ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

JOURNAL OF APPLIED LINGUISTICS AND TESOL

Vol.8. No.3.2025

DEVELOPING AND VALIDATING A METALINGUISTIC KNOWLEDGE TEST OF WRITTEN ACADEMIC DISCOURSE FOR UNIVERSITY UNDERGRADUATES IN PAKISTAN

Hira Hanif
MPhil Scholar, Department of Applied Linguistics
Government College University, Faisalabad, Pakistan
Email: hirahanif9876@gmail.com
Aleem Shakir (Corresponding Author)
Assistant Professor, Department of Applied Linguistics
GC University, Faisalabad, Pakistan
Email: almsha@yahoo.com

Abstract
Metalinguistic knowledge plays an important role in language learning. Its evaluation involves learners' ability to analyze their explicit knowledge of certain aspects of language. Various tests have been developed by researchers keeping in view the specific domain and the population. The present study aimed to develop and validate a Metalinguistic Knowledge Test (MKT) by focusing on a particular domain of written academic discourse. Before the test development, vocabulary knowledge of the undergraduate participants was noted using Vocabulary Size Test by Nation and Beglar (2007) as it helped in knowing the reading comprehension level of participants which was later used in the selection of the appropriate passages for the test. A test consisting of 132 MCQs was developed keeping in view the framework of Bialystok and Ryan (1985). Before test administration, initial piloting was done on 39 participants to notice any issues in the test. Then, actual data for the newly developed test was collected using convenience sampling from 385 undergraduate participants studying in 36 different universities in Pakistan. It was an untimed test, and this decision was made based on the administration of previously developed tests. After the test administration and data collection, it was analyzed to check validity and its measures, item difficulty, item discrimination, distractor effectiveness, reliability, and bias in terms of gender and subject of the participants. After performing calculations based on the participants' responses, it was revealed that the remaining 63-item test was valid and highly reliable, resulting in a value of .93. In addition, factors like gender and subject did not independently bias the test. This test was significant as no test was developed in the past on metalinguistic knowledge of written academic discourse.

Keywords: *Metalinguistic Knowledge Test, written academic discourse, test development, Pakistani undergraduates, test validity, test reliability*

1. Introduction

The complex phenomenon of language has always been a debatable subject matter among researchers as the investigation of the knowledge of a language (e.g. its structure and representation) is essential to be focused on, especially of the second language that involves the conscious awareness of its rules and structures that make up the language. This conscious knowledge of a language (explicit knowledge) can be verbalized (Paradis, 1994; Roehr, 2008) and it requires less effort to be investigated by the researchers than the investigation of knowledge that learners gain naturally without conscious efforts (implicit knowledge) (Roehr, 2008) and which cannot be verbalized (Kirsh, 1992). If the conscious knowledge of language is analyzed directly and explicitly by the language experts, it can be called metalinguistic knowledge of that language. The terms Metalinguistic knowledge and explicit knowledge can be used interchangeably (Alderson et al.,1997; Elder, 2009; Hu, 2002).

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

The evaluation of metalinguistic knowledge in any language is achievable considering different linguistic domains including phonology, morphology, semantics, syntax, and connected discourse (Roehr-Brackin, 2018). Although the major focus is on the ability to identify, correct, and explain grammatical errors, other domains have also been researched to some extent. The major focused grammatical aspects were explored mostly on undergraduates (Alipour, 2014; Aydin, 2018; Elder & Manwaring, 2004; Hu, 2002, etc.) while some studies have been conducted on children which especially looked for Morphological aspects of language (McBride-Chang et al., 2005) and phonemic aspects of language which the child goes through his early age (Alegria et al., 1982; Marsh & Mineo, 1977; Rosner, 1975; Yopp, 1988).

Due to the widespread necessity of English language, most of the metalinguistic knowledge tests have been developed so far in English (as it is being taught as a second language in institutions) by specifically focusing on sounds, words, and grammar. Despite assuming of having conscious knowledge of these domains which indicate the awareness of the kind of patterns that English language covers and their specific ways of representation, learners still lack in language skills because of their less attention to a bigger domain which focuses on the overall structure of the language. That's where the domain of connected discourse arises which deals with the idea of structuring the sentences to make them logical and understandable and that is where the students face challenges because of their major focus on learning phonemes, morphemes, and grammar and less focus on how to structure all of them together in an organized way.

Till today, no standard test covering only metalinguistic knowledge of written academic discourse for language teachers is available to access the students' knowledge of this aspect. So, the present study aims to develop and validate a test on metalinguistic knowledge of written academic discourse for university undergraduates of Pakistan.

2. Literature Review

2.1 Metalinguistic Knowledge

The term metalinguistic knowledge was defined by many theorists and experts in the past. Gombert (1992) defined the term metalinguistics as it is "concerned with linguistic activity which focuses on language" (p. 2) showing that the major focus is on the *language as an object.* Bialystok (2001) defined metalinguistic knowledge as it is the knowledge about language. She also distinguished metalinguistic knowledge from the linguistic knowledge based on greater generality of metalinguistic knowledge as it is broad and abstract and contains knowledge of general principles that can be applicable to more than one language. It can also be defined as "learners' explicit knowledge about language" (Alderson et al., 1997; Bialystok, 1979; Elder et al., 1999). Ellis (2004) defined this explicit knowledge as "the conscious awareness of what a language or language in general consists of and/or the roles that it plays in human life" (p. 229). These definitions show that metalinguistic knowledge can be evaluated as it is the explicit and conscious knowledge about the language.

Metalinguistic knowledge can be evaluated on different levels of linguistic domains which include phonology, morphology, vocabulary, grammar, and pragmatics (Ellis, 2004; Hu, 2002). Kurvers et al. (2006) broadened this list by including the domains of semantics and connected discourse. Kurvers et al. (2006, as cited in Roehr-Brackin, 2018, p. 15) stated "metalinguistic skills can be applied to a broad range of linguistic domains, including phonology, morphology, semantics, syntax, and connected discourse, with metalinguistic awareness in the latter domain referring to the explication of meaning relations between sentences". The metalinguistic knowledge of all these domains helps access the L2 knowledge of the learners as it sharpens their understanding of the target language (Ellis,

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

2004) leading it to verbalization which is responsible for the highest level of consciousness of language (Bialystok, 1990; Schmidt, 1990).

2.2 Theoretical Framework

One of the most important models of metalinguistic understanding was presented by Bialystok and Ryan (1985). According to them, all metalinguistic tasks require two skill components which are necessary for processing requirements. These include analysis of linguistic knowledge and control of linguistic processing. As these components are involved in metalinguistic understanding, a more elaborated definition of metalinguistic awareness can be the "conscious reflection on, analysis of, or intentional control over various aspects of language—phonology, semantics, morphosyntax, discourse, pragmatics—outside the normal unconscious processes of production or comprehension" (Karmiloff-Smith et al., 1996, p.198). With the help of this framework covering two components, development in learners' metalinguistic understanding can be described (Myhill & Jones, 2015).

The first component, analysis of linguistic knowledge, means the ability to mentally represent explicit and abstract structures in a very organized way (Bialystok, 2001). Tasks such as detection of errors, correction of errors, and explanation of detected errors are the ones that place greater demands on this component of analysis (Ricciardelli, 1993). Another component, the control of linguistic processing, means the ability to provide attention to specific aspects of mental representations and ignore those aspects which are irrelevant for that specific task (Bialystok, 1994, 2001). Some tasks require more or less one component than the other for successful completion. The relatively high and low demands for analysis and control components totally depend on the type of tasks.

2.3 A Review of Metalinguistic Knowledge Tests (MKT)

2.3.1 Types of Items Found in Different Metalinguistic Knowledge Tests

Tests based on phonemic awareness were developed by many researchers which include Phoneme blending test (Roswell-Chall, 1959), phoneme counting test (Liberman et al., 1974), phoneme segmentation test (Goldstein, 1974; Yopp-Singer, 1984), phoneme deletion test (Bruce, 1964; Rosner, 1975), rhyming test (Yopp, n.d.), and phoneme reversal test (Yopp, n.d.) consisting of 30, 42, 16, 22, 30, 13, and 20 items respectively. Participants were tested regarding blending of sounds and counting the number of phonemes. Other tests included pictures and participants guessed the names or items that the instructor uttered by breaking into parts or participants had to break the word into different sounds. In tests related to deletion, participants were provided with words and a phoneme that needed to be deleted to guess the word. The rhyming test tested for their identification. Lastly, in reversal tasks, words were presented, and participants were asked to reverse the order of phonemes.

A test based on morphological awareness was developed by McBride-Chang et al. (2005) consisting of 33 items. Pictures were presented, and words (target morphemes) were given to choose one picture that best relates to the meaning of the morpheme. In some items, scenarios were orally presented, and participants had to say a word based on the story of a few lines. Al Farsi (2008) adapted the test of McBride-Change et al. (2005) and added a few items to it giving a total of 28 items where words without context were given to test the participants regarding the division into as many segments as they could base on the meanings. In the remaining items, participants were also given frame sentences, and they were asked to solve another one keeping in view the target morpheme. The Third test, adapted by Tighe (2015) from different researchers, consisting of 61 items included three different tasks, namely *Derived Form Morphology, Derivational Suffix Choice, Morphological Analogy Real World Task.* The first task included 28 fill-in-the-blanks and participants responded with appropriate derived, complex word forms. In the second task, 18 items in the form of multiple-choice

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

questions were presented. In task 3, 15 items were given where participants were presented with A: B: C items and based on this they filled D.

The first test based on MKT grammar was developed by Alderson et al. (1997) comprising 100 items where participants were tested based on terms of grammar to know the difference between claimed familiarity versus actual. The second test was adapted from Bialystok (1979) by Renou (2001) who also developed a few items providing a total of 60 items related to the correction and description of error. They were also instructed to write the correct rule they used. The third test developed by Elder and Manwaring (2004) included 54 where items were about writing the names of grammatical terms being asked, and about the correction and explanation of errors. In addition, Ellis' (2005) adapted the test from Alderson et al. (1997) included 41 multiple choice questions about grammar. Tokunaga's (2010) test comprised 40 items presented in the form of sentences with few marked words which needed attention. The participants were asked to choose from the given terms and write those matched terms with the underlined part of the sentences. The last test by Sanosi (2022) consisted of 50 items in which grammatical items were given to match the terms with their Arabic equivalents as well as match them with their examples.

In most of the tests, the major evaluation was based on questions related to identification, correction, and explanation of errors, as can be seen in Alipour (2014) and Ayden's (2018) 30 item-based tests, Roehr's (2008) 54 item-based test, Correa's (2011) 15-item based test, Wistner's (2014) 39 item-based test, Zhang's (2015) test.

Tsuji and Doherty (2014) developed a test based on pragmatic awareness namely *politeness judgment task* consisting of six pairs of sentences which were presented orally to the participants to identify polite or impolite expressions. In the case of MKT of discourse, there is no complete and well-developed test available, but a few items could be found in other tests as well. For example, SAT (developed by College Board in the United States in 1926) writing and language section included some items related to discourse knowledge. These items were given in the form of multiple-choice questions. As there were a total of 44 questions, 24 were about certain aspects covered in discourse knowledge while the rest were about grammar and other concepts.

2.3.2 Population

Tests on phonemic knowledge have been developed by different researchers but due to the absence of validity and reliability measures in most of the cases, all these tests were compared to determine all the measures by applying them to the same population (Yopp, 1988). The selected participants (mostly white, but also included few black, Asians, and the ones with Spanish surnames) were studying in three public schools in California. The average age of the participants was 5 years and 10 months, and they belonged to the lower middle to upper middle class.

Morphological awareness tests were administered to different types of population. McBride- Chang et al. (2005) administered the test on kindergartens (mean age = 6.1 years) and second graders (mean age = 8 years) studying in local elementary schools. The participants were causations, Middle Eastern, African American, and Asian American. Al Farsi (2008) administered MKT of morphology on first semester middle-class EFL learners (mean age = 18.09 years) studying at Ibri College of Applied Sciences, Oman. Tighe (2015) administered the test to native speakers of English (mean age = 24 years) enrolled in ABE classes in Florida. They belonged to different ethnic groups namely African American, Hispanic, Asian, Caucasian, and mixed.

All the tests on metalinguistic knowledge of grammar (explained below) were developed for university students. Alderson et al. (1997) developed MKT of grammar for first-year

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

French undergraduate students at Lancaster University. It was pre-administered on sixth-year secondary school students as well to know the difficulty level of the test and more valuable information about other important aspects. Renou (2001) collected data from university students (mean age = 21 years) of Ottawa who were native speakers of English, but they were advanced-level French L2 learners. Elder and Manwaring's (2004) test investigated first- and second-year students who were studying Chinese in 2000 and 2001 as a foreign language at the University of Melbourne. Ellis (2005) administered the adapted test to undergraduates of arts or engineering courses, graduates, and former students at a university in New Zealand. Roehr (2008) included English-speaking learners (mean age = 20.1 years) enrolled as undergraduates at a British university.

In addition, Tokunaga (2010) collected data from students who were studying in their first class at a private university in 2009. The TOEIC bridge scores of the participants were also obtained. Correa (2011) applied the test to participants who were attending Spanish classes at a university in the United States. It was also noted that whether they were native speakers of Spanish. Alipour (2014) administered a test on EFL learners at the university level (first year & second year). This study did not control the factor of age as the researcher himself mentioned. Wistner (2014) administered the test to students (freshmen, juniors, seniors, and sophomores) of two Japanese universities. The age of 98% of the participants ranged from 18-23, two participants were 24 years old, and from the remaining two participants, one was 27 years old and the second was 33 years old. It was also noted whether the participants had ever gone abroad for studies or not.

Zhang (2015) administered the test to 1st year L2 students studying at a key university in China. The participants included in the study were from small towns as well as from big cities in China which showed that there was diversity in terms of the opportunities they gained from living in different areas. Aydin (2018) included Intermediate-level Turkish EFL learners studying at a Turkish university while Sanosi (2022) administered the test to undergraduate students (having good knowledge of grammar and language teaching) studying at Prince Sattam University, Saudi Arabia.

The test developed by Tsuji and Doherty (2014) was administered to nursery students in Japan to investigate their pragmatic knowledge. They were all native speakers (3 to 5 years old) of Japan, belonging to middle-class families. In the case of MKT of discourse, the developed SAT was applicable to participants who were interested in getting admission to universities. Most of the participants between the ages of 17 and 19 solve this test.

2.3.3 Sample Size and Sampling Strategy

A study by Yopp (1988) compared all the tests on phonemic knowledge by applying them to 104 children (50 boys, 50 girls) without mentioning the sampling strategy being used.

The three conducted studies on MKT of morphology by McBride-Chang et al. (2005), Al Farsi (2008), and Tighe (2015) included 220 (108 boys and 112 girls in kindergarten; 104 boys and 116 girls in second grade), 54 (29 girls, 25 boys) and 220 (108 girls, 112 boys) participants respectively. The second and third studies included those participants who were volunteers while in the first study, no information about the sampling strategy was available.

MKT grammar was administered by Alderson et al. (1997), Alipour (2014), Correa (2011), Elder and Manwaring (2004), Ellis (2005), Renou (2001), Roehr (2008), Tokunaga (2010), Wistner (2014), and Zhang (2015) to 128, 38, 177 (116 females and 61 males), 91, 11, 64 (59 females and 5 males), 60 (43 females, 17 males), 195 (the ratio of male and female participants was not discussed), 240 (123 females, 94 males), 100 participants (93 females, 7 males). Researchers mentioned that participants were volunteers while those who

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

did not mention the sampling technique included Alderson et al. (1997), Alipour (2014), Elder and Manwaring (2004), Ellis (2005), and Roehr (2008).

Th researchers who adopted different sampling strategies included Aydin (2018) and Sanosi (2022). Ayden (2018) included a sample of 38 participants using convenience sampling. Sanosi (2022) collected data from 124 participants (65 females, 59 males) using the cluster sampling method. The participants of this study were volunteers from the two universities.

Test of pragmatic knowledge was collected from 68 children (Tsuji & Doherty, 2014) 17 of them were three years old (12 males, 5 females), 23 participants were 4 years old (11 males, 12 females) and the remaining 28 were 5 years old (11 males, 17 females).

2.3.4 Item Difficulty

Wistner (2014) calculated item difficulty values using Rasch analysis. For the receptive metalinguistic knowledge test, the calculated item difficulty range was 4.67 logits (min = 2.27, max. = 2.40). In the case of the productive metalinguistic knowledge test, the easiest item had a difficulty value of -1.42 while the difficult one had a value of 2.00. Sanosi (2022) mentioned that difficulty ranged from 1% to 100% (.00-1.00). The values of test items ranged from .36-.84. No other researcher calculated item difficulty values.

2.3.5 Item Discrimination

Sanosi (2022) calculated discrimination values and interpreted them by following the criteria of good, fair, and poor. The results of the discrimination analysis indicated that all the items had a discrimination value above .30 which meant that all the items were good, and the test could easily distinguish between students.

2.3.6 Distractor Effectiveness

Distractor analysis is an effective way of knowing the quality of the distractors of all the constructed items but despite its necessity and importance, none of the studies conducted distractor analysis.

2.3.7 Validity

The construct and predictive validity of the phoneme awareness tests have been checked by Yopp (1988). Factor analysis evident that the tests on phonemic awareness were interrelated which indicated that the tests were measuring the same construct, thus proving their construct validity. In addition, the greatest predictive validity (r = .72) was noted for the Yopp (n.d.) modification of the sound isolation test. The closest to this was Goldstein's (1974) test of Phoneme segmentation (r = .71). After these two tests, Yopp-Singer's (1984) phoneme segmentation test and Bruce's (1964) phoneme deletion test obtained a similar value (r =.67).

In the case of MKT of morphology, McBride-Chang et al. (2005) and Al Farsi (2008), did not mention the validity of the instruments. Although Tighe (2015) did not check the validity of the adapted test, one of its developers, Nunes et al. (2006), discussed that the test had construct validity. Tong et al. (2011) also proved the construct validity of the developed test by discussing the correlations between the two measures included in the morphology test.

In the case of MKT of grammar, Elder and Manwaring (2004) ensured the content validity of the instrument by doing a systematic analysis of materials and textbooks used for item development. Ellis (2005) discussed the construct validity of the instrument by investigating several hypotheses whose results were in favour of the construct validity of the instrument. Wistner (2014) measured the construct validity of the instruments by using Rasch analysis which evidenced the construct validity of both the instruments. Zhang (2015) measured Component Factor Analysis (CFA) whose results supported the construct validity of the instrument. The construct validity was then checked by comparing the performance of

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

the sample of this study with the results that Ellis (2005) gained from another population. Aydin (2018) ensured validity by asking the opinions of the experts and necessary changes were made in the test after the consultation. The researchers who did not check the validity of MKT grammar were Alderson et al. (1997), Alipour (2014), Correa (2011), Renou (2001), Roehr (2008), Sanosi (2022), and Tokunaga (2010).

In the case of MKT of pragmatics, Tsuji and Doherty (2014) did not calculate the validity of the developed task. The developed test on discourse knowledge, SAT, has predictive validity and they maintain it by establishing its design based on the solid foundations of current research, testing the items and questions included in the test, and by looking at metrices on how students are performing.

2.3.8 Reliability

In case of phoneme awareness tests, the reliability of all the tests was checked using Cronbach's alpha. The Phoneme blending test developed by Roswell-Chall (1959) has the greatest reliability value (α = .96) followed by high alpha values of .95, .92, .88, .84, .83 by Yopp-Singer's (1984) phoneme segmentation test, Bruce's (1964) phoneme deletion test, Goldstein's (1974) phoneme segmentation test, Yopp's (n.d.) modified test of sound isolation, and Liberman's et al. (1974) phoneme counting test, respectively. Among these tests, there were two tests (Rosner's phoneme deletion test, 1975; Yopp's Rhyme test, n.d.) which obtained reliability values of .78 and .76 and were considered as moderate to high reliabilities. The lowest reliability value was in the case of Yopp's (n.d.) word-to- word matching test (α = .58). The Reliability value of Phoneme reversal test was missing as this test was dropped because it was too difficult when tried with a subset of 15 participants.

In the case of MKT of morphology, McBride-Chang et al. (2015) mentioned internal reliability values i.e. .71 morphological structure awareness task and .80 for morpheme identification task without mentioning the formula used for the calculation. In Al Farsi's (2008) study, the reliability of the analysis section was .87 and the reliability of the synthesis section was .93 using Cronbach's alpha resulting in an overall reliability value of .91. Tighe (2015) mentioned that the reliability of Morphological Analogy Real Word Task was .74, Derivational Suffix Choice Test was .82, and Derived Form Morphology (DMORPH) Task was .87 using Cronbach's Alpha.

In the case of MKT grammar, the reliability values calculated by Alderson et al. (1997) and Ellis (2005) were .89 and .90 using Cronbach's alpha. Renou (2001) calculated that reliability for written judgement task and oral judgement task was .77 and .87 using Cronbach's Alpha. Elder and Manwaring (2004) measured reliability which was .80 for grammatical terms, .90 for error correction, and .77 for rule explanation using Rasch analysis. Roehr (2008) measured reliability for each question type of the test. Reliability calculated by Roehr (2008) using alpha was .64 for correction of errors, .81 for description and explanation of errors, and .62 for language analytics.

Tokunaga (2010) used Winsteps (Linacre, 2007) for Rasch analysis to measure reliability. He mentioned that person reliability was .89 (good but not exceptional according to Hughes, 2020) while item reliability was .97 (extremely good). Correa (2011) mentioned that the reliability of the instrument was .81. As the test was in Spanish too, its inter-rater reliability was .92. Alipour (2014) did not measure the reliability of the developed test. Wistner (2014) calculated Rasch person reliability and item reliability of both the instruments. The values that the researcher calculated in the first test were .67 and 1.43 for Rasch person reliability and item reliability respectively. The internal reliability of the instrument was .67 using Cronbach's Alpha. For second test based on production Rasch person reliability and item reliability for technical terminology were .78 and 1.91

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

respectively. Its internal reliability was .82 using Cronbach's alpha. Rasch person reliability for the rule explanation scale was .81 while item reliability was 1.91. The internal reliability was .86 and it was also measured using Cronbach's Alpha.

In addition, Zhang (2014) mentioned .76 reliability value using Cronbach's Alpha. The values indicated high reliability of the test as the range fell between .70-.90 (according to DeVellis (1991). Aydin (2018) ensured reliability by asking the opinions of three experienced instructors of English. Sanosi (2022) added the reliability of scores which was .69 for MKT of grammar, However, the researcher did not mention how it was measured.

Tsuji and Doherty (2014) mentioned that the reliability of the pragmatic knowledge test was .71 using Cronbach's alpha. In the case of MKT of discourse, SAT was a reliable test and its reliability coefficient was .89-.91 for writing and language multiple-choice questions (College Board, 2013).

The review of various tests made it clear that various metalinguistic knowledge tests have been developed in the past concerning different linguistic domains for knowing the explicit knowledge of the learners of any language. Some of the domains were given more importance while some were not researched thoroughly. With the help of a thorough review, it was revealed that no test is available for the evaluation of metalinguistic discourse knowledge. Although a few items were included in SAT writing and language section, no well-developed test is available for administration. This type of test is possible as written by Kurvers et al. (2006) and Wistner (2014). So, the present study aims to fill this gap by developing a metalinguistic knowledge test of written academic discourse for university undergraduates. This test will help learners access their written academic discourse knowledge, that is, the relationship between the topic introduction, body, and conclusion, the relationship between different parts of the topic introduction which include hook, thesis statement, and signposting, the relationship between thesis statement, and topic sentences of each paragraph, and the relationship between topic sentence of each paragraph and supporting details of each paragraph, keeping in view the aspects of development, organization, and effective language use in different kinds of selected texts.

To develop and validate the Metalinguistic Knowledge Test of written academic discourse, the study aims to answer the following questions:

RQ 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?

RQ 2: To what extent do the MCQs align with the theoretical constructs of language proficiency?

RQ 3: How comprehensively do the MCQs cover the language content domains they are supposed to assess?

RQ 4: What is the difficulty index of the test items in the study?

RQ 5: How does the discrimination index vary across three sections of the test?

RQ 6: What is the effectiveness of each distractor in distinguishing between high-scoring and low-scoring groups?

RQ 7: What is the internal consistency of the test items, and how does it reflect the overall reliability of the test?

RQ 8: Are the MCQs free from bias related to test-takers' gender and subject of study?

3. Methodology

3.1 Instrumentation

3.1.1 Vocabulary Size Test

The comprehension level of the participants of the present study was necessary to be evaluated as the test developed in the present study was based on number of passages and

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

their selection was a difficult process because the level of comprehension must match with the level of the text included. According to Laufer and Sim (1985), vocabulary is the best gauge to understand whether the text will be understandable or not. So, to check the comprehension level of the participants, Vocabulary Size Test developed by Nation and Beglar (2007) consisting of 140 questions representing 14000-word families was used.

The test was administered to 91 undergraduate participants (diversified in terms of their gender, department, semester, and CGPA) from 14 departments of Government College university Faisalabad. As the students performed best from the first 1000 till the fifth 1000, these five levels evidenced the comprehension level of students (i.e., most of the students know up to fifth 1000-word families, so the text should include words that fall in those 5000-word families). However, their average vocabulary size was 6197 words because on sixth 1000 the average of known words was 5.69/10. After knowing their vocabulary size, the next step was to select a text based on the participants' vocabulary size.

3.1.2 Metalinguistic Knowledge Test of Written Academic Discourse

The test based on MK of written academic discourse was developed keeping in view the criteria followed in SAT writing and language section regarding the number of passages, topics covered in passages, types of passages, and features covered. The results of vocabulary size were also given major importance in the selection of passages. The details of the test are given below:

3.1.2.1 Selection of Passages

The test consisted of four passages (442-500 words per passage) covering topics from careers, humanities, social studies, science, and history. These selected passages which were informational, non-fiction narrative, and argumentative were taken from general online preparatory sides. These passages were run through AI software to rephrase the text. This initial transformation aimed to alter the surface structure and language of the passages while preserving the core ideas and concepts. Following the AI transformation, the rephrased texts were carefully reviewed and edited by human experts.

Apart from fulfilling all the requirements followed in the SAT, there was one more criterion that was kept in mind, that is, the texts should match the vocabulary level of the participants to ensure that students comprehend the passage easily. For this purpose, two methods were used which were later compared to get more reliable and accurate results. Firstly, three students (diversified in terms of departments and CGPAs) were provided with the selected texts, and they were asked to highlight the unknown words in the text carefully to ensure their honesty. Then based on the reviews provided by the experts and to ensure a broader and more representative evaluation of the participants' vocabulary knowledge, the selection of students for vocabulary checks was expanded beyond the initial three participants. In total, ten students were selected from different academic levels (first year, second year, third year, and fourth year) and various fields of study (sciences, humanities, and social studies), with an even distribution of gender. This allowed for a more accurate reflection of the diverse backgrounds of the participants and improved the reliability of the vocabulary check results. Secondly, the RANGE programme by Nation et al. (2002) was used, which made it possible to compare the vocabulary level of participants with the text comprehension level they needed to attain 95% comprehension of the text. The results given by participants and the RANGE programme were compared to know whether the students would be able to comprehend 95% of the text or not.

When the selected passages were run through RANGE one by one, the total number of words and the known number of tokens by participants were noted to calculate the comprehension percentage for each passage. When analyzing the proper nouns and tokens

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

"not in the list" in the RANGE programme, each word was individually reviewed based on two criteria: student feedback from the vocabulary check and its frequency of use in academic texts. If students did not mark these words as unknown and they appeared frequently in general academic use, they were deemed familiar and included in the "known tokens" list. Additionally, proper nouns and non-listed words were cross-checked with high-frequency word lists in academic writing to ensure that their inclusion reflected students' expected comprehension levels.

The comprehension percentage for Passage 1 was 93.03% as the known number of tokens were 441 (keeping in view the first five 1000-words) out of total 464 words. As the lists of words working under RANGE did not have certain words labeled as "not in the list", those words, mostly hyphenated, were reviewed based on two above-mentioned criteria. Those 12 tokens were added in known words and the percentage keeping in view the total number of known tokens (now 453) was 95.56%.

Passage 2 consisted of 457 known tokens out of total 500 words in the passage. The inclusion of 12 words of List 15 (proper nouns) and 10 (not in the list) tokens provided the required 95.8%. Passage 3 contained a total of 469 words out of which 447 were known tokens (95.30%). After the inclusion of 13 tokens of list 15 (proper nouns) and 3 'not in the list' tokens, the obtained percentage of known tokens reached 98.72%. It meant that the participants would easily understand the text. Passage 4 was an argumentative passage in which 431 out of 442 tokens were known according to the results gained by the participants in the Vocabulary Size Test. The obtained percentage was 97.51%. After the inclusion of 5 tokens from list 16 (proper nouns) and 5 'not in the list' tokens, the percentage was 99.77%.

3.1.2.2 Selection of Features for the Development of Items

Items for the test were developed according to SAT based on 10 features which belonged to three major categories falling in *expression of ideas.* The categories included organization, development, and effective language use. Organization included logical development of information, logical placement of information, transitions, and modifications in different parts of texts. The development included questions about main ideas only, supporting details, and main ideas. Effective language use covered questions regarding improving conciseness and removing wordiness. The test also added questions that demand 'command of evidence', that is, include the reason for making a change. For example, why a certain sentence needed to be added/deleted/revised. Some extra questions were added regarding data interpretation presented in the form of graphs, charts, or tables but they belonged to the same above-mentioned features.

These 10 features were added based on the metalinguistic framework of identification, correction, and explanation of errors in relationships. The identification section was quite easy while the explanation section was the most difficult where evidence-based questions were added.

3.1.2.3 Number of Questions and Their Division in Different Sections

There were a total of 132 multiple-choice questions divided into three sections, namely identification, correction, and explanation of relationship errors. Every section covered 10 questions per passage resulting in a total of 30 questions per passage. Two passages also covered items (12 items- 6 per passage) regarding data interpretation of graphs/tables because the same pattern was followed in the SAT. These questions were also based on the categories of identification, correction, and explanation of errors as included in the metalinguistic framework. Questions 1-36, 37-66, 67-96, 97-132 were about passage 1"career choices", passage 2 "Pakistan's journey of independence", passage 3 "assignment essays", and passage

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

4 “ban of mobile phones in schools” respectively. The last six questions in passage 1 and passage 4 were regarding graphs on the relevant topics.

3.1.2.4 Terms Used in the Development of the Test

As the test dealt with the relationship among different parts of the text, it was obvious to use terms like hook, thesis statement, signposting, and topic sentence in the development of stems and distractors but these terms were inly understandable for students of Applied linguistics. It was suggested by experts to replace them with their direct alternatives. Direct alternatives were found for the terms and were discussed with an expert to avoid any ambiguity in the comprehension of the terms. The term “hook” was replaced with “attention getter/attention grabber”, the term “thesis statement” was replaced with “central point of the whole passage”, the third term related to introduction section, “signposting”, was replaced with “preview of the main points”. The term whose presence is a must in every paragraph was “topic sentence” which was replaced with “main idea of the paragraph”. After the replacements, the test was sent to the participants, and they were able to comprehend the alternative terms. Another term called “redundant” confused them so its meaning “repetitive” was added in brackets with the term to ensure comprehension.

3.2 Improvements by Experts

After developing the test, it was sent to three experts, Ph.D. scholars of Applied Linguistics, to fulfill three purposes. They were first asked to find ambiguity in terms of the given instructions and the developed items. Experts reviewed the test and provided insights on instructions, stem, distractions, and formatting of the test. Secondly, they developed the answer key of the test which was later compared to see the ratio of agreement among experts and researcher regarding the correct options. Thirdly, they were asked to provide the difficulty level of every item to ensure the test was not too easy to solve by every participant or too difficult to be dealt with. It was done using a 5-point Likert scale adopted from Vagias (2006) which ranged from 1(*very easy*) to 5 (*very difficult*). None of the items were considered too easy or too difficult by any of the experts.

3.3 Piloting

Once the test was revised based on the improvements suggested by experts, it was piloted as it helps in modifying the instrument further according to the specific environment by identifying issues in it (Malmqvist et al., 2019). The revised test was piloted on 39 participants (28 females, 11 males) keeping in view the criteria given by Hertzog (2008) that it was preferable to take around 35-40 samples if the purpose was to check internal consistency or to revise the instrument based on item performance. The selected participants were studying in 17 departments belonging to the faculties of Arts and Social Sciences, Engineering and Technology, Life Sciences, Physical Science, and Medical Sciences at Government College University Faisalabad. The test was conducted online through Google Forms by configuring add-ons to ensure transparency. Reading passages were provided in a separate online document where participants were allowed to only view the document and switch tabs rather than scroll between passages and questions all the time. The active participation of participants was ensured by providing them with certificates of participation by the department of Applied Linguistics. The high achievers of the test were also presented with food deals as a reward for their dedication.

3.4 Advertisement for the Test

Keeping in view the necessity of having large number of participants for actual data collection, it was planned to advertise the test so that it could reach the maximum number of people. Participants from all over Pakistan were encouraged to register through a link to Google Forms by mentioning its benefits and its similarity with other language tests like the

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

SAT that most high schools require for admission purposes. The advertisement was shared on different social media platforms like LinkedIn, WhatsApp groups, Instagram, and Facebook. Apart from sharing advertisement on social media platforms, details about the test and its benefits were also announced at the university while interacting face-to-face with the students.

3.5 Participants

The data was collected from 385 undergraduate participants (241 females, 141 males) studying in 36 universities in Pakistan using convenience sampling. 18 of the universities were situated in Punjab, 7 universities were in Sindh, 6 universities were located in Islamabad Capital Territory, 3 universities were from Khyber Pakhtunkhwa, and the remaining 2 were in Baluchistan and Azad Kashmir. In addition, these participants belonged to 43 different degree programs. 19.74% of the participants were from software engineering, 19.22% were from Applied Linguistics, 8.57% were from English literature, 7.53% were from Physiology, and 7.01% were from Computer Science. The rest of the departments had a representation of less than 5%. These degree programs were divided into Health and Life Sciences, Engineering and Technology, and Humanities and Social Sciences.

3.6 Test Administration

The test was administered in 5 weeks in shifts through Google Forms. Previous studies on metalinguistic knowledge administered tests without any time restrictions as mentioned by Bowles (2011), Roehr-Brackin (2018), and Zhang (2015). So, keeping in view that the current test was also based on metalinguistic knowledge, there were no time restrictions for the completion of the test. During the data collection, the certificates of participation and scores were emailed to them daily to ensure their privacy. Incentives were provided to encourage the participants. As the types of incentives vary from age to age, keeping in view the choices, the participants were given food deals. This initiative not only encouraged them to participate but also motivated them to perform excellently to be among the high achievers.

3.7 Ethical Considerations

The research was conducted keeping in view the ethical considerations. Written consent was obtained from the participants of all the universities during the registration process. Participants were assured that the data was collected for research purposes only and their personal information would be kept confidential. Their privacy was also ensured by sharing the results of the test with the participants via email addresses.

3.8 Data Entry and Cleaning

After completing the test administration process, data was entered on the Excel sheet by assigning codes to universities, departments, semesters, genders, and participants. Answers to each test item were written in the form of A, B, C, D.

3.8.1 Abbreviations and Codes for Universities

Abbreviations and codes were assigned to each university to avoid lengthy names. The names of the 36 universities were abbreviated, for example, Cholistan University of Veterinary & Animal Sciences (CUVAS = 1), Fatima Jinnah Women University (FJWU = 2), Ghazi University (GU = 3) and Government College University Faisalabad (GCUF = 4) etc.

3.8.2 Abbreviations and Codes for Departments

Abbreviations and codes were assigned to 43 departments from which the data was collected, for example, Allied Health Sciences (AHS = 1), Applied Microbiology (AMB = 2), Applied Linguistics (AL = 3), and Business Administration (BBA = 4), etc.

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

3.8.3 Gender Codes

As the data was collected from both males and females, so they were assigned codes on the excel sheet. Males were given the code 1 while females were represented with 2.

3.8.4 Participants' ID Codes

IDs were allocated to all the participants of the study rather than analyzing the data with their names. For example, the first participant of the study was allocated 1 as the ID. Similarly, the ID of the last participant of the study was 385.

After the completion of the data entry process, the answers given by participants in the form of A, B, C, and D were changed to 0 (wrong answer) and 1 (right answer) for data analysis purposes.

3.9 Data Analysis

3.9.1 Validity

After the development phase, the test was analyzed to check its validity, which shows how accurately a method measures something. It is about the truthfulness of the results as mentioned by Altheide and Johnson (1994).

Face validity ensured whether the test appeared to measure what it meant to measure. It was the surface-level appropriateness of the instrument. On the other hand, Construct validity ensured that the test measured the construct that it was intended to measure. Content validity ensured the degree to which the instrument's questions and their scores were representative of all the possible questions that could be asked regarding the content (Creswell, 2005). Content validity was checked keeping in view the four elements namely domain definition, domain representation, domain relevance, and appropriateness of test development procedures (Sireci, 1998)

3.9.2 Item Difficulty Analysis

Difficulty analysis was conducted to see whether the test was appropriate according to the level of participants. This was calculated for each item separately by dividing the total number of participants who got the answer right with the total number of participants who attempted to solve it.

3.9.3 Item Discrimination Analysis

Discrimination analysis was conducted to see whether the items distinguished between the upper group and lower group of participants. For this purpose, the participants from the upper group who answered the item correctly were subtracted from participants of the lower group who answered the item correctly, and then the value was divided by the number of participants in each group. The obtained value was evaluated against a fixed criterion of item discrimination.

3.9.4 Distractor Analysis

Distractor analysis was conducted to see the effectiveness of distractors. It was checked by whether a distractor was chosen by 5% of the population or not. If it was not chosen by at least 5% of the participants, it meant the distractor was not good enough and need to be improved or removed to enhance the quality of the test.

3.9.5 Reliability Analysis

Reliability analysis was conducted to determine the consistency of the test. It was checked through Cronbach's alpha, popularized by Cronbach (1951). An alpha having a value of .70 was generally considered a sufficient value in science education (Taber, 2018).

3.9.6 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.

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

3.9.6.1 Sample Size. As the following table reveals, our study includes two independent variables, gender (2 levels) and subject area (3 levels).

Table 3.1
Distribution of Participants Across Levels of Independent Variables

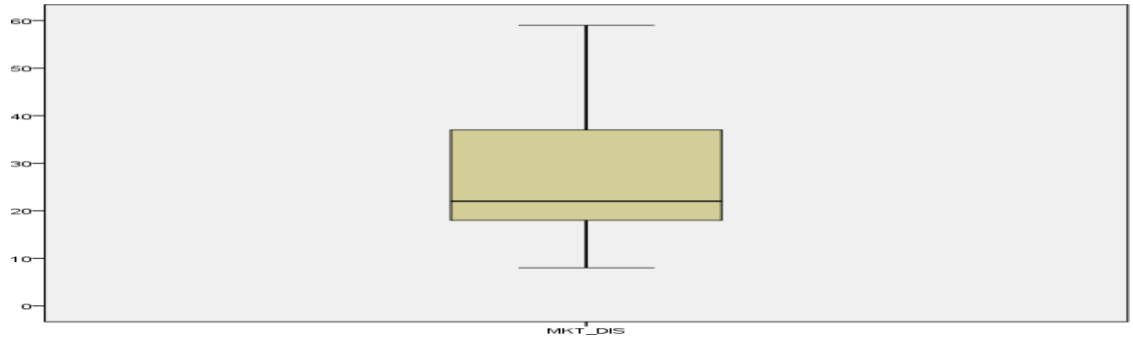
		Value label	N
Gender	1	Male	144
	2	Female	241
Subject	1	Health and Life Sciences	110
	2	Engineering and Technology	126
	3	Humanities and Social Sciences	149

Gender and Subject, with 144 males and 241 females, and 110 participants in Health and Life Sciences, 126 in Engineering and Technology, and 149 in Humanities and Social Sciences, resulting in a total sample size of 385 participants. The two-way ANOVA requires examining the interaction between these two independent variables, leading to a total of six groups (2 Genders × 3 Subjects). Each group should ideally have a sufficient number of observations to ensure the reliability of the statistical analysis. The average cell size in our study is approximately 64.17, which significantly exceeds the commonly recommended minimum of 20 observations per cell. Thus, our sample size is well-suited for the two-way ANOVA, providing a robust foundation for detecting potential main effects and interaction effects between Gender and Subject. Furthermore, preliminary checks have been conducted to confirm that the data meets the assumptions of two-way ANOVA, including independence of observations, normality of the dependent variable within groups, and homogeneity of variances across groups. These considerations support the appropriateness of our sample size and the validity of the subsequent analyses.

3.9.6.2 Independence of Observations. The data sufficiently meets the requirement of independence of observation because the data has been collected from 43 different departments belonging to Health and Life Sciences, Engineering and Technology, and Humanities and Social Sciences.

3.9.6.3 Absence of Significant Outliers in Dependent Variable. As the following box plot shows, the dependent variable does not involve any significant outlier.

Figure 3.1
Absence of Outliers



3.9.6.4 Normal Distribution of Dependent Variable for All Levels of Each of

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Independent Variable. The normal distribution of dependent variable for each level in each dependent variable was determined through skewness and kurtosis values which in all cases was within +/-2 range.

3.9.6.5 Outliers in Each Group. In total, 13 outliers, as shown in the following figures, were detected from the total sample of 385, which were before conducting two- way ANOVA. The emerging sample size was thus reduced to 372.

Figure 3.2

Outlier Analysis of Dependent Variable with Respect to Gender (Males) Belonging to Health and Life Sciences

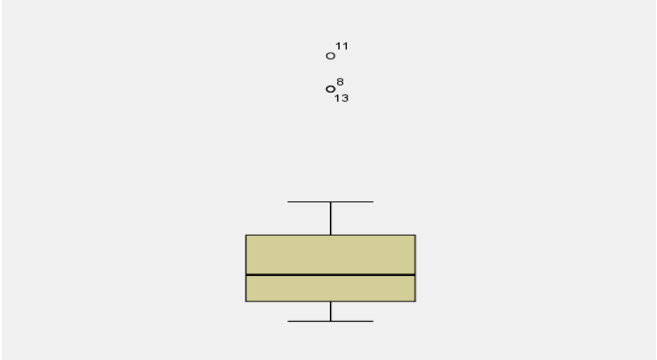


Figure 3.3

Outlier Analysis of Dependent Variable With Respect to Subject (Health and Life Sciences) Belonging to Males

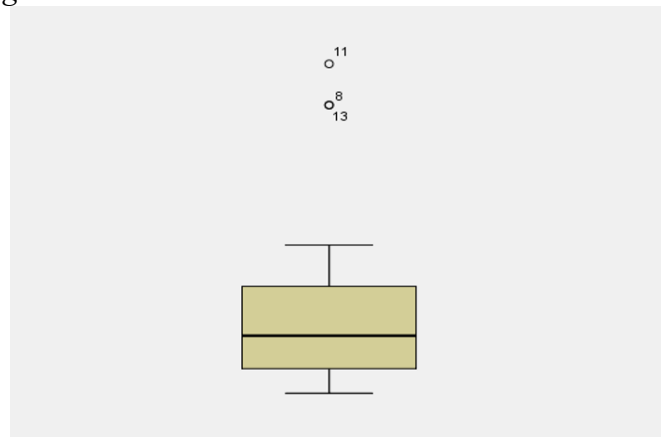
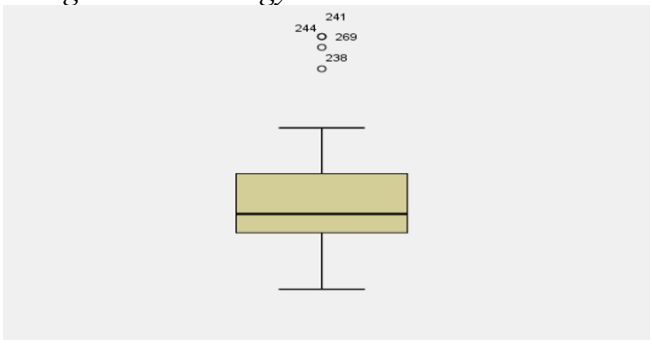


Figure 3.4

Outlier Analysis of Dependent Variable With Respect to Gender (Females) Belonging to Engineering and Technology



ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Figure 3.5

Outlier Analysis of Dependent Variable With Respect to Subject (Engineering and Technology) Belonging to Females

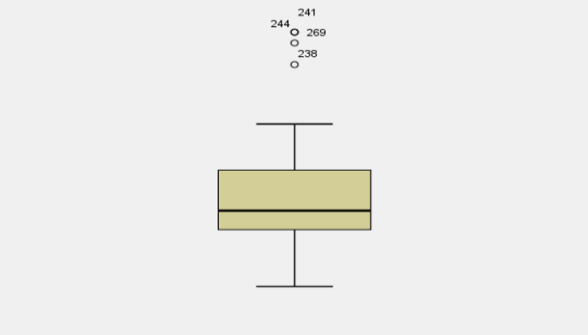
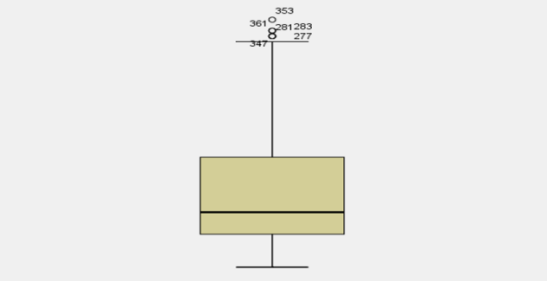


Figure 3.6

Outlier Analysis of Dependent Variable With Respect to Gender (Females) Belonging to Humanities and Social Sciences



Homogeneity of Variance of Dependent Variable for Each Group. Levene's Test of Equality of Error Variances was conducted to examine the assumption of homogeneity of variances for the dependent variable "MKT_DIS" across different groups based on “Gender” and “Sub”. The test results for gender revealed an F-value of 13.930 with a significance level of .000, indicating a significant difference in variances ($p < .001$). This led to the rejection of the null hypothesis, suggesting that the assumption of homogeneity of variances is violated for gender groups. Conversely, the test results for "Sub" showed an F-value of 2.236 with a significance level of .108, which was not statistically significant ($p > .05$). Consequently, the null hypothesis could not be rejected, indicating no significant difference in variances across "Sub" groups and supporting the assumption of equal variances for these groups. These findings underscore the necessity of verifying the homogeneity of variances assumption in statistical analyses, particularly when comparing groups based on gender, to ensure the validity of the results.

Table 3.3

Levene's Test of Equality of Error Variances[a] Across Subject

Dependent Variable: MKT_DIS

F	df1	df2	Sig.
2.236	2	369	.108

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Note. Tests the null hypothesis that the error variance of the dependent variable is equal across groups.
a. Design: Intercept + Sub

Table 3.4
Levene's Test of Equality of Error Variances[a] Across Gender

Dependent Variable: MKT_DIS

F	df1	df2	Sig.
13.930	1	370	.000

Note. Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

Design: Intercept + Gender

4. RESULTS AND DISCUSSION

After collecting responses from 385 participants, that data was run on IBM SPSS Statistics 21 to generate descriptive statistics of the data. The descriptive statistics were calculated based on 36 universities, 5 semesters, 3 faculties, and 2 genders, order of 4 passages, and 3 sections mentioned in the metalinguistic framework.

Statistics indicated the presence of the most number of participants from the universities located in Punjab, with most participants belonging to the software engineering department. The most prominent faculty was arts and social sciences with the inclusion of mostly female participants. in the test, passage 1 has the highest mean while passage 3 has the lowest mean. The range value was also the highest in passage 1, which showed that passage 1 was easy to comprehend for the participants. In addition, the mean values of three sections (covering 44 questions in each) were similar indicating a somehow similar level of average performance by participants across these three categories of questions. The median (central value) of the three categories was also close, which showed typically the participants scored 14 or 15 out of 44 in each category. The median value showed that half of the participants' scores were below 14 or 15 in each category, but the high average scores indicated the presence of outliers in the data.

RQ 1: To what extent do the MCQ items appear to measure the intended language skills from the perspective of both language experts and test-takers?

Face validity of the instrument was checked by getting perspectives from both the language experts and test takers. Three language experts ensured that the test appeared to measure the academic discourse knowledge of the participants. They also ensured that the test was fulfilling the purpose, and the items and the reading passages were according to the level of participants.

To get participants' opinions, a face validity form consisting of six questions was sent to the participants who were selected using convenience sampling and they were asked to respond to the questions by saying yes or no. Responses were compiled and % calculation was done.

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Table 4.6

Percentage agreement for face validity

		Participants' responses				
	Q1	Q2	Q3	Q4	Q5	Q6
% of per question agreement	96%	92%	88%	96%	80%	92%
% of overall agreement				90.6%		

These calculations were noted due to the remarks of Lynn (1986) and Zamanzadeh et al. (2015) who considered that face validity was unstandardized and unquantified approach so the best way to quantify was to calculate the percentages. Based on the criterion presented by Desai and Patel (2020) after an extensive literature search, it was concluded that overall agreement was above 90% which supported the presence of full strength of agreement ensuring that the test has face validity, so the participants will likely put less effort in it resulting in the inaccurate reflection of their abilities. However, despite its importance, none of the researchers checked the face validity of MKT in the past (e.g., Al Farsi, 2008; Alipour, 2014; Aydin, 2018; Ellis, 2005; Wistner, 2014).

RQ 2: To what extent do the MCQs align with the theoretical constructs of language proficiency?

For measuring the metalinguistic knowledge of the participants, a framework by Bialystok and Ryan (1985) consisting of two components, namely analysis of linguistic knowledge and control of linguistic processing, was used. The items of the current test were based more on the first component which included identification, correction, and explanation of relationship errors. The test showed the balanced and representative inclusion of the elements of the chosen theoretical construct.

The alignment of MCQs with the theoretical construct was also checked with the help of language experts. The experts analyzed the alignment and reported no misalignment as there was no emphasis on a certain element over the other.

Similar results were gained by Aydin (2018), Ellis (2005), Nunes et al. (2006), Tong et al. (2011), Wistner (2014), Yopp (1998), and Zhang (2015). These researchers also ensured the construct validity of their tests by adopting different methods leading to the generalization of its importance in test development.

RQ 3: How comprehensively do the MCQs cover the language content domains they are supposed to assess?

As the test was about knowledge of academic discourse, it was compulsory to include all the aspects mentioned to be a part of this test. To see whether the developed test contained all those aspects, the content validity of the instrument was checked was getting opinions from the language experts. The elements given by Sireci (1998) for the evaluation of test contents were used to check the content validity of the instrument.

Experts ensured that the test was fully covering the contents and aspects mentioned to be a part of test which led to the coverage of domain definition component. The domain representation was checked with the help of subject matter experts who mentioned that the test aspects were representative and balanced as they included an equal number of questions for each aspect of written academic discourse.

Domain relevance was examined by experts who verified that test items were relevant to the aspects being tested. The items were analyzed from one another aspect, that is, to

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

find out the presence of any irrelevant aspect in the test or the absence of any important aspect related to academic discourse. The subject matter experts concluded that all the items were relevant to the aspects, no aspect had been missed, and in addition, the presence of any irrelevant aspect was not reported. The last element to ensure content validity was the appropriateness of the test development process. Experts' reviews were obtained who ensured that the test had technical accuracy, and it also followed the standard principles required for quality item writing. It also included pilot testing (to find the ambiguous items related to certain aspects) and item analysis (to find the items related to certain aspects that were functioning poorly) which concluded that the test was comprehensively covering the language domains that it was supposed to access.

The result of the content validity of this study relates to Elder and Manwaring (2004), who also proved this type of validity by following the same way, that is, getting opinions from different sources during the test development process.

RQ 4: What is the difficulty index of the test items in the study?

Item difficulty was calculated based on participants' responses by using the following formula:

Item difficulty = Number of correct responses/ Total number of responses

The results were interpreted based on a criterion of very difficult (.00-.29), moderately difficult (.30-.70), and very easy (.71-1.00). As the values of item difficulty for the current test ranged from .12-.58, it showed that most of the items were very difficult and moderate and none of them fell into the category of very easy.

Table 4.8

Results and Interpretation of Item Difficulty Analysis

Item Difficulty value	Interpretation	Frequency	Percentage
.00-.29	Very Difficult	33	25
.30-.70	Moderately Difficult	99	75
.71-1.00	Very Easy	0	0

Keeping in view the criteria used by Adegoke (2013) and Date et al. (2019), 99 *moderately difficult* items that fell in the range of .30-.70 were kept. 33 items falling in *very difficult* out of 132 were deleted after calculating the difficult index.

RQ 5: How does the discrimination index vary across different sections of the test?

The remaining 99 items were further analyzed for calculating item discrimination. As the intention was to distinguish high achievers and low achievers, the total scores of all the participants were calculated and then based on the ranks, 104 (27%) out of 385 participants from the upper group and 104 (27%) from the lower group were taken. The scores of these two groups were used for discrimination analysis which was calculated using the formula:

Item Discrimination = Upper group - Lower group / No. of participants in each group

Table 4.13

Combined Result and Interpretation of Item Discrimination Analysis

Item discrimination index	Interpretation	Frequency
>.30	Good discrimination	63

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

.20-.30	Acceptable discrimination	10
< .20	Poor discrimination	26
Negative value	Defective item	4

According to the results, the values for the current test ranged from .00-.82. According to the criteria mentioned by Ebel and Frisbie (1972), items having a discrimination index >.30 were kept while all the other items were removed. So, a total of 63 out of 99 items were kept. The deleted items that fell between .20-.30, below .20, and negative value were 10, 26, and 4 items respectively. Only one item had a poor discrimination value of .00 in the whole test which meant that both the groups performed the same and it could not distinguish between the high achievers and low achievers. Four items of the test that had a negative discrimination index meant that the low-performing group performed better than the high-performing group in those items.

The test items were based on three sections, so discrimination values were calculated separately for each of them. The test was based on passages, and every passage had three sections. Section 1, Identification of relationship errors, included questions 1- 10, 31-32, 37-46, 67-76, 97-106, 127-128. Section 2, correction of relationship errors, included questions 11-20, 33-34, 47-56, 77-86, 107-116, and 129-130. Section 3, explanation of relationship errors, covered questions 21-30, 35-36, 57-66, 87-96, 117-126, and 131-132. The remaining 99 were separately analyzed based on their sections.

As far as the remaining 32 items out of 44 of section 1 are concerned, 19 items were considered to have good discrimination (>.30), 5 items had acceptable discrimination values (.20-.30), 6 items had poor discrimination values, and 2 items were considered defective items because of their negative values.

In section 2, the remaining items after difficulty analysis were 30 out of 40. 19 items had good discrimination values, 3 items had acceptable discrimination values, 6 items had poor discrimination values, and 2 items were considered defective.

Section 3 included 37 items out of 44 items after difficulty analysis. It had 25 items having good discrimination values, 2 items having acceptable discrimination, and 10 items having poor discrimination. There was no defective item in section 3. The information has been summarized in the following table:

Table 4.14

Result of Discrimination Analysis Across Categories

Category	Good discrimination	Acceptable discrimination	Poor discrimination	Defective item
Identification	19	5	6	2
Correction	19	3	6	2
Explanation	25	2	10	0

The above calculation indicated that the explanation section had the highest good discrimination values but also had the highest poor discrimination values as well. The remaining sections, identification and correction, are more balanced. A point worth noting was that the explanation section had the most difficult items as compared to the rest of the two sections, and the results of item discrimination showed that these items could distinguish between the groups.

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

The results of item discrimination of this test were compared with Sanosi's (2022), which indicated that all the items had a value above .3. The findings of the research were different from the results of the present study whose many items fell below .3 and some of the items even had negative values.

RQ 6: What is the effectiveness of each distractor in distinguishing between high-scoring and low-scoring groups?

The remaining 63 items were analyzed to calculate the effectiveness of distractors. For conducting distractor analysis, the parameters provided by Tarrant et al. (2009) were considered. The distractors needed to be chosen by 5% of the participants to consider them as functioning/effective distractors. Otherwise, they were named as non-functioning distractors. The percentage of all the selected distractors for each item was calculated to check whether it was selected by 5% of the participants or not. The table below indicated that all the distractors were chosen by at least 5% of the people exhibiting that all the distractors were good and functional and there were no non-functional or ineffective distractors. So, no item was deleted in the distractor analysis.

Table 4.16

Combined Results of Distractor Analysis

Category	Frequency	Percentage
Functional distractors (FD)	189	100
Non-functional distractors (NFD)	0	0

The results of distractor analysis could not be compared with the previous research because none of the researcher calculated distractor effectiveness in the past despite the importance of knowing the quality of distractors.

RQ 7: What is the internal consistency of the test items, and how does it reflect the overall reliability of the test?

Reliability analysis was conducted using Cronbach's alpha on 63 items of MKT of written academic discourse. Firstly, the reliability of sections was calculated separately. Then, the overall reliability of the test was analyzed. The following table showed the reliability of the identification section, correction section, and explanation section of the test:

Table 4.17

Reliability Analysis of the Test Across Categories

	Cronbach's alpha	N of items
Identification	.81	19
Correction	.80	19
Explanation	.85	25

The results indicated that the explanation section has the highest internal consistency which meant that the section was highly consistent in measuring the same construct. Other sections also had good internal consistency values. When the overall reliability of the test was

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

calculated, it resulted in a value of .93 which showed that the test was highly reliable in measuring the constructs of the metalinguistic knowledge of academic discourse.

Table 4.18

Reliability Analysis of the Whole Test		
Cronbach's alpha	Cronbach's alpha based on standardized items	N of items
.93	.93	63

As the overall reliability of the test was reflected in the internal consistency of each section, it was concluded that the internal consistency of all the sections was good, so they contributed positively to the homogeneity of items among different sections and the overall reliability of the test. In addition, item-total statistics indicated that the reliability value could not be increased by deleting any of the items. It showed that all the test items were positively contributing to the overall reliability.

The reliability of the newly developed test was compared with the already developed tests on metalinguistic knowledge. It could be compared with SAT results; the test being used for the development of MKT of academic discourse. Board (2013) mentioned that SAT had a reliability value of .89-.91 which is quite close to the results of the current study. Similar results were gained by Al Farsi (2008), Bruce (1964), Ellis (2005) as they obtained high reliability values of .91 .93, .90, respectively.

RQ 8: Are the MCQs free from bias related to test-takers' gender and subject of study?

To see whether the MCQs are free from bias, a hypothesis was considered that **t**here is no effect of gender and subject on performance in Metalinguistic Knowledge Test of Discourse.

A two-way ANOVA was conducted to examine the effects of gender and subject area on the Metalinguistic Knowledge Test of Discourse (MKT_DIS). The main effect of gender on MKT_DIS scores was not significant, $F(1,366) = 0.518$, $p = .472F(1, 366) = 0.518$, $p = .472F(1,366) = 0.518$, $p = .472$, suggesting that there was no difference in MKT_DIS scores between males and females when the subject area was not considered. Similarly, the main effect of subject area on MKT_DIS scores was not significant, $F(2,366) = 1.038$, $p = .355F(2,366) = 1.038$, $p = .355F(2,366) = 1.038$, $p = .355$, indicating that the subject area alone did not significantly affect the MKT_DIS scores. The hypothesis that there is no main effect of gender on MKT_DIS scores was supported by the data, as the p-value was greater than .05. Similarly, the hypothesis that there is no main effect of subject area on MKT_DIS scores was also supported, with the p-value being greater than .05.

Table 4.20

Tests of Between-Subjects Effects

Source	Type III sum of squares	df	Mean square	F	Sig.
Corrected Model	2631.71[a]	5	526.3	3.210	.008
Intercept	151806.70	1	151806.7	925.912	.000
Gender	84.89	1	84.8	.518	.472

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Sub	340.20	2	170.1	1.038	.355
Error	60007.06	366	163.9		
Total	334313.00	372			
Corrected Total	62638.78	371			

Note. a. R Squared = .042 (Adjusted R Squared = .029)

The results indicate that the MKT_DIS test is largely free from gender or subject bias in isolation, as there were no significant main effects of gender or subject area on the scores. This suggests that neither gender nor subject area alone unfairly influences the test results. However, the significant interaction between gender and subject area implies that the test's scores are influenced by the combination of these factors. This means that while gender and subject area do not independently bias the test, there is a differential impact when these factors are considered together. Therefore, the MKT_DIS test may not be entirely free from bias, as the scores vary depending on the interaction between gender and subject area.

It can be concluded that the test was first analyzed to ensure validity measures, which proved that the test was valid. Then, the test items based on a criterion of item difficulty, item discrimination, and distractor effectiveness were analyzed resulting in a final test consisting of 63 items. This test was highly reliable, and gender and the factors did not independently bias the test. Future researches may focus on this domain to develop such tests for different audiences and provide reliable and valid translated versions of the same test. The translated versions may be compared with each other to check their effectiveness.

5. Conclusion

The purpose of the present study was to develop and validate a metalinguistic knowledge of written academic discourse based on the framework of Bialystok and Ryan (1985) and items in SAT writing and language section. The test included four passages containing 132 items resulting into 63 items based on certain analysis. Lastly, the bias in terms of gender and subject showed the absence of bias when the factors were taken independently.

The test is subject to several limitations. Although the data was collected from 385 participants representing certain faculties and departments, some of the departments were under-representative and needed more participants but it was not possible due to time constraints. In addition, convenience sampling was used instead of simple random sampling. Despite the limitations, this test can help educators evaluate a specific skill in the context of writing which was difficult to access because of the availability of only vocabulary, morphology, and grammar-based tests to assess the writing skill of the students. The test can help students understand their strengths and gaps necessary for learning and development. In addition, the test can help students prepare for standardized tests like SAT which are necessary to get admission to certain colleges in the US. In addition, some of the universities in Pakistan accept SAT scores for securing admission as well. So, the writing and language section preparation level can be evaluated using the newly developed test which is valid and highly reliable.

The present study has laid the foundation for the development of a unique test that contributes to the writing skills of students. Still, there is a need for further investigation in certain areas of it. Future researchers can focus on this area and develop more tests considering the level of students. There is also a need to develop tests in different

languages due to the scope and need for translation in today's world. Different versions (based on languages) of a test can be used for comparative studies to check their effectiveness and other measures. Researchers can also investigate the relationship between MK of written academic discourse and the writing proficiency of university students because writing is a compulsory activity in language classrooms. So, it will help in understanding where the students lack in terms of organization in writing.

References

Adegoke, B. A. (2013). Comparison of item statistics of physics achievement test using classical test and item response theory frameworks. *Journal of Education and Practice*, *4*(22), 87-96.

Al Farsi, B. (2008). Morphological awareness and its relationship to vocabulary knowledge and morphological complexity among Omani EFL University students. *Unpublished Master's Thesis, University of Queensland, St Lucia, Australia.*

Alderson, J. C., Clapham, C., & Steel, D. (1997). Metalinguistic knowledge, language aptitude and language proficiency. *Language teaching research*, *1*(2), 93-121.

Alegria, J., Pignot, E., & Morais, J. (1982). Phonetic analysis of speech and memory codes in beginning readers. *Memory & Cognition*, *10*, 451-456.

Alipour, S. (2014).Metalinguistic and linguistic knowledge in foreign language learners.
Theory and Practice in Language Studies, 4(12), 2640-2645.

Altheide, D. L., & Johnson, J. M. (1994). Criteria for assessing interpretive validity in qualitative research.

Aydin, F. (2018). L2 metalinguistic knowledge and L2 achievement among intermediate-level adult Turkish EFL learners. *Journal of Language and Linguistic Studies, 14*(1), 28- 49.

Bialystok, E. (1979). Explicit and implicit judgements of L2 grammaticality. *Language learning*, *29*(1), 81-103.

Bialystok, E. (1990). Communication strategies: A psychological analysis of second language use. (No Title).

Bialystok, E. (1994). Analysis and control in the development of second language proficiency. *Studies in second language acquisition*, *16*(2), 157-168.

Bialystok, E. (2001). Metalinguistic aspects of bilingual processing. *Annual review of applied linguistics, 21,* 169-181.

Bialystok, E. (2001). *Bilingualism in development: Language, literacy, and cognition*.
Cambridge University Press.

Bialystok, E., & Ryan, E. B. (1985). A metacognitive framework for the development of first and second language skills. *Metacognition, cognition, and human performance*, *1*, 207-252.

Board, C. (2013). Test characteristics of the SAT®: Reliability, difficulty levels, completion rates, January 2012–December 2012. Retrieved from http://media.collegeboard.com/digitalServices/pdf/research/Test-Characteristics- of-SAT2013.pdf

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.

Bruce, D. J. (1964). The analysis of word sounds by young children. *British Journal of Educational Psychology, 34*(2), 158-170.

Correa, M. (2011).Subjunctive Accuracy and Metalinguistic Knowledge of L2 Learners of Spanish. *Electronic Journal of foreign Language teaching, 8*(1).

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Creswell, J. W. (2005). *Educational Research: Planning, Conducting and Evaluating Quantitative and Qualitative Research* (2nd Ed.). Pearson Merrill Prentice Hall

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, *16*(3), 297-334.

Date, A. P., Borkar, A. S., Badwaik, R. T., Siddiqui, R. A., Shende, T. R., & Dashputra, A. V. (2019). Item analysis as tool to validate multiple choice question bank in pharmacology. *Int J Basic Clin Pharmacol*, *8*(9), 1999-2003.

DeVellis, R. F. (1991). Scale development: Theory and practice (Vol. 26).

Elder, C., & Manwaring, D. (2004).The relationship between metalinguistic knowledge and learning outcomes among undergraduate students of Chinese. *Language Awareness, 13*(3), 145-162.

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. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in second language acquisition*, *27*(2), 141-172.

Goldstein, D. (1974). Learning to read and developmental changes in covert speech and in word analysis and synthesis skill (Doctoral dissertation, University of Connecticut). *Dissertation Abstracts International, 35,* IB-606B. (University Microfilms No. 67-4246)

Gombert, J. E. (1992). Metalinguistic development Hertfordshire: Harverster Wheatsheaf.

Hertzog, M. A. (2008). Considerations in determining sample size for pilot studies. *Research in nursing & health, 31*(2), 180-191.

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

Hughes, A. (2020). *Testing for language teachers*. Cambridge university press.

Karmiloff-Smith, A., Grant, J., Sims, K., Jones, M. C., & Cuckle, P. (1996). Rethinking metalinguistic awareness: representing and accessing knowledge about what counts as a word. *Cognition*, *58*(2), 197-219.

Kirsh, D. (1992). When is information explicitly represented?.

Kurvers, J. J. H., van Hout, R. W. N. M., & Vallen, A. L. M. (2006). Discovering features of language: Metalinguistic awareness of adult illiterates.

Laufer, B., & Sim, D. (1985). An attempt to measure the threshold of competence for reading comprehension. *Foreign Language Annals*, *18*(5), 405-411.

Liberman, I. Y., Shankweiler, D., Fischer, F. W., & Carter, B. (1974). Explicit syllable and phoneme segmentation in the young child. *Journal of experimental child psychology, 18*(2), 201-212.

Linacre, J. M. (2007). *Winsteps*. http://www.winsteps.com/

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing research*, *35*(6), 382-386.

Malmqvist, J., Hellberg, K., Möllås, G., Rose, R., & Shevlin, M. (2019). Conducting the pilot study: A neglected part of the research process? Methodological findings supporting the importance of piloting in qualitative research studies. *International journal of qualitative methods, 18,* 1609406919878341.

Marsh, G., & Mineo, R. J. (1977). Training preschool children to recognize phonemes in words. *Journal of Educational Psychology*, *69*(6), 748.

McBRIDE–CHANG, C. A. T. H. E. R. I. N. E., Wagner, R. K., Muse, A., Chow, B. W.

Y., &Shu, H. U. A. (2005). The role of morphological awareness in children's vocabulary acquisition in English. *Applied psycholinguistics, 26*(3), 415-435.

Myhill, D., & Jones, S. (2015). Conceptualizing metalinguistic understanding in writing/Conceptualización de la competencia metalingüística en la escritura. *Cultura y Educación*, *27*(4), 839-867.

Nation, I. S. P., Heatley, A., & Coxhead, A. (2002). Range: A program for the analysis of vocabulary in texts [software].

Nation, I., & Beglar, D. (2007). A Vocabulary Size Test. *The Language Teacher, 31*, 9- 13.

Nunes, T., Bryant, P., & Bindman, M. (2006). The effects of learning to spell on children's awareness of morphology. *Reading and writing*, *19*, 767-787.

Paradis, M. (1994). Neurolinguistic aspects of implicit and explicit memory: implication for bilingualism and second language acquisition. Implicit and explicit learning of languages.

Patel, N., & Desai, S. J. I. J. P. S. R. R. (2020). ABC of face validity for questionnaire. *Int J Pharm Sci Rev Res*, *65*(1), 164-168.

Renou, J. (2001). An examination of the relationship between metalinguistic awareness and second-language proficiency of adult learners of French. *Language awareness, 10*(4), 248-267.

Ricciardelli, L. A. (1993). Two components of metalinguistic awareness: Control of linguistic processing and analysis of linguistic knowledge. *Applied Psycholinguistics*, *14*(3), 349-367.

Roehr, K. (2008). 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.Routledge.

Rosner, J. (1975). Helping children overcome learning difficulties: A step-by-step guide for parents and teachers.

ROSWELL-CHALL AUDITORY BLENDING TEST. (1959). New York: Essay Press

Sanosi, A. B. (2022). Correlation of EFL learners' metalinguistic knowledge and grammatical accuracy. *Studies in English Language and Education*, *9*(3), 908- 925.

Schmidt, R. W. (1990). The role of consciousness in second language learning1. *Applied linguistics, 11*(2), 129-158.

Sireci, S. G. (1998). The construct of content validity. Social indicators research, 45, 83- 117.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in science education*, *48*, 1273-1296.

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*, 1-8.

Tighe, E. L. (2015). *Assessing the importance of metalinguistic skills to the word readingand reading comprehension abilities of adult basic education students* (Doctoral dissertation, The Florida State University).

Tokunaga, M. (2010).Metalinguistic knowledge of low-proficiency university EFL learners.*In JALT2009 Conference Proceedings. Tokyo: JALT* (pp. 140-150).

Tong, X., Deacon, S. H., Kirby, J. R., Cain, K., & Parrila, R. (2011). Morphological awareness: A key to understanding poor reading comprehension in English. *Journal of Educational Psychology*, *103*(3), 523.

ISSN E: 2709-8273
aa ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Tong, X., Deacon, S. H., Kirby, J. R., Cain, K., & Parrila, R. (2011). Morphological awareness: A key to understanding poor reading comprehension in English. *Journal of Educational Psychology*, *103*(3), 523.

Tsuji, H., & Doherty, M. J. (2014). Early development of metalinguistic awareness in Japanese: Evidence from pragmatic and phonological aspects of language. *First Language, 34*(3), 273-290.

Vagias, W. M. (2006). Likert-type scale response anchors. clemson international institute for tourism. *& Research Development, Department of Parks, Recreation and Tourism Management, Clemson University*, *4*(5).

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

Yopp, H. K. (1988). The validity and reliability of phonemic awareness tests. *Reading research quarterly*, 159-177.

Yopp, H., & Singer, H. (1984). Are metacognitive and metalinguistic abilities necessary for beginning reading instruction. In *Changing perspectives on research in reading/language arts processing and instruction, Thirty-third yearbook of the National Reading Conference* (pp. 110-116).

Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A. R. (2015). Design and implementation content validity study: development of an instrument for measuring patient-centered communication. *Journal of caring sciences*, *4*(2), 165.

Zhang, R. (2015). MEASURING UNIVERSITY-LEVEL L2 LEARNERS'IMPLICIT AND EXPLICIT LINGUISTIC KNOWLEDGE. *Studies in Second Language Acquisition*, *37*(3), 457-486.

Ebel, R. L., & Frisbie, D. A. (1972). Essentials of educational measurement.