Iran J Psychiatry 2025; 20: 3: 373-382

**Original Article** *https://doi.org/10.18502/ijps.v20i3.19043*

# Item Development and Psychometric Evaluation of the Scrambled Sentences Task: A Pilot Study in Persian

**Asiyeh A'lam Hakkakan, Setareh Mokhtari***

**Abstract**

**Objective:** One significant concern regarding the cross-cultural use of psychological instruments is their adaptation to the language of the target population. The Scrambled Sentences Task (SST) exemplifies this issue. As a well-known paradigm for assessing interpretation bias (IB), the SST has been widely used across different languages; however, it remains unclear whether the SST is a valid and reliable tool to be used in languages other than English. The current study aims to develop SST items for Persian language while controlling for word frequency, word length and morphological complexity— linguistic features critical to meet SST's unique demands. We also seek to examine the psychometric properties of the Persian version of the SST (P-SST).
**Method:** The verbal stimuli for the P-SST were developed by drawing on a pool of sentences from prior research as a guide. These sentences were adapted specifically to fulfill the task's unique demands, ensuring the appropriateness of the P-SST for detecting IB. Since the SST primarily measures depressogenic tendencies, the Beck Depression Inventory-II (BDI-II) was also employed as part of the assessment. The measures were administered to a sample of 222 female students, selected due to evidence suggesting higher susceptibility to depression among women. The reliability and validity of the P-SST were then calculated, along with an analysis of responses to identify areas for enhancing performance on this task.
**Results:** Our results provided evidence of the convergent validity of the P-SST, as it was significantly correlated with the BDI-II ($r_s$ = 0.43, 95% CI [0.29–0.57], N = 161), as well as evidence of its divergent validity ($r_s$ = 0.35, 95% CI [0.22–0.49], N = 161). Moreover, internal consistency analysis revealed a Cronbach's alpha value of $\alpha$ = 0.81 and a split-half correlation value of r = 0.9.
**Conclusion:** Findings from this research established the psychometric properties of the P-SST as a quick and easily administered assessment tool to be used in the context of depression in Persian. The paper concludes with both linguistic and methodological recommendations to guide future development of SST items in any language.

**Key words:** *Adaptation; Depression; Interpretation Bias; Linguistics; Psychometrics; Scrambled Sentences Task*

Institute for Cognitive and Brain Sciences (ICBS), Shahid Beheshti University, Tehran, Iran.

***Corresponding Author:**
Address: Institute for Cognitive and Brain Sciences (ICBS), Shahid Beheshti University, Tehran, Iran, Postal Code: 1983969411.
Tel: 98-21 22431617, Fax: 98-21 22431989, Email: s_mokhtari@sbu.ac.ir

**Article Information:**
Received Date: 2025/01/22, Revised Date: 2025/04/04, Accepted Date: 2025/05/03

The use of psychological instruments across different languages and cultures is a common practice in psychological research and application (1, 2). However, this practice poses significant challenges due to linguistic and cultural differences that may affect the validity of results when instruments are applied in a new cultural context (3). To address these challenges, the adaptation of psychological instruments is essential, and guidelines have been developed to provide methodological guidance and facilitate this process (4-7).

A critical aspect of the adaptation process is the translation of instruments. The translation of psychological instruments is a complex task that requires careful consideration to ensure that the adapted version captures the intended constructs accurately. Traditional methods of translation, such as translation and back translation, have been widely used but have faced criticism for focusing primarily on linguistic equivalence (i.e., correct grammar and syntax) while overlooking other essential aspects of developing a quality measure (8-10). Linguistic equivalence ensures that the words and phrases in the translated version match those in the original language. However, Peña (11) argues that this approach is insufficient for ensuring construct validity, as it fails to account for deeper issues such as functional equivalence. As defined by Peña, functional equivalence refers to ensuring that an instrument or its items elicit the same target behavior or construct across different cultural or linguistic groups. This goes beyond literal translation to focus on whether the instrument measures the same underlying concept in both contexts. For example, Peña highlights that achieving functional equivalence may require altering test items or procedures to align with culturally appropriate norms or behaviors. Additionally, addressing task demands in the translation of instruments can further necessitate shifting away from the source instrument's wording. For example, in a cross-cultural administration of the Stroop task to English- and Norwegian-speaking participants (12), the authors explained how differences between the two languages necessitated the use of different color words in each version. Specifically, the English word "purple" was replaced with "turkis" ("turquoise" in English) in the Norwegian version to maintain the same number of characters, as required by the study design (12). A similar issue may arise when adapting tasks with verbal stimuli into Persian if the task requires matching words by character count. In response to these limitations, new perspectives in psychological instrument adaptation highlight the importance of considering factors such as functional/cultural equivalence and task demands in both cross-cultural research and intracultural administration of instruments originally developed in a different language (13).

Another critical step in the adaptation process of psychological instruments is pilot testing and subsequent validity and reliability assessment of the adapted instrument for the target population (4, 6). Pilot data facilitate item analysis, which can lead to modifications in the initial version of the adapted instrument, enhancing its relevance and applicability in the target population and practical settings. Furthermore, it is essential to provide robust evidence regarding the psychometric quality of the adapted instrument to ensure its effectiveness in measuring psychological constructs.

***The Direction of the Current Study***

Challenges related to adapting psychological instruments can be particularly critical when psychological instruments rely heavily on verbal stimuli (14, 15). Among the various verbal tasks available, the Scrambled Sentences Task (SST) stands out as a valuable tool, particularly in the context of depression research. Originally developed in English by Wenzlaff and Bates (16), the SST is designed as a cognitive marker for evaluating how individuals interpret ambiguous information. This characteristic makes it uniquely suited for studying cognitive biases associated with depression, where negative interpretations of ambiguous stimuli, also known as interpretation bias (IB), are a hallmark feature. The task consists of scrambled 6-word sentences that respondents should quickly unscramble into meaningful and grammatically accurate sentences using five of the six words. Notably, two of the words (target words) have opposite meanings, requiring respondents to choose one. Depending on the respondents' choice, the resulting sentence will convey either a positive or a negative meaning.

Backed by a strong theoretical background, particularly Beck's cognitive theory of depression (17) and also later theories (18), the SST has been widely utilized in psychological studies of biased interpretation among both English-speaking and non-English-speaking populations. Specifically, the SST has been applied in diverse languages, including German, Dutch, Polish, Chinese, Serbian, and Spanish (19). However, the procedures involved in the adaptation process of the SST for these target languages, including item translation, pilot testing and establishing SST psychometric features, have rarely been thoroughly addressed.

In a study by Ren *et al*. (20), for example, the authors used the translation and back translation method to convert the SST from English to Chinese, but did not address potential language challenges of translating from English to Chinese. Novović *et al*. (21) illustrated one such challenge, showing that linguistic equivalence does not ensure functional or cultural equivalence between the original and the translated versions of the SST. In developing the Serbian version of the SST, these authors excluded four SST items due to their ambiguous connotations in Serbian. They also emphasized that, beyond item content, lexical features like word frequency and order are crucial for ensuring that SST items uniformly assess respondents' IB. Despite this, linguistic considerations that are necessary to meet task

demands have largely been overlooked in subsequent SST research, with only a few exceptions (22, 23).

Another critical factor often overlooked in reports on SST applications, both in English and non- English versions, is error analysis. Most studies have focused solely on correct trials, resulting in a paucity of information about the types and frequency of errors, which could inform item revisions. In educational assessments, for example, efforts are made to minimize errors that hinder learners from providing meaningful responses (24). Examples include multiple-choice questions with two correct options or ambiguous questions that confuse learners (24). While not entirely applicable, this concept can extend to SST item development, as lapses in item development can lead participants to make avoidable errors when unscrambling the sentences. Analyzing grammatically incorrect responses can help revise items — such as adjusting word order — and construct items that are less likely to lead participants into incorrectly unscrambling sentences in a fast-paced task like the SST. Such considerations necessitate pilot testing of the adapted SST, as pilot data can reveal how the SST functions in real-world settings and provide feedback for refining items before broader implementation.

In this study, we aimed to develop items for a Persian version of the SST (P-SST), moving beyond mere translation of original items. Following Cruchinho *et al*. (13), we focused on considering factors such as task demands, which were seemingly ignored or at least unreported in previous research using the SST with Persian-speaking individuals (25, 26). Our second objective was to conduct a pilot study to investigate the psychometric characteristics of the Persian version of the SST. Additionally, we intended to identify shortcomings in the adaptation process through item analysis, particularly focusing on errors made by participants when unscrambling the sentences.

Given that our SST is applied to the context of depression, we focused exclusively on female participants due to evidence of higher prevalence of depression among women (27, 28) and gender differences in SST, with men typically showing more negative interpretations (21, 29, 30); as many researchers investigating depression or interpretive biases have focused on female participants (31-34).

## Materials and Methods

To develop the P-SST, the following points were taken into consideration:

(1) The original version of the task included six words. However, a mere translation of the original SST items does not necessarily result in a six-word scrambled sentence in target languages. For example, the scrambled sentence "*others' cannot I can meet expectations*" cannot be translated into a 6-word sentence in Persian. To rectify this limitation, some researchers suggested to use the original SST as a guide and formulate sentences with a depressive semantic content in the intended language (21, 35, 36). For example, Sfärlea *et al*. (35) developed an extended version of the original SST with 70 scrambled sentences in German. We used this extended German version to develop SST items in Persian as it offered a bigger pool of scrambled sentences. Candidate sentences were those which could be translated into grammatically correct and meaningful six-word Persian sentences without altering their original semantic content.

(2) Target words in each item should be carefully selected to minimize differences in morphological structure within the same syntactic function. Adjectives used as target words in a given item, for instance, were checked for consistent word forms. This consideration primarily stems from psycholinguistic evidence suggesting that morphemes play a significant role in word processing (37, 38).

(3) Target words should be selected to exhibit a variety of linguistic forms. This approach not only covers various ways of expressing emotional concepts, but also prevents participants from becoming habituated to a constant linguistic pattern, thereby enhancing their engagement in the task. We noted that in the original SST (16), target words ranged from canonical antonyms like *loser/winner* to negations such as *can/cannot*, affixal negations like *possible/impossible*, and even forms that do not necessarily have opposite meanings but result in differently valenced interpretations. Examples are the words *helped* and *lost* in *I have helped/lost my friends*.

(4) The two sets of target words for emotionally positive and negative interpretations of the scrambled sentences should be matched in terms of frequency. However, negative words are generally far less frequent in languages compared to positive words (39). To address this issue, we intentionally included a negated verb in some items, so that choosing the more frequent positive words would lead to negatively unscrambled sentences . This helped maintain frequency balance between target words resulting in negatively unscrambled sentences and those yielding positively unscrambled ones in the whole test. To extract information on word frequency, we referred to *the Persian news corpus of fas_newscrawl_2011*, which is available online at http://wortschatz.uni-leipzig.de/de. Statistical analysis revealed no difference in frequency ($t(46) = 0.35$, $P = 0.73$) between target words yielding a positive interpretation and those resulting in a negative interpretation of the scrambled sentences.

(5) The two sets of target words for emotionally positive and negative interpretations of the scrambled sentences need to have matched lengths, too. Thus, we conducted a statistical analysis to explore any differences in length between the target words yielding a positive interpretation and those resulting in a negative interpretation of the scrambled sentences. Our results showed no significant difference in terms of the number

of characters included in positive and negative target words (t (46) = 1.22, P = 0.22).
(6) Since evidence suggests that function words (such as prepositions) are more likely to be skipped in reading (40) or misplaced in speech (41) compared to content words, we deliberately limited their inclusion. This approach could help minimize the exclusion of items caused by skipping or misplacing function words in a fast-paced task like the SST.
There are also other points that are related to the structure of the task in general. It seems that these basic points have been considered in most of the previous adaptations of the SST across languages and were also incorporated into the development of the P-SST items: (i) to control for parafoveal processing of adjacent words and wrap-up effects, target words were neither positioned next to each other nor placed at the beginning or end of the sentences; (ii) to keep the balance, in half of the emotional trials, positive target words preceded negative target words and vice versa; (iii) neutral trials, featuring neutral target words resulting in emotionally neutral sentences, were included, and the same word order restrictions were applied to them; and (iv) following Everaert *et al.* (22), neutral trials were interspersed among emotional ones, ensuring that no more than two emotional trials appeared consecutively to minimize priming effects.

***Participants***

Participants in this study included 222 graduate and undergraduate female students recruited via convenience sampling from Shahid Beheshti university. Specifically, we employed a volunteer sampling method, targeting students who were residing on campus. Participants were aged 18 to 45 (M = 24.8, SD = 6.2). G*power (version 3.1.9.7) was used to calculate the sample size for the current study. Setting α at 0.05, power (1-β) at 0.80 and assuming a correlation of ρ = 0.30 between IB and scores from the Beck Depression Inventory-II (BDI-II), power analysis indicated a sample size of at least 84 participants. Power analysis for a test comparing means between groups of high and low depressive symptoms revealed a total required sample size of at least 140 participants, with α set at 0.05, β at 0.80 and Cohen's d at 0.50.
The Biomedical Ethics Committee of Shahid Beheshti University approved the study design and its procedure (No.IR. SBU.REC.1400.263).

***Measures***

The P-SST included 36 trials presented in three blocks, each containing eight emotional and four neutral scrambled sentences. The task began with a practice block of three neutral trials, followed by the main blocks. Each trial began with a fixation cross on the right side of the screen (aligned with the Persian right-to-left writing system) for 500 milliseconds, followed by either an emotional or neutral scrambled sentence. Participants were instructed to mentally unscramble the sentence as quickly as possible and press the space bar when done, with their reaction time recorded. After pressing the space bar, the same scrambled sentence reappeared, this time with each word accompanied by a number. Participants then verbally reported the grammatical sentence they had formed using the numbers instead of the actual words. The scrambled sentence remained on the screen until reported.
A cognitive load procedure was applied by presenting a 6-digit number for 5,000 milliseconds at the beginning of each block and asking participants to recall it at the end of the block.
Since we utilized the depression-related version of the SST, the BDI-II was also administered to assess the convergent validity of the P-SST through its correlation with BDI-II scores. The BDI-II is a 21-item questionnaire developed to measure the presence and severity of depressive symptoms (42). In each item, four sentences are presented and participants are required to choose the one single sentence which best describes their mental state during the past two weeks. The scoring system assigns 0 to the first sentence and 1, 2, and 3 to the subsequent sentences. A higher score represents a greater severity of depressive symptoms. In a study of the psychometrics of the Persian version of the BDI-II in the Iranian population, an internal consistency value of α = 0.87 and test-retest reliability of r = 0.74 were reported (43). The internal consistency of the BDI-II for this study was: α [95% CI] = 0.91[0.89-0.93], SE = 0.01.

***Administration Procedure***

Each participant was tested individually. To avoid mood priming, participants first completed the P-SST on a 12-inch laptop screen and then filled out the BDI-II. The experimenter manually recorded participants' responses and the digit string they were asked to recall at the end of each block. After each block, participants were given a short break to regain focus and prepare for the next block. The whole procedure lasted for approximately 30 minutes.

***Data Preparation and Analysis***

Analysis was based on data from participants who had an accuracy rate of at least 80%. Accurate responses included sentences that were correctly unscrambled within a time limit of 8000 milliseconds (22). This led to the exclusion of 37 participants who failed to meet the accuracy requirement (16.6%) either because they were unable to form grammatically correct sentences (32%) or because they exceeded the time limit (68%). As previously mentioned, a cognitive load procedure was also applied to prevent the intrusion of response strategies. Following Gilbert and Hixon (44), we excluded participants who did not correctly remember at least three digits -in any order- across the three blocks (n = 24). This resulted in a sample size of 161 participants.
Statistical analyses were performed on an IB index calculated as the ratio of negatively unscrambled sentences to all correctly unscrambled emotional sentences. As Shapiro-Wilk normality test indicated that neither the IB (w = 0.90, P < 0.001) nor the BDI-II

scores (w = 0.93, P < 0.001) were normally distributed, Spearman's rank-order correlation coefficient was used to examine convergent validity. Divergent validity was explored using the Mann-Whitney U test and the Spearman's correlation coefficient. Since the data from the P-SST consisted of two categorical responses, i.e. either positive or negative, Kuder-Richardson 20 (KR-20) was used to measure the internal consistency of the P-SST. Split-half reliability was also calculated as another index of internal consistency. A significance level of 0.05 was used in all analyses.

## Results

### *Item Frequency for Linguistic Analysis*

Details on participants' responses to each item of the P-SST, as well as the cognitive load retention rate for each block, are provided in Table 1. The Table shows the number of times each item was positively unscrambled (e.g., the scrambled sentence "am winner born loser a I" unscrambled as "I am a born winner"), negatively unscrambled (e.g., "I am a born loser"), or unscrambled with grammatical errors (e.g., "I am a winner born") across participants. It also reports the number of times each item was missed and the number of times an item was excluded due to exceeding the time limit.

**Table 1. Cognitive Load Retention Rate and the Frequency of Responses to Items of the Persian Scrambled SentencesTask**

| | BLOCK 1 | | | | | BLOCK 2 | | | | | BLOCK 3 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 89.8 | | | | | 97.9 | | | | | 99.5 | | | | |
| **Items** | **PU** | **NU** | **GF** | **TL** | **M** | **PU** | **NU** | **GF** | **TL** | **M** | **PU** | **NU** | **GF** | **TL** | **M** |
| 1 | 110 | 18 | 4 | 28 | 1 | 108 | 21 | 3 | 27 | 2 | 76 | 41 | 2 | 41 | 1 |
| **2** | 82 | 43 | 17 | 18 | 1 | 103 | 52 | 4 | 1 | 1 | 136 | 17 | 2 | 6 | 0 |
| **3** | 86 | 74 | 0 | 0 | 1 | 144 | 14 | 1 | 2 | 0 | 152 | 7 | 1 | 1 | 0 |
| **4** | 116 | 24 | 17 | 4 | 0 | 102 | 50 | 4 | 4 | 1 | 91 | 64 | 1 | 4 | 1 |
| **5** | 143 | 10 | 4 | 4 | 0 | 110 | 35 | 3 | 13 | 0 | 117 | 25 | 4 | 15 | 0 |
| **6** | 106 | 27 | 9 | 18 | 1 | 145 | 11 | 2 | 2 | 1 | 122 | 22 | 13 | 4 | 0 |
| **7** | 130 | 13 | 16 | 1 | 1 | 149 | 10 | 1 | 1 | 0 | 85 | 68 | 5 | 3 | 0 |
| **8** | 123 | 23 | 15 | 0 | 0 | 125 | 33 | 0 | 2 | 1 | 138 | 16 | 6 | 1 | 0 |

Note: PU = Positively Unscrambled, NU = Negatively Unscrambled, GF = Grammatically False, TL = Time Limit Exceeded, M = Missed

As revealed in Table 1, the highest number of negatively unscrambled sentences belongs to the third item in block 1 (1-3) and the seventh item in block 3 (3-7). Table 3 also shows that participants made the highest number of grammatical mistakes in items 1-4, 1-7, 1-2, 1-8, and 3-6. Further, among items excluded because of exceeding the 8-second time limit, the first item in each block has taken the longest to get unscrambled.

### *Validity Analysis*

Analyses were conducted using R version 3.5.0. Table 2 depicts descriptive statistics for depressive symptoms and IB variables.

Results from the Spearman's correlation coefficient analysis revealed a positive small to medium correlation between the P-SST and the BDI-II ($r_s$ = 0.43, P < 0.001). Given that cognitive load is not a moderating factor of convergent validity of the SST (19), we conducted the correlation analysis again, this time including participants with cognitive load errors, too (no load condition in Table 2). This resulted in a larger sample size of 185 and the results showed a higher correlation between the P-SST and the BDI-II scores in our sample ($r_s$ = 0.47, P < 0.001), which probably better reflects the association between the P-SST and the BDI-II scores in our sample.

To assess divergent validity, participants were split into two groups of high (BDI-II ≥ 16) and low (BDI-II ≤ 15) depressive symptomology, and then the mean P-SST scores were compared between the two groups. The results, indicated a statistically significant difference between the mean IB scores of the two groups (w = 1820, P < 0.001). To further explore the relationship between IB and depressive symptoms severity, we calculated the correlation between the P-SST scores and group membership. The results showed a significant correlation between IB and depressive symptoms ($r_s$ = 0.35, P < 0.001).

**Table 2. Descriptive Statistics for Depressive Symptoms and Interpretation Bias under Load and No Load Conditions**

| Variable | N | M | SD | Min | Max |
|---|---|---|---|---|---|
| **Under Load** | | | | | |
| BDI-II | | | | | |
| High Depressive Symptom Group | 66 | 25.96 | 8.66 | 16 | 56 |
| Low Depressive Symptom Group | 95 | 7.92 | 4.36 | 0 | 15 |
| P-SST | | | | | |
| High Depressive Symptom Group | 66 | 0.27 | 0.17 | 0 | 0.8 |
| Low Depressive Symptom Group | 95 | 0.15 | 0.12 | 0 | 0.59 |
| **No Load** | | | | | |
| BDI-II | | | | | |
| High Depressive Symptom Group | 79 | 26.7 | 8.48 | 16 | 56 |
| Low Depressive Symptom Group | 106 | 7.88 | 4.35 | 0 | 15 |
| P-SST | | | | | |
| High Depressive Symptom Group | 79 | 0.29 | 0.17 | 0 | 0.8 |
| Low Depressive Symptom Group | 106 | 0.16 | 0.13 | 0 | 0.59 |

Note. BDI-II = Beck Depression Inventory-II; P-SST = Persian Scrambled Sentences Task

***Reliability Analysis***

The analysis revealed a Cronbach's alpha ($\alpha$) [95% CI] = 0.81 [0.74–0.83], SE = 0.024, indicating an acceptable level of internal consistency among the items. Cronbach's alpha is a measure of internal consistency and reliability, assessing the extent to which multiple items within a scale measure the same construct. A coefficient of 0.81 suggests that the items in the P-SST are highly intercorrelated and reliably measure the intended construct. Item-level statistics were also examined to identify any problematic item. With all alpha coefficients remaining above 0.76, dropping any item did not substantially increase the overall reliability of the P-SST. A split-half reliability analysis was also performed to further assess internal consistency. Split-half reliability evaluates how well the test items measure the same construct by dividing the test into two halves and correlating their scores. Several measures were calculated as shown in Table 3. These values offer different perspectives on internal consistency, reflecting various ways of estimating the reliability between the two halves. Specifically, Guttman's $\lambda_6$ represents a conservative and informative measure, showing the minimum reliability one can expect from a test. That said, with a $\lambda_6$ of 0.86 and other measures exceeding 0.7, the results suggest good internal consistency of the P-SST.

**Table 3. Split-Half Reliability Measures of the Persian Scrambled SentencesTask**

| | Maximum Split-Half Reliability ($\lambda_4$) | Guttman $\lambda_6$ | Average Split-Half Reliability | Guttman $\lambda_3$ | Guttman $\lambda_2$ | Minimum Split-Half Reliability (beta) |
|---|---|---|---|---|---|---|
| estimates | 0.9 | 0.86 | 0.81 | 0.81 | 0.82 | 0.71 |

## Discussion

In this study, we aimed to develop SST items tailored for the Persian language. Specifically, we moved beyond mere translation of the SST items to meet issues demanded by task characteristics. We also conducted a pilot study to establish the psychometric characteristics of the Persian version of the SST and identify areas for modification through error analysis. We first discuss the results of the validity and reliability tests, followed by recommendations for future SST adaptations in other languages based on our pilot study findings.

Findings from our study revealed a significant positive correlation between depression-related symptoms measured by the BDI-II and IB scores from the P-SST, supporting the convergent validity of the P-SST. The correlation value we observed was similar to those reported by Seeds (34) and Rude *et al.* (33). It is worth noting that due to the considerable heterogeneity in the SST versions (e.g., in terms of the number of trials,

stimuli or administration format), comparability across studies in terms of the SST psychometric indices can be problematic. However, our findings align with those of Würtz *et al*. (19), who investigated the psychometric properties of the SST via a meta-analysis while conducting moderation analysis to explain the high heterogeneity among studies. Results of their study revealed a wide range of correlational values between the SST and the BDI-II, ranging from 0.2 to 0.8, with an overall correlation of 0.46 for the SST convergent validity.

Our results also showed that the P-SST can differentiate between individuals with high and low depressive symptoms. Specifically, we found a statistically significant difference in mean IB scores of the two groups, measured by the P-SST. Correlational analysis between IB scores and group membership further supported the divergent validity of the P-SST. Importantly, the small divergent validity we found in our study is consistent with the literature, which shows that, in the context of depression, small and medium correlations of the SST are most frequent, with large correlations being rather rare (19). The results we found in terms of the divergent and convergent validity of the P-SST are consistent with those reported in prior research, demonstrating that our developed Persian version of the SST is comparable to the original SST and those developed in other languages. This contributes to research on the cross-linguistic applicability of the SST. Additionally, our findings on the internal consistency of the P-SST further reinforce this conclusion.

Our results from the reliability analysis of the P-SST agree with those reported in Würtz *et al*.'s meta-analysis, indicating an overall alpha coefficient of 0.79 across different disorders, with a slight difference in the internal consistency of $\alpha = 0.06$ when conducted in the context of depression. Internal consistency is particularly important in the SST because it ensures that all items on the test are measuring the same underlying construct; namely IB. If the items are not internally consistent, it would suggest that they are tapping into different cognitive processes or constructs, which would undermine the validity of the test as a measure of a single, coherent construct. Measures of split-half reliability also demonstrate the acceptability of the P-SST as a reliable test of depression-related IB, with most indices being above 0.81. However, split-half reliability has limitations. For example, the obtained estimate can vary depending on how the test is split, potentially leading to different reliability coefficients. To mitigate this limitation, we calculated several different measures of split-half reliability (Guttman's lambda coefficients, maximum split-half reliability, average split-half reliability), providing a more comprehensive assessment of internal consistency.

In developing items for the P-SST, we controlled for linguistic factors to minimize measurement errors. Results from the pilot study, however, provided a pool of experimental data that helped us detect linguistic challenges and offer modifications. First, our analysis indicated that the placement of non-target words within scrambled sentences affected outcomes. Specifically, items 1-3 and 3-7 were most frequently unscrambled negatively compared to other items. Notably, both items included the negative copular verb نیستم (meaning am not). Copular verbs in Persian, such as هستم (meaning am) and its negated form ,نیستم primarily function as links between the subject and predicate and often carry little independent meaning (45). Because copular verbs can be omitted in everyday speech without altering meaning, participants in the fast-paced P-SST probably tended to overlook ,نیستم leading to a negative response even if the participant did not intend to form a negative sentence.

Interestingly, the frequency of negatively unscrambled sentences was slightly higher for item 1-3 than for item 3-7. The key difference between the items is the position of the negated copula :نیستم it appeared last in item 1-3 but first in item 3-7. With the negated verb placed at the beginning— where visual attention was already focused due to the fixation cross— participants' attention was likely drawn to it, resulting in fewer negatively unscrambled sentences for item 3-7. This suggests that strategically placing non-target words in the initial position of scrambled sentences may help minimize errors by enhancing their visibility. Although this idea is speculative and calls for further investigation, the potential impact of word placement on participant responses is further illustrated in our subsequent analysis of grammatical errors for items 1-4 and 1-7.

The highest grammatical errors were found in items 1-4 and 1-7. In item 1-4, both the first word خانواده‌ام) meaning my family (and the fifth word من) meaning I ( can serve as subjects. However, using the first word leads to an agrammatical sentence due to subject-verb agreement violations, likely because it attracted participants' attention more quickly than the correct option. Similarly, in item 1-7, errors arose from subject-verb agreement violations that could have been avoided by placing the correct choice in the first position, where participants' visual attention was already focused.

Moreover, items 1-2 and 1-8 elicited many grammatical errors due to the displacement of the Persian case marker را.

We recommend minimizing the use of functional words like را and focusing on content words in SST items. Furthermore, items that consistently exceeded the time limit were the first item in each block across all participants. This is likely because participants were still engaged in memorizing the preceding 6-digit number. Future studies could benefit from initiating each block with a neutral filler item, which would not be subject to analysis, thereby reducing errors caused by exceeding the time limit.

Of note, a relatively high proportion of participants (16.6%) were excluded in this study due to failing to

meet accuracy requirements or exceeding the time limit. While these exclusion criteria ensured data integrity by including only participants who engaged adequately with the task, excluding a substantial portion of the sample may affect the representativeness of the findings. Unfortunately, exclusion rates are rarely reported in previous studies using the SST (19), making direct comparisons difficult. This underscores the need for greater transparency in reporting such data in future research. Importantly, our error analysis identified specific task-related factors that contributed to participant errors and timeouts. We recommend that addressing these factors can help reduce the exclusion rate in future studies, thereby improving inclusiveness and representativeness without compromising data quality.

## Limitation

First, the study sample was exclusively composed of non-clinical female participants. As a result, the current findings may not fully capture the complexity of IB present in clinical samples or in male individuals, and therefore caution should be exercised when applying these results to broader populations. Another limitation, which is commonly observed in research, is that the SST is not always easy to score. This issue is particularly pronounced in languages like Persian, which permit flexible word order, as opposed to languages with fixed word order. Administering a Persian version of the SST can result in data reflecting various ways of unscrambling each sentence, making it challenging to determine whether a given word string is acceptable. One potential approach to address this challenge involves adhering strictly to Persian grammar rules and rejecting responses that violate them. However, this approach can be problematic due to the observed mismatch between prescriptive grammar rules and what speakers of the language consider grammatically acceptable. One approach to tackle this issue in future research involves leveraging advancements in computational linguistics, which offer promising solutions through probabilistic models of language processing. These models use a symbolic component that is responsible for generating linguistic structures and a probabilistic component that, based on the likelihood of occurrence, assigns a probability value to these structures. By combining these two components, probabilistic models claim to be capable of predicting the extent to which any given sentence in a language is acceptable (but not merely grammatical). Implementing such models can alleviate SST scoring concerns. By setting a probability threshold, responses above it are simply accepted while those below are rejected.

## Conclusion

In closing, our findings established the psychometric properties of the P-SST as a quick and easily administered assessment tool to be used in the context of depression in Persian, demonstrated by a validity correlation of 0.43 with the BDI-II, and a Cronbach's alpha of 0.81. Moreover, the thorough examination of responses to the P-SST items revealed key considerations for developing SST items in any language. Specifically, our findings highlight: 1) the importance of the placement of non-target words within each scrambled sentence, particularly the unique role of the first position; 2) the need to minimize the use of functional words; and 3) the strategic positioning of neutral items among emotional ones, with a recommendation to begin each block with a neutral filler item rather than a target emotional item. Importantly, these suggestions are language-independent and can be applied broadly to SST development across different languages.

## Acknowledgment

We would like to thank the participants for their valuable cooperation in this study.

## Conflict of Interest

None.

## References

1. Epstein J, Santo RM, Guillemin F. A review of guidelines for cross-cultural adaptation of questionnaires could not bring out a consensus. J Clin Epidemiol. 2015;68(4):435-41.
2. Hambleton RK, Zenisky AL. Translating and adapting tests for cross-cultural assessments. 2011.
3. Shou Y, Chen HF, Takemura K, Wu J, Yang CT, Wang MC. Editorial: From West to East: Recent Advances in Psychometrics and Psychological Instruments in Asia. Front Psychol. 2022;13:875536.
4. Gudmundsson E. Guidelines for translating and adapting psychological instruments. Nordic Psychology. 2009;61(2):29-45.
5. Hambleton RK, Merenda PF, Spielberger CD. Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. Adapting educational and psychological tests for cross-cultural assessment: Psychology Press; 2004. p. 15-50.
6. Gregoire J. ITC guidelines for translating and adapting tests. Int J Test. 2018;18(2):101-34.
7. Van de Vijver F, Hambleton RK. Translating tests. European psychologist. 1996;1(2):89-99.
8. Van de Vijver F, Leung K. Methods and data analysis of comparative research. Handbook of cross-cultural psychology. 1997;1:257-300.
9. Geisinger KF. Cross-cultural normative assessment: translation and adaptation issues influencing the normative interpretation of assessment instruments. Psychol Assess. 1994;6(4):304.

10. Colina S, Marrone N, Ingram M, Sánchez D. Translation Quality Assessment in Health Research: A Functionalist Alternative to Back-Translation. Eval Health Prof. 2017;40(3):267-93.
11. Peña ED. Lost in translation: methodological considerations in cross-cultural research. Child Dev. 2007;78(4):1255-64.
12. Holmlund TB, Foltz PW, Cohen AS, Johansen HD, Sigurdsen R, Fugelli P, et al. Moving psychological assessment out of the controlled laboratory setting: Practical challenges. Psychol Assess. 2019;31(3):292-303.
13. Cruchinho P, López-Franco MD, Capelas ML, Almeida S, Bennett PM, Miranda da Silva M, et al. Translation, Cross-Cultural Adaptation, and Validation of Measurement Instruments: A Practical Guideline for Novice Researchers. J Multidiscip Healthc. 2024;17:2701-28.
14. Krach SK, McCreery MP, Guerard J. Cultural-linguistic test adaptations: Guidelines for selection, alteration, use, and review. School psychology international. 2017;38(1):3-21.
15. Kalfoss M. Translation and Adaption of Questionnaires: A Nursing Challenge. SAGE Open Nurs. 2019;5:2377960818816810.
16. Wenzlaff RM, Bates DE. Unmasking a cognitive vulnerability to depression: how lapses in mental control reveal depressive thinking. J Pers Soc Psychol. 1998;75(6):1559-71.
17. Beck A. Depression. Clinical, Experimental and Theoretical Aspects. New York (Hoeber) 1967. 1967.
18. Everaert J, Struyf S, Koster EH. Biased interpretation of ambiguity in depression and anxiety: Interactions with attention, memory, and cognitive control processes. Interpretational Processing Biases in Emotional Psychopathology: From Experimental Investigation to Clinical Practice. 2023:79-96.
19. Würtz F, Zahler L, Blackwell SE, Margraf J, Bagheri M, Woud ML. Scrambled but valid? The scrambled sentences task as a measure of interpretation biases in psychopathology: A systematic review and meta-analysis. Clin Psychol Rev. 2022;93:102133.
20. Ren Z, Li X, Zhao L, Yu X, Li Z, Lai L, et al. Effectiveness and mechanism of internet-based self-help intervention for depression: The Chinese version of MoodGYM. Acta Psychologica Sinica. 2016.
21. Novović Z, Mihić L, Biro M, Tovilović S. Measuring vulnerability to depression: The Serbian scrambled sentences test-SSST. Psihologija. 2014; 47(1): 33-48. https://doi.org/10.2298/PSI1401033N
22. Everaert J, Grahek I, Koster EH. Individual differences in cognitive control over emotional material modulate cognitive biases linked to depressive symptoms. Cogn Emot. 2017;31(4):736-46.
23. Everaert J, Grahek I, Duyck W, Buelens J, Van den Bergh N, Koster EH. Mapping the interplay among cognitive biases, emotion regulation, and depressive symptoms. Cogn Emot. 2017;31(4):726-35.
24. Suto I, Williamson J, Ireland J, Macinska S. On reducing errors in assessment instruments. Res Pap Educ. 2023;38(3):357-77.
25. Torkan H, Blackwell SE, Holmes EA, Kalantari M, Neshat-Doost HT, Maroufi M, et al. Positive Imagery Cognitive Bias Modification in Treatment-Seeking Patients with Major Depression in Iran: A Pilot Study. Cognit Ther Res. 2014;38(2):132-45.
26. Dolatshahi B, Naderi Rajeh Y, Pourshahbaz A, Zarghami M. Uncovering Negative Interpretation Bias in Remitted/Recovered Depression with Laboratory Task. Iran J Psychiatry. 2023;18(2):165-72.
27. Albert PR. Why is depression more prevalent in women? J Psychiatry Neurosci. 2015;40(4):219-21.
28. Slavich GM, Sacher J. Stress, sex hormones, inflammation, and major depressive disorder: Extending Social Signal Transduction Theory of Depression to account for sex differences in mood disorders. Psychopharmacology (Berl). 2019;236(10):3063-79.
29. Rude SS, Wenzlaff RM, Gibbs B, Vane J, Whitney T. Negative processing biases predict subsequent depressive symptoms. Cogn Emot. 2002;16(3):423-40.
30. Rude SS, Valdez CR, Odom S, Ebrahimi A. Negative cognitive biases predict subsequent depression. Cognit Ther Res. 2003;27:415-29.
31. Hammen C, Kim EY, Eberhart NK, Brennan PA. Chronic and acute stress and the prediction of major depression in women. Depress Anxiety. 2009;26(8):718-23.
32. Joormann J, Talbot L, Gotlib IH. Biased processing of emotional information in girls at risk for depression. J Abnorm Psychol. 2007;116(1):135-43.
33. Rude S S, Durham-Fowler J A, Baum E S, Rooney S B, Maestas K L. Self-report and cognitive processing measures of depressive thinking predict subsequent major depressive disorder. Cognitive Therapy and Research. 2010; 34(2): 107–115. https://doi.org/10.1007/s10608-009-9237-y
34. Seeds PM. Interpretive bias in the context of life stress and depression: An examination of stress generation and diathesis-stress models: The University of Western Ontario (Canada); 2012.
35. Sfärlea A, Löchner J, Neumüller J, Asperud Thomsen L, Starman K, Salemink E, et al. Passing on the half-empty glass: A transgenerational study of interpretation biases in children at risk for depression and their parents with depression. J Abnorm Psychol. 2019;128(2):151-61.
36. Viviani R, Dommes L, Bosch JE, Stingl JC, Beschoner P. A Computerized Version of the Scrambled Sentences Test. Front Psychol. 2017;8:2310.
37. Gagné CL. Psycholinguistic Approaches to Morphology: Theoretical Issues. Oxford research encyclopedia of linguistics2017.

38. Sandra D. Morphological units: A theoretical and psycholinguistic perspective. Oxford research encyclopedia of linguistics2020.
39. Taboada M, Trnavac R, Goddard C. On being negative. Corpus Pragmat. 2017;1:57-76.
40. Gao J, Suzuki H, editors. Long distance dependency in language modeling: an empirical study. International Conference on Natural Language Processing; 2004: Springer.
41. Garrett M. Levels of processing in sentence production. Language production Vol 1: Speech and talk: Academic Press; 1980. p. 177-220.
42. Steer RA, Clark DA, Beck AT, Ranieri WF. Common and specific dimensions of self-reported anxiety and depression: the BDI-II versus the BDI-IA. Behav Res Ther. 1999;37(2):183-90.
43. Ghassemzadeh H, Mojtabai R, Karamghadiri N, Ebrahimkhani N. Psychometric properties of a Persian-language version of the Beck Depression Inventory--Second edition: BDI-II-PERSIAN. Depress Anxiety. 2005;21(4):185-92.
44. Gilbert DT, Hixon JG. The trouble of thinking: Activation and application of stereotypic beliefs. J Pers Soc Psychol. 1991;60(4):509.
45. Mostafavi P. Predicative Clauses in Today Persian-A Typological Analysis. Asian Social Science. 2015;11(15):104.