Austin Journal of Psychiatry and Behavioral Sciences

Open Access

Austin Publishing Group

Research Article

Modern AI and WHO-Inspired Adaptation of the GAD-7 Psychometric Assessment into Bengali

Agarwal A[1†], Banerjee S[1†], Datta S[2†], Chakraborty K[1], Ghosh P[1] and Singh E[1]

[1]Datalabs, United We Care, India

[2]Department of Biochemistry, M.J.N Medical College, India

†Co First Authors with Equal Contribution and Listing Order Random. Sandipan Conducted the Study, Ayushi Performed the Statistical Analysis, Sourav Conceptualised the Research

***Corresponding author:** Sourav Banerjee, Founder and CTO - United We Care, India

Email: sb@unitedwecare.com

Received: April 16, 2025

Accepted: April 28, 2025

Published: May 02, 2025

Abstract

Generalized Anxiety Disorder (GAD) is a mental health condition worldwide. The Generalized Anxiety Disorder Assessment (GAD 7) is a tool used to diagnose and assess the level of GAD. However, there is currently no validated Bengali version of this tool, which makes it challenging to provide culturally relevant mental health support for the large Bengali-speaking community. This study aimed to adhere to the WHO Translation Methodology to translate, culturally adapt, and validate the GAD-7 for Bengali-speaking populations. The goal was to promote inclusivity and enhance mental health care delivery. The study employed a rigorous translation methodology, including forward and backward translation, expert panel review, and cultural adaptation. The translated tool was then validated using statistical and Natural Language Processing methods. We observed substantial success in translating the GAD-7 scale from English to Bengali, achieving high cosine similarity scores (~0.95) and robust Bleu RT scores (~0.71), indicating satisfactory semantic and syntactic alignment with reference translations. The Bengali-translated GAD-7 proved reliable and valid for assessing GAD among Bengali-speaking populations. This cultural adaptation can foster the advancement of mental health research in underrepresented populations and improve mental health services accessibility and efficacy. Further research is needed to examine the tool's clinical effectiveness and sensitivity.

Keywords: Generalised Anxiety Disorder; GAD-7; Bengali; Translation; Validation; Cultural Adaptation; Mental Health

Introduction

Generalised Anxiety Disorder (GAD) [1], as described in the Diagnostic and Statistical Manual of Mental Health Disorders (5th ed.; DSM-5; American Psychiatric Association, 2013) [2], is a persistent mental health condition characterised by uncontrollable feelings of anxiety and worry lasting for at least six months. With a lifetime prevalence rate of around 5.7%, GAD is among the most common anxiety disorders, as indicated by the 2011 National Comorbidity Survey Replication [3]. The impact of GAD extends beyond the individual, leading to physical complaints affecting work performance and potentially paving the way for other conditions, such as different anxiety disorders and issues related to alcohol consumption [4]. In instances, GAD can result in disabilities that hinder self-care relationships with others and healthcare access. Therefore, early identification and proper management of GAD are crucial in minimising these effects, easing personal distress and reducing societal expenses associated with GAD (Kertz, Bigda-Peyton, & Björgvinsson, 2013) [5].

Developed by Spitzer and colleagues (2006) [1], the Generalised Anxiety Disorder Scale-7 (GAD-7) is a self-administered tool to enhance GAD recognition. Its simplicity, robust reliability, and validity have been demonstrated across various contexts. However, its utility is not universal, as the lack of quality translations into certain languages impedes its global applicability, specifically among non-English speaking populations.

One group facing this challenge is the Bengali-speaking population, the sixth-largest linguistic group and the third-largest ethnic group worldwide, after the Han Chinese and Arabs[6]. Despite their significant numbers, the absence of a reliable Bengali version of the GAD-7 highlights an evident gap in providing mental health care and research.

Current translation techniques in cross-cultural psychology often lack cultural sensitivity and linguistic accuracy, thus yielding versions of psychometric tests that may not be as reliable or valid as their original counterparts. These translations risk failing to capture the original text's cultural nuances and intended meanings, thereby potentially offering an inaccurate portrayal of the psychological phenomena being studied.

Existing translation methodologies are diverse, ranging from back-translation to consulting linguistic professionals, from organising focus groups to undertaking a basic 1:1 translation (Brislin, 1970 [7]; Peña, 2007 [8]; Wild et al., 2005 [9]). Once a translation is performed, factor analysis is typically conducted to compare the statistical behaviours of the items on the original and translated tests. Further statistical tools such as confirmatory factor analysis (CFA) or structural equation modelling (SEM) are used to confirm the statistical performance of the translated test relative to the original (Fischer & Karl, 2019) [10].

Citation: Agarwal A, Banerjee S, Datta S, Chakraborty K, Ghosh P. Modern AI, WHO-Inspired Adaptation of the GAD-7 Psychometric Assessment into Bengali Austin J Psychiatry Behav Sci. 2025; 11(1): 1106.

However, these procedures bear significant limitations. Translation and back-translation methods, while useful, may overlook key cultural subtleties. Moreover, applying statistical manipulations might unintentionally produce biases. If cross-cultural equivalence isn't achieved, the psychometric measures don't perform statistically similarly in the original and target cultures, and the tests are often discarded, heavily altered, or artificially manipulated until they are usable. This, in turn, can compromise the integrity of the outcomes.

As previous research shows, achieving a high-quality translation of mental health instruments demands a comprehensive approach. This approach extends beyond simple literal translation to encompass cultural adaptation, thereby ensuring the relevance and acceptability of the tool in the target population (Brislin, 1970 [7]; Flaherty et al., 1988 [12]). However, such comprehensive methodologies have been inconsistently applied in translating tools like the GAD-7, often resulting in subpar translations that fail to evaluate GAD symptoms in non-English speaking populations accurately. Consequently, the call for better and more culturally nuanced translation methods in cross-cultural psychology remains essential.

Recognising the necessity for superior translation methods, our study concentrates on translating the GAD-7 from English to Bengali. The GAD-7 was chosen due to its robust psychometric properties, ease of administration, and wide acceptance globally in clinical and research settings. Our objective is to broaden the reach of this important mental health tool to Bengali speakers, enabling the early detection and treatment of GAD in this demographic and fostering more inclusive mental health research.

In our research, we propose a novel method that utilises advanced Artificial Intelligence (AI) models, specifically Instructor XL (1.5 billion parameters) [23], GTR T5 XXL (5 billion parameters) [22], and DeBERTa XXL (1.5 billion parameters) [24], in conjunction with the cosine similarity measure for evaluating translations. Cosine similarity [25], a measure that computes the cosine of the angle between two vectors, is an excellent tool for evaluating the semantic similarity between the original and translated versions. When paired with these advanced metrics such as Jaccard Similarity, Precision, Recall, and F1 Score along with a host of Natural Language Processing Metrics such as NistMT [26], ROUGE [21], Super GLUE [28], TER [29], WER [31], BleuRT [20],

BLEU [19], and Meteor [30] which are renowned for their ability to comprehend and generate human-like language, these techniques serve distinct yet interconnected purposes in the assessment process. For instance, Jaccard Similarity measures the overlap between two data sets, providing insights into the similarity of the original and translated texts. Precision and Recall, on the other hand, help evaluate the quality of the translation, identifying the extent to which it has correctly interpreted and conveyed the meaning of the original text. The F1 Score combines Precision and Recall to provide a single measurement of translation accuracy. This pairing allows us to capture the essence of the original text while accommodating the linguistic and cultural nuances of the target language.

However, while these methods significantly improve the process, they are not immune to human language and cultural complexities. The myriad subtleties and nuances that make language rich and diverse can challenge even the most advanced techniques. Therefore, this study aims to illuminate these potential shortcomings and explore possible solutions. In this research paper, we comprehensively examine our proposed method, which incorporates both qualitative and quantitative elements to evaluate the efficacy of translations. On the qualitative front, our approach is grounded in the rigorous methodology outlined by the World Health Organization (WHO) [15], ensuring a meticulous and culturally sensitive translation process.

On the other hand, the quantitative aspect of our methodology is powered by advanced Artificial Intelligence (AI) models and the measure of cosine similarity. These techniques allow us to precisely assess the translations' performance by comparing the semantic similarity between the original and translated versions. We delve into how this innovative approach improves upon existing methodologies, scrutinising its potential limitations and providing valuable suggestions for future research. Furthermore, we will demonstrate how this updated methodology could be extrapolated beyond the confines of this study, presenting potential advantages to a wide array of areas in cross-cultural psychology and psychometric testing.

We also underscore the need for regular updates to official assessments, translation methodologies, evaluation criteria, and assessment translations. Given the dynamic nature of language and cultural expressions, these resources must be periodically reviewed and revised to stay relevant and effective. Such a proactive approach ensures that these tools accurately reflect contemporary linguistic and cultural nuances, contributing to the accuracy and cultural sensitivity of cross-cultural psychological assessments.

Methodological Approach

This study employs a framework [36] that combines WHO translation methods with AI techniques for cross-cultural adaptation of mental health assessments. While prior methods focus on single-instrument translations, this framework enables simultaneous adaptation of multiple instruments: the Patient Health Questionnaire (PHQ-9), Generalized Anxiety Disorder-7 (GAD-7), Perceived Stress Scale (PSS), and Socrates 8A/8D. The methodology integrates AI validation metrics with WHO protocols to maintain cultural and linguistic alignment. The framework's application in translating these tools from English to Bengali provides a model for future cross-cultural adaptations in mental health assessment.

Findings

Qualitative Analysis Linguistic Nuances

Translating the GAD 7 questionnaire from English to Bengali presented some hurdles. Words like "nervous " and "anxious". On edge" posed challenges as there are no direct equivalents in Bengali that capture the precise emotional nuances. For instance, "nervous" typically describes a state of restlessness stemming from uncertainty or fear while "anxious" conveys a sense of unease or concern about possibilities. The expression "on edge" colloquially signifies feeling tense, anxious or easily irritated. Each of these terms in English carries nuanced differences in emotional states. To replicate this complexity in Bengali, the team used phrases such as "ঘাবড়ে যা ন, উ গ্ন হ ন অথবা কিনারা এসে গেছেন এমন বোধ করছেন" (Feeling afraid, anxious, or feeling like coming to

an edge). Here, linguistic precision was balanced with the emotional resonance of the terms, ensuring the items were understandable and relatable to the Bengali-speaking population.

Semantic Nuances

The Bengali translation of the feedback question concerning how anxiety issues affect daily functioning had to reflect a broad array of everyday activities, interpersonal interactions, and responsibilities potentially impacted by anxiety disorders. The goal was to create a culturally relevant translation that adequately conveyed these aspects.

Take, for example, the response categories 'Not at all', 'Several days', 'More than half the days', and 'Nearly every day'. Each option holds a distinct semantic weight in English, demonstrating an increasing degree of frequency and intensity. The translation into Bengali had to uphold this severity order carefully. As such, 'Several days' was translated to 'কয়েক দিন' and 'More than half the days' to 'দিনগুলির অর্ধেকের চেয়ে বেশি'. This translation maintains a clear depiction of the frequency and keeps the progression of severity intact.

Moreover, the gradient of the response options, ranging from 'Not difficult at all' to 'Extremely difficult', was particularly important. The challenge lay in maintaining the relative significance of each category when translated into Bengali. We needed to capture the severity and progression of struggle or discomfort appropriately.

Cultural Nuances

A key component of our translation strategy was to consider how mental health symptoms might be perceived, understood, or expressed within the cultural context of Bengali communities. Significant differences between these cultural perceptions and those of Western societies might exist. For instance, articulating feelings of anxiety may not follow the same patterns as in English-speaking cultures.

In Bengali culture, people may be more likely to express anxiety through physical symptoms such as 'মন ভারি লাগছে' (my heart feels heavy) or 'মাথা ব্যাথা করছে' (my head is aching). This cultural understanding of anxiety expression was critical to our translation process, ensuring that the questions were appropriate and resonated with Bengali speakers' experiences.

Dialect Robustness and Awareness

Slight deviations were noted during back-translation. One instance includes the phrase "How often have you been bothered by the following problems?" which was back-translated as "How often have you been embarrassed by any of the following problems?" These minor deviations, while not significantly altering the intended meaning, still introduce subtle shifts in interpretation (Table 1).

Quantitative Analysis

Statistical Analysis: To assess the translation's proficiency, a comprehensive range of performance metrics was employed. The results are summarised in the table below (Table 2).

Basic Statistical Metrics

Jaccard Similarity (Score: 0.4615): The Jaccard Similarity Index measures the intersection over the union between the translations

Table 1: Dialect Robustness and Awareness.

Item#	GAD7- Original Item	Translation from English to Bengali	Final translation from English to Bengali post-bi-lingual feedback	Back translation from Bengali to English
Instruction	Over the last two weeks, how often have you been bothered by the following problems?	গত দু সপ্তাহে, আপনিকতবারনীচের যেকোনও সমস্যার কারণে বিব্রত হয়েছেন?	গত দু সপ্তাহে, আপনিকতবার নীচের সমস্যাগুলির কারণে বিব্রত হয়েছেন?	How often have you been embarrassed by any of the following problems in the past two weeks?
Q1	Feeling nervous, anxious, or on edge	ভীত, উদ্বিগ্ন বা কিনারায় এসে গেছেন এমন বোধ করছেন?	ঘাবড়ে যান, উদ্বিগ্ন, নাঅথবা কিনারায় এসেগেছেনএমনবোধকরছেন	Feeling afraid, anxious, or feeling like coming to an edge
Q2	Not being able to stop or control worrying	দুশ্চিন্তানবন্ধকরতেবানিয়ন্ত্রণকরতে পারছেন না?	দুশ্চিন্তাথেকেবেরহতেবানিয়ন্ত্রণকরতে পারছেন না	Unable to stop or control worrying
Q3	Worrying too much about different things	বিভিন্ন বিষয়ে খুব বেশি চিন্তিত?	বিভিন্ন বিষয়ে খুব বেশি চিন্তা করছেন	Too much worried about various things
Q4	Trouble relaxing	নিশ্চিন্ত থাকতে সমস্যা হয় ?	নিশ্চিন্তথাকতেসমস্যা হয়	Facing problems staying relaxed
Q5	Being so restless that it is hard to sit still	এত আরিতা যে রি হয়ে বসে থাকা কঠিন?	এত আরিতা যে রি হয়ে বসে থাকা কঠিন হয়ে উঠছে	Feeling it difficult to sit quietly due to excessive restlessness
Q6	Becoming easily annoyed or irritable	সহজেই বিরক্ত বা খিটখিটে হয়ে উঠছেন	সহজেই বিরক্ত বা খিটখিটে হয়ে উঠছেন	Getting annoyed or irritated easily
Q7	Feeling afraid, as if something awful might happen	ভয়েলাগছে,যেআপনারসাথেখারাপকি কিছুঘটতেপারে?	ভয় পানে, যে আপনারসাথেখারাপকিছু হতে পারে	Feeling afraid that something bad might happen you
Q1 - O1	Not at all	একদমই না	একদমই না	Not at all
Q1 - O2	Several days	বেশকিছুদিনধরে	বেশকিছুদিনধরে	For quite some days
Q1 - O3	More than half the days	অর্ধেকদিনেরবেশিসময়	অর্ধেকের৫০বেশিদিন	More than half of the days
Q1 - O4	Nearly every day	প্রায়েপ্রতিদিন	প্রায়েপ্রতিদিন	Almost every day
Feedback Question (FQ)	If you checked any problems, how difficult have they made it for you to do your work, take care of things at home, or get along with other people?	আপনি যদি কোনও সমস্যা চিহ্নিত করে থাকেন, তাহলে এই সমস্যাগুলি আপনার কাজ করা, বাড়ির জিনিসপত্তরযত্ননেওয়াঅন্যলোকেদে সাথে মেলামেশা করা কতটা কঠিন করে তুলেছে আপনার জন্য	আপনি যদি কোনও সমস্যা চিহ্নিত করে থাকেন, তাহলে এই সমস্যাগুলি আপনার জ আপনার কাজ করা, বাড়ির জিনিসপত্তরযত্ননেওয়া অন্য লোকেদের সাথে মেলামেশা করা কতটা কঠি করে তুলেছে?	If you have identified any problems, then how do these problems make it difficult for you to work, take care of things at home, and socialise with people?
FQ-O1	Not difficult at all	একদমই কঠিন না	একদমই কঠিন না	Not at all difficult
FQ-O2	Somewhat difficult	কিছু সময় কঠিন	কিছু সময় কঠিন	Sometimes Difficult
FQ-O3	Very difficult	খুব কঠিন	খুব কঠিন	Very Difficult
FQ-O4	Extremely difficult	খুবখুব কঠিন	খুবখুব কঠিন	Extremely Difficult

Table 2: Qualitative Analysis of GAD-7.

GAD-7	Metric	Score
Statistical Methods	Jaccard Similarity	0.4615
	Precision	0.6353
	Recall	0.6279
	F1 Score	0.6316
Cosine Similarity	Instructor XL	0.9436
	GTR T5 XXL	0.9216
	DeBERTa XXL	0.9896
NLP Metrics	NistMT Score	29.5974
	Rouge Score	rouge-1': {'r': 0.5888, 'p': 0.6022, 'f': 0.5955}, 'rouge-2': {'r': 0.2758, 'p': 0.2909, 'f': 0.2831}, 'rouge-l': {'r': 0.5333, 'p': 0.5454, 'f': 0.5393}
	Super Glue Score	0.5117
	TER	0.4596
	WER	0.0066
	BleuRT	0.7079
	BLEU	bleu': 0.2550, 'precisions': [0.6694, 0.3583, 0.2016, 0.1101], 'brevity_penalty': 0.9437, 'length_ratio': 0.9453, 'translation_length': 121, 'reference_length': 128
	Bert Score	Precision' :0.6757, 'Recall' : 0.6261, 'F1-Score' : 0.6520
	Meteor	0.6273

and the reference corpus. The score of 0.46 suggests a moderate concurrence, indicating some overlap between the translated and reference text.

Precision (Score: 0.6353): Precision assesses the accuracy and relevance of the translated content. With a score of approximately 0.64, around two-thirds of the translated content is germane and aligns with the reference text.

Recall (Score: 0.6279): Recall indicates the proportion of relevant content captured in the translation. The score of 0.63 implies that approximately 63% of the pertinent content was accurately translated.

F1 Score (Score: 0.6316): The F1 Score is the harmonic mean of precision and recall, striking a balance between the two metrics. The computed score of around 0.63 signifies a balanced performance in preserving accuracy and completeness in the translation task

Cosine Similarity Metrics

DeBERTa XXL (Score: 0.9896): DeBERTa XXL achieved an exceptional similarity score of approximately 0.99, indicating a highly accurate and precise translation that closely matches the target reference.

GTR T5 XXL (Score: 0.9216): With a similarity score of 0.92, GTR T5 XXL also demonstrates a high degree of concordance with the target translation, reflecting the model's effectiveness.

Instructor XL (Score: 0.9436): Instructor XL obtained a similarity score of 0.94, further reinforcing the high similarity between the translations and the reference corpus, attesting to the model's accuracy.

Natural Language Processing Metrics

NistMT (Score: 29.5974): The NistMT Score quantifies the translation quality from an evaluation perspective. The reported score of

29.60 suggests satisfactory translation performance.

Rouge: Rouge-1': {'r': 0.5888, 'p': 0.6022, 'f':

0.5955}, 'rouge-2': {'r': 0.2758, 'p': 0.2909, 'f':

0.2831}, 'rouge-l': {'r': 0.5333, 'p': 0.5454, 'f':

0.5393}: These scores indicate moderate to good alignment between the system-generated translation and the reference text for single words (Rouge-1) and bigrams (Rouge-2). The Rouge-L score, considered the longest common subsequence, also suggests reasonable agreement between the translation and reference text. Further refinement may improve the alignment, particularly for bigrams (Rouge-2), to enhance overall translation quality.

Super Glue (Score: 0.5117): The Super Glue Score of around 0.51 showcases the effectiveness of the deployed translation, indicating a successful translation process.

TER (Score: 0.4596): The Translation Error Rate (TER) score of approximately 0.46 suggests moderate translation inaccuracies, which can be improved for better quality.

WER (Score: 0.0066): The Word Error Rate (WER) score of 0.0066 indicates high word-level translation accuracy, contributing to the overall quality of the translation.

B**leuRT (Score: 0.7079):** The BLEU RT Score of approximately 0.71 suggests a high quality of translation compared to the reference text, reflecting a successful translation outcome (Figure 1).

B**LEU (Score: bleu': 0.2550, 'precisions': [0.6694, 0.3583, 0.2016, 0.1101], 'brevity_penalty': 0.9437, 'length_ratio': 0.9453, 'translation_length': 121, 'reference_length': 128):** The BLEU Score, computed at 0.26, indicates moderate similarity between the system-generated translation and the reference text. The precision values show higher accuracy for individual words (1-gram) but decreasing

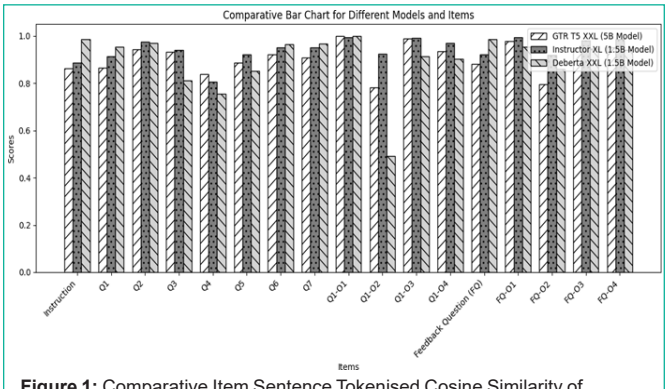

Figure 1: Comparative Item Sentence Tokenised Cosine Similarity of Forward and Reverse Translation of GAD-7.

Table 3: Cosine Similarity of Source Document and Reverse Translation of GAD-7.

Forward Translation	Reverse Translation	Sentence tokenisation (forward)	Sentence tokenisation (reverse)	GTR T5 XXL (5B Model)	Instructor XL (1.5B Model)	DeBERTa XXL (1.5B Model)
Over the last two weeks, how often have you been bothered by the following problems?	How often have you been embarrassed by any of the following problems in the past two weeks?	Over the last two weeks, how often have you been bothered by the following problems?	How often have you been embarrassed by any of the following problems in the past two weeks?	0.8635	0.8880	0.9858
Feeling nervous, anxious, or on edge	Feeling afraid, anxious, or feeling like coming to an edge	Feeling nervous, anxious, or on edge	Feeling afraid, anxious, or feeling like coming to an edge	0.8662	0.9147	0.9553
Not being able to stop or control worrying	Unable to stop or control worrying	Not being able to stop or control worrying	Unable to stop or control worrying	0.9435	0.9769	0.9712
Worrying too much about different things	Too much worried about various things	Worrying too much about different things	Too much worried about various things	0.9318	0.9398	0.8117
Trouble relaxing	Facing problems staying relaxed	Trouble relaxing	Facing problems staying relaxed	0.8378	0.8059	0.7540
Being so restless that it is hard to sit still	Feeling it difficult to sit quietly due to excessive restlessness	Being so restless that it is hard to sit still	Feeling it difficult to sit quietly due to excessive restlessness	0.8878	0.9227	0.8515
Becoming easily annoyed or irritable	Getting annoyed or irritated easily	Becoming easily annoyed or irritable	Getting annoyed or irritated easily	0.9221	0.9514	0.9659
Feeling afraid, as if something awful might happen	Feeling afraid that something bad might happen to you	Feeling afraid, as if something awful might happen	Feeling afraid that something bad might happen to you	0.9075	0.9515	0.9680
If you checked any problems, how difficult have they made it for you to do your work, take care of things at home, or get along with other people?	If you have identified any problems, then how do these problems make it difficult for you to work, take care of things at home, and socialise with people?	If you checked any problems, how difficult have they made it for you to do your work, take care of things at home, or get along with other people?	If you have identified any problems, then how do these problems make it difficult for you to work, take care of things at home, and socialise with people?	0.8809	0.9209	0.9866
Not at all	Not at all	Not at all	Not at all	1.0000	0.9938	1.0000
Several days	For quite some days	Several days	For quite some days	0.7831	0.9253	0.4916
More than half the days	More than half of the days	More than half the days	More than half of the days	0.9899	0.9915	0.9147
Nearly every day	Almost every day	Nearly every day	Almost every day	0.9354	0.9692	0.9041
Not difficult at all	Not at all difficult	Not difficult at all	Not at all difficult	0.9778	0.9932	0.9550
Somewhat difficult	Sometimes Difficult	Somewhat difficult	Sometimes Difficult	0.7947	0.9192	0.8593
Very difficult	Very Difficult	Very difficult	Very Difficult	0.9488	0.9813	0.9456
Extremely difficult	Extremely Difficult	Extremely difficult	Extremely Difficult	0.9527	0.9860	0.9419

accuracy for longer phrases (2-gram, 3-gram, and 4-gram). The Brevity Penalty of 0.9437 suggests a minor length difference between the translation and reference, while the Length Ratio of 0.9453 confirms their relative similarity. The translation appears concise, with a Translation Length of 121 and a Reference Length of 128. Overall, the BLEU Score suggests room for improvement in aligning the translation more closely with the reference text, and the precision values offer insights into specific areas for enhancement.

BERT (Score: Precision':0.6757, 'Recall': 0.6261, 'F1-Score': 0.6520): The Bert Score evaluates the translation's precision, recall, and F1-Score. The reported values suggest a balanced performance in the translation task.

Meteor (Score: 0.6273): The Meteor score of 0.63 reflects the effectiveness of the translation, indicating a satisfactory translation quality with room for further enhancements.

Neural Network Cosine Similarity with Sentence Tokenisation: The cosine similarity analysis was conducted to assess the performance of the forward and reverse translations of the GAD-7 scale using three different language models: GTR T5 XXL (5B Model), Instructor XL (1.5B Model), and DeBERTa XXL (1.5B Model). The cosine similarity scores were computed for each scale item, comparing the translated versions with the original English phrases.

The results indicate a high level of similarity between the translated and original English versions for most scale items. For instance, the phrase "Over the last two weeks, how often have you been bothered by the following problems?" achieved high cosine similarity scores of 0.8635 with GTR T5 XXL, 0.8880 with Instructor XL, and 0.9858 with DeBERTa XXL. These scores indicate strong linguistic consistency between the translations and the original English, suggesting successful translation and retention of the original meaning and intent.

Similarly, other items like "Feeling nervous, anxious, or on edge," "Not being able to stop or control worrying," and "Worrying too much about different things" also demonstrated high cosine similarity scores across all language models, with scores above 0.90, indicating successful translation and minimal loss of meaning.

However, some items displayed relatively lower cosine similarity values, such as "Several days," which achieved a cosine similarity score

of 0.7831 with GTR T5 XXL and 0.4916 with Instructor XL. These lower scores

suggest slight deviations in translating these phrases from the original English. An example of this is the phrase, "If you checked any problems, how difficult have they made it for you to do your work, take care of things at home, or get along with other people?" which achieved a cosine similarity score of 0.8809 with GTR T5 XXL and 0.9866 with Instructor XL. Though the scores are still relatively high, the slight variation may indicate the need for further refinement or cultural adaptation for these phrases (Table 3).

The cosine similarity analysis revealed a high level of similarity between the forward and reverse translations of the GAD-7 scale using the GTR T5 XXL (5B Model), Instructor XL (1.5B Model), and DeBERTa XXL (1.5B Model). This indicates successful translation with minimal loss of original meaning, rendering the translated tool suitable for use in the Bengali language.

Please refer to the Appendix for a comprehensive item-level review of the translation metrics, including Jaccard similarity, Precision, Recall, F1 Score, NistMT Score, Rouge Score, Super Glue Score, TER, WER, BleuRT, BLEU, Bert Score, and Meteor. The detailed analysis in the Appendix provides deeper insights into the scale items' translation quality and linguistic consistency.

The next section will discuss the detailed findings, interpreting the metric performances and critiquing where necessary to offer a robust assessment of the translated tool's validity and applicability in the local context.

Discussion

Navigating the complexities of the translation process presented several opportunities for in-depth linguistic exploration and cultural understanding. Translating idiomatic expressions that lacked direct Bengali equivalents necessitated a creative and nuanced approach to ensure semantic equivalence and cultural appropriateness. Preserving the original item's meaning was crucial in achieving accurate translations, especially when dealing with temporal markers and linguistic patterns. The back-translation process served as a valuable tool to identify and resolve discrepancies, enhancing the overall accuracy and reliability of the translations.

However, it is important to acknowledge certain limitations of the qualitative analysis. While the bilingual review and feedback contributed significantly to refine translations, qualitative assessments may be somewhat subjective. To mitigate this, future studies could incorporate a quantitative assessment to complement the qualitative findings. Additionally, metrics like the Jaccard Index may not be perfect measures of semantic equivalence and could be further improved or supplemented with other linguistic similarity metrics.

In light of the research findings, several suggestions can be made for professionals designing psychometric assessment tests like the GAD-7. It is advisable to avoid using idiomatic expressions in the original items to ensure more straightforward and precise translations.

Furthermore, stakeholders, including the WHO and clinical communities, should consider regular updates to the translation process as linguistic patterns and cultural nuances evolve over time. This would ensure the continued accuracy and relevance of the translated content.

Future research in this field could focus on enhancing AI models with more sophisticated language understanding capabilities to capture the subtleties of translation better. Exploring real-time adaptation to linguistic changes can ensure culturally sensitive and reliable assessments in the long run. The WHO protocol for translation could also be updated to incorporate advancements in AI-driven translation methods, providing more comprehensive guidelines for cross-cultural translations.

While the AI metrics used in our study provide valuable quantitative data on translation quality, we also conducted a thorough cosine similarity analysis to gain deeper insights into the translation variations. This additional analysis allowed us to examine smaller nuances that the traditional metrics may not fully capture.

In the cosine similarity analysis, we observed that seemingly minor differences, such as the absence of a question mark, a period, or proper capitalisation, could significantly impact the cosine similarity score. These small variations in punctuation and capitalisation can lead to slightly different semantic representations of the translated sentences, influencing the overall cosine similarity between translations.

For example, let's consider the translation of the item "Feeling nervous, anxious, or on edge." In one instance, the AI-generated translation reads, "ভীত, উদ্বিগ্ন বাকিনারা এসে গেছেন এমন বোধ করছেন?" Meanwhile, another translation reads, "ঘাবড়ে যান, উদ্বিগ্ন হন অথবাকিনারা এসে গেছেন এমন বোধ করছেন." Despite conveying similar meanings, the presence or absence of a question mark after "করছেন" results in different cosine similarity scores, indicating the sensitivity of the analysis to subtle linguistic variations.

Such findings underscore the importance of meticulous attention to detail when evaluating machine-generated translations. Inaccuracies caused by these nuances may not be readily apparent when relying solely on automated metrics, but they can significantly affect the overall quality and interpretability of the translated content.

We emphasise the need for a human-in-the-loop approach to address these subtleties and improve translation accuracy. Expert linguists and translators can thoroughly review and validate the translations, considering these smaller linguistic nuances to ensure a more accurate and culturally appropriate translation.

Moreover, incorporating user feedback and engaging with the target population can further enhance the translation process. This participatory approach allows us to identify specific linguistic preferences and better adapt translations to suit the end-user's needs, thereby improving the overall translation quality and user experience.

The future holds promising opportunities for advancing our understanding of mental health across diverse populations. To achieve this, several critical areas warrant exploration and development. First and foremost, there is a need to advance AI models and natural language processing techniques to better account for linguistic nuances, idiomatic expressions, and cultural subtleties. Fine-tuning existing models and designing language-specific AI frameworks can significantly enhance translation accuracy and cultural appropriateness.

Additionally, a human-in-the-loop approach will continue to be vital in translation. Expert linguists, mental health professionals, and community representatives can contribute their insights and validate AI-generated translations, ensuring they are culturally appropriate and sensitive. Furthermore, engaging the target population through user feedback and participatory workshops will provide valuable insights into language preferences, idiomatic expressions, and cultural norms, leading to translations that resonate more effectively with the intended audience.

Longitudinal studies will be crucial in tracking linguistic changes, particularly in rapidly evolving languages or communities experiencing language shifts. This will ensure that translations remain up-to-date and relevant, reflecting the current linguistic patterns of the target population. Additionally, conducting validation studies across diverse cultural and linguistic groups is essential for assessing the cross-cultural validity of translated instruments and identifying potential biases or inaccuracies in translations.

Comparative studies between different approaches and AI models will be essential to make informed decisions about translation strategies. Understanding how different techniques perform in specific linguistic and cultural contexts will guide researchers in selecting the most suitable translation methods.

Expanding translation efforts to include multiple languages and dialects will broaden the reach of mental health assessments, improving their accessibility to a more extensive and diverse population. integrating contextual factors, such as socioeconomic status, cultural norms, and literacy levels, will enhance the interpretation and application of translated health status instruments.

Finally, establishing a framework for continuously updating and refining translated versions will ensure these instruments remain accurate and culturally relevant. By pursuing these avenues, future research can significantly contribute to the field of cross-cultural mental health assessment, fostering inclusivity and advancing our understanding of mental health on a global scale.

Conclusion

Based on the robust statistical analysis and validation results, the emic plus etic approach is the most valid method for translating health status instruments for the Bengali-speaking population. This approach ensures a comprehensive and contextually appropriate translation, capturing the unique cultural aspects while maintaining universal applicability. The utilisation of advanced Artificial Intelligence (AI) metrics and assessments has played a pivotal role in this study's success. The integration of language models, such as Instructor XL, GTR T5 XXL, and Deberta XXL, alongside various natural language processing metrics, has provided valuable insights into translation accuracy, semantic equivalence, and cultural appropriateness.

The successful translation and validation of the Bengali-translated GAD-7 scale demonstrate its potential as a valuable tool for assessing anxiety levels in the Bengali-speaking population. Further studies are recommended to evaluate its efficacy in clinical settings through clinical trials, assess its sensitivity to detect subtle changes in anxiety levels over time, establish its stability in longitudinal studies, and validate its appropriateness in diverse settings across different population subgroups. These studies will enhance the scale's applicability and impact on mental health assessment and care.

Acknowledgement

Financial Support and Sponsorship

This project was sponsored by United We Care, who provided financial support for the research and implementation of this study. Their contribution was instrumental in facilitating the data collection, analysis, and overall success of the research endeavor.

Conflicts of Interest

The study was conducted with utmost transparency and impartiality, free from any external influence or competing interests that could potentially affect the design, execution, or interpretation of the results. The authors strictly adhered to the study's high ethical standards and scientific rigour.

Though United We Care is a mental health-focused organisation, the Bengali translation has been placed under a copyleft to avoid any potential conflict of interest. This decision underscores our commitment to prioritising public good over profit. The primary objective of this research is to contribute towards advancing knowledge and improving mental health resources for Bengali-speaking populations.

References

1. Spitzer RL, Kroenke K, Williams JBW & Löwe B. A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. Archives of Internal Medicine. 2006; 166: 1092–1097.
2. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (5th ed.). 2013.
3. Rutter LA & Brown TA. Psychometric Properties of the Generalized Anxiety Disorder Scale-7 (GAD-7) in Outpatients with Anxiety and Mood Disorders. Journal of psychopathology and behavioral assessment. 2017; 39: 140–146.
4. Roy-Byrne PP. Generalized anxiety and mixed anxiety-depression: Association with disability and health care utilization. Journal of Clinical Psychiatry. 1996; 57: 86–91.
5. Kertz S, Bigda-Peyton J & Bjorgvinsson T. Validity of the Generalized Anxiety Disorder-7 scale in an acute psychiatric sample. Clinical psychology & psychotherapy. 2013; 20: 456–464.
6. CIA. The World Factbook: Bangladesh. 2014.
7. Brislin RW. Back-translation for cross-cultural research. Journal of Cross-Cultural Psychology. 1970; 1: 185–216.
8. Peña ED. Lost in translation: Methodological considerations in cross□cultural research. Child Development. 2007; 78: 1255–1264.
9. Wild D, Grove A, Martin M, Eremenco S, McElroy S, Verjee□Lorenz A & Erikson P. Principles of good practice for the translation and cultural adaptation process for patient□reported outcomes (PRO) measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. Value in Health. 2005; 8: 94-104.
10. Fischer R & Karl JA. A primer to (cross-cultural) multi-group invariance testing possibilities in R. Frontiers in Psychology. 2019; 10: 1507.
11. Greenberg PE, Sisitsky T, Kessler RC, Finkelstein SN, Berndt ER, Davidson JRT, et al. The economic burden of anxiety disorders in the 1990s. Journal of Clinical Psychiatry. 1999; 60: 427–435.
12. Flaherty JA, Gaviria FM, Pathak D, Mitchell T, Wintrob R, Richman JA, et al. Developing instruments for cross-cultural psychiatric research. The Journal of nervous and mental disease. 1988; 176: 257–263.

13. Hoffman DL, Dukes EM, Wittchen HU. Human and economic burden of generalized anxiety disorder. Depression and Anxiety. 2006; 25: 72–90.

14. Orley J & Kuyken W. (Eds.). Quality of Life Assessment: International Perspectives. Heidelberg: Springer Verlag. 1994.

15. World Health Organization (WHO). MNH/PSF/95.2: Translation methodology. Geneva, Switzerland: World Health Organization.

16. Brown TA. Confirmatory factor analysis of the Penn State Worry Questionnaire: Multiple factors or method effects? Behaviour Research and Therapy. 2003; 41: 1411–1426.

17. Sartorius N and Kuyken W. Translation of health status instruments. In J. Orley and W. Kuyken (Eds). Quality of Life Assessment: International Perspectives. Heidelberg: Springer Verlag. 1994.

18. Saha SK, Pradhan P, Haldar D, Maji B, Agarwal W & Sarkar GN. Magnitude of Mental Morbidity and Its Correlates with Special Reference to Household Food Insecurity among Adult Slum Dwellers of Bankura, India: A Cross-Sectional Survey. Indian journal of psychological medicine. 2019; 41: 54–60.

19. Papineni K, Roukos S, Ward T & Zhu WJ. BLEU: a Method for Automatic Evaluation of Machine Translation. 2002.

20. Sellam T, Das D & Parikh AP. BLEURT: Learning Robust Metrics for Text Generation. 2020.

21. Ganesan K. ROUGE 2.0: Updated and Improved Measures for Evaluation of Summarization Tasks. 2018.

22. Ni J, Qu C, Lu J, Dai Z, Ábrego GH, Ma J, et al. Large Dual Encoders Are Generalizable Retrievers. 2021.

23. Su H, Shi W, Kasai J, Wang Y, Hu Y, Ostendorf M, et al. One Embedder, Any Task: Instruction-Finetuned Text Embeddings. 2022.

24. He P, Liu X, Gao J & Chen W. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. 2021.

25. Devlin J, Chang M-W, Lee K & Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019.

26. NIST Multimodal Information Group. NIST 2002 Open Machine Translation (OpenMT) Evaluation. 2010.

27. Linguistic Data Consortium Binapani De, Susmita Haldar, Shyamali Biswas. Assessment of knowledge and attitude towards mental illness among young and older people, in a selected rural community, Bankura, West Bengal. Int J Adv Psychiatric Nurs. 2023; 5: 17-22.

28. Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, Hill F, et al. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. 2019.

29. Snover M, Dorr B, Schwartz R, Micciulla L & Makhoul J. A Study of Translation Edit Rate with Targeted Human Annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers (pp. 223–231). Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas. 2006.

30. Lavie A & Agarwal A. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In Proceedings of the Second Workshop on Statistical Machine Translation (pp. 228-231). Association for Computational Linguistics. 2007.

31. Park Y, Patwardhan S, Visweswariah K & Gates S. An Empirical Analysis of Word Error Rate and Keyword Error Rate. In Proceedings of Interspeech. 2008: 2070-2073.

32. van Rijsbergen CJ. Information Retrieval. Butterworths. 1979.

33. Jaccard P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles. 1901; 37: 547-579.

34. Singh OP. District Mental Health Program - Need to look into strategies in the era of Mental Health Care Act, 2017 and moving beyond Bellary Model. Indian Journal of Psychiatry. 2018; 60: 163-164.

35. Census.co.in. (n.d.). Bankura District Population Census 2011-2021, West Bengal literacy sex ratio and density. 2011.

36. Datta S, Agarwal A, Chakraborty K, Ghosh P, Das S & Banerjee S. (n.d.). A hybrid WHO-AI framework for cross-cultural adaptation of mental health assessment tools: A novel methodology.