladyna et al.'s (2002) conclusion that, over five decades, classroom assessment practices have shown remarkable continuity in their preference for objective formats. Likewise, Williams (1991) maintains that while objective tests effectively measure cognitive comprehension and foster higher-order thinking, subjective tests remain indispensable for assessing creative and organizational skills vital for written communication. Taken together, these findings portray a balanced yet incomplete picture: JRMSU instructors demonstrate competence in constructing objective assessments aligned with intended learning outcomes but display limited engagement with evaluative and creative testing formats. For a more holistic evaluation of student learning, future assessment design should integrate both objective rigor and subjective depth, ensuring that students are not only accurate and analytical but also expressive and reflective in their intellectual performance.

4.4 Problem 4. Overall Collective Analysis of Teacher-Made Tests

The overall analysis of the teacher-made tests administered by JRMSU instructors and professors, encompassing 5,320 items across English 11, Mathematics 11, and Natural Sciences 11, reveals a complex intersection of cognitive emphasis, item construction quality, and alignment with instructional objectives. The findings from the aggregated data illuminate both commendable practices and recurrent areas for pedagogical refinement in test development. A major trend observed across all test types is the predominance of lower-order cognitive assessments, particularly in multiple-choice and alternative response items. Nearly half (46.15%) of the 663 multiple-choice items and 26.61% of the 1,800 alternative response items were designed to assess Knowledge, which is the most basic level of Bloom's taxonomy. These items primarily measured students' ability to recall factual information, recognize definitions, and retrieve previously learned material from memory. While this focus reflects teachers' desire to ensure mastery of foundational concepts, it also highlights a persistent overreliance on recall-based questioning.

This finding aligns with Frey's (2005) conclusion that multiple-choice questions most often operate at the Knowledge level rather than stimulating higher-order thinking. Conversely, it diverges from Williams (1991), who contended that well-constructed multiple-choice tests can engage higher cognitive skills such as Application, Analysis, Synthesis, and Evaluation. The discrepancy underscores that while the multiple-choice format has the potential to assess complex reasoning, its actual implementation frequently emphasizes memory recall. Walstad and Becker (1994) similarly noted that multiple-choice tests provide broad content sampling and efficient scoring but at the cost of encouraging surface learning and guessing.

The present findings suggest that JRMSU instructors may still exceed this proportion, emphasizing content coverage over cognitive depth. The data further reveal that all select-type

objective tests (multiple choice, alternative response, and matching type) and one supply-type test (completion) measured Synthesis least. This indicates that opportunities for students to generate original responses, create new ideas, or solve problems with multiple valid solutions were limited. In contrast, most matching, completion, and short-answer items assessed Application, comprising 30.77%, 63.16%, and 39.01% of their respective categories.

These findings suggest that objective tests at JRMSU tend to measure procedural understanding and applied knowledge more effectively than creative or evaluative reasoning. Among the subjective test types, only the Restricted-Response form was observed, representing a mere 1.13% of all test items and exclusively assessing Synthesis. No Extended-Response items appeared in any subject area. This suggests that teachers seldom required students to engage in original composition, argumentation, or creative problem-solving. Nevertheless, the restricted-response items that were present effectively targeted students' ability to integrate information and express understanding in written form. Reise's (2012) multi-domain review supports the value of such constructed-response tasks, noting that while they are more challenging to score objectively, they yield richer insights into learners' reasoning processes and demonstrate stronger systemic validity than fixed-response formats. Beyond cognitive emphasis, the overall quality of the tests was assessed through the instructors' Tables of Specifications (TOS). The analysis revealed that the majority of TOS documents clearly specified content areas and instructional objectives, along with their relative emphasis. However, not all learning outcomes were proportionately represented by at least ten corresponding objective test items, particularly in subjects where subjective questions were included. Despite this limitation, most test formats corresponded logically to the intended outcomes, demonstrating a degree of content validity. The test directions were generally concise, grammatically correct, and written at a readability level appropriate for students. While most directions specified the procedural aspects of the tasks, few indicated the time allotment or scoring procedures. The individual test items were independent and presented clear, well-defined tasks that matched students' comprehension level. They were largely free from mechanical errors, grammatical clues, and misleading determiners. Nevertheless, as discussed earlier, several grammatical inconsistencies persisted in a minority of items, particularly in teacher-made directions. Structurally, each objective test item featured one correct answer, with multiple-choice options properly aligned on a single page and completion blanks standardized in placement and length. Short-answer and essay-type questions provided sufficient space for responses, though tests were not consistently sequenced from easy to difficult. Sample items were rarely provided, suggesting that teachers assumed student familiarity with standard test formats. Overall, the tests were sufficiently comprehensive to measure course content but not so lengthy as to assess speed rather than mastery- a positive indicator of balanced test design.

The analysis also revealed some inconsistencies between the targeted and actual cognitive levels indicated in the TOS, reaffirming earlier observations in the preceding problems. Moreover, although certain types of tests demonstrated high-quality item construction, others exhibited minor deviations from established test-writing standards. Importantly, despite occasional misalignments in lesson coverage between syllabi and examinations, where preliminary lessons sometimes extended into midterm or final scopes, the tests still adequately represented the prescribed curricular content in Communication Arts/Skills I, College Algebra, and Physical Sciences.

Synthesizing these findings, the study concludes that JRMSU instructors demonstrate commendable effort in constructing teacher-made tests that are systematically organized, aligned with learning objectives, and increasingly oriented toward higher-order cognitive skills.

However, gaps remain in achieving balanced assessment coverage across cognitive domains. The general reliance on objective formats, the underrepresentation of synthesis-level tasks, and the persistence of minor linguistic errors indicate the need for sustained faculty training in both test construction and language use. Consistent with Stiggins et al. (2004), the findings affirm that constructing first-rate teacher-made tests is a continuous professional challenge. As classrooms become more diverse and cognitively demanding, educators must develop a refined understanding of test validity, linguistic precision, and assessment design. Ultimately, strengthening teachers' competence in these areas will not only improve the reliability and fairness of their tests but also ensure that assessments serve as authentic instruments of learning rather than mere measures of recall.

5. Conclusions

Based on the findings of the study, the following conclusions were drawn:

5.1 Accuracy in Language Use: The teacher-made tests of JRMSU instructors and professors demonstrated consistent mechanical accuracy but exhibited occasional grammatical lapses. While technical precision in punctuation, spelling, and formatting was generally maintained, syntactic inconsistencies, particularly in function words, prepositions, and modifier placement, indicate the need for greater linguistic rigor in test construction.

5.2 Emphasis on Higher-Order Thinking Skills: Overall, the teacher-made tests across English 11, Mathematics 11, and Natural Sciences 11 reflected a deliberate pedagogical orientation toward the assessment of higher-order thinking skills (HOTS). The predominance of items targeting application and analysis levels suggests that instructors designed their tests not merely to measure recall but to encourage interpretation, reasoning, and problem-solving.

5.3 Disciplinary Variations in Assessment Design: Distinct disciplinary patterns emerged in test construction. Instructors of English 11 favored select-type objective items, requiring students to recognize the correct answer among options provided. By contrast, instructors of Mathematics 11 and Natural Sciences 11 predominantly employed supply-type objective tests, which required students to generate correct responses independently. These differences reflect each discipline's characteristic approach to measuring mastery- recognition-based assessment in language learning versus production-based assessment in quantitative and scientific reasoning.

5.4 Cognitive Levels and Test Type Alignment: The study established a clear hierarchy of cognitive targeting among test types. Objective tests seldom assessed synthesis or creative reasoning, with multiple-choice and alternative response formats functioning primarily at the knowledge level. Conversely, matching-type, completion, and short-answer tests most frequently measured application skills. The restricted-response subjective tests, though limited in number, uniquely assessed synthesis- the ability to generate, integrate, and express original

ideas. This pattern underscores both the structured precision and the cognitive constraints inherent in objective testing formats.

In sum, the findings affirm that JRMSU instructors possess the foundational competence to design valid and structured teacher-made tests aligned with learning outcomes. However, enhancing grammatical precision, expanding the range of cognitive levels assessed, and integrating more open-ended and creative response formats would further strengthen the validity and instructional value of classroom assessments.

6. Recommendations

6.1. Establish Departmental Item Review Panels. Adopt pre-administration peer review focused on cognitive alignment, language-in-use, and scoring clarity. Use brief checklists and exemplars to streamline workflow.

6.2. Build a Calibrated Item Bank. Curate vetted items tagged by cognitive level, item type, content strand, and common misconceptions. Include annotated rationales and model keys.

6.3. Provide Assessment Literacy Workshops. Offer short cycles on item writing, blueprinting, and basic psychometrics (difficulty, discrimination, reliability), integrating hands-on revision of existing items.

6.4. Strengthen TOS Practices. Require evidence of alignment between intended cognitive distribution and sample items before test approval; implement spot-checks post-administration using item statistics where available.

6.5. Enhance Authenticity in Language Testing. Increase tasks that assess language-in-use (discourse comprehension, argumentation, editing in context) and integrate brief constructed responses to triangulate multiple-choice evidence.

6.6. Standardize Directions and Formatting. Adopt templates for stems, options, and direction lines to reduce ambiguity and language error.

7. Limitations and Future Work

The documentary design constrains causal inference: the data-set reflects one semester and a defined set of courses. Future work can incorporate item statistics from operational administrations, student think-alouds to probe construct representation, and experimental comparisons between redesigned and legacy items to estimate gains in discrimination and validity. Cross-campus faculty communities of practice can be studied longitudinally to examine how calibration routines change item quality and student outcomes over multiple terms.

8. Acknowledgement

The researcher expresses deep heartfelt gratitude to his MA Thesis Adviser Dr. Rizza B. Bagalanon, who is currently (2026) the Vice President for Research, Development and Extension of JRMSU System. She has consistently believed in and helped develop the full potential of the researcher since day one up until the present.

9. References

Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition. Addison Wesley Longman, Inc.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain (pp. 1103-1133). New York: Longman.

Brown, H. D. (2010). Language Assessment: Principles and Classroom Practices.

Corder, S. P. (1982). Error analysis and interlanguage (Vol. 198, No. 1). London: Oxford university press.

Earl, L. M. (2012). Assessment as learning: Using classroom assessment to maximize student learning. Corwin press.

Embretson, S. E., & Reise, S. P. (2013). Item response theory for psychologists. Psychology Press.

Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. Teaching and Teacher Education, 21(4), 357-364.

Gareis, C. R., & Grant, L. W. (2015). Teacher-made assessments: How to connect curriculum, instruction, and student learning. Routledge.

Gronlund, N. E. (1998). Assessment of student achievement. Allyn & Bacon Publishing, Longwood Division, 160 Gould Street, Needham Heights, MA 02194-2310; tele.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. Applied measurement in education, 15(3), 309-333.

Mertler, C. A. (1999). Assessing Student Performance: A Descriptive Study of the Classroom Assessment Practices of Ohio Teachers. Education, 120(2).

Reise, S. P. (2012). The rediscovery of bifactor measurement models. Multivariate behavioral research, 47(5), 667-696.

Sharkness, J., & DeAngelo, L. (2011). Measuring student involvement: A comparison of classical test theory and item response theory in the construction of scales from student surveys. Research in Higher Education, 52(5), 480-507.

Stiggins, R. J., Arter, J. A., & Chappuis, J. (2004). Classroom assessment for student learning: Doing it right, using it well. Assessment Training Institute.

Tejero, E., & Catchillar, G. (2004). Teaching reading in the elementary grades. Mandaluyong City, PH: National Book Store.

Van der Linden, W. J., & Glas, C. A. (Eds.). (2000). Computerized adaptive testing: Theory and practice (Vol. 13). Dordrecht: Kluwer Academic.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. Applied Measurement in Education, 6(2), 103-118.

Walstad, W. B., & Becker, W. E. (1994). Achievement differences on multiple-choice and essay tests in economics. The American Economic Review, 84(2), 193-196.

Williams, J. M. (1991). Writing Quality Teacher-Made Tests: A Handbook for Teachers.