التحقق من صحة اختبار الوعي بالاستعارة (MAT) المطور بستخدام الذكاء الاصطناعي لاستكشاف كفاءة اللغة المجازية لدى طلاب الاكراد المتعلمين للغة الإنجليزية كلغة أجنبية

مدرس مساعد: ديار لطيف عمر مزوري
قسم اللغة الأنجليزية، كلية تربية الأساس، جامعة صلاح الدين–أربيل، أربيل، أقليم كردستان العراق
diyar.mzury@su.edu.krd
بروفيسور د. علي محمود جوكل
قسم اللغة الأنجليزية، كلية تربية الأساس، جامعة صلاح الدين–أربيل، أربيل، أقليم كردستان العراق
ali.jukil@su.edu.krd

الكلمات المفتاحية: الكناية، الأستعارة، اللغة المجازية، اختبار الوعي بالأستعارة.

كيفية اقتباس البحث

مزوري ، ديار لطيف عمر ، علي محمود جوكل ، التحقق من صحة اختبار الوعي بالاستعارة (MAT) المطور بستخدام الذكاء الاصطناعي لاستكشاف كفاءة اللغة المجازية لدى طلاب الاكراد المتعلمين للغة الإنجليزية كلغة أجنبية،مجلة مركز بابل للدراسات الانسانية، آب 2026،المجلد:16 ،العدد:8.

مفهرسة في Indexed
IASJ

مسجلة في Registered
ROAD

Validating an AI-Developed Metaphor Awareness Test (MAT) to Exploring Figurative Language Competence of Kurdish EFL Students

Assistant Lecturer: Diyar Lateef Omar Mzury
Department of English Language, College of Basic Education, Salahaddin University-Erbil, Erbil, Kurdistan Region of Iraq
diyar.mzury@su.edu.krd
Professor Dr. Ali Mahmood Jukil
Department of English Language, College of Basic Education, Salahaddin University-Erbil, Erbil, Kurdistan Region of Iraq
ali.jukil@su.edu.krd

Keywords : figurative language, metaphor, metaphor awareness, metonymy, Metaphor Awareness Tool (MAT)

How To Cite This Article
Mzury, Diyar Lateef Omar, Ali Mahmood Jukil, Validating an AI-Developed Metaphor Awareness Test (MAT) to Exploring Figurative Language Competence of Kurdish EFL Students , Journal Of Babylon Center For Humanities Studies, Auguts 2026,Volume:16, Issue8.

الملخص

تُعدّ الكفاءة في اللغة المجازية حجر الزاوية في إتقان اللغة على مستوى عالٍ، حيث تُعتبر الاستعارة والكناية المرسل من أهمّ الأساليب اللغوية، ويجعل انتشار اللغة المجازية في الخطاب الطبيعي إتقانها أمرًا بالغ الأهمية للمتعلمين لكي يتواصلوا بفعالية، أما الذين يفتقرون إلى هذه القدرة فينقطعون فعليًا عن كم هائل من المدخلات اللغوية، لذا، يجب أن تخضع أدوات قياس الوعي بالاستعارة للتحقق الإحصائي والمنهجي اللازم. تُطوّر هذه الدراسة اختبارًا للوعي بالاستعارة (MAT) باستخدام الذكاء الاصطناعي التوليدي، وتحديدًا ChatGPT. بعد تجربة

أولية شملت 15 طالبًا للتحقق من صحة الاختبار وموثوقيته، طُبّق الاختبار المُحسّن على 92 طالبًا جامعيًا تتراوح مستويات إتقانهم بين ما قبل المتوسط والمتوسط، وتضمّن تطبيق الاختبار تعليمات وتعريفات كافية قبل إجرائه لقياس البنية المستهدفة (الوعي)، يُشكّل هذا الاختبار أساسًا لدراسة تجريبية لاحقة تتضمن اختبارات قبلية وبعدية للاستعارة والمجاز المرسل في مهارات القراءة والكتابة، إلى جانب مواد تعليمية نصية ومتعددة الوسائط مُولّدة بواسطة الذكاء الاصطناعي، وأظهر المشاركون فرقًا ثابتًا في التعرّف على الاستعارة والكناية، كما واجهوا صعوبات في تحديد عناصر اللغة المجازية، مع أداء أقوى بشكل ملحوظ في الاستعارة مقارنة بالكناية، وتحقق الأداة الموثوقية السيكومترية من خلال التحسين والتعديل المستمرَّين، وتُعَدّ النتائج ذات صلة بتدريس اللغة المجازية لمتعلمي اللغة الإنجليزية الكرد كلغة أجنبية.

Abstract

Figurative language competence is a cornerstone of higher-level language proficiency, with metaphor and metonymy operating as master tropes. The pervasion of figurative language in natural discourse makes it crucial to the learners to communicate effectively, and those who lack such ability essentially become cut off from enormous amount of input. Tools to measure metaphor awareness must undergo proper statistical and methodological validation. This study adapts and develops a Metaphor Awareness Test (MAT) using generative AI, specifically ChatGPT. After a pilot conducted with 15 students checking validity and reliability, the refined MAT was administered to 92 undergraduate students at pre-intermediate to intermediate proficiency. The administration of the test involved sufficient directions and definitions prior to the test to measure the intended construct (awareness). The instrument serves as the baseline for a subsequent experimental study involving pre- and post-tests for metaphor and metonymy in reading and writing skills, alongside AI-generated textual and multimodal instructional materials. Participants demonstrated a consistent difference in recognizing metaphor and metonymy. They also showed challenges identifying elements of figurative language, with markedly stronger performance on metaphor than on metonymy. The tool achieves psychometric reliability through continuous improvement and adjustments, and the results are relevant for teaching figurative language to Kurdish EFL.

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 8

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 8

Introduction

Figurative language is no longer considered as an ornamental or peripheral feature of communication, but as an essential dimension of how speakers conceptualize the world through experiencing, structuring discourse, and negotiating meaning. Since the work on conceptual metaphor by Lakoff and Johnson (1980) and the subsequent consolidation of Cognitive Linguistics, metaphor and metonymy have been repositioned as cognitive operations pertinent to everyday cognition rather than deviations from a literal norm. This shift yields direct consequences to teaching second languages. Thus, the pervasion of figurative language in natural discourse makes learners who lack the awareness to detect and interpret it being cut off from a significant amount of authentic input. For EFL learners, the inability to identify and interpret figurative meaning is reflected in misreadings of texts, flattened written production, and difficulty being engaged with culturally embedded expressions whose force is precisely non-literal.

Despite this increasing appreciation, explicit formal instruction and systematic evaluation of figurative language competency remain variable or lacking in EFL contexts and virtually nonexistent in Kurdish institutions of higher education. Educators often postpone instruction on figurative language until an advanced level based on the belief that figurative language is a stylistic component of language rather than an integral cognitive tool (Liopis-Garcia 2024). Assessment procedures reflect this lapse in figurative language, as there are available adaptations of the Metaphor Awareness Test (MAT) for various populations (Chen and Lai 2012; Hilliard 2017); however, no validated MAT exists for Kurdish EFL learners, whose linguistic and cultural backgrounds may produce factors that influence figurative language comprehension in ways that imported instruments cannot properly assess.

The development of such instruments constitutes the second gap concerns. Earlier MATs have relied on native-speaker intuition for the construction of items, a practice that, while linguistically grounded, which adds subjective judgement into sentence structure, figurative–literal balance, and the calibration of difficulty. Rapid and recent developments in Natural Language Processing (NLP) and the emergence of large language models (LLMs) provide a complementary approach: AI-assisted item generation that can generate regulated and parametrically consistent sentence pairs at scale while still requiring human expert validation. It is yet an empirical question whether such AI-developed items can produce a psychometrically sound assessment instrument in an EFL context.

These overlapping gaps are addressed in the present study. Its purpose is to adapt, develop, and validate an AI-developed MAT using ChatGPT-generated items, and to use this instrument to explore the figurative language competence of Kurdish EFL undergraduates at the Department of English Language, College of Basic Education at the University of Salahaddin. In addition, the validated MAT is planned not only to serve as a stand-alone diagnosis phase, but as the empirical baseline for a subsequent experimental study involving pre- and post-tests in reading and writing skills and AI-generated educational materials that focus on teaching metaphor and metonymy.

Literature Review

Figurative Language, Metaphor, and Metonymy

Defining figurative language first requires distinguishing 'figurative' from 'literal' and 'non-figurative'. Traditionally, literal language refers to the standard meaning provided in dictionaries, while deviation from this truth would be figurative language (Glucksberg 2001). However, contemporary developments treat the distinction as a cognitive continuum rather than a rigid dichotomy. Recent work argues that figurative language is more dynamic and emotionally deeper than literal language, generating structural varieties in human perception that literal language cannot duplicate (Dori and Durgin 2025, p. 2). The idea that figurative languages create more extreme emotional responses than literal (Kövecses, 2020; Citron et al, 2020) has been supported Neuroscientific DATA by increasing numbers of research studies conducted utilizing modern neuroimaging technology; Generally, When Experience/knowledge relates to ones linguistic facility, More can be expressed through figurative constructions (Kövecses, 2020) that carry almost only the basic forms of meaning - generally speaking, by their basic-level forms (objects), and either Primary Emotional content (Kövecses, 2020 pp. 28-33).

According to Giora's Graded Salience Hypothesis (2003), the distinction between literal and figurative is not always a fixed line but a cognitive distinction in terms of salience as it impacts how we process information. For example, when your brain attempts to interpret an expression it does not always pick the literal/figure first. Rather, it will choose the most salient meaning of an expression in terms of how familiar or frequent that expression is to you — i.e., some expressions will be rated highly on a scale of salience; other expressions will be rated less so. In keeping with this premise, Glucksberg's findings from his 2001 research indicate that both figurative and literal expressions are processed in parallel and largely at the same rate as a literal expression would be processed; this

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 8

challenges the traditional Standard Pragmatic Model (Searle, 1979) of figurative expression processing as being serial (e.g., after going through the literal expression first), and subordinate to literal expressions.

Although figurative language encompasses a wide range of tropes, metaphor and metonymy are the most frequent, serving as the main organizers of cognitive architecture and conceptual system. Littlemore (2024) explains that both are dealt with as figurative and cognitive operations: metaphor involves a comparison relationship between two discrete domains (source and target), while metonymy expresses a contiguity relationship between foreground and background within a single domain. Dancygier and Sweetser (2014, p. 4) define figurative as "a usage motivated by a metaphoric or a metonymic relationship to some other usage, a usage which might be labeled literal," where literal refers not to everyday talk but to meaning not dependent on figurative extension. According to Littlemore (2024, p. 125), metaphors and metonymy are "two major cognitive processes at the centre of most human thought and communication". This importance of metaphor and metonymy is also seen within translation research, with Denroche (2023, p. 173) proposing that metaphors and metonymy be considered to be 'master tropes'; thus, when other tropes are examined, it is with reference to metaphors and metonymy.

Teaching Metaphors and Metonymies

The pedagogical imperative of metaphor and metonymy results from their class as fundamental components of everyday communication rather than ornaments of speech. This view grounds the shift from a literal-first approach to the integration of figurative language teaching. Ma, Zhang and Parr (2025) argue that raising metaphoric awareness contributes to figurative language competence, which fosters broader language skills. Explicit teaching strengthens learners' conscious recognition of the conceptual motivation behind an expression—the why behind how it is conceptualized and articulated. Boers (2021) shows that learners' retention of newly encountered lexical items improves when they grasp this conceptual motivation: absorbing TIME IS MONEY enables learners to logically categorize save time, waste time, and spend time.

Despite this, the view persists in teaching communities that figurative language instruction should be reserved for advanced proficiency, with claims that "metaphor reflects an advanced use by a minority of speakers" (Littlemore & Low 2006, p. 269). Yet metaphor is pervasive and embedded in daily experience regardless of speaker level, and metaphoric competence is a "basic feature of native-speaker competence" (Danesi 1993, p. 493). Language learners process figurative language differently

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 8

from native speakers, and without explicit instruction it is unlikely that learners will actively engage with figurative language (Liopis-Garcia 2024).

Metaphor Awareness Test (MAT)

The Metaphor Awareness Test (MAT) is an assessment tool intended to examine students' capability to identify metaphorical terms in their target language (Chen and Lai 2012). It asks students to evaluate expressions on a scale reflecting whether they view the language as figurative or literal. The test comprises 40 sentences—20 figurative and 20 literal—randomized in order. Hilliard (2017) adapted the MAT and administered it to assess metaphorical competence. For the present study, it was adapted to fit the proficiency level of second-year university undergraduates.

The purpose of the test is to examine learners' comprehension of figurative expressions and, more broadly, to surface pedagogical insights and variances in response to figurative language. Results from the MAT inform the subsequent development of (a) pre- and post-tests on metaphor and metonymy in reading and writing, and (b) AI-generated instructional materials—both textual and multimodal—to be evaluated in an experimental study for difficulty, applicability, item selection, cultural sensitivity, and other pertinent issues.

Unlike previous MATs (Hilliard 2017; Chen and Lai 2012), the sentences in the present instrument were generated using ChatGPT through specific descriptive prompts and iterative modifications. The prompts controlled for English language proficiency level, sentence-structure variety, word-choice difficulty, and sentence length, constrained to 9–11 words per item. This integration of AI signifies a turning point in the way that language will be assessed, as the previous method of assessment using MATs relied solely upon judgement by a native speaker, which provides a level of subjectivity, while ChatGPT as a well-established LLM provides authentic sentences written in English with a greater level of methodological consistency. Recent advancements made in the field of Natural Language Processing (NLP) support the position that AI can create metaphors in language; as demonstrated by Stowe et al. (2021), AI models can produce metaphoric expressions that are both coherent and acceptable in context by way of conceptual mappings rather than through the sole use of surface patterns, which was also evident through the work of Gargett and Branden (2015) in the creation of brand new metaphors from the combination of corpus-based modelling and AI modelling. Even with this strong reliance on AI, the principle of human augmentation was intact as all generated examples were shared with a jury of experts for

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 8

determining content validity and provided insights regarding the distinction between metaphor and simile, examples where an item fit into both categories of metaphor and metonymy, and finally, in terms of vocabulary, they provided examples of vocabulary that may present an overload for the students.

Methodology

Participants

The participants were second-year students from the Department of English Language at the College of Basic Education, Salahaddin University. The sample was made up of 92 (N=92) year two EFL students aged between 19-21 years, who were located in two intact student populations. All participants are assumed to possess similar proficiency in English because they have the same educational history based on the same curriculum. All of the participants will receive ten weeks of direct instruction in metaphor and metonymy, with one receptive skill (reading) and one productive skill (writing).

Instrument (Metaphor Awareness Test)

The MAT is composed of 40 randomized sentences (20 of each type) where each literal and figurative sentence in the set is composed of the same vocabulary. For example, the literal use of the verb ‘to boil’ (The water boiled in the pot creating steam.) was developed alongside its figurative use (His anger boiled within him until he stormed out). A Likert scale (1-5) was given to each sentence with 1= completely literal and 5= completely figurative.

The construction of these items was completed by ChatGPT based upon prompt specificity of proficiency level, sentence structure, difficulty of word choice, and length (9 to 11 words). The items generated by ChatGPT underwent expert jury review to ensure that they did not suffer from there being metaphor/simile confusion, that they did not allow for both metaphorical and metonymical use of the same vocabulary, and that they did not use vocabulary that was beyond the range comprehended by learners.

Procedure

A brief introduction to the concept of figurative language was provided to all participants before they took the test. Participants received directions and definitions for individual words in Kurdish, the participants' native language, prior to the test in order to measure the intended construct (awareness). This way, testing was free from external sources of error, including ambiguous or misreading of words or not understanding word meanings. Access to online dictionaries and mobile devices was not allowed.

Results: Statistical Analysis

Piloting MAT

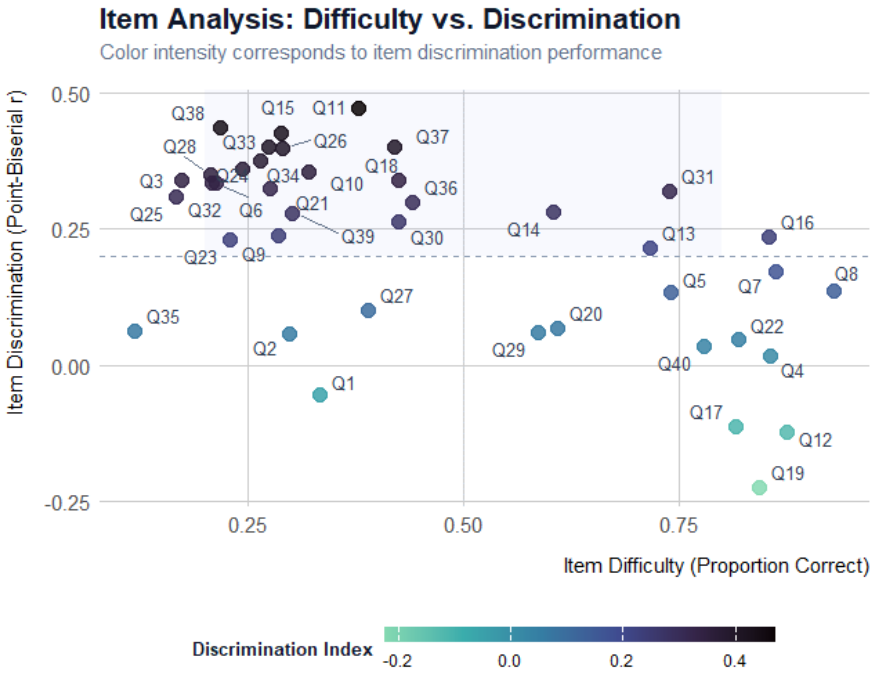
Earlier pilot testing was done with a sample of 15 second-year students at the College of Basic Education — the evening program sample resembles the day class samples in cultural/educational background so they can provide appropriate feedback regarding content validity of the tool and any necessary modifications will be made.

Pilot Study Results *(Sample Size: 40 items – 20 figurative and 20 literal)*

Item Analysis: Difficulty and Discrimination

Item difficulty values ranged from 0.12 to 0.93, reflecting a widespread student performance. Some items (e.g., Q3, Q25, Q35) were very difficult, while others (Q8, Q7, Q12) were very easy. Items of moderate difficulty (0.3–0.7) are preferable, since they discriminate ability best.

Figure (1) Item Analysis: Difficulty vs. Discrimination



Q11 (0.47), Q10 (0.35), Q38 (0.44) showed good discrimination amongst high and low achievers. A few (Q1, Q12, Q19) were negatively discriminating and may need to be revised or removed.

Reliability Analysis

The use of Cronbach's alpha to measure the internal consistency of the 40-item scale produced an alpha value of .615. This figure is below the usual standard of .70 for strong reliability; however, as a result of being

used early on in exploratory studies measuring attitudes related to psychological constructs, this value is acceptable.

Some of the items that had negative and very low correlations based on the item-total correlation data will not provide a helpful contribution to the overall structure of the measurement scale. Negative corrected item-total correlations were observed for Q1, Q4, Q5, Q7, Q8, Q12, Q17, Q19, and Q40. Detaching such items should enhance Cronbach's alpha with the largest gain (to 0.652) if Q19 is erased. Conversely, several items contributed strongly to internal consistency: Q6 (0.417), Q11 (0.459), Q25 (0.475), Q26 (0.384), Q32 (0.362), Q35 (0.370), and Q38 (0.495).

Table (2) Pilot reliability statistics.

Cronbach's Alpha			**Number of Items**		
0.615			**40**		
Item	**Corrected Item-Total Correlation**	**Cronbach's Alpha if Item Deleted**	**Item**	**Corrected Item-Total Correlation**	**Cronbach's Alpha if Item Deleted**
Q1	-0.141	0.636	Q21	0.234	0.603
Q2	0.299	0.597	Q22	-0.176	0.635
Q3	0.320	0.597	Q23	0.400	0.587
Q4	-0.296	0.639	Q24	0.343	0.592
Q5	-0.126	0.633	Q25	0.475	0.584
Q6	0.417	0.588	Q26	0.384	0.587
Q7	-0.074	0.625	Q27	0.209	0.604
Q8	-0.307	0.637	Q28	0.299	0.598
Q9	0.333	0.593	Q29	0.135	0.611
Q10	0.342	0.591	Q30	0.170	0.608
Q11	0.459	0.580	Q31	0.063	0.617
Q12	-0.223	0.633	Q32	0.362	0.591
Q13	0.046	0.618	Q33	0.338	0.595
Q14	0.101	0.614	Q34	0.348	0.593
Q15	0.310	0.597	Q35	0.370	0.593
Q16	-0.033	0.622	Q36	0.248	0.600
Q17	-0.297	0.645	Q37	0.271	0.599
Q18	0.231	0.602	Q38	0.495	0.579
Q19	-0.447	0.652	Q39	0.261	0.600
Q20	0.036	0.620	Q40	-0.213	0.638

These findings suggest that certain items may be misaligned or not functioning effectively. As a result, it is advisable to refine the tool by

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 8

changing or eliminating items that show negative correlations. Additionally, it's possible that the instrument addresses several underlying concepts (such as figurative versus literal language or metaphor versus metonymy), which could account for the moderate alpha value observed.

Construct Validity

Confirmatory Factor Analysis (CFA) was carried out on the pilot sample data to evaluate whether the items being measured are valid; there are two key latent constructs assessed with this CFA: figurative (20 items) and literal (20 items).

Several figurative items had consistent substantial sub-factors, including Q5 (0.680), Q12 (0.918), and Q19 (0.891), so all three of these items effectively measured understanding figurative language. Unfortunately, several other figurative items including Q14, Q24, and Q34 had weak sub-factor loadings or negative sub factor loadings so therefore cannot be used to measure understanding figurative language. In addition, strong sub-factor loadings were found for four literal items (Q9, Q10, Q25, and Q27) at 0.881, 0.952, 0.787, and 0.722 respectively. Many literal items would not have had any statistical significance because of low variability and/or large sampling error for those items.

Table (3) Pilot CFA factor loadings.

Figurative			Literal		
Item	**Factor Loadings**	**SE (P-Value)**	**Item**	**Factor Loadings**	**SE (P-Value)**
Q4	0.379		Q1	0.099	
Q5	0.68	0.922 (0.052)	Q2	-0.109	2.826 (0.696)
Q7	0.393	0.572 (0.070)	Q3	0.557	10.395 (0.587)
Q8	0.48	0.463 (0.006)	Q6	0.598	10.203 (0.552)
Q12	0.918	1.245 (0.052)	Q9	0.881	15.631 (0.568)
Q13	0.451	0.678 (0.079)	Q10	0.952	16.824 (0.566)
Q14	-0.023	0.435 (0.888)	Q11	0.55	9.297 (0.549)
Q15	-0.171	0.521 (0.385)	Q18	0.648	11.437 (0.566)
Q16	0.467	0.909 (0.175)	Q23	0.645	11.640 (0.574)
Q17	0.631	0.616 (0.007)	Q25	0.787	14.388 (0.579)

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 8

Q19	0.891	1.177 (0.046)	Q26	0.378	7.073 (0.588)
Q20	0.336	0.676 (0.190)	Q27	0.722	12.575 (0.561)
Q21	-0.527	0.907 (0.125)	Q29	0.1	2.039 (0.621)
Q22	0.559	0.833 (0.077)	Q30	0.533	9.400 (0.565)
Q24	-0.346	0.600 (0.128)	Q32	0.166	4.516 (0.710)
Q28	-0.207	0.552 (0.323)	Q33	0.123	2.953 (0.672)
Q31	-0.126	0.563 (0.554)	Q35	0.651	1.177 (0.566)
Q34	-0.473	0.638 (0.050)	Q37	-0.101	11.485 (0.710)
Q36	-0.062	0.481 (0.732)	Q38	0.279	2.759 (0.586)
Q40	0.719	0.897 (0.034)	Q39	0.586	5.184 (0.342)

In conclusion, the pilot analysis indicates that the measurement tool has a solid starting point to measure both figurative and literal language but would benefit from additional refinement of items that have weak reliability characters (e.g., low reliability, poor discrimination, low construct loading). Therefore, the items that performed well for both constructs will remain intact for future use.

Results

The final versions of the MAT were given to individuals after they had completed the refinement described above. The methods used in the pilot to introduce figurative and literal language were also used to introduce the use of figurative and literal language in these MAT items. Where necessary, instructions and word meanings were provided in Kurdish.

On average, the participants were able to identify figurative vs literal language 67% of the time with an accuracy level of 0.67 across all MAT items. Almost half of the participants scored above 70% on these items based on the slightly greater than average median scores achieved by participants. Thus, the results show that participants have a moderately developed ability to understand figuratively vs literally and that there is considerable variability in this understanding across participants as evidenced by the standard deviation of 0.13.

Figure (4) Descriptive accuracy statistics.

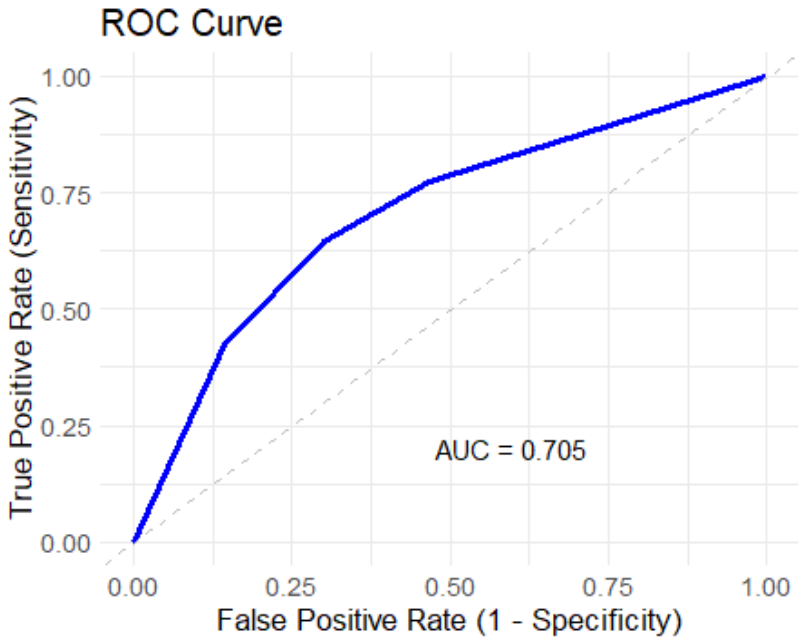
Measurements	Value
Mean Accuracy	0.67
SD (Standard Deviation)	0.13
Median Accuracy	0.70
Min Accuracy	0.33
Max Accuracy	0.9
Skewness	-0.39
Kurtosis	-0.57
Standard Error	0.01

Participants with varying recognition ability in the sentences, ranging from very poor (0.33) to very good (0.90), also indicated considerable variability measured by a range of 0.57. A kurtosis value of -0.57 shows a distribution that is flatter than normal would be expected to look, reflecting an expected absence of extreme outliers. The skewness score of -0.39 shows a small to moderate left skew, with above average values received by a greater number of participants than below average value scores.

The mean is highly reliable as demonstrated by very low standard error (0.01).To investigate the degree to which the instrument can differentiate between figurative and literal use of language, an AUC was obtained through Receiver Operating Characteristic (ROC) analysis. The resulting AUC of 0.7052 indicates moderate discrimination, and is therefore considered acceptable for educational settings as values greater than or equal to 0.70 are regarded as acceptable.

Figure (5) ROC curve and discrimination analysis

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 8



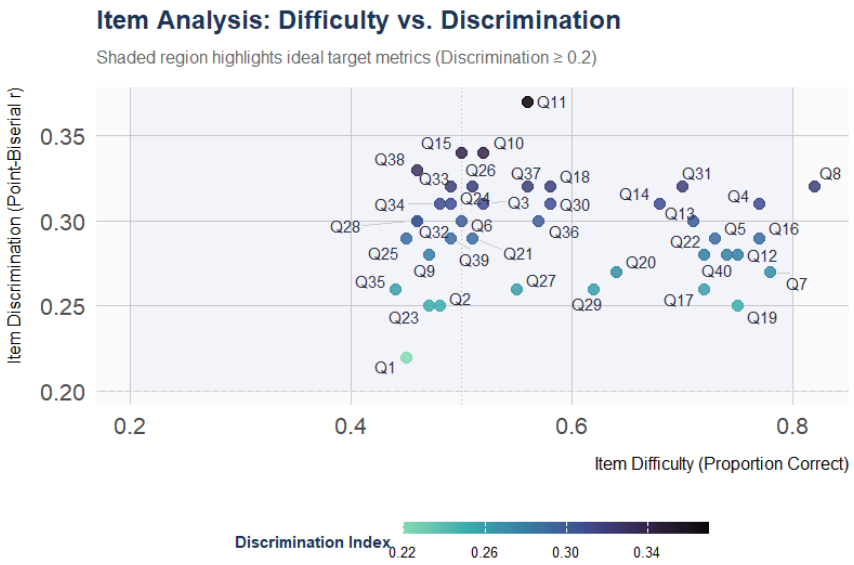
.

Item Difficulty and Discrimination

Assessing how difficult each item is for participants is called item difficulty; determining how effectively an item discriminates between test-takers who have a high level of ability and one who does not is termed item discrimination.

The MAT items had difficulty values that were appropriate and moderate (0.44 to 0.78), indicating that there was an appropriate balance throughout the assessment. This appropriate range of difficulty values supports continuous engagement with the assessment and allows the assessment to measure figurative and literal comprehension at several levels. The item discrimination indexes from the MAT assessment were from .22 to 0.37 and were classified as good to excellent, meaning that the items were able to successfully discriminate between those test-takers that have a clear understanding of the constructs, and test-takers that do not; thereby further demonstrating that the MAT assessment is both valid and reliable.

Figure (6) Item Analysis: Difficulty vs Discrimination.



Chi-Square Test

To see if differences between the responses were statistically significant (i.e., test-takers' responses were more than just random guessing and were indicative of having recognised the difference between figurative and literal meanings), we ran a Chi-square analysis on the response data. As a result of the analysis, we found a significant and strong relationship between the participants' responses and the type of meaning of the sentences completed by the participants ($\chi^2 = 349.73$, $p < 0.001$). Participants identified the type of meaning in sentences with far greater frequency than would have occurred by chance alone. This further demonstrated the ability of the MAT to accurately assess the understanding of figurative language and figurative expressions.

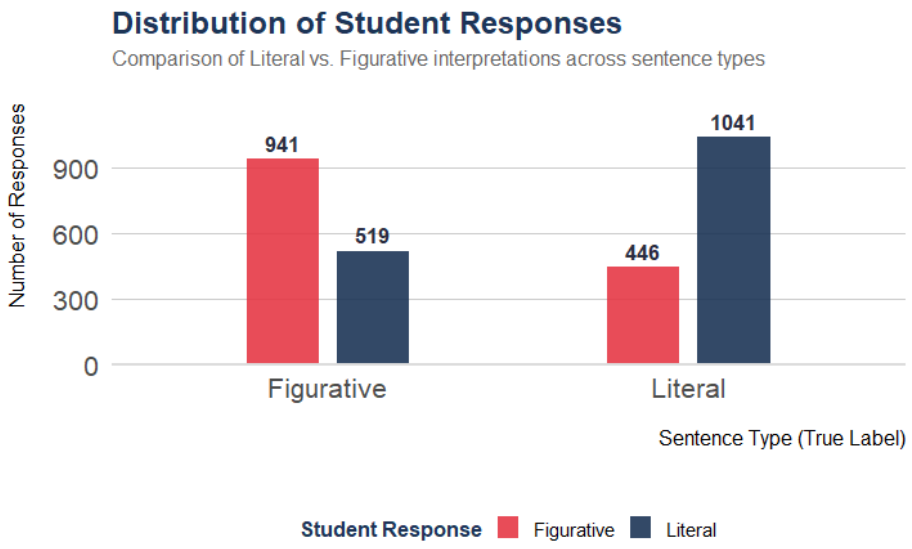
Table (7) Chi-square contingency of responses.

True Sentence Type	Student Responded Figurative	Student Responded Literal	Total
Figurative	941	519	1,460
Literal	446	1,041	1,487
Total	1,387	1,560	2,947

As previously stated, students provided higher levels of accuracy when interpreting and identifying metaphors (76.0% accuracy) than when interpreting metonymies (52.7% accuracy), leading to an implication that students perceive metaphors as being easier to interpret and identify than

metonymies. There could be several explanations for this difference. However, the most common explanation relates to the implicit/context dependent, rather than the explicit/objective, nature of metonymy compared to metaphor. Results of this observation were then considered for selection and placement of metonymies in future administration of pre- and post-tests and instructional materials, with aim to increase frequency, visibility and pedagogical appropriateness in metonymic items.

Figure (8) Participants' Responses vs. True Responses.



Testing Reliability

The MAT's reliability testing guarantees that the internal measurement used by the MAT produces accurate and consistent data that reflects the construct measured with reliability and consistency across 40 items. Cronbach's alpha indicated a high level of internal consistency for these 40 items (0.823), indicating that participants' responses were stable. The corrected item-total correlations for the items were between 0.40 and 0.53, and all items contributed positively to the total score. Furthermore, none of the items had near-zero or negative correlations with the total; thus, all of the items had an appropriate relationship with the construct measured.

Table (9) Final reliability statistics.

Cronbach's Alpha			Number of Items		
0.823			40		
Item	**Corrected**	**Cronbach's**	**Item**	**Corrected**	**Cronbach's**

	Item-Total Correlation	Alpha if Item Deleted		Item-Total Correlation	Alpha if Item Deleted
Q1	0.425	0.812	Q21	0.461	0.81
Q2	0.512	0.807	Q22	0.434	0.811
Q3	0.487	0.808	Q23	0.513	0.806
Q4	0.431	0.811	Q24	0.488	0.808
Q5	0.462	0.81	Q25	0.525	0.804
Q6	0.489	0.807	Q26	0.5	0.807
Q7	0.402	0.813	Q27	0.45	0.81
Q8	0.439	0.81	Q28	0.476	0.809
Q9	0.517	0.805	Q29	0.422	0.812
Q10	0.523	0.804	Q30	0.437	0.811
Q11	0.49	0.807	Q31	0.415	0.812
Q12	0.438	0.81	Q32	0.49	0.807
Q13	0.416	0.812	Q33	0.475	0.809
Q14	0.4	0.813	Q34	0.482	0.808
Q15	0.482	0.808	Q35	0.478	0.809
Q16	0.423	0.812	Q36	0.455	0.81
Q17	0.455	0.81	Q37	0.442	0.81
Q18	0.498	0.807	Q38	0.53	0.803
Q19	0.445	0.81	Q39	0.44	0.81
Q20	0.412	0.812	Q40	0.428	0.812

Validity and Comparative Fit Index (CFI)

The purpose of this study was to evaluate the MAT as an accurate measure of figurative language awareness. A Comparative Fit Index (CFI) score was computed for both literal and figurative items, based on how well they clustered together in the two-factor model. The CFI results indicated that the sentences adequately assessed both literal and figurative understanding as intended, with evidence of construct validity related to the use of CFA. The factor loadings of the individual items with their respective latent variables were statistically significant ($p < 0.01$) and well above the conventionally acceptable lower limit of 0.40, thus supporting the accuracy of the measurement of the constructs for which the items have been developed.

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 8

Table (10) Final CFA factor loadings.

Figurative			Literal		
Item	**Factor Loadings**	**SE (P-Value)**	**Item**	**Factor Loadings**	**SE (P-Value)**
Q4	0.512		Q1	0.423	
Q5	0.681	0.075 (0.002)	Q2	0.528	0.072 (0.002)
Q7	0.576	0.068 (0.001)	Q3	0.579	0.063 (0.001)
Q8	0.59	0.071 (0.001)	Q6	0.603	0.060 (0.001)
Q12	0.738	0.058 (0.001)	Q9	0.711	0.058 (0.001)
Q13	0.544	0.073 (0.001)	Q10	0.726	0.057 (0.001)
Q14	0.392	0.066 (0.004)	Q11	0.576	0.062 (0.001)
Q15	0.431	0.065 (0.003)	Q18	0.645	0.060 (0.001)
Q16	0.613	0.063 (0.001)	Q23	0.633	0.061 (0.001)
Q17	0.652	0.061 (0.001)	Q25	0.692	0.058 (0.001)
Q19	0.71	0.057 (0.001)	Q26	0.512	0.065 (0.001)
Q20	0.503	0.069 (0.002)	Q27	0.703	0.059 (0.001)
Q21	0.534	0.071 (0.001)	Q29	0.429	0.070 (0.002)
Q22	0.588	0.068 (0.001)	Q30	0.585	0.062 (0.001)
Q24	0.476	0.066 (0.002)	Q32	0.463	0.067 (0.001)
Q28	0.438	0.062 (0.001)	Q33	0.417	0.069 (0.003)
Q31	0.385	0.072 (0.004)	Q35	0.71	0.057 (0.001)
Q34	0.579	0.061 (0.001)	Q37	0.654	0.060 (0.001)
Q36	0.41	0.068 (0.003)	Q38	0.435	0.068 (0.002)
Q40	0.693	0.059 (0.001)	Q39	0.484	0.066 (0.001)

To evaluate the validity component, both the Root Mean Square Error of Approximation (RMSEA) and the Standardized Root Mean Square Residual (SRMR) were measured. The RMSEA score was 0.038, which is well below the normal cutoff of 0.05, and the 90% confidence interval was [0.025, 0.048] indicating the model fit the data well. The SRMR was 0.054, which is also below the acceptable limit (< 0.08), indicating the model fit the data well based on how it reproduced the covariances.

Table (11) Model fit indices.

Fit Index	Value	Threshold / Interpretation
Chi-Square (scaled)	735.121	p = 0.083 → Acceptable
Degrees of Freedom	701	–
Comparative Fit Index (CFI)	0.921	Good (> 0.90)

Tucker-Lewis Index (TLI)	0.913	Good (> 0.90)
RMSEA (scaled)	0.038	Good (< 0.05)
RMSEA 90% CI	[0.025, 0.048]	Indicates close model fit
SRMR	0.054	Acceptable (< 0.08)

These findings suggest that the MAT is a reliable tool backed by strong psychometric validation. It can be effectively utilized in the Kurdish EFL context to assess understanding of both literal and figurative language among the targeted group.

Discussion

The findings highlight three interconnected aspects: a) Kurdish EFL learners demonstrate a consistent difference in recognizing metaphor and metonymy; b) an AI-generated MAT can achieve psychometric reliability through continuous improvement; and c) these results are relevant for teaching figurative language to Kurdish students learning English as a foreign language.

The striking discrepancy between results of metonymy identification (52.7%) and metaphor identification (76.0%) is the most noticeable pattern. The near-chance performance on metonymy items is consistent with a long-standing finding in cognitive linguistics that, despite being similarly prevalent, metonymy is less cognitively observable than metaphor. While metaphor depends on a perceptible cross-domain mapping—a source domain projected onto a target domain—metonymy operates within a single domain through contiguity, foregrounding one element to access another. This single-domain action tends to be treated as ordinary referential language, particularly when the metonymic relation is highly conventional, for instance, container-for-content, institution-for-people. Many metonymic expressions are likely to go entirely unnoticed by EFL learners as they do not feel nonliteral and are therefore not marked as figurative on a Likert-scaled judgement task. A parallel explanation is being provided by the Graded Salience Hypothesis (Giora 2003) as conventional metonymies are frequently very salient and may be immediately interpreted as literal statements, leaving learners with no perceptual cue that figurative processing is required. The unevenness in our data is thus not coincidental; rather, it reflects a true cognitive distinction that figurative language instruction is rarely openly addressed in EFL contexts.

Given that previous MATs (Chen and Lai 2012; Hilliard 2017) grounded items in actual native usage while leaving room for subjective variation in difficulty, length, and figurative density, a second consideration related

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 8

the methodological shift from native-speaker intuition to AI-assisted item generation. Compared to intuition-based development, ChatGPT produced a more evenly calibrated instrument by enabling parametric control along precisely these dimensions—9–11-word sentences, paired literal–figurative items sharing the same core vocabulary, and stylistic consistency across the 40-item set. Nevertheless, it is vital to explain that AI generation cannot perform separately. The expert jury identified issues that the model was not successful to surface, for example, metaphor–simile confusion, terminologies that were beyond the specified lexical range of the learners, statements admitting both metaphoric and metonymic readings. A merely feasible approach is AI supplementing human expertise rather than replacing it, which is a principle of human augmentation in which the LLM speeds up and standardizes generation as domain experts arbitrate conceptual and pedagogical adequacy.

For Kurdish EFL classes, pedagogical implications follow directly. Firstly, the moderate overall accuracy (mean = 0.67) and the wide score range (0.33–0.90) show that figurative language competence of second-year undergraduates is neither consistently developed nor reliably acquired through incidental exposure, thus, challenging the common presumption that figurative language belongs to advanced levels of instructional phases. Secondly, the asymmetric performance of metaphor and metonymy show that the two tropes cannot be treated interchangeably in pedagogical materials. In particular, metonymy requires targeted and explicit instruction that articulate and highlight the single-domain contiguity relations, for example through foregrounding conventional patterns such as PLACE FOR INSTITUTION or PART FOR WHOLE across both writing and reading. Third, the validated MAT is a useful and insightful diagnostic instrument to determine which students require specific figurative-language scaffolding prior to encountering authentic reading texts receptively or using them productively in writing tasks. These implications directly inform the next phase of the larger research study, in which the MAT serves as baseline to inform the level and quality of AI-generated instructional materials and for pre- and post-tests in reading and writing, and subsequently the material that would be taught in the experimental study.

It is important to acknowledge the limitation of the instrument. Although the number of participants (92) are sufficient for presented study here, it should be cautiously acknowledged that the drawn sample come from a single institution and cohort, and generalization to other populations of Kurdish EFL groups will not be rational. Additionally, MAT also assesses recognition rather than production; a learner who can recognize a

metaphor on a Likert scale may still be unable to employ one in writing, the subsequent experiment is designed to address this distinction.

Conclusion and Suggestions

This research investigates the adaptation, development, and validation of an AI-generated Metaphor Awareness Test (MAT) to assess the figurative language proficiency of Kurdish EFL undergraduate students. The psychometric analysis, from a pilot Cronbach's alpha of .615 to a robust final value of .823, demonstrated that the MAT was a stable measure of figurative awareness for this population after minor revisions through expert review and item analysis. The final MAT yielded a mean accuracy of .67, a discriminative area under the curve (AUC) of .7052, and a statistically significant deviation from chance (χ^2=349.73, p<.001) when administrated to 92 second-year undergraduate students at the University of Salahaddin. The data showed that the participants were able to identify metaphors with greater accuracy (76.0%) than metonymy (52.7%), demonstrating that metonymy is a less cognitively transparent and a less-studied trope than metaphor.

In addition to these key findings, this study represents a methodological case for using generative AI to create language assessment tools. The use of ChatGPT provided greater consistency in the item development process than is commonly found in traditional, intuitive methods to develop assessment instruments, yet maintained the ability of human experts to create conceptual and pedagogical measures. As a consequence, the MAT can now be used for two purposes: as a stand-alone tool to assess figurative awareness within a Kurdish EFL context and as a baseline measure for a planned experimental phase, in which AI-generated texts and multimodal materials will be used to teach reading and writing using metaphors and metonymies.

Suggestions for future directions based on this research include: (1) develop early, explicit EFL instruction in Kurdish programs to teach figurative language and especially metonymy; (2) aid in the development of additional MATs for similar populations through the use of AI-generated items—larger item pools, item control for conceptual metaphor type, or matched parallel forms for pre-and post-testing; (3) replicate this study with larger, multi-institutional sample sizes across a range of EFL proficiency; and (4) create complementary assessment tools to measure productive figurative competence through writing and speaking, in conjunction with recognition-based assessment tools such as the MAT. Finally, the broader integration of generative AI into language assessment merits sustained methodological attention: principles of human augmentation, expert validation, and transparent reporting of prompt

Journal of Babylon Center for Humanities Studies: 2026, Volume: 16, Issue: 8

design should be foregrounded if AI-developed instruments are to earn the same scrutiny and trust as traditionally constructed counterparts.

References

Boers, F. (2021) 'Glossing and vocabulary learning', *Language Teaching*, 55(1), pp. 1–23.

Chen, Y., & Lai, H. (2012) EFL Learners' Awareness of Metonymy–Metaphor Continuum in Figurative Expressions. Language Awareness, 21(3), 235–248.

Citron, F. M. M., Michaelis, N., & Goldberg, A. E. (2020) Metaphorical language processing and amygdala activation in L1 and L2. Neuropsychologia, 140, 107381.

Danesi, M. (1992) 'Metaphorical competence in second language acquisition and second language teaching: The neglected dimension', in Alatis, J.E. (ed.) *Georgetown University Round Table on Languages and Linguistics: Language, Communication and Social Meaning*. Washington, DC: Georgetown University Press, pp. 489–500.

Dancygier, B., & Sweetser, E. (2014) *Figurative Language*. Cambridge: Cambridge University Press.

Denroche, C. (2023). *Metonymy and Language: A New Theory of Linguistic Processing*. London: Routledge.

Dori, S. and Durgin, F.H. (2025) 'The dynamics of figurative language processing', *Language and Cognition*, 36(1), 1–22.

Gargett, A., and Barnden, J. (2015) Gen-Meta: Generating novel metaphorical expressions from corpus models', Web Intelligence, 13, pp. 103-114.

Giora, R. (2003). *On Our Mind: Salience, Context, and Figurative Language*. Oxford: Oxford University Press.

Glucksberg, S. (2001) *Understanding Figurative Language: From Metaphors to Idioms*. Oxford: Oxford University Press.

Hilliard, A. (2017) *Using Cognitive Linguistics to Teach Metaphor and Metonymy in an EFL and ESL Context*, Unpublished PhD thesis, University of Birmingham).

Kövecses, Z. (2020) Extended Conceptual Metaphor Theory. Cambridge: Cambridge University Press.

Lakoff, G., and Johnson, M. (1980) *Metaphors We Live By*. Chicago: University of Chicago Press.

Liopis-Garcia, R. (2024) *Figurative Language Pedagogy in EFL: Theory and Practice*. Bristol: Multilingual Matters.

Littlemore, J. (2024) *Metaphor and Metonymy: An Introduction*. Edinburgh: Edinburgh University Press.

Littlemore, J., and Low, G. (2006) *Figurative Thinking and Foreign Language Learning*. Basingstoke: Palgrave Macmillan.

Ma, T., Zhang, L.J. and Parr, J.M. (2025) 'Developing and validating instruments for measuring English-as-a-second/Foreign-Language (L2) learners' metaphor awareness', *Language Awareness*, 34(1), pp. 195–230.

Searle, J.R. (1979) *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge: Cambridge University Press.

Stowe, A.M., Utama, P., & Klakow, D. (2021) Metaphor generation with conceptual mappings', P*roceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 1234-1245.