Linguistic Forum

Published by
MARS Publishers

ISSN (Online) 2707-5273
Volume 7, Issue 1, 2025
http://doi.org/10.5281/zenodo.16967940
Pages 108-133

Developing and Validating a Metalinguistic Knowledge Test of Grammar for University Undergraduates in Pakistan

Research Article

	Zulaikha Nadeem <zulaikhanadeem@gmail.com>	MPhil Scholar, Department of Applied Linguistics, Government College University, Faisalabad, Pakistan.
Corresponding Author:	Aleem Shakir <almsha@yahoo.com>	Assistant Professor, Department of Applied Linguistics, Government College University, Faisalabad, Pakistan.

Publication Details

Received: July 25, 2025 **Accepted:** August 20, 2025 **Published:** August 30, 2025

Abstract

The exploration of metalinguistic awareness has gained significant importance in the realm of second language acquisition. This research aimed to develop and validate a Metalinguistic Knowledge Test (MKT) of grammar for undergraduates in Pakistan, addressing the gap in such resources for Pakistani undergraduate learners. The theoretical framework of Bialystok and Ryan (1985) supported the test's construct and content. A convenience sampling strategy was employed to administer the MKT to 440 participants from diverse academic and regional backgrounds. Experts assessed the test's face, content, and construct validity. SPSS, a statistical software, facilitated the exploration of descriptive statistics and item analysis, including facility value, discrimination analysis, and distractor analysis. The internal consistency was high, with a Cronbach alpha of *.967*. From an initial pool of 150 items, 87 were retained in the final version. A two-way ANOVA revealed a significant gender effect but no interaction between gender and subject, supporting the MKT's validity and reliability across different subjects. Limitations include the small sample size and a limited number of experts on the panel. This study has broad implications for ESL education and assessments at higher education levels.

Keywords: Metalinguistic Knowledge Test (MKT), grammar, instrument development, validity, ESL, Pakistani undergraduate learners, reliability, distractor analysis, discrimination index, SPSS, Cronbach alpha formula, two-way ANOVA, Levene's test

1. Introduction

Second language acquisition (SLA) theory suggests, that linguistic knowledge can be assessed explicitly (Krashen, 1981). Researchers in SLA have focused on understanding the structure and representation of second language (L2) knowledge. In the recent years, research on metalinguistics has gained significant attention in language acquisition.

Metalinguistics refers to the ability to consciously analyze and reflect on language. Bialystok (2001) defines metalinguistic ability as the capacity to understand and think about language rather than simply using it. The MKT is an assessment tool that evaluates an individual's awareness of language structures, rules, and their ability to reflect on or manipulate language. It assesses key linguistic domains, including phonology, morphology, syntax, semantics, pragmatic, and discourse. Bloor's study (1986) reported that the metalinguistic knowledge of first year undergraduates is very low, particularly in English as a foreign language. Several studies (Alipour, 2014; Wistner, 2014; Tokunaga, 2014; Amrani, 2015; Ayden, 2018; Sanosi, 2022) have adapted and developed grammar focused MKTs to explore their relationship with other linguistic variables among undergraduate learners.

The grammatical component of the MKT is essential in language research, especially for evaluating participants' metalinguistic awareness and their explicit, conscious understanding of form/meaning relationships in the target language. By providing a structured and standardized assessment, the MKT enables researchers to explore the complexities of SLA and bilingualism. This is particularly relevant in understanding how individuals, especially those engaged in learning a second language, navigate the complexities of grammar. The test serves as a lens through which researchers can gain insights into the participants' grammatical competence, shedding light on their ability to comprehend and manipulate the formal structures of the language.

Despite the significance of metalinguistic knowledge in language learning, there exists a substantial research gap in investigating the linguistic challenges faced by Pakistani undergraduate students. English being the second language in Pakistan is compulsory in official and academic settings. However, no comprehensive instrument has been developed to assess the metalinguistic knowledge of this demographic. Consequently, there is a pressing need for a reliable and valid MKT, specifically designed for Pakistani undergraduate learners. The lack of such an instrument impedes the systemic evaluation of their linguistic challenges, thereby restricting the development of the targeted language instructions and academic support. This study addresses this gap developing and validating a reliable Metalinguistic Knowledge Test of grammar for Pakistani undergraduates.

2. Literature Review

Metalinguistic knowledge, also known as metalinguistic awareness, is a multifaceted concept that has been defined by various scholars in the field of linguistics. Roehr (2007) briefly defined it as "The learner's ability to correct, describe, and explain L2 errors." Kurvers et al. (2006) define it as

the conscious reflection on various aspects of language, covering domains such as phonology, morphology, semantics, syntax, and connected discourse. Metalinguistic knowledge, as highlighted by Roehr (2007), is closely associated with explicit knowledge of L2 categories and their interrelations. Explicit metalinguistic knowledge aids learners to identify error sources. By recognizing the impacts of L1 interference, overgeneralization, and the complexities of English grammar, learners can develop more targeted and effective learning strategies.

The MKT of grammar includes a variety of item types aimed at evaluating individuals' grammatical awareness. Steel (1997) developed a metalinguistic knowledge test that incorporates self-assessment of familiarity with grammatical terms, knowledge of English and French grammatical terminology, and the ability to apply these terms in identifying parts of speech. Adapted from Bloor (1986), the test included items like "auxiliary verb," "adjective," "infinitive," "preposition," "past participle," Although Steel's (1997) study included additional tests to form a test battery, the metalinguistic test alone comprised 100 items.

Renou (1998, 2001) used a Grammaticality Judgment Test (GJT) to evaluate L2 learners' oral and written proficiency. Participants were asked to determine whether a sentence was grammatical, identify and correct incorrect forms, and explain the relevant rules. This test included 30 items, comprising nine grammatically correct sentences and 21 ungrammatical ones, with 18 ungrammatical items sourced from Bialystok's (1979) study and the remainder created by Renou (1998). The ungrammatical items included errors related to adjectives, direct/indirect object pronouns, and verbs.

Hu (2002) utilized a unique approach by creating an explanation-based task rather than relying on pre-existing metalinguistic tests. This written test featured 12 items focused on grammatical rules, with ungrammatical sentences underlined for participants to explain the rules necessary for correction. Students were permitted to provide their explanations in Chinese. The test items were designed specifically to evaluate L2 learners' metalinguistic knowledge.

Elder and Manwaring (2004) designed a Chinese MKT based on the formats of Alderson et al. (1997) and Elder et al. (1999). The test, structured to evaluate students' knowledge of the Chinese language, was divided into two sections. The first section included 34 items related to parts of speech. Similarly, Ellis (2005) adapted an MKT from Alderson et al. (1997), incorporating multiple-choice questions with distinct components. The first part presented 17 ungrammatical sentences requiring participants to select the rule explaining each error, while the second part was further divided into two sections.

Roehr (2007) employed the MKT to evaluate participants' abilities to correct, describe, and explain language-related concepts. This test consisted of two sections, each containing 15 items. In the correction and explanation section, learners corrected errors in L2 sentences and provided explanations for their corrections. Additionally, they analysed and justified why certain paraphrases of L2 passages were incorrect, evaluating their ability to apply grammar rules and analytical skills. In the language analysis section, learners identified grammatical roles in highlighted sections of L2 sentences. Tokunaga (2010) investigated metalinguistic knowledge of grammar, requiring participants to assess part of speech, sentence structures, tense, mood, and other grammatical features.

This study included a total of 40 multiple-choice items arranged according to difficulty. The most challenging items involved complex concepts such as the subjunctive and causative forms, followed by structures like the past perfect and passive voice. In contrast, more fundamental elements, such as verbs, nouns, and subjects, were among the easiest items.

Bowles (2011) validated Ellis's (2005) adaption of Alderson et. al., (1997) test battery. The test comprising of two sections was structured as an untimed MCQs assessment where section 'A' focused on identification of reasons behind an error, and section 'B' presented a Spanish text for grammatical item identification.

Lan's (2011) modelled the Metalinguistic Awareness Test (MAT) on Andrew (1999). This test included four sections addressing production, recognition, error correction and explanation. The metalanguage production part comprised of 12 sentences, while the recognition section had 18 sentences, and both error correction and explanation included 15 sentences each. The production task was focused on the ability of the learners to provide grammatical forms for the underlined words in the sentences.

Tokunaga (2014) implemented a comprehensive metalinguistic knowledge test comprising 36 items across four sections: parts of speech, parts of sentences, tenses, and other grammatical aspects. Designed for L2 learners, the test presented sentences in English with underlined features. An additional section was later included for higher-proficiency students, focusing on error identification with options provided in L1 (Japanese). The second section, based on Elder's (2009) study, assessed metalinguistic knowledge through a purely receptive approach, using vocabulary from the General Service List and items from the TOEIC Bridge official workbook and preparation guide (Educational Testing Service, 2002, 2008).

Wistner (2014) employed two instruments: The Receptive MKT and the Productive MKT. The receptive test, initially comprising 18 items, was expanded to 22 items in multiple-choice format. Tasks included selecting Japanese terms for linguistic rules, identifying metalinguistic terms related to English structures, and matching English examples to given terms. Conversely, the productive test assessed participants' knowledge of technical terms and grammatical rules, including the identification of ungrammatical parts of sentences and error correction in English, comprising 17 items adapted from Ellis (2005). This comprehensive approach evaluated participants' comprehension and application of metalinguistic knowledge in both receptive and productive contexts.

Alipour (2014) implemented MKT to assess the explicit knowledge. The test consisted of 30 sentences with grammatical errors that required participants to correct, describe, and explain. A few words were underlined in each sentence and students were instructed to tick the incorrect answer. Each sentence contained underlined words that participants needed to evaluate, with examples such as 'One group was satisfied with the explanation whereas the other group wanted to explore the subject farther.'

Sanosi (2022), implemented MKT with 55 items distributed across two categories. The first part included 30 matching items connecting grammatical terms to their Arabic translations, while the second part featured four sub-parts, each containing five items that matched grammatical terms with examples related to parts of speech, sentence structures, tenses, and sentence cases.

Zhang (2015a, 2015b) utilized an instrument developed by Ellis (2005), incorporating a pool of test items from various assessments, including the elicited imitation test (EIT), timed grammaticality judgement test (TGJT), untimed grammaticality judgment test (UGJT) for ungrammatical items, and the metalinguistic knowledge test. The MKT had two sub divisions. The first part had multiple choice questions related to identification of rules explaining errors and the second part of the test comprised of a text passage from which the examples grammatical features of the text were to be identified.

2.1 Theoretical Frameworks

The theoretical foundation for this study on developing and validating a metalinguistic knowledge test of grammar is Bialystok and Ryan's (1985a) 'Analysis and Control' model. This model emphasizes the cognitive aspects of metalinguistic knowledge, particularly the influence of analysed processes and controlled knowledge in language processing. It is crucial in the field of language proficiency and cognitive development, with applications in both linguistic and metalinguistic knowledge (Roehr, 2018).

The key contributions of this model are in language processing, metalinguistic awareness, and explicit knowledge of an individual's language structure. The model consists of two major components: analysed knowledge, defined as conscious knowledge, and control process, defined in terms of three functions: the selection of informational items, coordination of the items, and the extent of automation in selection and coordination, and inhibition of irrelevant tasks (Bialystok, 1994a, 2001).

The development of analysed knowledge and control is proposed to influence the ability to meet task demands satisfactorily. Bialystok and Ryan (1985c) also claimed that analysed knowledge precedes control in terms of order. However, it can be possible that a person is more 'fluent' (controlled) as compared to 'accuracy' (analysed knowledge) depending upon certain factors like learning approaches (Renou, 2001). This model has been applied in numerous studies in their research of metalinguistic knowledge, particularly in grammar judgment tests (Alderson & Steel, 1994; Ellis, 1987; Gombert, 1992; Tarone, 1987; Ricciardelli, 1993). This model emphasizes the importance of considering how individuals process linguistic information and highlights the differences in metalinguistic abilities among bilingual individuals.

2.2 Population

Research on metalinguistics, particularly regarding explicit knowledge and second language acquisition (SLA), has involved a diverse range of populations. Key studies have included undergraduate students learning Chinese (Elder and Manwaring, 2004), advanced French L2 learners (Renou, 2001), non-English majors at a private university (Tokunaga, 2010), and English majors in Beijing (Zhang, 2015b).

Additionally, the studies have focused on different language proficiency levels. Amrani (2015) examined first-year L.M.D Master's students in Literature and Linguistics, while Alipour (2014) targeted EFL learners in their initial university years. Tokunaga (2014) and Wistner (2014) explored metalinguistic knowledge among low to intermediate proficiency EFL learners in Japan.

Several studies also addressed adult education, such as Ayden's (2018) focus on intermediate Turkish EFL learners, and Correa's (2011) investigation of Spanish learners from various academic levels at a large U.S. university. Roehr (2007) analyzed L1 English speakers at a British university across different year levels, while Lan (2011) studied in-service primary school teachers in Hong Kong.

Furthermore, Alderson et al. (1997) explored the relationship between L2 proficiency, L1 and L2 metalinguistic knowledge, and language-analytic ability among L1 English learners of L2 French. They operationalized grammatical sensitivity using the MLAT's words-in-sentences subtest. Elder et al. (1999) employed similar metalinguistic tests and measures of language-analytic abilities on advanced learners in Australia.

2.3 Sample Size and Sampling Strategy

The sample sizes in the reviewed studies vary significantly. Alipour (2014), Bowles (2011), and other studies had smaller samples, consisting of 38, 28, and 30 participants, respectively. In contrast, Amrani (2015), Ayden (2018), Correa (2011), Hu (2002), Renou (2001), Roehr (2007), Steel (1997), Wistner (2014), and Zhang (2015b) included larger samples of 60, 38, 171, 64, 64, 60, 128, 240, and 49 participants, respectively. Elder and Manwaring (2004), Tokunaga (2010), and Zhang (2015a) reported sample sizes of 91, 195, and 100, respectively. Notably, Tokunaga's (2014) research featured a large sample of 1,180 participants with intermediate proficiency. Additionally, Ellis (2005) conducted a study with 91 students from a New Zealand university, providing insights into the metalinguistic knowledge of this moderate-sized group.

Most studies did not specify the age range of participants. However, Amrani (2015) divided participants into Arabic and English language groups with age ranges of 22 to 35 and 22 to 30, respectively. Correa (2011) reported a mean age of 21.28 years for participants, while Roehr (2007) found a mean age of 20 years. Lan (2011) conducted a similar study involving 20 participants. Sampling strategies varied among studies. Convenience sampling was used by Ayden (2018) and Elder and Manwaring (2004), while random selection was employed by Hu (2002) and cluster sampling by Sanosi (2022). In contrast, volunteer-based recruitment was the common strategy for Correa (2011) and Wistner (2014), though some studies did not specify their sampling strategies.

2.4 Validity

Validity refers to measures relevant to the applicability and precision of the purpose of the employed measuring device (Taherdoost, 2016). Several factors go into study validity in order to guarantee precise measurement. While the validity of various Metalinguistic Knowledge Tests of grammar is not explicitly addressed, studies by Ellis (2005) and Wistner (2014) focus on construct validity. Content validity is established through analyses of textbooks and syllabi (Elder and Manwaring, 2004; Sanosi, 2022), while theoretical frameworks support construct validity (Renou, 2001; Zhang, 2015a).

2.5 Reliability

Reliability is a crucial measure of consistency in metalinguistic knowledge tests, typically evaluated using Cronbach's alpha. Ellis (2005) reported a Cronbach's alpha of *.9* for the Metalinguistic

Knowledge Test, suggesting high internal consistency. Steel (1997) presented various reliability coefficients for components like MLAT, Grammar, and French Reading. Roehr (2007) provided coefficients of *.640* for correction, *.818* for description/explanation, and *.624* for language analysis. Wistner (2014) reported multiple reliability measures, including Rasch person reliability of *.67* for the receptive test and Cronbach's alpha values of *.82* and *.86* for different scales in the productive test. Elder and Manwaring (2004) indicated reliability for subcomponents, such as grammatical terms (*.80*), error correction (*.90*), and rule explanation (*.77*). Renou (2001) found a Cronbach's alpha of *.87* for the written judgment test and *.77* for the oral judgment test. Tokunaga (2010) reported person reliability of *.89* and item reliability of *.97* through Rasch analyses, while Bowles (2011) found high reliability at *.93*. In contrast, Sanosi (2022) reported a lower reliability of *.69*.

2.6 Research Questions

1. To what extent do the MCQ items appear to measure the intended language skills from the perspective of both test-takers and language experts?
2. To what extent do the MCQs align with the theoretical constructs of language proficiency?
3. How comprehensively do the MCQs cover the language content domains they are supposed to assess?
4. What is the difficulty index of the test items in the study?
5. How does the discrimination index vary across different sections of the test?
6. What is the effectiveness of each distractor in distinguishing between high-scoring and low-scoring groups?
7. What is the internal consistency of the test items, and how does it reflect the overall reliability of the test?
8. Are the MCQs free from bias related to test-takers' gender and subject of study?

3. Methodology

3.1 Pre-Administration

3.1.1 Instrument Development

The Metalinguistic Knowledge Test (MKT) for grammar was developed based on the model by Bialystok and Ryan (1985a, 1985b). Items were adapted from well-known tests such as HAT, GAT, and NTS, as well as the British Council's B1, B2, and C1 grammar levels, ensuring they matched the learners' proficiency. The test was designed to comprehensively cover grammatical concepts, consisting of 150 multiple-choice questions divided into three sections: Identification, Correction, and Error Explanation, each containing 50 items. These sections tested 25 different grammatical features, with two items per feature evenly distributed across the sections. After the test items were finalized, each question had three distractors and one correct option. These distractors were generated using AI technology (ChatGPT 3.5) and were based on prior studies by Wistner (2014), Tsang Wai Lan (2011), and Amrani (2015) enhancing the quality and relevance of the distractors.

3.1.2 Improvements by Experts

3.1.2.1 Ambiguity in the Test

During the test development phase, five doctoral researchers with backgrounds in Applied Linguistics and instrument development, investigated the test items in great detail. To improve clarity, unclear items were carefully identified and removed. Three rounds of revisions were conducted to ensure the balanced representation of grammatical elements in all three sections of the test.

3.1.2.2 Item Difficulty

A five-point Likert scale, adapted from Vagias (2006), was employed to assess the difficulty level of test items for undergraduate learners. The scale was divided into three sections corresponding to the test's structure, with each section containing fifty items. Ph.D. scholars teaching undergraduates evaluated the difficulty level of the items by selecting responses based on their perception. The anchors were as follows: 1= Very Easy, 2= Easy, 3= Moderate, 4= Difficult, and 5= Very Difficult. Items rated as very easy or very difficult were eliminated from the test. Specifically, item 30 from Section 1 (Identification), items 51 and 55 from Section 2 (Correction), and item 95 from Section 3 (Explanation) were deemed very easy and were removed. Similarly, items rated as very difficult, including items 11, 14, 15, 16, 31, 33, 36, and 37 from Section 1, as well as items 44 and 46 from Section 2, were also eliminated. The removed items covered various grammatical features, such as Rhetorical Questions (Item 11), Clauses (Item 14), Negatives (Items 15 and 37), Prepositions (Item 16), Tense (Item 30), Subject-Object Complement (Items 31 and 36), Case (Item 44), Modifiers (Item 46), Verbs (Item 51), and Adverbs (Item 55). Following the revision, 136 items remained in the test.

3.1.2.3 Key Development

The test key was developed by the researcher, following a thorough review by PhD scholars in Applied Linguistics Department of Government College University Faisalabad. Correct responses were determined based on the grammatical rules outlined in High School English Grammar & Composition by Wren and Martin (2015). The key was documented in an Excel sheet, with correct answers coded as "A," "B," "C," or "D." Discrepancies in scholars' judgments were resolved by considering the majority consensus among the evaluators.

3.2 Face and Content Validity

In developing the MKT of grammar, both content and face validity were evaluated. Content validity, essential for confirming the effectiveness of an instrument in measuring the intended construct (Anastasi, 1988), was ensured by having a panel of PhD researchers meticulously review the test content for appropriateness and relevance. Face validity, as noted by Turner (1979) emphasized that face validity reflects whether a test appears valid to the participant, underscoring the significance of the test's outward presentation. To uphold face validity, relevant grammatical items were selected, and the test pattern was designed to align with existing grammar assessments for undergraduate learners.

3.3 Initial Piloting

Item trialling was conducted to evaluate and refine the quality and clarity of test items before large-scale administration. The main goal of the pilot study was to determine necessary modifications in the questions and procedures that could hinder effective data collection, as emphasized by Gudmundsdottir and Brock Utne (2010) and Kim (2010). Following the development of the instrument, the pilot study facilitated the refinement of the test and addressed practical issues, thereby enhancing the overall quality of the instrument.

3.4 Sampling for Item Trialling

The piloting strategy the test was guided by Hertzog's (2008) recommendations, emphasizing feasibility, instrumentation, and intervention goals instead of fixed sampling factors. The initial pilot included 10 participants from both arts and science faculties, with suggestions indicating that 10-15 participants per group would suffice for feasibility and 25-40 would enhance precision.

Ultimately, the study involved 41 participants from different departments and universities across Pakistan. This diverse representation strengthened the test's validity and generalizability, making it a credible tool for assessing metalinguistic knowledge among Pakistani undergraduates.

3.5 Test Administration

3.5.1 Settings

The MKT of Grammar was conducted online using Google Forms, with *Quilgo* software ensuring transparency through features like a timer, AI proctoring, and camera monitoring for audio and facial movements. Initially, the test was timed for three hours, but the limit was later removed, allowing students to complete it without time restrictions. Research on MKT, particularly for grammar, suggests that untimed tests better reflect explicit knowledge. Timed Grammar Judgement Tests (GJTs) tend to draw on unconscious decision-making, while untimed tests allow more deliberate reflection of explicit, consciously accessible knowledge (Bialystok, 1979; Elder, 2009). Some students were disqualified by *Quilgo's AI* due to behaviours like low confidence, eye movements, or attempts to use unauthorized resources, such as *ChatGPT* or web searches.

3.5.2 Sampling

The MKT of grammar was administered on undergraduate students across Pakistan, with 440 participants selected through convenience sampling. This method was used because stratified random sampling by gender and semester was not feasible in an online setting. The sample included 188 male and 252 female students. Participants were drawn from 64 universities across Pakistan.

3.5.3 Population

The study recruited 440 participants from all provinces of Pakistan, representing various semesters and academic fields, which were classified into four faculties including Health and Life Sciences,

Engineering and Technology, Humanities and Social Sciences and Agricultural and Veterinary Sciences.

3.5.4 Incentives

Andrew et al. (2014) explored various recruitment incentives for online research, including cash payments, gift cards, lottery draws, and tangible goods. However, Fan (2010) argued that cash incentives were impractical online, favouring digital alternatives. In this study on the MKT of grammar, participants received e-certificates, while high achievers and those who promoted the test were rewarded with goodie bags, nationwide food delivery and recognition on social media platforms to boost their professional profiles.

3.5.5 Ethical Considerations

The study received formal approval from the Department of Applied Linguistics, and participants provided informed consent. Clear instructions were communicated prior to testing, fostering transparency and trust. Explicit permission was obtained to record videos and access participants' screens. Data collection, scoring, and analysis were conducted with privacy protections in place.

3.6 Data Entry and Data Cleaning

3.6.1 Code Assignment

3.6.1.1 Student IDs.

To ease the task of data analysis, the data was converted into numeric to be entered into Excel sheets and SPSS software. Each student was assigned a student ID code in the form of numbers i.e., 1, 2, to 440.

3.6.1.2 Gender Coding

The data were collected from participants of all genders in this research. To distinguish the participants of the test, coding technique was implemented where "0" represented "Others", the females were allocated code "1" whereas males were allocated as "2".

3.6.1.3 University Code

A total of 61 institutes from Punjab, Sindh, KPK, Baluchistan, Gilgit Baltistan and Azad Jammu Kashmir took part in the study. The universities were alphabetically sorted and provided an abbreviation along with a code e.g., Abasyn University Peshawar (AUP 1), Abdul Wali Khan University Mardan (AWKU 2), Women University Mardan (WUM 3), Allama Iqbal Medical College (AIMC 4), and Air University Islamabad (AIR 5) etc.

3.6.1.4 Department Code

Students from diverse faculties participated, enhancing the test's reliability and generalizability. Data was categorized into four faculties: Health and Life Sciences (1), Engineering and Technology

(2), Humanities and Social Sciences (3) and Agricultural and Veterinary Sciences (4). A total of 62 departments were coded, such as "Allied Health Professionals (AHP 1), Animal Husbandry (AH 2), Applied Linguistics (AL 3), and Applied Psychology (AP 4), etc.

3.6.2 Data Collection and Entry

Student responses were initially recorded in alphabetical format (A for option 1, B for option 2, etc.) and compiled in an Excel spreadsheet. These were then converted into binary format (1 for correct, 0 for incorrect) using a CSV file. This numeric format enabled the calculation scores, facility value and discrimination value. For distractor analysis responses from high- and low-scoring groups were reorganized alphabetically, empty rows were removed, and questions were relabelled as Q1, Q2, etc.

For further analysis, the data was divided into two Excel sheets, with one transposed to facilitate response comparison between high- and low achieving students.

3.7 Data Analysis

3.7.1. Descriptive Analysis

The data responses from undergraduate students across Pakistan, covering diverse universities, departments, semesters and gender representation were analysed to ensure sample diversity.

3.7.2 Validity

Ghaui and Gronhaug(2005) define validity as the extent to which the collected data accurately represents the intended area of investigation. Essentially, validity ensures that a measure assesses what it is intended to assess. (Field, 2005). The four major types include: Criterion, Face, Content, and Construct validity.

3.7.2.1 Face Validity

Face validity was evaluated using a ten-item survey questionnaire, developed according to the guidelines of Desai and Patel (2020). The questionnaire collected yes or no responses from test takers, and professionals assessing whether the measurement tool appeared logical, clear, and relevant based on subjective judgments by non-experts and field experts (Oluwatayo, 2012).

3.7.2.2 Construct Validity

To assess construct validity, the alignment of the test with the theoretical construct was evaluated. Since convergent and discriminant validity are typically used for quantitative analysis, experts' opinions and judgments were used in this study to qualitatively analyze the instrument's construct validity (Cronbach & Meehl, 1955).

3.7.2.3 Content Validity

Content validity was examined through expert evaluation, ensuring the test items were representative and relevant to the construct being measured. The format, wording, and presentation were reviewed in line with Thorndike and Thorndike-Christ's (2014) definition of content validity. The assessment ensured that the instrument effectively captured the intended construct, addressing potential gaps or excess elements (Messick, 1981).

3.7.3 Item Analysis

3.7.3.1 Difficulty Analysis

The Difficulty Analysis also known as facility value (FV) assessed the challenge level of test items to enhance the assessment's credibility in distinguishing between easy and hard test items. The difficulty index was calculated using the criteria that values between *.3* and *.7* are acceptable, with *.5* as the ideal level (Osterlind, 2006). Specifically, a value close to *.5* indicates a well-designed item, while values below *.3* suggest excessive difficulty and values above *.7* indicate excessive ease. The difficulty index was computed in Excel using the formula:

$$(FV = \frac{Correct\ responses}{Total\ responses})$$

3.7.3.2 Discrimination Analysis

The discrimination index (DI) was employed to differentiate high-performing students from low-performing ones. After eliminating 27 items that did not meet the facility value criteria, 109 test items were retained. The DI was calculated using the formula:

$$Discrimination\ Index\ = \frac{Correct\ Responses\ in\ High\ Group - Correct\ Responses\ in\ Low\ Group}{Number\ of\ Participants\ in\ each\ group}$$

The index ranges from -1.0 to +1.0, with values near *.0* indicating poor discrimination and values close to *1.0* indicating strong discrimination, where a score of *.3* or greater is deemed highly discriminating (Osterlind, 2006). In this study, 440 participants were analyzed, with 119 in each group based on the 27% criterion (Downing, 2006). The analysis was performed in Excel, transposing the data to arrange test items as rows and participants as columns.

3.7.3.3 Distractor Analysis

Distractor analysis examines the effectiveness of incorrect options in MCQs by analyzing the selection frequency, especially among students who lack knowledge of the correct answer. Well-designed distractors improve the test's validity and reliability (Haladyna & Downing, 1989).

According to Tarrant, Ware, and Mohammed (2009), a distractor is considered effective if selected by a significant proportion of students, especially those with lower overall test scores (p. 375). Responses from higher and lower groups were analyzed using the formula: DA

$= \left(\frac{(Higer\ Group\ Response + Lower\ Group\ Response)}{Total\ Number\ of\ Participants}\right)$ X 100. Distractors were deemed functional if selected by at least 5% of respondents; those with less than 5% were considered non-functional.

3.7.4 Reliability Analysis

Reliability refers to the consistency of measurements, indicating the degree of agreement among items within an instrument. It reflects how consistently an assessment yields results when administered to the same participants (Sullivan, 2011). Reliability can be evaluated using methods, such as test-retest reliability and internal consistency, which measures how well items assess the same construct (Tavakol & Dennick, 2011). Cronbach's alpha was used to assess overall and inter-item correlation, with values above *.3* considered ideal. The internal consistency was categorized as follows: $\alpha \geq 0.9$ *(Excellent),* $0.7 \leq \alpha < 0.9$ *(Good),* $0.6 \leq \alpha < 0.7$ *(Acceptable),* and $\alpha < 0.6$ *(Poor).*

3.7.5 Bias Analysis

Bias analysis was conducted to determine fairness of test across genders and subject areas through a two-way ANOVA. Before conducting a two-way ANOVA, assumptions for the analysis were checked as elaborated below.

3.7.5.1 Sample Size

To assess the suitability of our data for a two-way ANOVA, the distribution of participants across gender and subject groups was examined. The sample comprised 252 females and 188 males, distributed across the following subject areas: 138 participants in Health and Life Sciences, 64 in Engineering and Technology, 205 in Humanities and Social Sciences, and 33 in Agricultural and Veterinary Sciences. Notably, the Agricultural and Veterinary Sciences group, although the smallest, met the minimum recommended threshold of 20 observations (Cohen, 1988), thus ensuring adequate power and robustness for the ANOVA analysis.

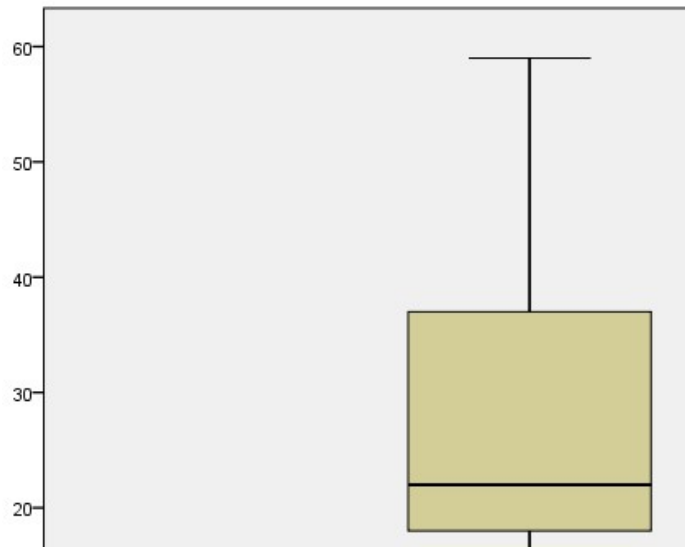
3.7.5.2 Independence of Observations

Data collection encompassed 62 distinct departments, thereby affirming the independence of observations.

3.7.5.3 Absence of Significant Outliers in Dependent Variable

A comprehensive outlier analysis of the dependent variable (MKT) revealed no significant outliers, indicating that the data were suitable for further statistical analysis.

Figure 3.1: Absence of outliers in the dependent variable



3.7.5.4 Normal Distribution of Dependent Variable for All Levels of Each of Independent Variable

The normality of the dependent variable was assessed through skewness and kurtosis calculations, which were found to be within the acceptable range of ±2 (Gravetter & Wallnau, 2014).

3.7.5.5 Outliers in Each Group

Prior to conducting the two-way ANOVA, the data were screened for outliers using boxplots for each combination of the independent variables (gender and subject) and the dependent variable (MKT scores). The analysis revealed no outliers across any combinations, indicating that the data are suitable for further analysis without requiring additional transformations or exclusions.

Figure 3.2: Outlier Analysis of Dependent Variable with Respect to Gender

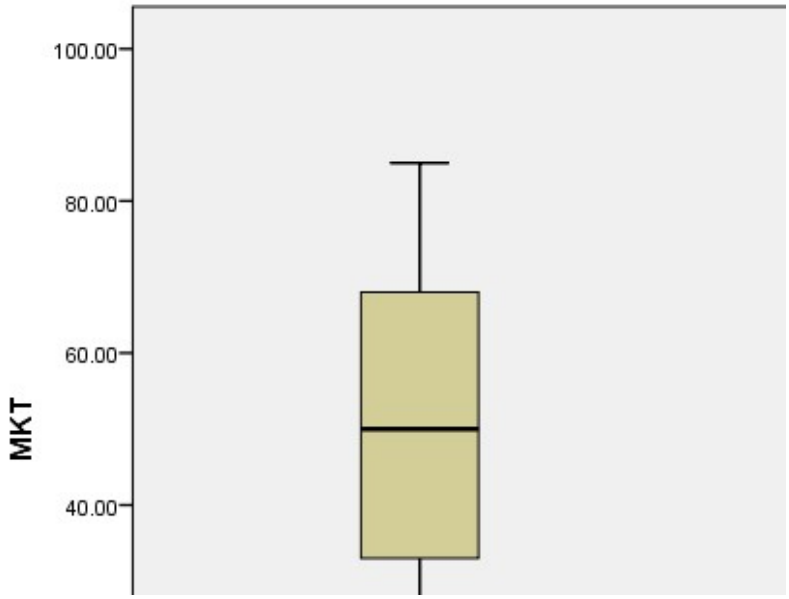
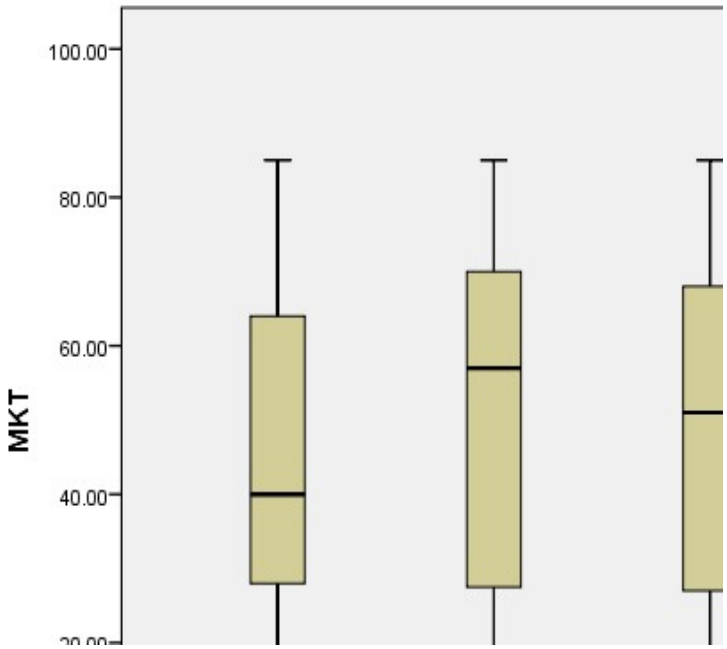


Figure 3.3: Outlier Analysis of Dependent Variable with Respect to Subject



3.7.5.6 Homogeneity of Variance of Dependent Variable for Each Group

The Levene's Test of Equality of Error Variances was performed to assess the assumption of homogeneity of variances for the dependent variable, MKT, across groups defined by the interaction of gender and subject.

Table 3.1 Levene's Test of Equality of Error Variances[a]

F	df1	df2	Sig.
1.900	7	432	.068

a. Design: Intercept + Gender + Subject + Gender * Subject

The test produced an F-value of *1.900* with 7 and 432 degrees of freedom, yielding a significance value of *.068*. This non-significant p-value, exceeding the conventional threshold of *.05*, indicates that we fail to reject the null hypothesis, confirming that error variances are equal across groups. Consequently, the assumption of homogeneity of variances is satisfied, suggesting consistent variances that support the robustness of subsequent analyses, such as ANOVA or regression. All necessary assumption tests for conducting a two-way ANOVA for bias analysis were thoroughly executed, confirming that the data were suitable for analysis. Consequently, the two-way ANOVA was performed with confidence in its validity.

4. Results and Discussion

4.1 Descriptive Statistics

4.1.1 Universities' Representation

The study included 440 participants from 61 universities across Pakistan, with most from Faisalabad-based institutions. Despite uneven regional representation, the diverse sample enhances the test's generalizability.

4.1.2 Semesters' Representation

The highest participation came from the 8th semester (142 students), followed by the 4th (117). The mean semester value was 5.56, with a standard deviation of 2.498.

4.1.3 Gender Distribution

Female comprised 57% and males 43%. Despite convince sampling, gender representation remained balanced.

4.1.4 Descriptive Statistics of Test Items

The test had 440 items, with a mean score of *220.50* and a standard deviation of *127.161*, indicating significant variability.

4.2 Validity

4.2.1 Face Validity

The face validity of the test was evaluated by the PhD scholars. They assessed whether the test items effectively measured the grammatical knowledge, aligned with language skills, and maintained clarity in structure, instructions, and visual presentation. To validate the test further, feedback was collected from the ten test-takers through a survey. Most of the participants found the test relevant, appropriate and well-structured. However, one participant found the difficulty level high, and another had concerns about the clarity of instructions. The percentage of the agreement among the test-takers was calculated using Desai and Patel's (2020) method, yielding 96%, which falls under the *"Full Agreement"* category. Consequently, all the test items were retained.

4.2.2 Construct Validity

Construct validity was assessed by the experts, ensuring the test accurately measured the metalinguistic competencies according to the model of Bialystok and Ryan (1985). The test comprised three sections: identification, correction, and explanation, emphasizing analysis and control over explicit grammar knowledge. Experts confirmed the test's alignment with theoretical constructs, effectively assessing the intended skills. The omission of a production component was intentional, as it is not a core element of the theoretical model.

4.2.3 Content Validity

To evaluate the content validity the experts from the field of Applied Linguistics were involved throughout the test development. They evaluated the test's alignment with theoretical frameworks, domain definitions, and grammatical aspects such as tenses, subject-verb agreement, structure, active, and passive, etc. A rigorous review process including literature review and item analysis, ensured domain representation and relevance. Initially, 150 items were developed; after validity analysis, 136 items retained. A pilot study confirmed the test's content validity, ensuring its appropriateness in terms of language, content, and format.

4.3 Item Analysis

4.3.1 Facility Value

The difficulty analysis also known as facility value (FV) was calculated to assess item difficulty in the test**.** The formula applied in Excel classified items as "Too Difficult" ($FV < 0.3$), "Appropriate/Ideal" ($0.3 \leq FV \leq 0.7$) or "Too Easy" ($FV > 0.7$). Based on the criterion, 27 items including 2, 6, 12, 15, 17, 19, 20, 27, 39, 40, 44, 46, 51, 53, 57, 82, 86, 90, 94, 95, 96, 99, 105, 112, 115, 122, and 135 were dropped before conducting the discrimination index analysis due to need for refinement. Results indicate that 80% of the test items were in the "Appropriate/Ideal" range. 10% were "Too Easy" and 10% were "Too Difficult." This distribution suggests that the test effectively distinguishes varying proficiency levels among undergraduate participants in Pakistan.

Descriptive statistics further confirm a well-balanced difficulty distribution. The mean FV of *.527* suggests moderate difficulty, with 52.7% of participants answering correctly. The median FV (*.555*) aligns closely with the mean, indicating a relatively symmetrical distribution. A standard deviation of *.155* reflects a reasonable spread of item difficulties, with the most challenging item ($FV = .120$) answered correctly by only 12% of the test-takers and the easiest item ($FV = .820$) by 82%. Comparisons with prior studies support these findings. Wistner (2014) reported item difficulty ranging from *-2.27* to *2.40* logits using Rasch analysis, while Sanosi (2022) observed difficulty values between *0.36* and *0.82*, indicating a balanced assessment. The present study aligns with these patterns, underscoring the necessity of item calibration to maintain an appropriate difficulty level.

4.3.2 Discrimination Analysis

The discrimination analysis was conducted after excluding the items with extreme difficulty levels. From the 109 items, 85 were classified as very good, as they were effectively distinguishing between high and low performers. Five items were categorized as good, four as moderate, and eleven as poor/ questionable. Four items were marginal, showing limited discrimination, while eighteen were excluded for failing to meet the *.3* discrimination index (DI) threshold. The remaining 92 items were retained for further analysis.

The Identification section exhibited a mean DI of *.51* ($SD=.17$), with values ranging from *.16* to *.85*, indicating consistent discrimination. The Correction section had the highest mean DI (*.60,* $SD=.20$), but also included negatively discriminating items (*-.19* to *.82*), suggesting the need for refinement. The Explanation section demonstrated moderate discrimination ($M=.46, SD=.20$), with indices ranging from *-.15* to *.76*.

Hence, the Correction section showed the strongest differentiation between high and low performers, while the Identification and Explanation sections maintained relatively high discrimination levels. Consistent with Sanosi (2022), who set *.3* as the benchmark for effective discrimination, all sections exceeded the threshold, affirming their effectiveness. However, items with negative discrimination require revision to enhance validity.

4.3.3 Distractor Analysis

Distractor analysis assessed the effectiveness of the incorrect answer choices in differentiating between high- and low-scoring students. While most distractors functioned well, some exhibited moderate effectiveness, indicating a need for revision. Strengthening weaker distractors can enhance test reliability and validity.

Each test item contained one correct answer (CA) and three distractors, analyzed across upper and lower groups. The results showed that only four distractors (option D in items 5,25,37, and 120) were non-functional, selected by fewer than 5% of the test-takers. These often-included vague choices like "none of the above" or "no error", which may reduce test effectiveness (Mahjabeen et al., 2017). All the other distractors were functional, reinforcing their role in maintaining the test quality. Revising the non-functional distractors will further improve the discriminatory power to the test items.

4.4 Reliability Analysis

After conducting the item analysis, a reliability analysis was conducted on the test, initially consisting of 92 items, which resulted in a Cronbach's Alpha of *.967*. To enhance internal consistency, four items (92, 118, 133, and 134) with low corrected item-total correlation were removed. After elimination these items, the test's reliability was reassessed with 87 remaining items, divided into three sections: Identification, Correction, and Explanation.

The Identification section, which included 26 items, recorded a Cronbach's Alpha of *.908*, indicating a strong internal consistency. The Correction section, reflecting strong coherence among the items. The Explanation section, containing 26 items, yielded a Cronbach's Alpha of *.891*, which, while slightly lower than the other sections, still indicated a substantial level of reliability.

Previous research has shown variations in internal consistency across different sections of metalinguistic tests. Sakai (2008) found that the Identification section had a high reliability of .9 due to strong rater agreement, while Elder and Manwaring (2004) reported lower reliability for rule Explanation (*.77*) compared to Correction (*.9*). Roehr (2007) found higher internal consistency for Explanation (*.818*) than for Correction (*.640*). Suggesting that conceptual explanations tend to be assessed more reliably than error identification and correction. Similarly, Wistner (2014) found that rule explanation had a high internal consistency of *.86*, demonstrating coherence in assessing metalinguistic competence.

Table 4.1 Cronbach Alpha Coefficient of Reliability

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	No of Items
.967	.967	87

The final reliability analysis, conducted using SPSS on 87 items with responses from 440 participants, resulted in a Cronbach's Alpha of *.967*, confirming an exceptionally high level of internal consistency. This finding validates the test's stability and effectiveness in measuring university undergraduates' metalinguistic knowledge of grammar across various disciplines in Pakistan.

Compared to prior studies, the reliability score of this test is notably high. Ellis (2005), Bowles (2011), and Steel (1997) reported Cronbach's Alpha values of *.90*, *.93*, and *.897*, respectively, demonstrating similarly strong reliability. In contrast, research by Sanosi (2022) and Zhang (2014, 2015) yielded moderate reliability scores between *.69* and *.76*, while Correa (2011) recorded a coefficient of *.812*, indicating reasonable consistency. According to established language testing standards (Ayden, 2018; Elder & Manwaring, 2004; Renou, 2001), the Metalinguistic Knowledge test of grammar used in this study is validated as a highly reliable instrument for evaluating student's cognitive skills in grammar.

4.5 Bias Analysis

A two-way ANOVA was conducted to examine the effects of gender and academic subject on MKT scores. The results, presented in the table 4.2, indicate a significant overall model effect ($F = 2.829$, $p = .007$), suggesting that the included factors contribute to variations in MKT scores. However, the R-squared value (0.044, Adjusted $R^2 = 0.028$) indicates that the model explains only a small portion of variance.

Table 4.2 Tests of Between-Subjects Effects

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	8612.337[a]	7	1230.334	2.829	.007
Intercept	526759.066	1	526759.066	1211.120	.000
Subject	2627.793	3	875.931	2.014	.111
Gender	2060.884	1	2060.884	4.738	.030
Subject * Gender	1797.129	3	599.043	1.377	.249
Error	187892.063	432	434.935		
Total	1240048.000	440			
Corrected Total	196504.400	439			

a. R Squared = .044 (Adjusted R Squared = .028)

The main effect of academic subject was not significant (*$F = 2.014$, $p = .111$*), indicating that MKT scores did not vary substantially across different subjects. However, gender had a significant effect (*$F = 4.738$, $p = .030$*), suggesting performance differences between male and female participants. The interaction between subject and gender was not significant ($F = 1.377$, $p = .249$), meaning that the relationship between subject and MKT scores did not differ by gender.

While gender differences in performance suggest some variability, the absence of a subject effect and interaction indicates that the test measures metalinguistic knowledge consistently across subjects. Despite the observed gender differences, the results support the test's validity and reliability as an unbiased assessment tool suitable for diverse participant groups.

4.6 Limitations of the Study

This study's limitations include the use of convenience sampling, which limits generatability. A stratified random sampling would have ensured better representation across demographic variables. Additionally, the sample size of 440 participants was insufficient for nationwide generalizability. Given the 150 test items, a more appropriate sample size would be around 1500 participants. Online administration, despite proctoring, introduced potential biases. The untimed format restricted insights into test difficulty, and an in-person setting would have ensured more accuracy.

4.7 Pedagogical Implications

The validated Metalinguistic Knowledge Test (MKT) offers a reliable assessment tool for Pakistani undergraduates, supporting grammatical proficiency in academia and competitive exams (PPSC. FPSC, HAT, GAT, GRE). Its finding contributes to curriculum development and policies aimed at enhancing metalinguistic competence in ESL education.

4.8 Recommendations for Future Research

Further research should conduct Exploratory Factor Analysis (EFA) to examine the factorability and dimensions of the test, providing insights into its structural validity. Correlational research could explore links between metalinguistics knowledge and other linguistic skills, particularly writing proficiency. A larger, more diverse sample would improve generalizability and refine metalinguistic assessment tools for more effective language instruction.

5. Conclusion

This study validated the Metalinguistic Knowledge Test (MKT) for university undergraduates, demonstrating strong psychometric properties. Face validity was confirmed through expert evaluations and test-taker feedback, with a high survey reliability (*.96*). Construct validity aligned with Bialystok and Ryan's (1985) model, ensuring focus on analysis and control. Item analysis refined the test by removing 27 overly easy or difficult items, ensuring 80% fell within an optimal range. The correction section showed the highest discrimination index (.5986), followed by identification (.5644) and explanation (.4565). Distractor analysis confirmed strong item functionality. Reliability analysis of 91 items yielded a high Cronbach's Alpha (.967), with 87 items retained after removing low-performing ones. Bias analysis found no significant impact of gender or field of study. The findings support MKT's use in higher education for competitive exams, curriculum development, and ESL assessment. Future research should expand the sample and explore alternative test formats to enhance generalizability and assessment precision.

Funding: This study was not funded in any shape or form by any party.

Conflict of Interest: The author declares that he has no conflict of interest.

Bio-note:

Zulaikha Nadeem is an MPhil Scholar in the Department of Applied Linguistics, Government College University, Faisalabad. Her areas of interest include Applied Linguistics, English Language Teaching (ELT), Corpus Linguistics, and Academic Writing.

Dr. Aleem Shakir is an Assistant Professor in the Department of Applied Linguistics, Government College University, Faisalabad. His areas of interests are ELT, Phonetics and Phonology, ESP, Testing and Evaluation and Corpus Linguistics.

References

Alderson, J. C., & Steel, D. (1994). Metalinguistic knowledge, language aptitude and language proficiency. In D. Graddol & S. Thomas (Eds.), *Language in a Changing Europe* (pp. 93-103). Clevedon: Multilingual Matters.

Alderson, J. C., Clapham, C., & Steel, D. (1997). Metalinguistic knowledge, language aptitude and language proficiency. *Language Teaching Research, 1*, 93-121.

Alipour, S. (2014). Metalinguistic and linguistic knowledge in foreign language learners. *Theory and Practice in Language Studies, 4(*12), 2640-2645. doi:10.4304/tpls.4.12.2640-2645

Amrani, A. (2015). Testing the Role of Metalinguistic Awareness and the Use of Grammatical Terminology in Promoting Students' Proficiency in Second Language Learning: The Case of 1st Year L.M.D Master Students of Both English Literature and Applied Linguistics at the Department of English, University of Constantine 1. [Master's thesis, University of Constantine 1]

Bialystok, E. (1979) Explicit and implicit judgments of L2 grammaticality. Language Learning 29, 81–103.

Bialystok, E. (1990). *Communication Strategies*. Oxford: Blackwell.

Bialystok, E. (1992). Attentional control in children's metalinguistic performance and measures of field.

Bialystok, E. (1994). Analysis and control in the development of second language proficiency. *Studies in Second Language Acquisition, 16*, 157–168.

Bialystok, E. (1999). Levels of Bilingualism and Levels of Linguistic Awareness. *Developmental Psychology, 24*, 560-567.

Bialystok, E. (2001). *Bilingualism in Development: Language, Literacy, and Cognition*. Cambridge: Cambridge University Press.

Bialystok, E., & Bouchard Ryan, E. (1985a). Toward a Definition of Metalinguistic Skill. *Merrill-Palmer Quarterly, 31*(3), 229-251.

Bialystok, E., & Ryan, E. B. (1985b). Metacognitive Framework for the Development of First and Second Language Skills. In D. L. Forrest-Pressley, G. E. MacKinnon, & T. G. Waller (Eds.), *Metacognition, Cognition, and Human Performance* (pp. 207-252). New York: Academic Press.

Bialystok, E., & Ryan, E. B. (1985c). On precision and virtue of simplicity in metalinguistics: A reply to Menyuk. *Merrill-Palmer Quarterly, 31*, 261–264.

Bloor, T. (1986a). 'What Do Language Students Know about Grammar?' *British Journal of Language Teaching, 24*, 157-160.

Bowles, M. A. (2011). Measuring Implicit and Explicit Linguistic Knowledge: What Can Heritage Language Learners Contribute? *Studies in Second Language Acquisition, 33*(2), 247-271.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Correa, M. (2011). Subjunctive Accuracy and Metalinguistic Knowledge of L2 Learners of Spanish. *Electronic Journal of Foreign Language Teaching, 8*(1), 39–56.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Downing, S. M. (2006). Twelve steps for effective test development. In *Handbook of test development* (pp. 3-25). Lawrence Erlbaum Associates.

Educational Testing Service. (2002). *TOEIC Bridge koushiki gaido & mondaishu [TOEIC Bridge official preparation guide]*. IIBC.

Educational Testing Service. (2008). *TOEIC Bridge koshiki work book [TOEIC Bridge official prep guide and workbook]*. IIBC.

Elder, C., & Manwaring, M. (2004). The relationship between metalinguistic knowledge and learning outcomes among undergraduate students of Chinese. *Language Awareness, 13*(3), 145-162. https://doi.org/10.1080/09658410408667092

Elder, C. (2009). Validating a test of metalinguistic knowledge. In R. Ellis, S. Loewen, C. Elder, J. Erlam, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 113-138). Multilingual Matters.

Elder, C., Warren, J., Hajek, J., Manwaring, D., & Davies, A. (1999). Metalinguistic knowledge: How important is it in studying a language at university? *Australian Review of Applied Linguistics, 22*(1), 81–95.

Ellis, R. (1987). *Second Language Acquisition in Context*. London: Prentice-Hall International.

Ellis, R. (2004). The definition and measurement of L2 explicit knowledge. *Language Learning, 54*(2), 227–275. https://doi.org/10.1111/j.1467-9922.2004.00255.x

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language. *Studies in Second Language Acquisition, 27*(1), 141-172. https://doi.org/10.1017/S0272263105050096

Ellis, R. (2008). Investigating grammatical difficulty in second language learning: Implications for second language acquisition research and language testing. *International Journal of Applied Linguistics, 18*(1), 4-22. https://doi.org/10.1111/j.1473-4192.2008.00158.x

Ellis, R. (2009). Implicit and explicit learning, knowledge and instruction. In R. Ellis (Ed.), *Implicit and explicit knowledge in second language learning, testing, and teaching* (pp. 3-26). Multilingual Matters.

Fan, W., & Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review. *Computers in Human Behavior, 26*(2), 132–139. https://doi.org/10.1016/j.chb.2009.10.015

Field, A. P. (2005). *Discovering statistics using SPSS*. Sage Publications Inc.

Ghauri, P., & Gronhaug, K. (2005). *Research methods in business studies*. FT/Prentice Hall.

Gombert, J. E. (1992). *Metalinguistic Development*. Chicago: University of Chicago Press.

Gravetter, F. J., & Wallnau, L. B. (2014). *Essentials of statistics for the behavioral sciences* (8th ed.). Wadsworth.

Gudmundsdottir, G. B., & Brock-Utne, B. (2010). An exploration of the importance of piloting and access as action research. *Educational Action Research, 18*(3), 359-372. https://doi.org/10.1080/09650792.2010.491415

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), 37-50. https://doi.org/10.1207/s15324818ame0201_4

Hertzog, M. (2008). Considerations in determining sample size for pilot studies. *Research in Nursing & Health, 31*(2), 180-191. https://doi.org/10.1002/nur.20247

Hu, G. (2002). Psychological constraints on the utility of metalinguistic knowledge in second language production. *Studies in Second Language Acquisition, 24*(3), 347–386.

Kim, Y. (2011). The Pilot Study in Qualitative Inquiry: Identifying Issues and Learning Lessons for Culturally Competent Research. *Qualitative Social Work, 10*(2), 190-206.

Krashen, S. (1981). *Second Language Acquisition and Second Language Learning*. Oxford. England: Pergamon.

Kurvers, J. (2006). Discovering Language: Metalinguistic Awareness of Adult Illiterates. In I. van de Craats, J. Kurvers, & M. Young-Scholten (Eds.), *Low-Educated Adult Second Language and Literacy Acquisition* (pp. 69–88). Utrecht: LOT.

Mahjabeen, W., Alam, S., Hassan, U., Zafar, T., Butt, R., Konain, S., & Rizvi, M. (2017). Difficulty index, discrimination index and distractor efficiency in multiple choice questions. *Annals PIMS-Shaheed Zulfiqar Ali Bhutto Medical University, 13*(4), 310–315.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher, 10*(9), 9–20. https://doi.org/10.3102/0013189X010009009

Osterlind, S. J. (2006). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats*. Kluwer Academic Publishers.

Patel, N., & Desai, S. (2020). ABC of face validity for questionnaire. *International Journal of Pharmaceutical Sciences Review and Research, 65*(1), 164–168. https://doi.org/10.47583/ijpsrr.2020.v65i01.025

Renou (2001) An Examination of the Relationship between Metalinguistic Awareness and Second-language Proficiency of Adult Learners of French. *Language Awareness, 10*(4), 248-267. DOI: 10.1080/09658410108667038.

Ricciardelli, L. A. (1993). Two components of metalinguistic awareness: Control of linguistic processing and analysis of linguistic knowledge. *Applied Psycholinguistics, 14*, 349–367.

Ricciardelli, L. A., Rump, E. E., & Proske, 1. (1989). Metalinguistic Awareness as a Unitary Construct and Its Relation to General Intellectual Development. *Rassegna Italiana di Linguistica Applicata, 21*, 19-40.

Roehr, K. (2005). *Metalinguistic Knowledge in Second Language Learning: An Emergentist Perspective.* Unpublished PhD thesis, Lancaster University.

Roehr, K. (2007). Metalinguistic Knowledge and Language Ability in University-Level L2 Learners. *Applied Linguistics, 29*(2), 173-199.

Roehr-Brackin, K. (2018). *Metalinguistic Awareness and Second Language Acquisition* (1st ed.). Routledge. https://doi.org/10.4324/9781315661001

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104*(3), 192–233. https://doi.org/10.1037/0096-3445.104.3.192

Rosch, E. (1977). Human categorization. In N. Warren (Ed.), *Studies in cross-cultural psychology* (Vol. 1; pp. 1–49). San Diego, CA: Academic Press.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Erlbaum.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. Cognitive Psychology, 7(4), 573–605. https://doi.org/10.1016/0010-0285(75)90024-9

Sanosi, A. B. (2022). Correlation of EFL Learners' Metalinguistic Knowledge and Grammatical Accuracy. *Studies in English Language and Education, 9*(3), 908-925.

Sullivan G. M. (2011). A Primer on the Validity of Assessment Instruments. *Journal of Graduate Medical Education, 3*(2), 119–120. https://doi.org/10.4300/JGME-D-11-00075.1.

Taherdoost, H. (2016). Validity and reliability of the research instrument; How to test the validation of a questionnaire/survey in a research. *International Journal of Academic Research in Management (IJARM), 5*(3), 28-36. http://dx.doi.org/10.2139/ssrn.3205040

Tarone, E. (1987). Methodologies for Studying Variability in Second Language Acquisition. In Ellis, R. (Ed.), *Second Language Acquisition in Context*. London: Prentice-Hall.

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Medical Education, 9*(1), 40.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53-55. doi:10.5116/ijme.4dfb.8dfd

Thorndike, R. M., & Thorndike-Christ, T. (2014). Measurement and evaluation in psychology and education. *Journal of the American Statistical Association, 56*(296), 1029. https://doi.org/10.2307/2282039

Tokunaga, M. (2010). Metalinguistic Knowledge of Low-Proficiency University EFL Learners. In A. M. Stoke.

Tokunaga, M. (2014). Exploring Metalinguistic Knowledge of Low to Intermediate Proficiency EFL Students in Japan. *SAGE Open, 5*(2), 1–10. DOI: 10.1177/2158244014553601.

Tsang, W. L. (2011). English metalanguage awareness among primary school teachers in Hong Kong. *GEMA Online Journal of Language Studies, 11*(1). ISSN: 1675-8021.

Turner, S. P. (1979). The Concept of Face Validity. *Quality and Quantity, 13*(1), 85–90.

Wistner, B. (2014). *Effects of metalinguistic knowledge and language aptitude on second language learning* [Doctoral dissertation, Temple University]. Temple University.

Zhang, R. (2015a). Measuring university-level L2 learners' implicit and explicit linguistic knowledge. *Studies in Second Language Acquisition, 37*(3), 457-486. https://doi.org/10.1017/S0272263114000370

Zhang, R. (2015b). Examining the role of implicit and explicit L2 knowledge in general L2 proficiency. *International Journal of English Linguistics, 5*(3), 12-24. https://doi.org/10.5539/ijel.v5n3p12