Exploring the Reliability of Grammar Test Tasks' Scoring Methods Among High Schools English Teachers in Morocco

Saber Abou El Fadl

PhD Assistantship, Mohammed V university, Faculty of Education, Morocco.

Abstract

The current study addresses the growing importance of grammar assessment in educational settings and the diverse scoring methods employed by teachers for grammar test tasks. Through a questionnaire administered to 40 high school teachers, the research aims to determine the reliability of these scoring methods and identify the grammatical areas prioritized for evaluation. The study reveals that gap-filling, matching, dialogue completion, and multiple-choice tasks are frequently used by teachers. Notably, multiple-response and limited production tasks often coexist in evaluations, and both singular and multiple grammatical areas are targeted. Overall, the study demonstrates that teachers opt for structured score guidelines on exam sheets, resulting in highly consistent scoring methods.

Keywords: Language assessment, Reliability, Grammatical knowledge, Scoring methods, Consistency.

1. Introduction

Within the paradigm of communicative methodology, a significant challenge arises in the way teachers assess the grammatical abilities of their students. Scores are utilized to aid in making decisions about individuals, and the methods employed to derive these scores constitute a crucial component of the measurement process. Teachers employ a set of grammar tasks to assess the grammatical knowledge of their students. The scoring process plays a pivotal role in ensuring the reliability of the test scores (Palmer and Bachman, 1996). In some instances, considerations related to scoring may influence task use or the intended responses included in the test. The way the test tasks are defined will determine the nature of the intended responses, and this holds clear implications for scoring (1996).

Scoring method stands as a pivotal element in language testing. Teachers' conceptualizations of scoring methods often vary based on the types of tasks they employ. The selection of an appropriate scoring method presupposes adherence to several conditions necessary for achieving reliable scores. When evaluating grammar, teachers might opt for a scoring method that does not align with the task's inherent characteristics. Each task necessitates a specific method. In cases where a teacher neglects to consider the task's suitable application, the resultant scores may fail to accurately represent the student's grammatical knowledge. With this foundation, an investigation is undertaken to explore the reliability of high school teachers' scoring methods in grammar tests, aiming to unravel their approaches to assessing their students. Assessing grammar poses a challenge for high school teachers as they employ various grammar tasks to gauge their students' grammatical proficiency. Each grammar task necessitates a specific scoring typology. It is imperative for teachers to strive for comprehensive results that accurately reflect students' grammatical knowledge, thereby enhancing their learning process and ensuring the attainment of teaching objectives. Drawing from personal experience as an English teacher, I remained acutely mindful of my scoring methodology. Upon reviewing my students' scores, I discerned inconsistencies, signaling a flaw in my approach. The discordance between task types and scoring methods became evident. Through this research, I anticipate gaining valuable insights into language testing, particularly in the realm of grammar assessment. Moreover, I aim to refine my understanding of the appropriate utilization of grammar tasks within the context of classroom testing.

This paper aims to investigate the reliability of high school teachers' scoring methods for grammar test tasks. It delves into the process by which teachers establish criteria for correctness and scoring procedures for frequently used tasks in the classroom. The study also aims to illustrate the degree to which teachers enhance the reliability of their tests. Through this project, insights will be gained into how teachers extract information about the grammatical abilities of test takers.

2. Research Questions

This study seeks to address the following questions:

1. What are the most commonly employed grammar test tasks by teachers?
2. What procedures do teachers utilize for scoring the responses?
3. To what extent do teachers standardize their scoring methods?

3. Research Hypothesis

The research hypothesis underpinning this study asserts the reliability of grammar test scoring methods among high school teachers.

4. Review of the Literature

This section of the article is divided into two parts, each addressing distinct facets of the research domain. The initial part offers an in-depth typology of scoring methods associated with grammar tasks, specifically delving into the nuances of multiple responses and limited production scenarios, whereas the second part represents a comprehensive overview unfolds, encompassing diverse grammar task classifications, the very essence of grammar assessment, and the indispensable theme of test reliability. This layered approach not only elucidates the spectrum of grammar test typologies but also delves into the critical consideration of how reliably these assessments gauge language proficiency.

4.1 Scoring Methods

At the core of language testing lies the quantification of student responses into scores, a quantifiable metric guiding consequential educational decisions. This intricate process, intricately tied to the accuracy and robustness of test scores, entails a three-fold procedure, meticulously outlined by Palmer and Bachman (2004). This process comprises three key steps (2004):

1. The theoretical definition of the construct.
2. The operational definition of the construct.
3. The establishment of a method to quantify responses to test tasks

Furthermore, the specific test modality inherently prescribes the characteristics of anticipated responses, thereby significantly influencing the scoring approach to be adopted. It is noteworthy to emphasize that the scope of language testing encompasses a myriad of task typologies, each necessitating distinct scoring method to be tailored to its unique attributes. These methods fall into two broad approaches (Brown, 2009, 2010). In one approach "the score defined as the number of the test tasks successfully completed, so that the number of correct responses is added up" (Bachman & palmer,1996, p.194). This approach concerns with the items that required selected or limited production responses (1996). However, the second approach tends to "define several levels on one or more rating scales of language ability, and then to rate responses to test tasks in terms of these scales" (Bachman & Palmer.1996, p.195). The latter approach is applicable to produce extended responses like prompt-type tasks (1990, 1996).

4.2 Specification of Criteria for Correctness

Within the realm of both selected and limited production responses, the selection of criteria for correctness holds a variable spectrum contingent upon the specific facet of language knowledge under scrutiny. The meticulous articulation of criteria for correctness assumes paramount significance, as it entails the delineation of factors constituting an accurate response to a given task (Bachman & Palmer, 1996, p. 196).

Apparently, many languages tests use a single criterion for correctness to end up with a 'pure' measure of a specific area of language knowledge (1996). Therefore, if an item intended to measure only grammatical knowledge, one might use grammatical accuracy as the only criterion of correctness. For instance, 'when ones use selected response items, by providing only one alternative that is grammatically correct, while the other choices, are grammatically incorrect, even though they may be appropriate in term of lexical choice (Palmer & Bachman,1996).

1.My neighbor asked me …..away the tall weeds in my yard.

a. A clear
b. *To clear
c. Cleared
d. Cleaning

Further, Bachman and Palmer stated "in items designed to measure lexical knowledge, the test developer might not consider grammatical accuracy at all, but use meaningfulness as the sole criterion, which would be counted as a correct answer if grammatical accuracy were not considered"(1996, p.197). This may work effectively with selected responses, but may create problems with limited production responses if we measure a single area of language knowledge in isolation, we may end up giving credit for an answer that would be perceived to some extent incorrect or inappropriate (1996). Evidently, this view of language use as interactive process involving multiple area of language knowledge, so in order to make inferences about test takers language ability on the basis of their responses in test tasks, multiple criteria for correctness should be involved in the scoring of these responses (grammatical accuracy, meaning, and appropriateness) (1996,p.197).

4.3 Procedures for Scoring Test Takers' Responses

Selected and limited production responses can be scored in one of the following: right /wrong or partial credit. As Bachman and Palmer explains (1996), by using right/wrong scoring a response receives a score of '0' if it is wrong and '1' if it is correct. By using partial credit, responses can be scored on several levels, ranging from no credit ('0') to full credit. Right/wrong method of scoring works if ones want to measure a single area of language knowledge, thus decide to score responses in terms of a single criterion for correctness (1996, p.199). However, Bachman and Palmer recommend that "using partial credit method is more suitable if ones is interested in assessing responses instances of language use, in which several areas of language knowledge are likely to be involved. Thus, responses may be scored according to several criteria for correctness. As a result, test takers might give a number of different answers" (1996, p.200) as follows:

1. Some answers are grammatical accurate.
2. Some answers are semantically appropriate.
3. Some answers meet both of these criteria for correctness.

Table 1: Multiple right/wrong scores for responses to a single item

Response/Criterion	Grammar	Meaning
"hitted"	0	0
'avoided'	0	1
'preserve'	1	0
'avoid'	1	1

Henceforth, once multiple criteria is applied for correctness and right/wrong as scoring method, the score of a wrong response fails to reflect the specific area(s) of language knowledge development that the test taker is subject to. For instance, all answers in the table above receive a score of 0 : hitted, avoided, preserve, even though they are wrong for different reasons(1996).

Hitted: is wrong according to both grammatical and semantic criterion.
Avoided: is grammatically inaccurate but semantically appropriate.
Preserve: is grammatically accurate but semantically inappropriate.

Table 2: Single partial credit scores for responses to a single item.

Response/primary	Grammar	Meaning	Both
"hitted"	0	0	0
'avoided'	0	1	1
'preserve'	1	0	1
'avoid'	1	1	2

Bachman and palmer (1996, p.200)

In contrast, the adoption of partial credit scoring as multiple criteria for correctness permits to report scores for different area of language ability under investigation. In the above example, the grammar scores for all the items could be accounted to be the total score for grammar accuracy, and similarly for meaning appropriateness (1996).

To sum up Bachman and Palmer concluded that "multiple scores or partial credit offers the test users the potential for capturing more information about responses, and hence more information about the test takers' area of strength and weakness, than does giving a single right/wrong score" (1996, p. 200).

4.3 Grammar Test Tasks

Two distinct facets constitute grammatical knowledge: explicit knowledge (Ellis, 2005; Rahimpour, & Salimi, 2010), synonymous with conscious awareness of grammatical structures and their semantic implications (Dekyser, 1995); and implicit knowledge, which manifests through various forms of linguistic performance, such as conversation (Ellis, 2005; Han & Ellis, 1998; Rahimpour & Salimi, 2010). Consequently, in the realm of grammar assessment, the endeavor encompasses the evaluation of both explicit and implicit grammar knowledge, either individually or collectively (Brown, 2009, 2010). Therefore, within the classroom context, "the evaluation of explicit (2005) knowledge might involve employing tests featuring multiple-choice questions probing grammatical forms. Nevertheless, it is imperative to recognize that this form of assessment does not guarantee the students' internalization of grammatical form and use adequately" (Purpura, 2014, pp. 45-46).

Consequently, Purpura (ibid) advocates for a comprehensive assessment test tasks that incorporates the evaluation of students' implicit and explicit grammar knowledge (2005, 2010). From the aforementioned, the theoretical conceptualization of what is intended to be measured presents a real challenge for language testing, since the objective of assessment is to use test scores to make inferences about what learners know (Purpura, 2014, p. 45-46). In nutshell, Language educators (Purpura, 2014, p.47) "should be capable to define the language grammatical knowledge they are measuring and separate them from other facets of language".

4.4 Types of Grammar Test Tasks

4.4.1 Selected-Response Tasks

a) Multiple-Choice(MC) Task

Purpura defined it as "the task that presents input with gaps or underlined words or phrases. Examinees have to choose the correct answer from the response options given" (2014, p.127). It is designated to measure the followings:

1) Designed to test grammatical form (morphosyntax-word order).
2) Designed to test grammatical form and meaning (cohesive-ellipsis).
3) Designed to test grammatical form and meaning (multiple areas).

b) Multiple-Choice Error Identification Task

Purpura defined it as "this task presents test –takers with an item that contains one incorrect, unacceptable, or inappropriate feature in the input" (214, p.131). It is designated to measure grammatical form.

c) The Matching Task

Purpura 2014.P131 defined as "This task presents input in the form of two list of words, phrases or sentences. One list can also be in the form of visual cues. Examinees match one list with the other". Designed to measure grammatical meaning (denotation) (Purpura, 2014)

4.4.2 Limited-Production Tasks

a) The Gap-Filling Task

Purpura defined it as "This task presents input in the form of a sentence, passage or dialogue with a number of words deleted. The gaps are specifically selected to test one or more areas of grammatical knowledge. Examinees are required to fill the gap with an appropriate response for the context. Gap-filling tasks are designed to measure the learning's knowledge of grammatical forms and meaning" (2014, p.135).

Designed to measure grammatical form and meaning.

b) The Short Answer Task

Purpura defined it as “This task presents input in the form of a question, incomplete sentence or some visual stimulus. Test-takers are expected to produce responses that range in length from a word to a sentence or two. The range of acceptable responses can vary considerably” (2014, p. 136). The task is designed to measure both grammatical form and meaning.

c) The Dialogue Completion Task (DCT)

Purpura defined it as “present input in the form of a short exchange or dialogue with an entire turn or part of a turn deleted. Examinees are expected to complete the exchange with a response that is grammatically accurate and meaningful” (2014, p.138). The task is designed to measure grammatical form and meaning on the discourse level.

4.5 Reliability of Grammar Test Tasks

Reliability of test scores is determined by their consistency across various test administrations. This consistency remains unaffected by factors such as test timing, format variations, or the evaluator’s role in scoring responses (Purpura, 2014). The test stability in measurement is a key element of its reliability (Purpura, 2014). As example, if students are tested on a Monday and the same test is repeated on a Friday, the resulting scores should exhibit minimal divergence (Purpura, 2014). Through meticulous test design, efforts are made to mitigate the potential impact of manageable sources of inconsistency (Bachman & Palmer, 1996; Brown, 2009, 2010).

Objective scoring techniques are needed to neutralize the scoring method and avoid any kind of decision-making in the scoring process (Purpura, 2014). Moreover, Purpura explains (2014) that scoring method can be objectified by training raters to score consistently according to an agreed-upon rubric, and by having more than one independent rater judging performance. Furthermore, simultaneously, reliability of test-task can be ascertained if ones increase the number of tasks on a test (2014).

In essence, the reliability of a test hinges upon the extent of consistency and precision sought during the measurement and scoring of performance. This is achieved through the implementation of a standardized rubric and the inclusion of a diverse range of task types and quantities within the test.

5.Methodology

Within the realm of educational assessment, the nuances of grammar evaluation assume a pivotal role in shaping pedagogical practices and gauging students' linguistic proficiency. This study delves into the landscape of English education in Moroccan public high schools within Rabat, with a focused exploration into the reliability and objectivity inherent in teachers' scoring approaches for grammar test tasks. Specifically, the research seeks to identify prevalent types of grammar test tasks employed by educators and to measure the degree to which these methods align with impartial assessments of student developmental progress. Guided by meticulous research questions and hypotheses, this study employs a quantitative approach centered around a purpose-designed questionnaire, shedding light on the practices of high school teachers and enhancing our understanding of the intricate dynamics surrounding grammar assessment strategies.

5.1 Sample

The sample includes 40 English instructors from public high schools in Rabat. These educators teach at 15 distinct schools throughout the city. The age of the participants ranges from 30 to 57 years, and they include both male and female individuals. All participants have significant experience in teaching English and in scoring English grammar tasks.

Table 3: The High Schools Selected for the Scope of the Study

Name of High Schools	Number of visits	Number of questionnaires which are collected
MOULAY YOUSSEF	2	3
LALLA AICHA	4	2
COLLECTION OF SCHOOLS MOHAMMED V	4	2
LALLA NAZHA	4	2
LAYMOUN	3	4
ABDELKARIM EL KHATABI	1	1
MALKI	2	2
ABIDAR KHAFARI	2	5
ABDELLAH GUENNOUN	2	3
MOLLAY ABDELLAH	2	3
DAR ESSALAM	2	3
ALLAL EL FASSI	3	2
IBRAHIM RODANI	2	3
HASSAN 2	4	0
IBN ROCHD	5	5
The Total Number	42	40

5.2 The Research Instrument

To address the research questions, a questionnaire consisting of five items was distributed to public high school English teachers. The decision to employ a questionnaire as the sole instrument for data collection stemmed from its perceived efficacy in elucidating teachers' scoring methods regarding grammar test tasks. This methodological choice aimed to ascertain the extent to which these tasks reliably measure students' linguistic proficiency.

The questionnaire encompassed four items presented as a checklist and one as a multiple-choice question with an accompanying open-ended component. Importantly, the selection of these question types was deliberate, designed to maximize data acquisition from the participants.

The items within the questionnaire sought to probe specific facets related to the reliability of grammar test task scoring methods. These items were categorized into four sections. Pertinently, the first section, encompassing the initial item, endeavored to examine the frequency with which teachers employ selected and limited-response grammar tasks in classroom assessment scenarios.

5.3 Administration

When administering the questionnaire, many participants took a long time to complete it, mainly because some faced pressures from end-of-term examinations. Some instructors declined to participate in the study for unspecified reasons. During our visits, the high schools welcomed the research team, and the administrative staff was of great help.

5.4. Data Analysis Procedure

This research embraced an exploratory methodology. Subsequent data analysis employed descriptive statistical techniques via the SPSS software platform. Consequently, findings were quantitatively represented through calculated percentages and visualized in graphical formats, details of which will be expounded upon in the ensuing section.

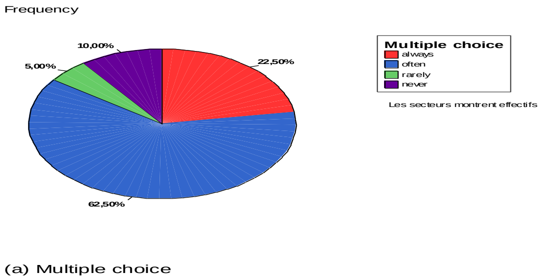
6. Data Analysis

This study seeks to scrutinize the reliability of high school teachers' scoring methods for grammar tests. Accordingly, a series of questions were designed to align with the research objectives. Each item in the questionnaire is purposefully designed to yield data earmarked for in-depth analysis in the subsequent section. The forthcoming section is dedicated to presenting the items from the questionnaire in a quantified manner. Each question will be considered individually. An array of graphs is integrated to provide readers with a quantitative depiction of the collated data

6.1 Frequency of Grammar Task Utilization

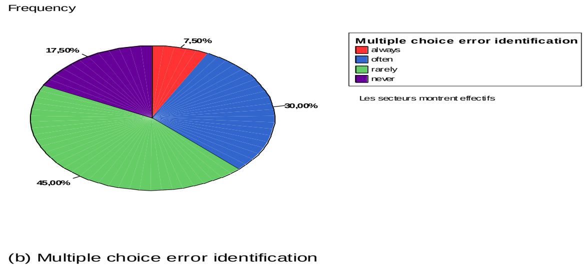
The tasks included in the study's questionnaire are categorized into two primary classifications: Selected Response Tasks and Limited Production Tasks. Each task's frequency of utilization by teachers is illustrated distinctly through separate graphs.

Multiple Choice Tasks: Graph (a) illustrates the frequency with which high school teachers employ the multiple-choice task. A substantial 62% of teachers tend to use the multiple-choice task "often." In contrast, 22.5% invariably "always" gravitate towards this format. Only a small fraction, 5%, "rarely" implement it, while 10% abstain from deploying the multiple-choice task entirely. Cumulatively, the data indicates that an approximate 85% of teachers integrate the multiple-choice task, albeit with varying regularity.



(a) Multiple choice

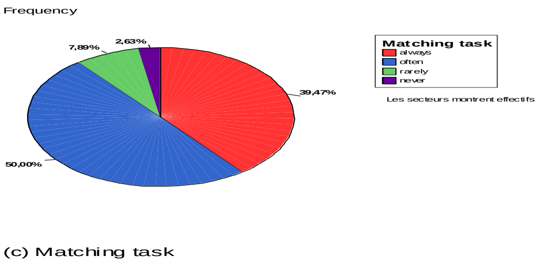
Graph 1. Multiple Choice Task Use Frequency



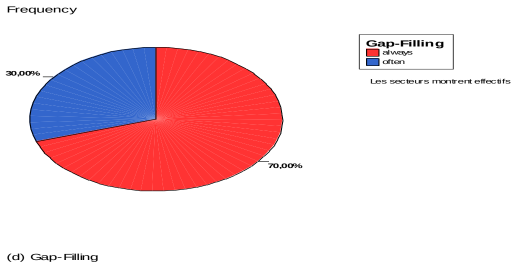
Graph 2. Multiple Choice Error Identification Use Frequency

Multiple Choice Error Identification (MCEI) Task: **Graph (b)** elucidates the utilization frequency of the Multiple-Choice Error Identification task among high school teachers. Notably, the predominant segment, representing 45%, indicates teachers who "rarely" employ the MCEI task. This is followed by 30% who "often" resort to the MCEI task. A mere 7.5% "always" favor this task, whereas 17.5% abstain from using the MCEI task altogether. Consequently, the consolidated data reveals that a substantial 62.5% of teachers display a reticence towards using the MCEI task.

Matching Task (MT): Graph (c), presented below, delineates the utilization frequency of the Matching task by high school teachers. A significant 50% of teachers "often" employ the MT task, followed closely by 39.47% who "always" prefer this format. Conversely, a marginal 7.89% "rarely" incorporate the MT, and a mere 2.63% abstain from its usage entirely. Cumulatively, the data indicates that an overwhelming 89.47% of teachers integrate the Matching task, though their frequency of employment varies.



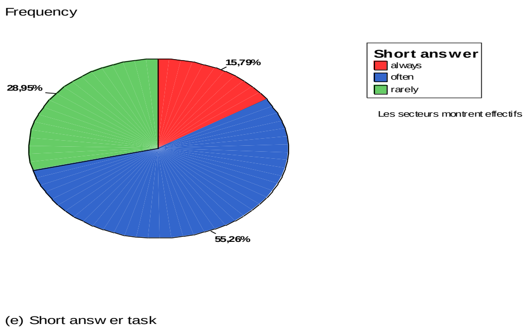
Graph 3: Matching Task Use Frequency



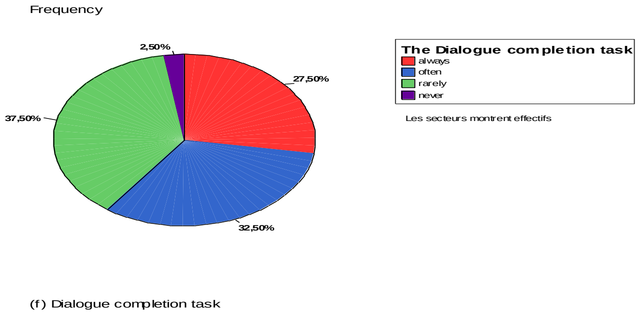
Graph 4: Gap-Filling Task Use Frequency

Gap-filling (GF): Graph (d) delineates the utilization frequency of the Gap-filling task among high school teachers. A pronounced majority, approximately 70%, consistently "always" employ the Gap-filling task. Subsequently, just over 30% indicate a preference for this task "often". Thus, the cumulative data underscores that the entirety of the respondents, or 100%, incorporate the Gap-filling task, albeit with varying regularity.

Short Answer Task (ShA): Graph (e) elucidates the frequency with which the short answer task is employed by high school teachers. A predominant 55.26% of teachers "often" gravitate towards the short answer task, while 15.79% consistently "always" employ it. In contrast, 28.95% "rarely" integrate the ShA task into their assessments. Collectively, the data highlights that a substantial 71.05% of teachers utilize this task format.



Graph 5. Short Answer Task Use Frequency

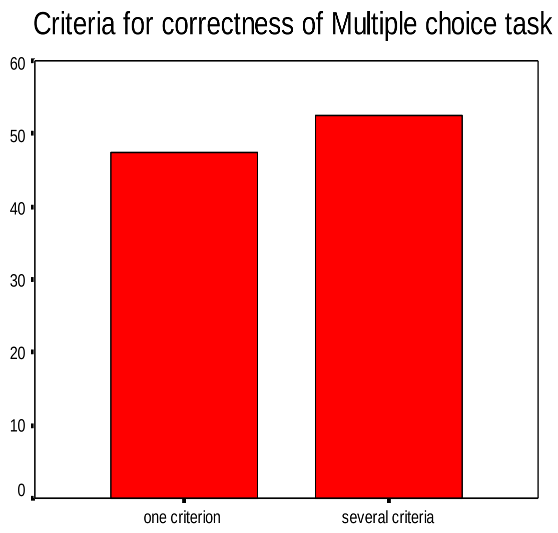


Graph 6: The Dialoque Completion Task's Use Frequency

Dialogue Completion Task (DCT): Graph (f) delineates the frequency of utilization of the dialogue completion task by high school teachers. Notably, 37.5% of teachers "rarely" incorporate the dialogue completion task, while a slightly smaller proportion, 32.5%, "often" employ it. In addition, 27.5% "always" favor the DCT, whereas a marginal 2.5% abstain from its use entirely. Cumulatively, the data indicates that approximately 60% of teachers regularly (either "often" or "always") implement this task format.

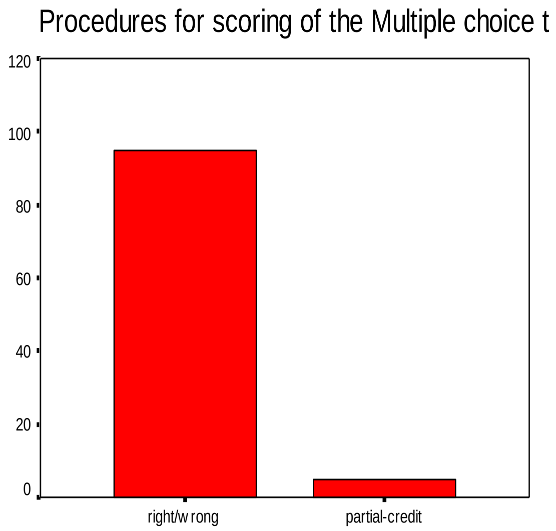
6.2 Criteria for correctness and scoring methods.

Analysis of Multiple-Choice Task: Referring to Graph (a), it is evident that 47.5% of high school teachers utilize multiple choice tasks to evaluate a singular dimension of grammatical knowledge, be it meaning or form. Conversely, 52.5% opt to concurrently assess both these grammatical facets. Moreover, **Graph (b)** elucidates the scoring methods for this task. A significant majority, 95%, employ the unequivocal 'right/wrong' criterion, in contrast to those who adopt the partial credit approach.



a) Criteria for correctness of Multiple choice task

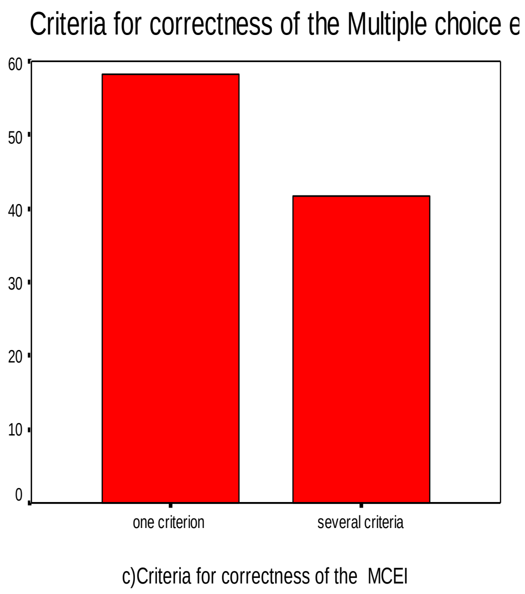
Graph 7. Multiple Choice Task Criteria for Correctness

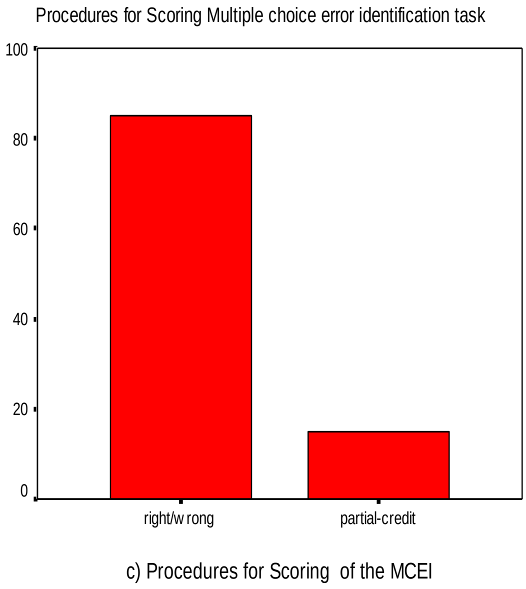


b) Procedures for scoring of the Multiple choice task

Graph 8. Multiple Choice Task Procedures for Scoring

Analysis of Multiple-Choice Error Identification: With reference to Charts (c) and (d), it is observed that 58% of high school teachers employ the multiple-choice error identification method to evaluate a singular dimension of grammatical knowledge, either form or meaning. In contrast, 42% favor a comprehensive assessment, encompassing both meaning and form. Chart (d) further delineates the preferred scoring procedures for this task. An overwhelming 85% adhere to the binary 'right/wrong' evaluation, while a mere 15% adopt the more nuanced partial credit scoring method.

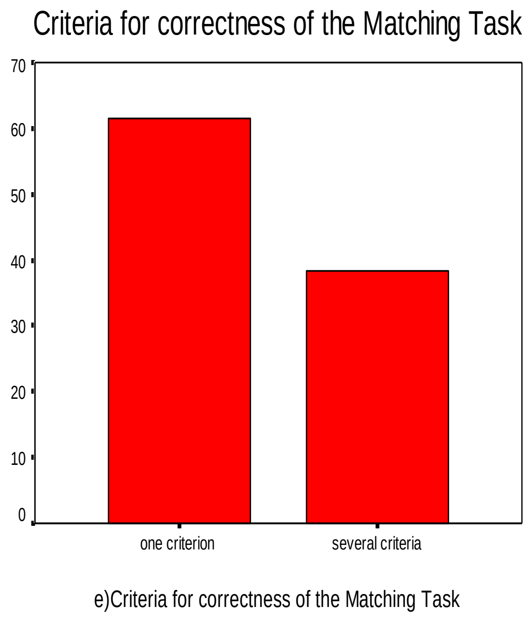


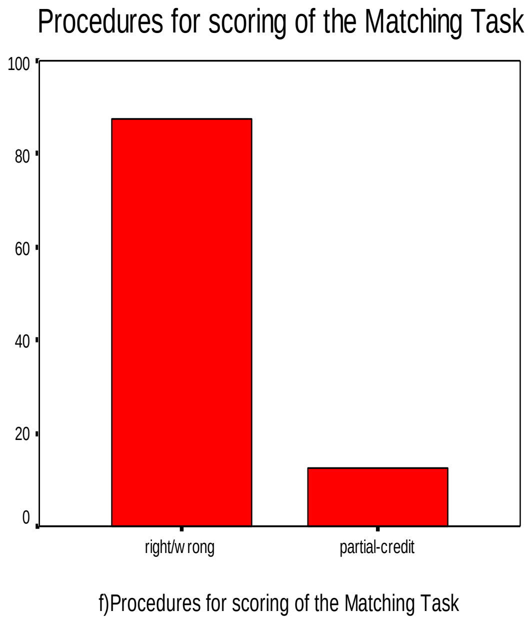


Graph 9. MCEI Criteria for Correctness

Graph 10. MCEI Procedures for Scoring

Analysis of the Matching Task: As delineated in Charts (e) and (f), 62.5% of high school teachers employ the Matching task to evaluate either form or meaning as distinct components of grammatical knowledge. On the other hand, 37.5% integrate a holistic approach, examining both meaning and form concurrently. Further insights from Chart (f) reveal that a predominant 87.5% of educators gravitate towards the straightforward 'right/wrong' scoring method, while a smaller faction, 12.5%, utilizes the partial credit system for evaluation.

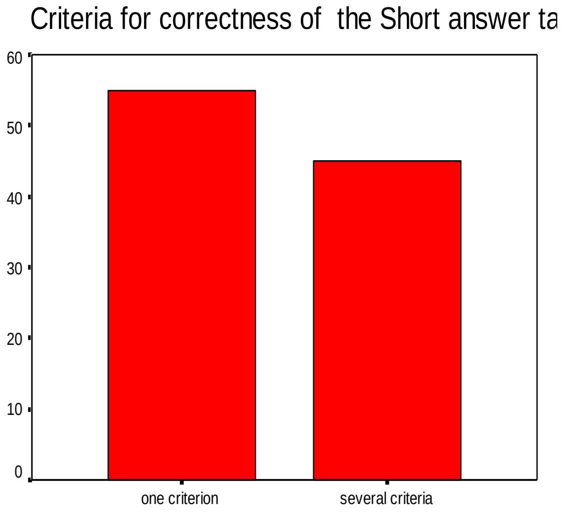




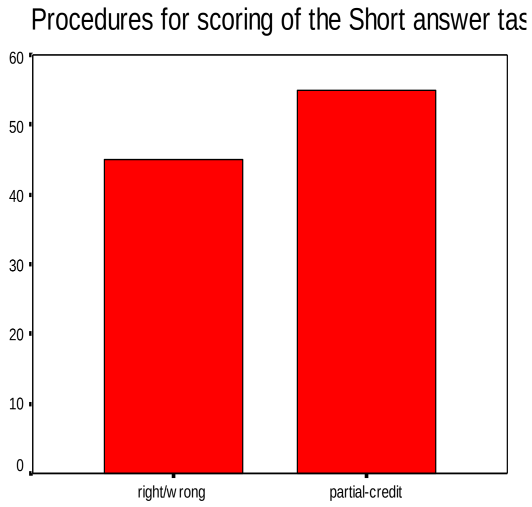
Graph 11. Matching Task Criteria for Correctness

Graph 12. Matching Task Procedures for Scoring

Analysis of the Short Answer Task: As per the data presented in Charts (g) and (h), 55% of high school teachers utilize the Short Answer task to focus on a singular facet of grammatical knowledge, be it form or meaning. Conversely, 45% opt for a comprehensive evaluation, assessing both form and meaning in tandem. A further examination of Chart (h) elucidates that 45% of educators favor the unequivocal 'right/wrong' scoring criterion. In contrast, a slightly larger proportion, 55%, embraces the nuanced approach of partial credit scoring.



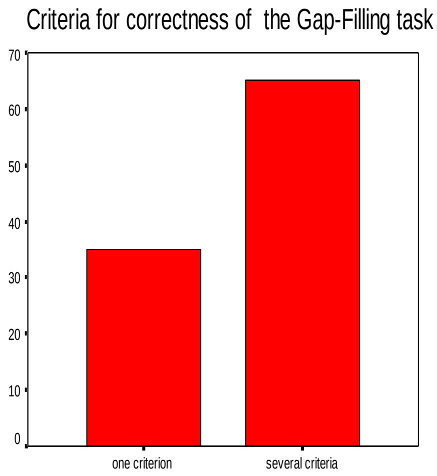
g) Criteria for correctness of the Short answer task



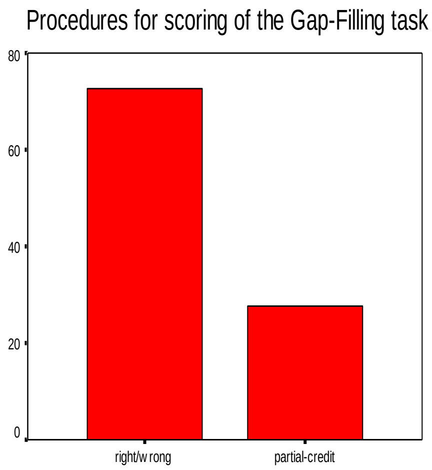
h)Procedures for scoring of the Short answer task

Graph 13. Short-Answer Task Criteria for Correctness **Graph 14.** SAT Procedures for Scoring

Analysis of the Gap-Filling Task: According to Charts (i) and (j), 65% of high school teachers employ the Gap-Filling task to evaluate multiple facets of grammatical knowledge, either form or meaning. In juxtaposition, 35% focus their assessment on a singular dimension, specifically targeting either form or meaning. Chart (j) further provides insight into the scoring modalities preferred: a significant 72% of educators adhere to the clear-cut 'right/wrong' metric, while a minority of 27.5% incorporate the partial credit scoring method.



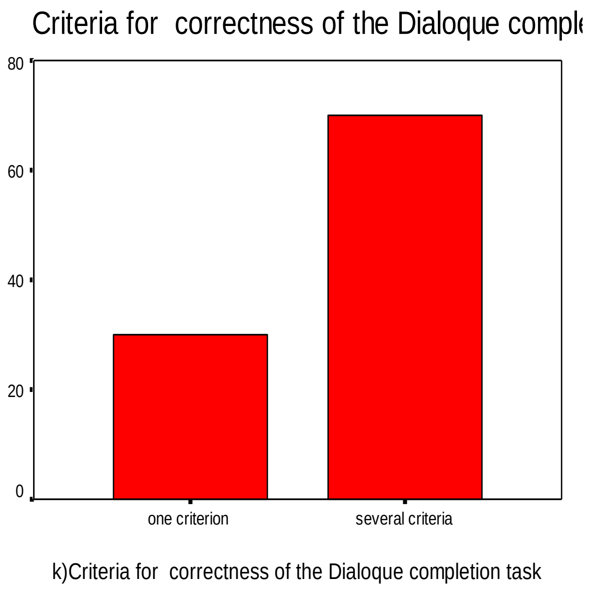
i)Criteria for correctness of the Gap-Filling task



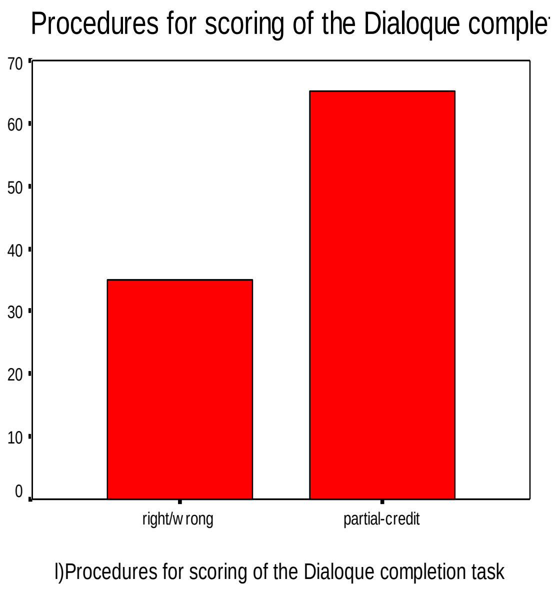
j)Procedures for scoring of the Gap-Filling task

Graph 15. Gap-Filling Task Criteria for Correctness **Graph 16.** GFT Procedures for Scoring

Analysis of the Dialogue Completion Task: Referencing Chart (k), it emerges that 70% of high school teachers employ the Dialogue Completion task to holistically evaluate multiple aspects of grammatical knowledge, encompassing both form and meaning. Conversely, 30% narrow their assessment to a single domain, either form or meaning. Chart (l) further demarcates the preferred scoring strategies: notably, 65% of educators gravitate towards the partial credit scoring approach, while the remaining 35% adhere to the unambiguous 'right/wrong' metric.



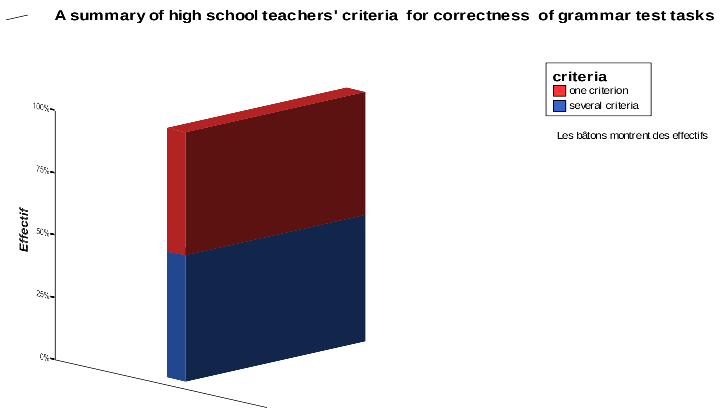
Graph 17. DCT Criteria for Correctness



Graph 18. DCT Procedures for Scoring

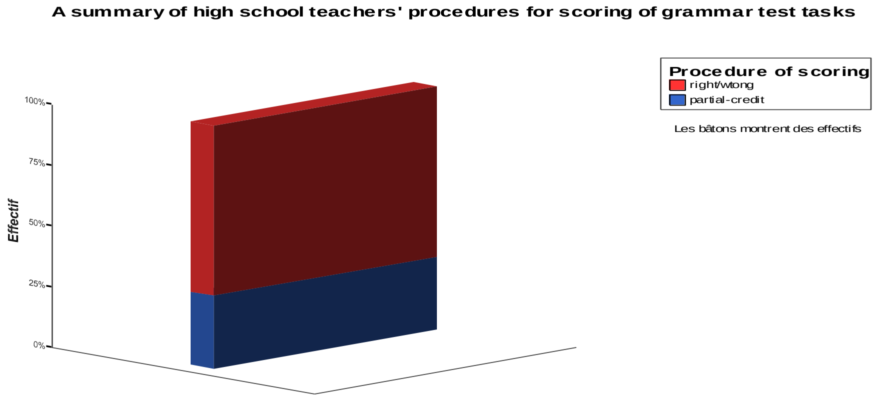
6.3 An Overview of High School Teachers' Scoring Methods

Grammatical Assessment Criteria in High Schools: Within high school contexts, educators frequently base their assessments on distinct grammatical dimensions: either form or meaning, or an integration of both. As delineated in the subsequent data, 49.4% of teachers align with a singular criterion for correctness, while a marginally larger proportion, 50.6%, encompass multiple criteria in their evaluations.



Graph 19. High School Teachers' Criteria for Correctness across their Grammar Test Tasks

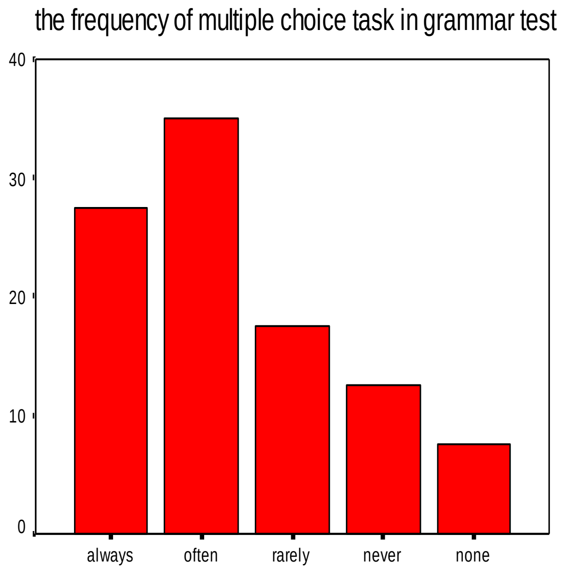
Scoring Procedures in Grammar Tasks Analysis: As depicted in the subsequent chart, a significant 70% of high school teachers adhere to the binary 'right/wrong' scoring approach for evaluating responses in grammar tasks. In contrast, a minority of 30% employ the more nuanced partial-credit scoring method.



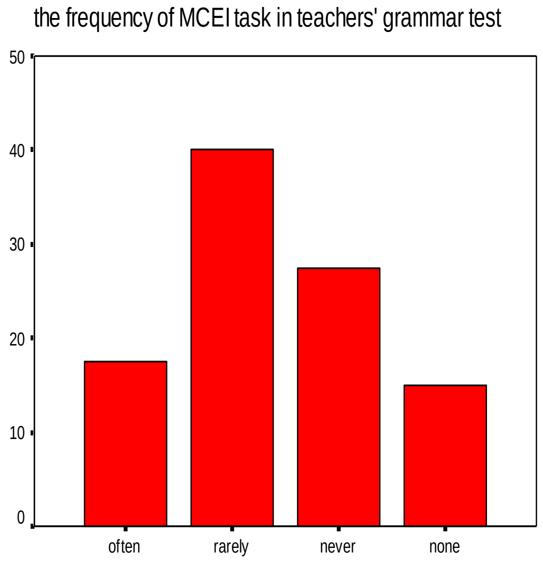
Graph 20. High School Teachers' Procedures for Scoring of their Grammar Test Tasks

6.4 The Occurrences of Tasks in the Same Grammar Test .

Analysis from Charts (a, b): Drawing upon the insights provided by Charts (a) and (b), it is discernible that the multiple-choice format frequently manifests as either "always" or "often" in assessments. In juxtaposition, the multiple-choice error identification is less predominant, tending to appear "rarely" or, in some instances, might "never" feature in tests.



a) the frequency of multiple choice task in grammar test

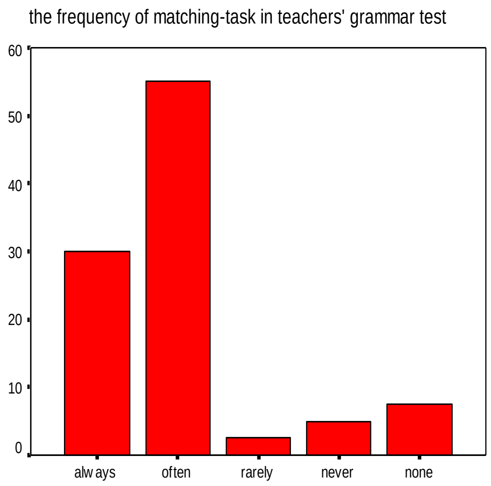


b) The frequency of MCEI task in teachers' grammar test

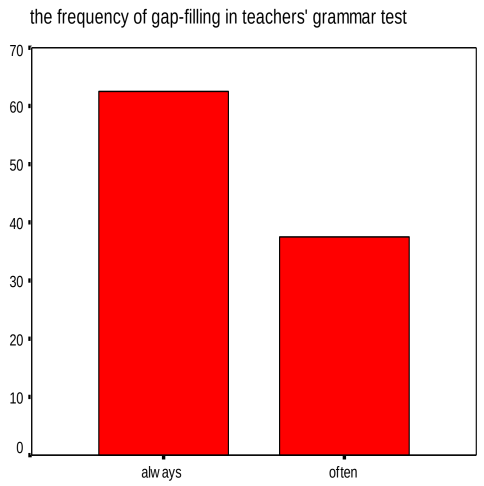
Graph 21. Frequency of MCT in the Teachers' Test

Graph 22. Frequency of MCEIT in T's Test

Evaluation of Test Formats: Analogously, both the Matching-task and the Gap-filling task recurrently manifest in grammar assessments, typically categorized as being used "always" or "often".



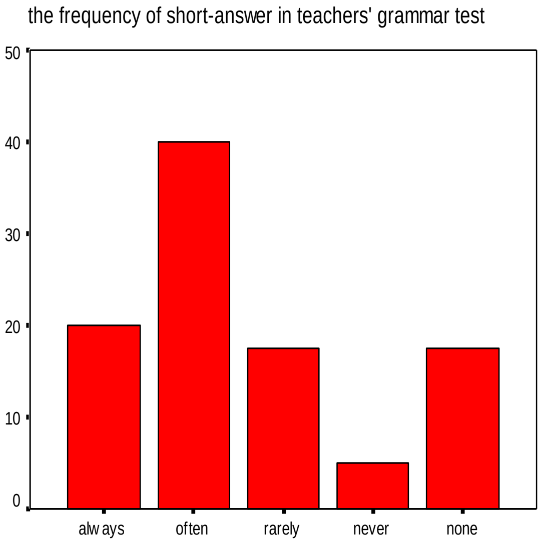
c) the frequency of matching-task in teachers' grammar test



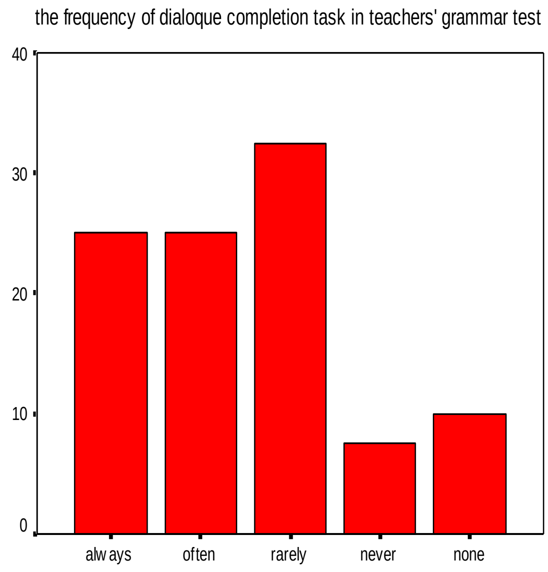
d) the frequency of gap-filling in teachers' grammar test

Graph 23. Frequency of MT in the Teachers' Test **Graph 24.** Frequency of GFT in T's Test

According to the juxtaposed charts, 19.5% of educators consistently ("always") employ the Short Answer task, while 24.4% exhibit a similar preference for the Dialogue Completion task. Additionally, the Short Answer task is "often" favored by 39% of teachers, paralleled by 24% for the Dialogue Completion task. However, it's noteworthy that a majority, surpassing 50%, might infrequently ("rarely") or abstain ("never") from utilizing these specific tasks.



e) the frequency of short-answer in teachers' grammar test

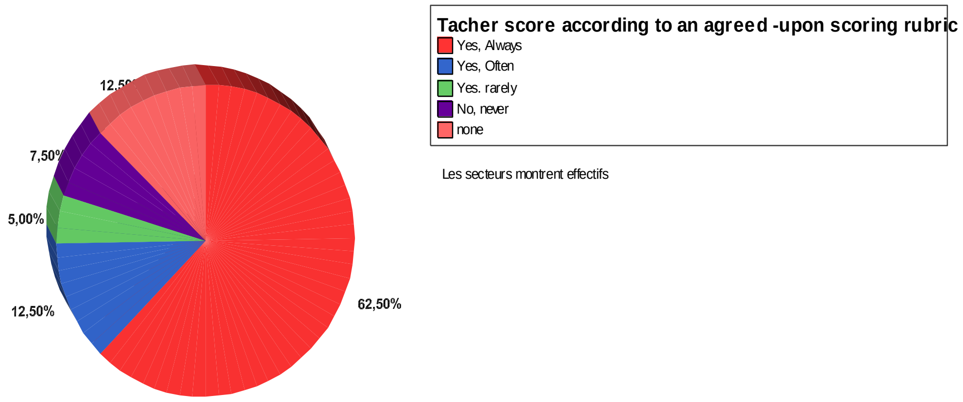


f) the frequency of dialoque completion task in teachers' grammar tes

Graph 21. Frequency of SAT in the Teachers' Test **Graph 22.** Frequency of DCT in T's Test

6.5 Teachers' Identification of Test Items' Scores on the Exam Sheet.

Analysis of Scoring Rubrics from the Graph: The subsequent graph delineates the proportion of educators who standardize their test-item scores by allocating a fixed score to each item, aiming to objectify their scoring methodology. A notable 62% of these teachers consistently ("always") adhere to a predefined scoring rubric within the examination sheet.



Graph 23: Distribution of Teachers' Specification of Test Item Scores on Examination Sheets

Variations are evident in teachers' selections of grammar tasks. A significant proportion amalgamate multiple response tasks, such as multiple choice, with limited production tasks, typified by the Gap-Filling task. In terms of correctness criteria, a close juxtaposition exists: 49.4% of educators gravitate towards a singular correctness criterion to gauge students' grammatical prowess, contrasted by 50.6% who employ multiple correctness criteria. Pertaining to scoring approaches, a substantial 70% adopt the binary right/wrong method, while the remaining 30% embrace the partial credit technique. Predominantly, the multiple choice, gap-filling, and matching tasks emerge as the most recurrent, frequently woven into a majority of teachers' assessment tools. Conclusively, a noteworthy 62.5% of educators offer explicit score indicators on examination sheets, aimed at apprising students of their accurate responses.

7. Discussion

The objective of this research is to scrutinize the reliability of scoring methods in grammar test tasks. The investigation seeks to discern the degree to which high school educators employ multiple response and limited production tasks in their evaluations of students, as addressed by the subsequent set of inquiries.

7.1 The Prevalence of Task Types in High School Grammar Assessment.

The results of this study accentuate the hybrid methodology adopted by high school educators. They seamlessly blend both selected response tasks and limited production tasks to gauge students' grammatical prowess. This blend not only signifies a comprehensive assessment technique but also underscores educators' intricate comprehension of the complexities inherent in linguistic acquisition.

The gap-filling task distinctly emerges as a preeminent tool in grammar assessments. Its pervasive use across many evaluations underscores its utility in probing students’ aptitude to contextualize grammatical norms, thereby mirroring authentic linguistic usage. The widespread utilization of the gap-filling task elucidates its robustness as a tool for gauging grammatical understanding while emphasizing the pragmatic application of linguistic competencies. Moreover, the research unveils a considerable inclination towards other task variants such as the matching task, dialogue completion task, and multiple-choice tasks. This eclectic selection attests to educators’ commitment to probe diverse facets of the learners’ linguistic proficiency, asserting that language proficiency is not a monolithic construct; rather, it encompasses a diverse array of competencies. The salience of the gap-filling task in pedagogical evaluations deserves special emphasis. It insinuates educators’ cognizance of its academic pertinence in efficaciously evaluating students' command over grammar conventions and their aptitude in practical applications. This task’s prominence reiterates its quintessential role in harmonizing pedagogical practices with educational objectives.

An in-depth evaluation of grammar task predilections offers a window into the coherent propensities of high school educators. These trends indicate a deliberate approach to task choice that is grounded in pedagogical principles, which both aligns with instructional objectives and reinforces the robustness of their assessment techniques. The diverse array of assessment tools employed is emblematic of educators' pedagogical acumen, suggesting a calibrated approach to scoring, and thus cementing the reliability of their evaluative paradigms.

To sum up, the insights garnered highlight educators' steadfast dedication to robust linguistic assessment strategies, cognizant of the nuanced labyrinth of language education. The eclectic approach in task selection embodies a deliberate and nuanced assessment paradigm, fortified by educators' pedagogical prowess and unwavering allegiance to comprehensive evaluative techniques.

7.2 Analyzing Scoring Trends in Multiple Response Tasks

The empirical data sheds light on discernible tendencies in the evaluation techniques chosen by high school educators for **multiple response tasks**. In particular, the scrutiny of the multiple-choice task offers salient insights into educators’ predilections regarding criteria of accuracy and the associated scoring procedures.

Amongst the educators incorporating the multiple-choice task into their assessment repertoire, nearly 48% prioritize a singular accuracy criterion, centering on one distinct grammatical domain. Concurrently, an overwhelming majority, approximately 95% of these educators, implement a binary right/wrong scoring paradigm. Such a nexus between the chosen accuracy criterion and the scoring paradigm epitomizes a systematic methodology in gauging students' grammatical proficiencies.

On the contrary, 52% of educators delineate multiple accuracy criteria, integrating both morphological (form) and semantic (meaning) dimensions. Within this demographic, a niche segment, representing 5% of educators, gravitates towards the partial-credit scoring paradigm. It is compelling to note an inconsistency in the congruence between the accuracy criteria and the scoring mechanism amongst this subgroup. Specifically, 47% of educators within this stratification deviate from the prototypical alignment by endorsing a monolithic scoring procedure, notwithstanding their adoption of diverse accuracy criteria.

This examination accentuates the dichotomy in educators' inclinations when scoring multiple choice tasks. Whilst a substantial fraction exemplifies a harmonized alignment between accuracy criteria and scoring paradigms, an alternate fraction showcases potential discrepancies in their evaluative paradigms.

7.3 Analyzing Scoring Trends in Limited-Production Tasks:

Limited-production tasks, as demarcated by Purpura (2014), present unique opportunities and challenges for language assessment. Requiring students to produce limited and specific responses, they inherently gauge nuanced facets of grammatical understanding. The evolving patterns of teacher scoring methods for these tasks furnish significant insights into contemporary classroom assessment practices.

Gap-Filling Task: Alignment and Divergence The Gap-Filling task serves as a quintessential limited-production task, compelling learners to exercise both form and meaning. Yet, the assessment practices surrounding it reveal a dichotomy. A substantive 35% of educators operate within the classical testing paradigm, emphasizing singular criterion assessment coupled with the dichotomous right/wrong scoring. This lends itself to clarity but may miss nuanced understandings. More intriguing is the 37% subgroup that, despite embracing a multifaceted assessment lens, reverts to the dichotomous scoring model. This suggests potential tension between the intricacy inherent in gap-filling tasks and the pragmatic ease of singular scoring. Nevertheless, the combined 63% adhering to singular criterion methods underscores a broader pedagogical commitment to reliability.

Short Answer Task: Predominance of Alignment. In the realm of the Short Answer task, educators showcase remarkable alignment between their chosen criteria for correctness and their scoring methods. With 90% of educators opting for methods that match their correctness criteria, there's an evident consensus on the pivotal role of coherent evaluation. Such robust alignment not only aids in consistency but reaffirms the trustworthiness of the assessment, ensuring that learners' grammatical competencies are gauged accurately.

Dialogue Completion Task: A Trend Towards Multifaceted Assessment. The Dialogue Completion task further illuminates educators' proclivity towards multi-criteria assessment, with 65% embracing both form and meaning considerations and aligning this with the partial-credit approach. This suggests a prevailing understanding of the task's inherent complexity and the importance of nuanced feedback. The scant 5% who deviate from the norm, employing multiple criteria yet defaulting to single scoring, points to occasional discrepancies in assessment methodologies.

In sum, high school educators' scoring choices for limited-production tasks provide a window into the pedagogical landscape of grammar assessment. While clear patterns of coherence and consistency emerge, instances of divergence underscore the dynamic, multifaceted nature of language testing in classroom contexts. As educators continue to navigate the intricate interplay between assessment criteria and scoring methods, these findings serve as a testament to their dedication to aligning instructional goals with evaluation practices.

7.4 On the Objectification of Scoring Protocols and the Reliability of High School Teachers' Grammar Test Tasks Assessment Techniques.

Research findings indicate that 62.5% of educators prioritize objectivity in their assessment techniques, minimizing dependence on expert subjectivity. This approach is largely achieved by allocating predetermined scores to each item, setting clear benchmarks for correct answers. Educators have several reasons for this approach, emphasizing the importance of providing students with clear feedback and highlighting high-value tasks.

Reliability is a foundational principle in language assessment. Consistent and stable assessment outcomes are essential for meaningful interpretations of the skills being assessed. High school educators use various grammar tasks to evaluate students' grammatical skills. The study reveals consistent scoring methods among educators in this context.

Among the tasks educators commonly use, Gap-Filling is most prevalent, followed by Matching, Dialogue Completion, and Multiple-Choice. These tasks assess both the structure and meaning of grammar, measuring students' ability to link form and function effectively.

Many educators aim for objective and consistent scoring methods. Remarkably, about 62% of high school educators use pre-set scoring criteria for their tests. This inclination towards clear assessment criteria ensures a more standardized and unbiased evaluation, as noted by Purpura (2014).

To enhance the reliability of grammar assessments, educators often include a variety of tasks in one test. Data suggests that, on average, four different tasks are incorporated in each test, blending both multiple response and limited production tasks. This diversified approach offers a more comprehensive evaluation of students' grammatical skills, which in turn informs subsequent teaching and assessment strategies.

Conclusion

From the comprehensive data and findings discussed in this study, it is clear that educators make rigorous efforts to bolster the reliability of their assessments. One notable strategy is the integration of diverse task types, including both multiple-response and limited-production tasks, within a single assessment. This combination allows for an in-depth evaluation of students' grammatical skills, ensuring a rigorous and accurate assessment mechanism. Teachers' meticulous attention to both grammatical structure and semantics demonstrates their dedication to crafting holistic evaluations. Depending on the specific nature of the task, educators judiciously focus on one or both of these dimensions, ensuring their assessment criteria align with the inherent characteristics of each task. This intentional alignment enhances the accuracy and breadth of the evaluation process.

A consistent theme evident throughout this study is the emphasis on uniformity in scoring techniques. This commitment is manifested in the deliberate alignment between the standards of correctness and the chosen scoring procedures. By coherently harmonizing these elements, educators aim to diminish subjective biases, championing impartiality in their evaluations.

To conclude, this research comprehensively addresses the initial queries: 1) the predominant grammar test tasks preferred by educators; 2) the scoring methods teachers utilize; and 3) the degree to which objectivity permeates their scoring practices. The results corroborate the preliminary hypothesis, confirming that the grammar test scoring strategies used by Moroccan EFL high school educators indeed possess a high level of reliability. Through a thoughtful integration of diverse tasks, a balanced focus on grammatical structure and semantics, and unwavering adherence to consistent scoring guidelines, educators establish a framework for assessments that produce reliable and insightful results.

Pedagogical Implications

The results of this study emphasize the necessity for educators to incorporate a range of task types within a single assessment framework to augment the reliability of the evaluative process. Integrating a mix of tasks, including both multiple responses and limited production tasks, allows for a holistic appraisal of students' grammatical competence. The advantage of this methodology is its ability to provide a thorough examination of students' varied linguistic competencies, leading to more trustworthy evaluation outcomes.

Furthermore, the study sheds light on a critical aspect concerning the congruence between scoring mechanisms and the predefined criteria for correctness. In situations where teachers employ a single standard for accuracy, a dichotomous scoring approach appears most apt. However, when multiple criteria are in focus, the dichotomous method's suitability wanes. In such cases, the partial credit system emerges as a preferable option, granting educators the flexibility to recognize and assess the subtle variations within the multiple dimensions of grammatical knowledge being evaluated.

In summary, the findings highlight the intricate decisions educators face when formulating and administering grammar assessments. By diversifying assessment tasks and aligning scoring techniques with the intricacies of correctness benchmarks, educators can enhance the robustness and reliability of their evaluation methods. Such measures ensure a richer insight into students' linguistic proficiencies.

Limitations and Recommendations

In educational settings across different contexts. As the linguistic and grammatical demands of various tasks vary, the exclusion of certain task types may result in a less comprehensive view of assessment practices and their associated outcomes.

Another potential limitation pertains to the reliance on self-reported measures from educators, which could be influenced by recall biases or the desire to portray oneself in a particular light. Objective observation of classroom practices or the analysis of actual graded tests might provide a more unvarnished view of assessment methodologies and scoring rubrics in action. Further, while the study offers insights into the scoring practices of Moroccan EFL teachers in Rabat, cultural and contextual variations in teaching and assessment might differ in other regions or countries. The specific challenges, priorities, and resources available to educators in different contexts can significantly influence assessment decisions.

In conclusion, while the present study offers valuable perspectives on prevalent grammar assessment practices, there remains an expansive territory in the realm of language testing yet to be charted. By broadening the scope of future research to encompass lesser-studied task types and by probing deeper into their associated scoring challenges and potentials, we can hope to foster a more holistic and nuanced understanding of grammar assessment practices. Such endeavors will undoubtedly pave the way for more informed, effective, and equitable assessment strategies across varying educational landscapes.

References

1. Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
2. Bachman, L. F., Bachman, P. O. a. L. L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and Developing Useful Language Tests*. Oxford University Press.
3. Brown, H. D. (2018). *Language assessment: Principles and Classroom Practices*. Pearson Education ESL.
4. Brown, H. D. (2009). *Understanding language testing*. Routledge.
5. Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and Classroom Practices*. Pearson Longman.
6. DeKeyser, R. (1995). Learning second language grammar rules. *Studies in Second Language Acquisition, 17*(3), 379–410. https://doi.org/10.1017/s027226310001425x

7. Ellis, R. (2005). Measuring Implicit and Explicit Knowledge of A Second Language: A Psychometric Study. *Studies in Second Language Acquisition*, *27*(02). https://doi.org/10.1017/s0272263105050096
8. Han, Y., & Ellis, R. (1998). Implicit knowledge, explicit knowledge and general language proficiency. *Language Teaching Research*, *2*(1), 1–23. https://doi.org/10.1177/136216889800200102
9. Purpura, J. E. (2004). *Assessing grammar*. https://doi.org/10.1017/cbo9780511733086
10. Rahimpour, M., & Salimi, A. (2010). The impact of explicit instruction on foreign language learners' performance. *Procedia - Social and Behavioral Sciences*. https://doi.org/10.1016/j.sbspro.2010.03.976