Exploring ChatGPT-Based Development of Computerized Adaptive Testing Item Banks for Emotional Stability

Authors: Gao Yaojie, Qi Yunxiao, Ma Yuanqiu, Liu Tuo, Liu Tuo

Date: 2024-02-01T18:18:56+00:00

Abstract

The substantial human and material resources required for traditional item development constrain the development and application of computerized adaptive testing, while automatic item generation based on the latest natural language processing technology holds promise for addressing this challenge. With advances in generative pre-trained models based on the Transformer architecture, automatically generating test items according to specific measurement objectives (particularly non-cognitive tasks) and establishing computerized adaptive item banks based on them has become feasible. This study aims to employ the latest version of ChatGPT to generate a large number of Chinese-language personality items measuring emotional stability, and through item bank construction procedures including unidimensionality testing, IRT model selection, item analysis, item bank quality analysis, as well as simulated computerized adaptive testing, to explore the suitability of these items for computerized adaptive testing and compare their performance with widely-used emotional stability items, ultimately forming a high-quality emotional stability item bank.

Full Text

Exploration of Computerized Adaptive Item Bank Development for Emotional Stability Based on ChatGPT

Gao Yaojie[2], Qi Yunxiao[2], Ma Yuanqiu[2], Liu Tuo[1],[2],[3] (Corresponding author)

[1]Key Research Base of Humanities and Social Sciences of the Ministry of Education, Academy of Psychology and Behavior, Tianjin Normal University, Tianjin 300387
[2]Faculty of Psychology, Tianjin Normal University, Tianjin 300387

[3]Tianjin Social Science Laboratory of Students' Mental Development and Learning, Tianjin 300387

Abstract

To obtain a high-quality large-scale item bank, the extensive manpower and resources required for traditional project development have been constraining the development and application of computerized adaptive testing. However, automatic item generation based on the latest natural language processing technology holds promise in addressing this challenge. With the advancements in generative pre-trained models based on the Transformer architecture, the generation of items tailored to specific measurement objectives (especially non-cognitive tasks) becomes feasible. This study aimed to utilize ChatGPT to generate a large number of Chinese version personality items measuring emotional stability and to establish a computerized adaptive item bank based on this premise.

We utilized ChatGPT based on GPT-4 Turbo to generate 114 items measuring emotional stability. Following expert review, 75 items were retained and formed the GPT item bank, while 42 widely-used items were selected to form the classic item bank. Testing was conducted on the aforementioned items, yielding 479 valid participants. Additionally, sample data from two separately administered measures, CBF-PI-B and BFI-2, were going to be used for subsequent cross-sample reliability comparisons. Procedures for item bank construction including unidimensionality test, IRT model selection, item analysis, and item bank quality analysis, as well as simulated computerized adaptive testing, were employed to assess the quality and CAT performance of the item bank.

After the above analysis steps, it was found that all items in the classic item bank and the GPT item bank passed the unidimensionality test, showing no differential item functioning, and had good discrimination parameters and reasonable difficulty distribution. Both item banks provided high test information and marginal reliability for most trait levels of the examinees, with low measurement error. The overall item bank formed by combining all items remained of good quality. Simulation results of computerized adaptive testing showed that all three item banks achieved high validity with fewer items compared to traditional tests for the same level of precision. Under the same testing length, the GPT item bank exhibited higher reliability and demonstrated stability across samples.

Additionally, comparison revealed that the CAT performance of the GPT item bank even exceeded that of the classic item bank, while the overall item bank performance was slightly better than that of the GPT item bank.

This study innovatively explores the development of a computerized adaptive item bank using the latest version of ChatGPT, validating the feasibility of this user-friendly project generation tool. Through comparison with previous research results, it reconfirms the excellent quality of projects generated by GPT-4. The study showcases the immense potential and possibilities of large language

models in project development, particularly in the creation of large-scale item banks, while also indicating a shift in the responsibilities of psychologists in future project development.

Keywords: Computerized Adaptive Test, Automatic Item Generation, Natural Language Processing, emotional stability, Item Response Theory

Corresponding author: Liu Tuo, E-mail: mikebonita@163.com

Introduction

Computerized Adaptive Testing (CAT) represents a computer-based assessment methodology that dynamically matches test items to examinees' ability levels during the testing process, thereby enhancing measurement efficiency and score validity (Fliege et al., 2005; Jiao & Lissitz, 2020). Compared to traditional tests grounded in Classical Test Theory (CTT), CAT offers two principal advantages. First, examinees need not respond to all items in a scale, which reduces cognitive burden and prevents accuracy degradation from boredom during prolonged testing while improving measurement efficiency. Second, CAT achieves "tailored testing" that minimizes measurement error arising from mismatches between item difficulty and examinee trait levels. The comparability of results despite different items administered to each examinee stems from CAT's foundation in Item Response Theory (IRT). Within IRT, as long as all item parameters are calibrated on a common scale, ability estimates (trait levels) for each examinee remain comparable even when the number and difficulty of administered items differ. Consequently, an item bank comprising items with shared measurement objectives and parameters on the same scale constitutes the fundamental prerequisite for CAT. Achieving this requires developers to initially collect numerous items and conduct large-scale testing to obtain stable item parameters for constructing the initial bank, followed by ongoing management and supplementation, such as controlling overexposed items and eliminating low-quality ones. Since traditional item development relies heavily on domain experts and demands substantial human, material, and financial resources, establishing a rich and continuously updatable item bank represents a major bottleneck constraining current CAT development and application (Gierl & Haladyna, 2013; Gierl & Lai, 2018).

To address this challenge, the integration of Automatic Item Generation (AIG) with CAT holds considerable promise (Hommel et al., 2022). Gierl and Lai (2018) summarized AIG as a three-step process. First, subject-matter experts (SMEs) use cognitive models—representations emphasizing the knowledge, skills, and abilities required to solve problems within specific domains—to organize and structure content needed for item generation. Second, experts develop item templates that specify where content from the cognitive model can be placed, enabling item generation through substitution of specific content within these templates. Third, computer algorithms place cognitive model

content into templates and generate new items through content replacement. While this template-based AIG has been widely applied in educational testing contexts (Gierl et al., 2012; Gierl & Haladyna, 2013; Lai et al., 2016; Kurdi et al., 2020), it proves less suitable for non-cognitive assessments such as personality tests common in psychological measurement. These instruments often involve more complex semantics, contexts, and nuanced distinctions (Hernandez & Nie, 2022; Hommel et al., 2022; Lee et al., 2023). For instance, in the BFI-2 Chinese version (Zhang et al., 2022), the neuroticism dimension includes the item "I am someone who can control my emotions." If "control" or "emotion" were treated as replaceable template elements, meaningful substitution candidates would be extremely limited, and synonym replacement would yield only marginally useful new items, constraining the scale's capacity to measure neuroticism across diverse contexts. Fortunately, advances in Natural Language Processing (NLP) technology have prompted researchers to explore algorithmic generation of non-cognitive items (Von Davier, 2018; Hernandez & Nie, 2022; Hommel et al., 2022; Götz et al., 2023; Lee et al., 2023).

NLP constitutes a subfield of Artificial Intelligence (AI) that develops quantitative models enabling computers to understand, analyze, and generate human language (Goldberg, 2017; Lee et al., 2023). The introduction of neural networks—computational models simulating biological neural networks that transform input data through successive mappings before output (Goldberg, 2017)—has yielded substantial progress in NLP language models (Götz et al., 2023). Von Davier (2018) pioneeringly applied Long Short-Term Memory (LSTM) network-based language models, then state-of-the-art, to AIG for non-cognitive items. However, this technology could only mimic the syntactic structure of example items, lacked capacity to generate items targeting specific constructs, and suffered from extreme computational demands and inconsistent grammatical correctness (Hernandez & Nie, 2022; Hommel et al., 2022). Moreover, LSTM fundamentally represents a supervised machine learning model that learns to predict labels for new inputs through training on labeled data (e.g., emails marked as spam), requiring large volumes of manually annotated sample data—precisely the resource scarcity that motivates AIG use (Goldberg, 2017; Lee et al., 2023). The release of OpenAI's Generative Pre-trained Transformer (GPT) models and their iterative versions, based on the Transformer architecture, has attracted AIG researchers focused on non-cognitive item generation due to their superior performance.

Self-attention mechanisms represent a key Transformer innovation, enabling any word in an input sequence (e.g., a sentence) to interact with all other words, such that each word's output incorporates information about its relationships to others. This eliminates the need to store all preceding words as in earlier recurrent models (e.g., LSTM), instead capturing subtle word relationships through mathematical computation using key contextual information to predict content (see Vaswani et al., 2017 for mathematical details). Additionally, Transformers can process different words in text sequences in parallel, reducing computational

load compared to recurrent models and enabling training of larger, more powerful models (Götz et al., 2023). GPT was developed based on Transformer decoders as an autoregressive model that predicts the (n+1)th word using only the first n words of a sequence (with subsequent information masked), then trains the model by comparing predicted words against actual text. In essence, it predicts subsequent information based on preceding context, cycling until completing the entire text sequence. This autoregressive approach has proven more suitable for text generation than the "cloze test" format that simultaneously uses contextual information to infer missing middle words (Hommel et al., 2022). As a pre-trained model, GPT-3 contains 175 billion parameters and was trained on massive text corpora (Brown et al., 2020), enabling users to perform specific tasks (e.g., generating personality test items) with only a few examples (Few-Shot, typically 10-100), eliminating the need for large annotated training datasets or task-specific fine-tuning. Subsequent OpenAI releases of ChatGPT based on GPT-3.5 and GPT-4 have garnered widespread attention due to their conversational interface requiring no coding. Despite limited technical disclosure, GPT-4's superior performance across tasks, particularly its perfect score on AP Psychology exams (Achiam et al., 2023), makes its potential for generating non-cognitive items highly promising. Moreover, GPT-4 demonstrates excellent Chinese comprehension and generation capabilities, making validation of its capacity to generate quality items in Chinese contexts essential given the domestic research gap.

Therefore, this study aims to generate Chinese personality items using the latest version of ChatGPT and preliminarily explore the feasibility of constructing CAT item banks with these items, ultimately establishing a high-quality computerized adaptive item bank. Given the nascent stage of exploration in this domain, this study follows previous paradigms by continuing to use the widely adopted Big Five personality framework (John et al., 2008) as the item generation target. Since classic CAT requires a single measurement objective, we identified generating items for the emotional stability (ES) dimension as our final task, both because ES has garnered increasing societal attention and because it plays crucial roles in organizational management, mental health, school education, and decision-making behavior, being considered the most important predictor of psychological health (Bajaj et al., 2019; Bec & Becken, 2021; Margetić et al., 2022; Park et al., 2022; Wettstein et al., 2021).

Methodology

2.1 Measurement Instruments

To enable ChatGPT to generate appropriate items, we conducted a series of prompt engineering steps in Chinese: (1) Instructed GPT to role-play as an experienced psychologist; (2) Introduced definitions of emotional stability from renowned domestic and international scholars; (3) Described the relationship between ES and the Big Five model, explaining that ES as a personality trait is typically measured through the neuroticism subscale in Big Five inventories,

thereby tasking GPT with creating novel measurement items by referencing existing ones; (4) Specified fundamental principles for item development, including originality (avoiding excessive semantic similarity), non-repetition (avoiding duplication among new items and with reference items), clear measurement targets (emotional stability), rich item formats (avoiding structural similarity), and diversity and comprehensiveness. The final principle warrants particular attention, originating from prior research identifying excessive situational similarity in GPT-generated items (Lee et al., 2023). Therefore, when articulating this rule, researchers should provide GPT with as many situational contexts and relevant factors related to the measurement target as possible, as these explanations enhance generated item quality (Lampinen et al., 2022); (5) Finally, provided GPT sequentially with ES items, instructions, and scoring methods from different Big Five questionnaires as references (rather than imitation templates), requesting generation of equivalent numbers of items (including scoring methods).

During this process, if GPT generated items clearly violating development principles—such as directly copying reference items or duplicating previously generated items—we prompted GPT to review the fundamental principles and revise or regenerate the batch.

Following these procedures, we obtained 114 emotional stability items generated by GPT. We then invited 10 psychology graduate students to evaluate item grammar and content validity. Based on their evaluations, we first eliminated 1 improperly formulated item. Regarding content validity, the 10 experts rated each item's representativeness of emotional stability on a 4-point scale. Using these ratings, we calculated corrected kappa coefficients (k*; Polit et al., 2007). Following Polit et al.'s evaluation criteria, items with* $k > 0.74$ were considered high-quality, leading to elimination of all items not meeting this standard. Ultimately, 75 items were retained for formal testing. Notably, these 75 items underwent no human modification or editing, representing purely GPT-generated content.

Additionally, we selected 42 emotional stability (neuroticism) items from four widely-used Big Five personality scales for inclusion in the testing battery: 8 items from the Chinese Big Five Personality Inventory Brief Version (CBF-PI-B; Wang et al., 2011), 12 items from the BFI-2 Chinese version (Zhang et al., 2022), 2 translated items from TIPI-10 (Gosling et al., 2003), and 20 items from IPIP-BFAS (DeYoung et al., 2007). In this study, the Cronbach's α coefficients for the emotional stability dimension of these four scales were 0.921, 0.935, 0.799, and 0.955, respectively.

Items in the bank included multiple scoring formats, with each item retaining its original scoring method.

2.2 Participants

This study included three sample datasets. The primary Sample 1 employed convenience sampling on 117 emotional stability items, yielding 479 valid participants (163 males) with a mean age of 22.82 years (SD = 6.52). Additionally, we separately administered the emotional stability items from CBF-PI-B and BFI-2 Chinese versions for cross-sample reliability comparisons, with valid samples designated as Sample 2 and Sample 3. Sample 2 comprised 2,484 participants (820 males), while Sample 3 included 655 participants (197 males) with a mean age of 28.58 years (SD = 10.58).

2.3 Analytical Methods

Before implementing CAT, a high-quality item bank must first be established. For clarity, this study designates the 42-item bank from existing scales as the Classic Bank, the 75 GPT-generated items as the GPT Bank, and conducts principal component analysis and unidimensionality testing, IRT model selection, item analysis, and overall information and marginal reliability analysis on both banks using Sample 1 data. This process evaluates and compares the quality of both banks, eliminating non-compliant items. The final retained items from both sources are combined to form the Total Bank, which undergoes identical quality verification procedures to obtain a richer final bank and validate the combinability of items from both sources. Subsequently, we conduct simulated CAT on the three final banks to compare GPT-generated versus classic item performance and validate CAT's advantages over traditional testing. Specific procedures follow:

2.3.1 Principal Component Analysis and Unidimensionality Testing First, to ensure bank quality, items with low correlation to measurement objectives should be removed, necessitating principal component analysis (PCA) to eliminate items with primary component loadings below 0.4.

Subsequently, unidimensionality testing is performed on retained items. Unidimensionality represents a fundamental IRT assumption (Hambleton et al., 1991). Previous research indicates that in exploratory factor analysis (EFA), a ratio of first-to-second eigenvalues exceeding 4 with first-factor variance explanation above 20% satisfies the unidimensionality assumption (Reckase, 1979; Andrich, 1996; Reeve et al., 2007).

2.3.2 IRT Model Selection Since all items in this study employ polytomous scoring, available IRT models primarily include the Generalized Partial Credit Model (GPCM; Muraki, 1992) and Graded Response Model (GRM; Samejima, 1969). This study compares model fit indices, specifically AIC (Akaike, 1974) and BIC (Schwarz, 1978), selecting the better-fitting model for subsequent parameter estimation.

2.3.3 Item Analysis Word embedding, a crucial component of NLP models, has attracted research attention due to its consistent and pervasive gender bias (Gonen & Goldberg, 2019; Lee et al., 2023). Word embedding represents each word as a multi-dimensional vector, capturing relationships between words through vector differences in semantic space (Bolukbasi et al., 2016; Caliskan & Lewis, 2020). Real-world stereotypes may cause word embedding techniques to erroneously capture these biases (Garg et al., 2018). To prevent systematic gender bias in ES measurement, we conduct Differential Item Functioning (DIF) analysis. DIF occurs when examinees from different groups exhibit different statistical properties on an item despite matching on the latent trait being measured (Zumbo, 1999). This study employs logistic regression to test for gender-related DIF, with McFadden's $R^2 > 0.02$ indicating DIF requiring item removal (Choi et al., 2011). The calculation formula is:

$$\text{McFadden's } R^2 = 1 - \frac{\ln L}{\ln L_0}$$

Model 1: logit $P(u_i \geq k) = \alpha_k + \beta_1 \cdot \theta$
Model 2: logit $P(u_i \geq k) = \alpha_k + \beta_1 \cdot \theta + \beta_2 \cdot \text{gender}$
Model 3: logit $P(u_i \geq k) = \alpha_k + \beta_1 \cdot \theta + \beta_2 \cdot \text{gender} + \beta_3 \cdot \theta \cdot \text{gender}$

Where L_0 represents the baseline model's likelihood value and L represents the likelihood after adding predictor variables. $P(u_i \geq k)$ denotes the cumulative probability of responding at or above category k for item i ($1 \leq k \leq$ total response options), α represents the regression intercept, β denotes regression coefficients, and θ represents the trait level being measured (emotional stability in this study). For testing uniform DIF, Model 1 serves as the baseline and Model 2 adds predictors; for non-uniform DIF, Model 2 is the baseline and Model 3 adds the interaction term.

Item discrimination represents another crucial quality indicator, referring to an item's capacity to differentiate among examinees with varying ability levels, where higher values indicate better discrimination.

2.3.4 Item Bank Information and Marginal Reliability Item information represents the certainty level provided by an item in evaluating examinee trait levels, with higher values indicating greater reliability. Test information, the sum of all item information values, is inversely proportional to the square of the standard error. The calculation formula is:

$$SE(\theta) = \sqrt{\frac{1}{\sum_{i=1}^{m} I_i(\theta)}}$$

Where θ represents the examinee's latent trait level (emotional stability in this study), m denotes the total number of test items, and $I_i(\theta)$ indicates the information provided by item i for examinees with trait level θ.

Researchers have used Marginal Reliability (MR) to represent overall test reliability (Liu, 2022; Xu et al., 2020), calculated from the average measurement standard error across all examinees. The formula is:

$$\overline{SE(\theta)} = \frac{\sum_{i=1}^{N} SE(\theta_i)}{N}$$

$$MR = 1 - \overline{SE(\theta)}^2$$

Where N represents total examinee count, i indexes each examinee, and $SE(\theta_i)$ denotes the measurement standard error for examinee i's final θ estimate.

2.3.5 Simulated CAT After finalizing the item banks, we conducted simulated CAT based on 479 examinees' actual responses to all items across the three banks. CAT employed the maximally informative item selection method (MFI), currently the most widely used item selection strategy, with Expected A Posteriori (EAP) for ability estimation. For termination rules, we first implemented a rule requiring completion of all items in the total bank, treating resulting ability estimates as true values. We then employed fixed measurement precision rules, terminating tests when ability estimation standard errors reached specific values (using SE = 0.447/0.387/0.316/0.224 corresponding to marginal reliability = 0.8/0.85/0.90/0.95, and SE = 0.34/0.27/0.46/0.21 matching four existing measures' actual precision). We examined CAT performance by comparing required item numbers, ability estimation standard errors, and correlations with true ability values across three banks under different termination conditions. Subsequently, we employed fixed-length termination rules, stopping after administering predetermined numbers of items (corresponding to ES item counts in four existing scales = 12/8/20/2), estimating abilities and measurement errors as examinees would experience in traditional testing with these four scales. We compared measurement errors, marginal reliability, and ability correlations between CAT and traditional formats, then further compared reliability with Samples 2 and 3 traditional test results to comprehensively explore GPT bank feasibility and CAT performance improvements over traditional testing.

Validity constitutes another crucial CAT performance indicator (Xu et al., 2020). This study uses criterion validity as a reference, considering CAT valid only when its results correspond to criterion scale measurements. We used ES dimensions from CBF-PI-B, BFI-2, TIPI, and BFAS as criteria, calculating correlations between ability estimates from administering all items in each of the three banks and scores on these four scales to validate bank validity.

2.4 Research Tools

This study used ChatGPT based on OpenAI's GPT-4 Turbo version released in early November 2023 for item generation. SPSS 26.0 conducted principal

component analysis and unidimensionality testing, while R packages performed remaining analyses: the mirt package for IRT model selection, item discrimination analysis, and bank information calculation; the lordif package for DIF detection; and the catR package for simulated CAT.

Results

3.1.1 Unidimensionality Testing

Principal component analysis (PCA) on both banks' items revealed all items had primary component loadings exceeding 0.4 (details in Figure 1 [FIGURE:1]), thus all items were retained.

Subsequent Kaiser-Meyer-Olkin (KMO) tests confirmed data suitability for factor analysis, with results of 0.979 and 0.986 for Classic and GPT banks respectively ($\chi^2(861) = 17662.03, p < 0.001$; $\chi^2(2775) = 35225.09, p < 0.001$). Principal Axis Factoring (PAF) with oblique rotation yielded EFA results showing all factor loadings exceeded 0.4. The Classic bank's first and second eigenvalues were 22.75 and 2.29 (ratio = 9.92), with the first factor explaining 53.28% of variance. The GPT bank's first and second eigenvalues were 42.93 and 1.97 (ratio = 21.78), with the first factor explaining 56.79% of variance. Both banks thus satisfied the unidimensionality assumption.

3.1.2 IRT Model Selection

As shown in Table 1 , both banks exhibited smaller AIC and BIC values for GRM, indicating superior fit. Therefore, GRM was selected for subsequent analyses.

Table 1 Model Fit Indices

Bank	Loglik	AIC	BIC
Classic Bank	-	-	-
GPT Bank	-	-	-

3.1.3 Item Analysis

DIF analysis revealed all items had McFadden's R^2 values below 0.02, indicating no gender-related DIF and thus all items were retained.

Generally, discrimination values above 0.8 indicate high-quality items (Liu, 2022), warranting removal of items below this threshold. Refitting GRM to both banks showed all items exceeded 0.8 discrimination (details in Figure 2

), with means of 2.27 (SD = 0.51) and 2.44 (SD = 0.58) for Classic and GPT banks respectively. Difficulty ranges of [-3.1, 2.2] and [-4.3, 2.1] indicated broad

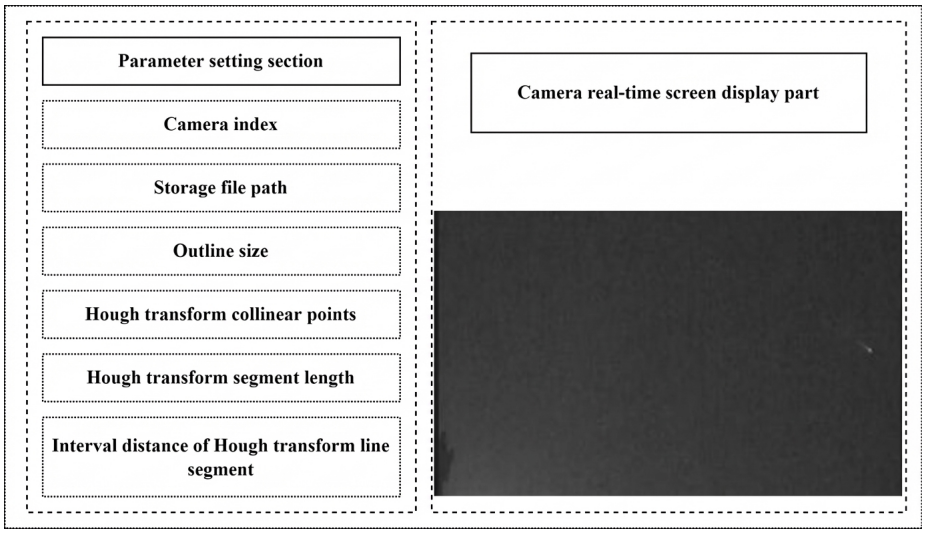

Figure 1: Figure 2

difficulty coverage. Overall, both banks demonstrated high quality, with GPT-generated items showing slightly superior discrimination and broader difficulty distribution than existing items.

3.1.4 Item Bank Information and Marginal Reliability

Figure 3

presents test information and standard errors for both banks. Generally, standard errors not exceeding 0.39 define low measurement error boundaries (Xu et al., 2020). Both banks provided high information and low measurement error for most examinees, with limitations only for those with extremely high emotional stability levels. Additionally, Figure 4 [FIGURE:4] shows marginal reliability, with average values reaching 0.96 and 0.98, indicating excellent overall reliability for both banks. Comparatively, the GPT bank demonstrated higher quality than the Classic bank, particularly for high trait-level examinees, maintaining low standard errors and higher reliability.

3.1.5 Total Bank Construction

Combining the 117 retained items from both banks formed the Total Bank. PCA on Total Bank items again showed all primary component loadings exceeding 0.4 (see Appendix Figure 6 [FIGURE:6]). KMO test results were 0.986 ($\chi^2(6786) = 57274.54, p < 0.001$). EFA using identical methods showed all factor loadings exceeded 0.4, with first and second eigenvalues of 62.98 and 4.47 (ratio = 14.10) and first-factor variance explanation of 53.83%, confirming the Total Bank's unidimensionality.

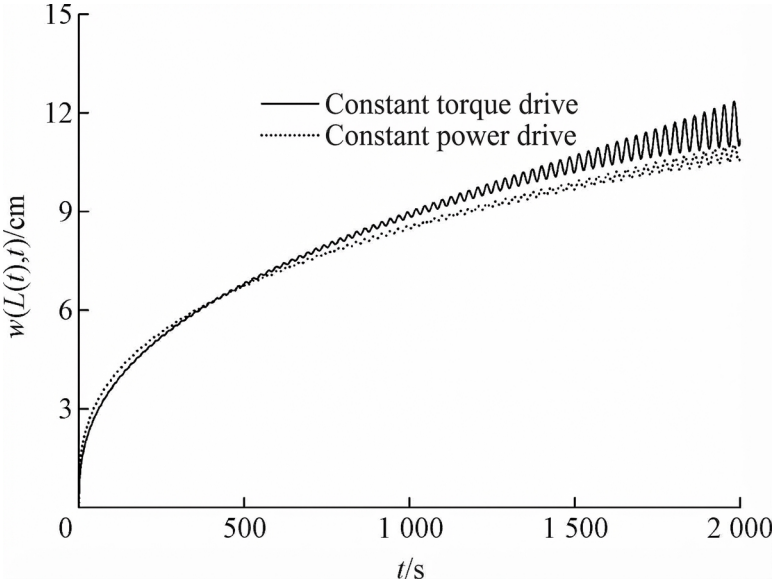

Figure 2: Figure 3

Fit indices again favored GRM (see Appendix Table 8), which was used for parameter estimation. All items maintained discrimination above 0.8 (mean = 2.25, SD = 0.50; see Appendix Figure 7 [FIGURE:7]). Total Bank information, standard error, and marginal reliability calculations (see Appendix Figures 8 [FIGURE:8] and 9 [FIGURE:9]) revealed smaller measurement errors and higher reliability than either constituent bank, providing precise measurement even for high emotional stability individuals, with average marginal reliability reaching 0.99.

Figure 3 Information and Standard Error for Classic Bank (left) and GPT Bank (right)
Figure 4 Marginal Reliability Curves for Classic Bank (left) and GPT Bank (right)

3.2.1 CAT Performance Under Fixed Precision Conditions

Table 2 displays simulated CAT results for three banks under different fixed-precision stopping rules. Even with SE( ) $\leq$ 0.447, all three banks required fewer than three items on average to achieve correlations between estimated and true abilities exceeding 0.9 (n = 479, p $<$ 0.001), though marginal reliability was suboptimal at this level. Generally, reliability coefficients reaching 0.85 indicate high reliability (Zhang et al., 2020; May et al., 2006). Achieving 0.87 marginal reliability required only 3.06–3.77 items, while 0.91 reliability needed merely 4.79–5.87 items. Higher reliability demands more than double the items (10.11–12.90) but yields marginal reliability of 0.95 and true-ability correlations

exceeding 0.97 (n = 479, p < 0.001). To illustrate CAT efficiency, Table 3 compares item numbers required to achieve traditional test precision levels, showing overall savings of 24.4%–47.5% across banks, with only minimal increase compared to TIPI's 2 items, likely due to CAT's initial random item selection when ability estimates are unavailable.

Comparing Classic and GPT banks reveals that under identical precision stopping rules, the GPT bank required fewer items than the Classic bank while achieving similar or slightly superior marginal reliability and true-ability correlations. When matching traditional test precision, the GPT bank required significantly fewer items than the Classic bank—for BFAS precision, the GPT bank reduced item count by 23.13% compared to the Classic bank ($t(478) = 17.20$, $p < 0.001$, Cohen's $d = 0.79$). The Total Bank showed CAT performance similar to the GPT bank.

In summary, under fixed measurement precision, the GPT bank's CAT performance exceeded the Classic bank, while CAT format enhanced testing efficiency, further reducing required items while maintaining measurement accuracy.

3.2.2 CAT Performance Under Fixed-Length Conditions

Tables 4 and 5 present measurement error reductions and reliability improvements when using traditional test lengths as CAT termination conditions. CAT based on the Classic Bank significantly reduced measurement error by 13.83%–22.69% compared to traditional tests (except TIPI) and increased reliability by 1.28%–5.40% ($p < 0.001$, Cohen's $d > 1$), with no significant difference from TIPI ($p > 0.05$). CAT from GPT and Total Banks significantly outperformed all four traditional tests, reducing measurement error by 6.50%–31.44% and 7.91%–31.65% respectively ($p < 0.001$), while increasing reliability by 2.20%–7.14% and 2.30%–7.17% ($p < 0.001$).

Table 6 shows cross-sample reliability improvements under fixed-length conditions, revealing generally consistent patterns with Sample 1 results. Even compared to traditional tests with larger sample sizes, all three CAT banks demonstrated significant reliability improvements ($p < 0.001$, Cohen's $d > 1$).

Comparisons revealed that under identical test lengths, the GPT bank significantly reduced measurement errors ($t(478) = 26.38, 30.35, 12.07, 37.13$, $p < 0.001$) and increased reliability ($t(478) = 22.53, 24.68, 11.13, 29.28$, $p < 0.001$) compared to the Classic bank, particularly converting negative to positive improvements over TIPI's 2-item length. The Total Bank showed marginally better performance than the GPT bank.

We further compared correlations among ability estimates from traditional methods and three CAT banks under fixed-length conditions, visualizing results through heatmaps. Figure 5 [FIGURE:5] shows correlations between CAT and traditional methods ranging from 0.79–0.97 (mean = 0.90). All three CAT banks demonstrated higher correlations with true ability values than traditional meth-

ods at equivalent lengths, confirming CAT's high accuracy. Notably, the GPT bank's true-ability correlations generally exceeded the Classic bank's, demonstrating superior measurement accuracy.

3.2.3 Validity Verification

Validity verification results appear in Table 7 , showing all three banks significantly correlated with all four criterion scales ($p < 0.001$) with coefficients exceeding 0.83, indicating ideal validity.

Table 7 Criterion-Related Validity of Three Banks

Bank	CBF-PI-B	BFI-2	TIPI	BFAS
Classic Bank	0.832***	0.913***	0.860***	0.914***
GPT Bank	0.831***	0.917***	0.861***	0.918***
Total Bank	0.836***	0.921***	0.862***	0.919***

Discussion

This study aimed to generate numerous Chinese emotional stability personality items using the latest ChatGPT version and explore the suitability of these item banks for CAT. Results demonstrate the GPT bank's high quality and superior CAT performance compared to classic items. The final Total Bank achieved adequate validity while substantially improving measurement precision and efficiency over traditional testing.

This study innovatively employs GPT to construct Chinese computerized adaptive item banks. Traditional item writing requires experts to leverage experience and knowledge to create new items through iterative review, revision, and refinement until meeting quality standards—a process both time-consuming and costly (Gierl et al., 2012). Our method provides an efficient and economical pathway for bank construction. Moreover, ChatGPT's conversational interface, unlike older versions requiring solid NLP and machine learning foundations for task-specific fine-tuning or programming (Hernandez & Nie, 2022; Götz et al., 2023; Lee et al., 2023), dramatically lowers the entry barrier for item developers.

ChatGPT based on GPT-4 maintains user-friendly operation while producing superior items. Compared to prior research, our generated items demonstrate higher validity. For example, correlating GPT-generated items with BFI-2 yielded a coefficient of 0.917 ($p < 0.001$) for emotional stability, substantially higher than the 0.786 ($p < 0.001$) reported for GPT-2 (Götz et al., 2023). Additionally, GPT-4 produced fewer invalid items, with a 65.78% retention rate after human quality judgment. In contrast, Götz et al. (2023) reported that 60% of 10,000 iteratively generated items completely duplicated example items, with only 92 items ultimately retained after expert review. Similarly, Hommel et al. (2022) found 1,077 of 1,360 generated items completely duplicated examples, with only 53.4% of remaining items receiving expert content validity

approval. Screening thousands of items proves both cumbersome and contrary to AIG's cost- and time-saving purpose, a pain point largely resolved by NLP advances.

Our exploration reveals GPT's enormous potential for generating non-cognitive items. The emerging substitutability of automatic item generation for human item writing signals a transformation in psychologists' responsibilities. Since GPT's item generation process for specific measurement targets remains complex and opaque, introducing unpredictability (Hommel et al., 2022), future efforts must emphasize accurate and detailed measurement target definitions to maximize structural validity, as Bejar (2013) noted: "Item generation and construct representation are complementary" (p. 43). Moreover, item quality still requires expert-defined evaluation criteria and review of generated items for content alignment, grammatical errors, and bias. Standardizing generation procedures, establishing comprehensive GPT prompting principles, and accurately defining measurement targets represent important future directions. OpenAI's recent GPTs feature, enabling users to create specialized ChatGPT versions with custom rules, demonstrates that customizing professional models using expert knowledge for specific tasks represents an inevitable trend.

Despite providing strong evidence for combining latest NLP-based AIG with CAT, this study has limitations. First, GPT-generated items still exhibit deficiencies, such as lacking reverse-scored items and variable item quality. Second, optimal CAT algorithm selection remains open for discussion—for instance, global Kullback-Leibler information-based item selection (Chang & Ying, 1996) may offer more robustness than Fisher information, and exposure control methods like maximum priority index (Cheng & Chang, 2009) warrant exploration.

Overall, NLP technology development provides accessible and effective tools for non-cognitive AIG and CAT. This study leverages the latest ChatGPT version, demonstrating large language models' enormous potential for computerized adaptive item bank generation. Future research should explore GPT's applicability to broader measurement targets, especially constructs lacking existing measures, and investigate whether more diverse formats like forced-choice items (Brown & Maydeu-Olivares, 2011) can be advanced through GPT.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, I., Akkaya, I., Aleman, F. L., … & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Akaike, H. (1974). Stochastic theory of minimal realization. *IEEE Trans. Automat. Control* 19, 667–674. doi: 10.1109/cdc.1976.267680

Andrich, D. (1996). A general hyperbolic cosine latent trait model for unfolding polytomous responses: reconciling thurstone-likert methodologies. *Br. J. Math. Stat. Psychol.* 10.1177/014662169301700307

Bajaj, B., Gupta, R., & Sengupta, S. (2019). Emotional Stability and Self-Esteem as Mediators Between Mindfulness and Happiness. *Journal of Happiness Studies*, 20(7), 2211–2226. https://doi.org/10.1007/s10902-018-0046-4

Bec, A., & Becken, S. (2021). Risk perceptions and emotional stability in response to Cyclone Debbie: An analysis of Twitter data. *Journal of Risk Research*, 24(6), 721–739. https://doi.org/10.1080/13669877.2019.1673798

Bejar, I. (2013). Item generation: Implications for a validity argument. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 40–55). Routledge.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Brown, A., & Maydeu-Olivares, A. (2011). Item Response Modeling of Forced-Choice Questionnaires. *Educational and Psychological Measurement*, 71(3), 460–502. https://doi.org/10.1177/0013164410375112

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.

Caliskan, A., & Lewis, M. (2020). Social biases in word embeddings and their relation to human cognition [Preprint]. *PsyArXiv*. https://doi.org/10.31234/osf.io/d84kg

Chang, H.-H., & Ying, Z. (1996). A Global Information Approach to Computerized Adaptive Testing. *Applied Psychological Measurement*, 20(3), 213–229. https://doi.org/10.1177/014662169602000303

Cheng, Y., & Chang, H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 62(2), 369–383. https://doi.org/10.1348/000711008X304376

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An R Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations. *Journal of Statistical Software*, 39(8). https://doi.org/10.18637/jss.v039.i08

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, 93(5), 880–896. https://doi.org/10.1037/0022-3514.93.5.880

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a Computer-adaptive Test for Depression (D-CAT). *Quality of Life Research*, 14(10), 2277–2291. https://doi.org/10.1007/s11136-005-6651-9

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings National*

Academy Sciences, 115(16). https://doi.org/10.1073/pnas.1720347115

Gierl, M. J., & Haladyna, T. M. (Eds.). (2013). *Automatic item generation: Theory and practice*. Routledge.

Gierl, M. J., & Lai, H. (2018). Using Automatic Item Generation to Create Solutions and Rationales for Computerized Formative Testing. *Applied Psychological Measurement*, 42(1), https://doi.org/10.1177/0146621617726788

Gierl, M. J., Lai, H., & Turner, S. R. (2012). Using automatic item generation to create multiple-choice test items: Automatic generation of test items. *Medical Education*, 46(8), 757–765. https://doi.org/10.1111/j.1365-2923.2012.04289.x

Goldberg, Y. (2017). Neural Network Methods for Natural Language Processing. *Synthesis Lectures on Human Language Technologies*, -, 1-309.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them (arXiv:1903.03862). arXiv. http://arxiv.org/abs/1903.03862

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Götz, F. M., Maertens, R., Loomba, S., & Van Der Linden, S. (2023). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods*. https://doi.org/10.1037/met0000540

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications Inc. doi: 10.2307/2075521

Hernandez, I., & Nie, W. (2022). The AI-IP: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology*, peps.12543. https://doi.org/10.1111/peps.12543

Hommel, B. E., Wollang, F.-J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-Based Deep Neural Language Modeling for Construct-Specific Automatic Item Generation. *Psychometrika*, 87(2), 749–772. https://doi.org/10.1007/s11336-021-09823-9

Jiao, H., & Lissitz, R. W. (Eds.). (2020). *Application of artificial intelligence to assessment*. Information Age Publishing, inc.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research*, 3(2), 114-158.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A Systematic Review of Automatic Question Generation for Educational Purposes.

International Journal of Artificial Intelligence in Education, 30(1), 121–204. https://doi.org/10.1007/s40593-019-00186-y

Lai, H., Gierl, M. J., Touchie, C., Pugh, D., Boulais, A.-P., & De Champlain, A. (2016). Using Automatic Item Generation to Improve the Quality of MCQ Distractors. *Teaching and Learning in Medicine*, 28(2), 166–173. https://doi.org/10.1080/10401334.2016.1146608

Lampinen, A. K., Dasgupta, I., Chan, S. C. Y., Matthewson, K., Tessler, M. H., Creswell, A., McClelland, J. L., Wang, J. X., & Hill, F. (2022). Can language models learn from explanations in context? (arXiv:2204.02329). arXiv. http://arxiv.org/abs/2204.02329

Lee, P., Fyffe, S., Son, M., Jia, Z., & Yao, Z. (2023). A Paradigm Shift from "Human Writing" to "Machine Generation" in Personality Test Development: An Application of State-of-the-Art Natural Language Processing. *Journal of Business and Psychology*, 38(1), 163–190. https://doi.org/10.1007/s10869-022-09864-6

Liu, X., Lu, H., Zhou, Z., Chao, M., & Liu, T. (2022). Development of a computerized adaptive test for problematic mobile phone use. *Frontiers in Psychology*, 13, 892387.

Margetić, B., Peraica, T., Stojanović, K., & Ivanec, D. (2022). Spirituality, Personality, and Emotional Distress During COVID-19 Pandemic in Croatia. *Journal of Religion and Health*, 61(1), 644–656. https://doi.org/10.1007/s10943-021-01473-6

May, S., Littlewood, C., & Bishop, A. (2006). Reliability of procedures used in the physical examination of non-specific low back pain: A systematic review. *Australian Journal of Physiotherapy*, 52(2), 91–102. https://doi.org/10.1016/S0004-9514(06)70044-7

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *ETS Res. Rep.* 16, i–30. doi: 10.1177/014662199201600206

Park, I.-J., Shim, S.-H., Hai, S., Kwon, S., & Kim, T. G. (2022). Cool down emotion, don't be fickle! The role of paradoxical leadership in the relationship between emotional stability and creativity. *The International Journal of Human Resource Management*, 33(14), 2856–2886. https://doi.org/10.1080/09585192.2021.1891115

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, 30(4), 459–467. https://doi.org/10.1002/nur.20199

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *J. Educ. Stat.* 4, 207–230. doi: 10.3102/10769986004003207

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi,

J. A., et al. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: plans for the patient-reported outcomes measurement information system (PROMIS). *Med. Care* 45, S22–S31. doi: 10.1097/01.mlr.0000250483.85507.04

Samejima, F. (1969). Estimation of latent ability using a pattern of graded responses. *Psychometrika* 34:100. doi: 10.1007/BF03372160

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Von Davier, M. (2018). Automated Item Generation with Recurrent Neural Networks. *Psychometrika*, 83(4), 847–857. https://doi.org/10.1007/s11336-018-9608-y

Wang, M.C., Dai, X.Y., Yao, S.Q. (2011). The Application of CAT on Emotional Intelligence with Item Response Theory. *Chinese Journal of Clinical Psychology*, 19(4), 454-457.

Wettstein, A., Ramseier, E., & Scherzinger, M. (2021). Class- and subject teachers' self-efficacy and emotional stability and students' perceptions of the teacher–student relationship, classroom management, and classroom disruptions. *BMC Psychology*, 9(1), 103. https://doi.org/10.1186/s40359-021-00606-6

Xu, L., Jin, R., Huang, F., Zhou, Y., Li, Z., & Zhang, M. (2020). Development of Computerized Adaptive Testing for Emotion Regulation. *Frontiers in Psychology*, 11, 561358. https://doi.org/10.3389/fpsyg.2020.561358

Zhang, B., Li, Y. M., Li, J., Luo, J., Ye, Y., Yin, L., Chen, Z., Soto, C. J., & John, O. P. (2022). The Big Five Inventory–2 in China: A Comprehensive Psychometric Evaluation in Four Diverse Samples. *Assessment*, 29(6), 1262–1284. https://doi.org/10.1177/10731911211008245

Zhang, L. F., Liu, K., Song, G., & Tu, D. B. (2020). The application of cat on emotional intelligence with item response theory. *Journal of Jiangxi Normal University (Natural Science)*, 44(5), 454–461.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. Ottawa: National Defense Headquarters, 160.

Appendix

Figure 6 [**FIGURE:6**] Factor loadings for all items in Total Bank
Table 8 Model fit indices for Total Bank
Figure 7 [**FIGURE:7**] Discrimination parameters for all items in Total Bank
Figure 8 [**FIGURE:8**] Information and standard error for Total Bank
Figure 9 [**FIGURE:9**] Marginal reliability curve for Total Bank