Journal of Pacific Rim Psychology

www.cambridge.org/prp

Original Article

Cite this article: Li Z., Cai Y., and Tu D. (2020) A new approach to assessing shyness of college students using computerized adaptive testing: CAT-Shyness. *Journal of Pacific Rim Psychology*, Volume 14, e20. https://doi.org/10.1017/prp.2020.15

Received: 26 December 2019
Revised: 7 August 2020
Accepted: 16 August 2020

Keywords:
computerized adaptive testing; shyness assessment; item response theory

Author for correspondence:
Dongbo Tu, Email: tudongbo@aliyun.com

© The Author(s), 2020. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is unaltered and is properly cited. The written permission of Cambridge University Press must be obtained for commercial re-use or in order to create a derivative work.

CAMBRIDGE
UNIVERSITY PRESS

A new approach to assessing shyness of college students using computerized adaptive testing: CAT-Shyness

Zifei Li[1,2], Yan Cai[2] and Dongbo Tu[2]

[1]Faculty of Psychology, Beijing Normal University, Beijing, China and [2]School of Psychology, Jiangxi Normal University, Nanchang, Jiangxi Province, China

Abstract

Assessing shy symptoms via computerized adaptive testing (CAT) provides greater measurement precision coupled with a lower test burden compared to conventional tests. The computerized adaptive test for shyness (CAT-Shyness) was developed based on a large sample of 1400 participants from China. Item bank development included the investigation of unidimensionality, local independence, and exploration of differential item functioning (DIF). CAT simulations based on the real data were carried out to investigate the reliability, validity, and predicted utility (sensitivity and specificity) of the CAT-Shyness. The CAT-Shyness item bank was successfully built and proved to have excellent psychometric properties: high content validity, unidimensionality, local independence, and no DIF. The CAT simulations needed 14 items to achieve a high measurement precision with a reliability of .9. Moreover, the results revealed that the proposed CAT-Shyness had acceptable and reasonable marginal reliability, criterion-related validity, and sensitivity and specificity. It not only had acceptable psychometric properties, but also had a shorter but efficient assessment of shyness, which can save significant test time and reduce the test burden for individuals with less information loss.

Shyness is a temperamental characteristic that is typically expressed in unfamiliar social situations and in feelings of social assessment, and includes feeling uncomfortable, excessively cautious, and sensitive (Crozier, 1995). From the perspective of social motivation, shy individuals have conflicts between social approach and social avoidance motivation. Although they are eager to participate in social interactions, they feel nervous and anxious in the face of communication (Asendorpf, 1990). Individuals in a collectivism culture have a stronger sense of self-blame, depression and loneliness than those in an individualistic culture (Zhao, Kong, & Wang, 2012). Previous studies have shown that the shame experience of Chinese college students was more common and more serious than that found in other countries, and it has seriously hindered the development of college students' social skills (Ban, 2010).

During a critical period of self-consciousness and identity establishment, college students understand themselves through interaction with others or society. However, shy or evasive behavior hinders this kind of interaction, which has different degrees of negative impact on the interpersonal communication and self-growth development of college students. Shy college students are prone to maladaptation and may not get along with teachers and classmates. The college student community is an important builder of the future society, and a sound personality and good social skills are of paramount importance to students. Therefore, analyzing college students' shyness and helping them to overcome shyness may have important value and significance to improve their mental health.

Shyness is the discomfort and inhibition state of an individual when others appear (Cheek & Buss, 1981). In the past few years, several scales have been developed to measure shyness, such as the Social Avoidance and Distress Scale (SAD; Watson & Friend, 1969), the Revised Shyness Scale (RCBS; Cheek & Buss, 1981), the Interaction Anxiety Scale (IAS; Leary, 1983a), the Shyness Syndrome Inventory (SSI, Melchior & Cheek, 1990) and the Stanford Shyness Questionnaire (Shy-Q, Bortnik, Henderson, & Zimbardo, 2002). However, none of these scales reveal the whole picture of shyness. For example, the IAS measures the cognitive component and emotional components of shyness, while the SSI measures the cognitive, somatic, and behavioral components of shyness (Su & Wu, 2008). These scales were compiled according to a classical test theory (CTT) framework and have fixed lengths. The only way to cover all aspects of shyness is to increase the number of items, but this would enlarge the test burden and reduce the test motivation (Forkmann et al., 2009). Besides, the scales usually contain items corresponding to various levels of shyness. A large number of items may deviate from respondents' symptoms of shyness, in that they are commonly required to answer each item of a questionnaire, which may increase individuals' measurement burden and prolong test time. A more rapid, convenient and accurate

https://doi.org/10.1017/prp.2020.15 Published online by Cambridge University Press

measurement method is needed to reduce the burden of shyness measurement and solve the problem of low measurement accuracy.

Computerized adaptive testing (CAT) is a new measurement technology based on item response theory (IRT) that has been developed over the past two decades. It is considered to be a suitable measurement method for various types of psychological assessment (Meijer & Nering, 1999). CAT selects an appropriate item based on the participant's trait (theta) from an item pool and then updates the trait according to the responses to this item. Compared with traditional paper-and-pencil testing, CAT has many advantages. First, in CAT administrations, IRT models enable selection of the most informative items for a particular range of shyness and allow for estimation of comparable test scores from any combination of items, along with individual assessment of measurement precision (Walter et al., 2005). Second, a CAT participant's motivation to respond increases, because the selected items correspond highly to their trait (Gibbons et al., 2008), and they may think that the test is tailored for their own condition. Third, while paper-and-pencil testing fixed the number of items, CAT can flexibly select items based on the participant's trait (theta); therefore, CAT greatly reduces the amount of test items and reduces the test burden of participants (Tonidandel, Quinones, & Adams, 2002). However, CAT also has disadvantages, such as being technically complex, having high initial costs, and requiring a substantial amount of human and financial resources to organize a CAT program. However, the advantages significantly outweigh the disadvantages (Meijer & Nering, 1999).

At present, measurement of shyness is mainly conducted with a questionnaire, and the existing shyness questionnaires have been mostly developed based on classical test theory, which is not conducive to the measurement and evaluation of shyness. Therefore, a new approach to assessing shyness in college students using CAT is worth exploring.

In this study, we aim to solve the problems mentioned above by establishing a new, more effective and accurate measurement for shyness using CAT (hereby referred to as the CAT-Shyness). The items in the initial CAT-Shyness bank were selected from seven widely used shyness scales according to the definition of shyness by Cheek and Buss (1981). Otherwise, the graded response model (GRM; Samejima, 1969) and the generalized partial credit model (GPCM; Muraki, 1992), which are polytomously scored IRT models, were compared based on test-level, model-fit checks to choose one optimal model to fit the data of the CAT-Shyness. Then, several statistical analyses, including unidimensionality, local independence, item fit, item discrimination, and differential item functioning (DIF) analyses were conducted to create the final item bank of the CAT-Shyness (see Appendix). Finally, both a CAT simulation study and a real data study were carried out to investigate the marginal reliability, convergent-related validity, and predictive utility (sensitivity and specificity) of the proposed CAT-Shyness.

Methods

Participants

About 1400 participants were recruited from four universities in Jiangxi Province, China. Before the survey, participants were informed that their personal information would be kept confidential and the test would take about 20 minutes. Participants volunteered to take part in the survey. After excluding some invalid data due to large missing responses, 1278 participants remained. The mean age was 20.06 years ($SD = 1.57$, ranging from 18 to 29 years). Table 1 contains the detailed demographic information. This study was approved by the Research Center of Mental Health, Jiangxi Normal University and the Ethics Committee of the Department of Psychology of Jiangxi Normal University. Written informed consent was obtained from all of the participants in accordance with the Declaration of Helsinki.

Table 1. Demographic characteristics ($N = 1278$)

Variables	Category	Frequency	Percent (%)
Gender	Male	611	47.81
	Female	667	52.19
Grade	Freshmen	358	28.01
	Sophomores	296	23.16
	Juniors	328	25.67
	Senior Students	296	23.16
Region	City	498	39.00
	Rural	780	61.00
Major	Arts	440	34.43
	Science	522	40.85
	Engineering	316	24.73

Measures

The initial item bank was determined by referring to previous studies and consisted of 117 items from seven widely used shyness scales, including the Revised Shyness Scale (RCBS; Cheek & Buss, 1981), Social Avoidance and Distress Scale (SAD; Watson & Friend, 1969), Brief Fear of Negative Evaluation Scale (BFNS; Leary, 1983a), Interaction Anxiousness Scale (IAS; Leary, 1983b), Shyness Scale (SS; Su & Wu, 2008), McCrosky Shyness Scale (MSS; McCrosky & Richmond, 1982), and the Shyness Syndrome Inventory (SSI; Melichor & Cheek, 1990). As different shyness scales may contain very similar or even the same topics, in order to avoid overlapping of the items in the item bank, those items with the same topics were removed. Based on previous studies, items from the seven selected scales could be classified into nine domains (Cheek & Buss, 1981; Leary, 1983a, 1983b; McCrosky & Richmond, 1982; Melichor & Cheek, 1990; Su & Wu, 2008; Watson & Friend, 1969): shyness, social avoidance, social distress, cognitive component of shyness, somatic component of shyness, emotional component of shyness, behavioral component of shyness, fear of negative evaluation, and interaction anxiousness. Table 2 contains detailed information about these scales.

All of the seven chosen scales are self-reported scales. The RCBS contains 13 items with a 5-point Likert-type scale (*Very uncharacteristic or untrue, strongly disagree* to *Very characteristic or true, strongly agree*). The SAD contains 19 items and each item has two levels (yes and no). The BFNS and the IAS both contain 12 items with a 5-point Likert-type scale (*Not at all characteristic* to *Extremely characteristic*). The SS contains 36 items with a 5-point Likert-type scale (*Not at all characteristic* to *Extremely charac*teristic). The SSI contains 13 items with a 5-point Likert-type scale (*Very uncharacteristic or untrue, strongly disagree* to *Very characteristic or true, strongly agree*). The MSS contains 12 items with a 5-point Likert-type scale (*Strongly disagree* to *Strongly agree*).

https://doi.org/10.1017/prp.2020.15 Published online by Cambridge University Press

Table 2. Sources and proportions of items

Scales	Item quantity	Proportion of total item bank	Dimension	Cronbach's alphas
RCBS	13	11.11%	1	.81
SAD	19	16.24%	2	.78
SSI	13	11.11%	3	.70
BFNS	12	10.26%	1	.81
IAS	12	10.26%	1	.77
MSS	12	10.26%	1	.80
SS	36	30.77%	4	.87
Total	117	100%		.96

Note: RCBS, Revised Shyness Scale; SAD, Social Avoidance and Distress Scale; SSI, Shyness Syndrome Inventory; BFNS, Brief Fear of Negative Evaluation Scale; IAS, Interaction Anxiousness Scale; MSS, McCrosky Shyness Scale; SS, Shyness Scale.

Except for the MSS and the SSI, the other five scales have a Chinese version. The RCBS was revised into Chinese for college students (Xiang, Ren, Zhou, & Liu, 2018). The results demonstrated that the Cronbach's alpha and retest reliability of the Chinese version of RCBS were .88 and .58 respectively. As for validity, the Chinese version of the RCBS had a close association with the Social Interaction Anxiety Scale (r = .77, p < .01). Peng, Fan, and Li (2003) modified the SAD in China and the results showed that the Cronbach's alpha and retest reliability of the Chinese version of the SAD were .85 and .76 respectively, and the subscale reliabilities of the Chinese version of SAD were .77 and .73 respectively. Regarding its validity, the Chinese version of SAD had a significant correlation with the IAS (r = .67, p < .01). The Chinese versions of the BFNS and IAS were developed by Wang, Wang, and Ma (1999). The Chinese version of the BFNS had a Cronbach's alpha of .90 and a retest reliability of .75. Regarding validity, the Chinese version had a close correlation with the SAD (r = .51, p < .01). The Chinese version of the IAS had a Cronbach's alpha of .87 and a retest reliability of .80. Regarding validity, the Chinese version of the IAS had a close correlation with the RCBS (r = .60, p < .01). The SS is a Chinese scale developed by Su and Wu (2008), with a Cronbach's alpha of .95 and a retest reliability of .90. Their findings indicated that the subscale reliabilities of the Chinese version of the SS were .80 ~ .87. Regarding validity, the SS had a close correlation with the RCBS (r = .88, p < .01).

Melchior and Cheek (1990) reported that in the SSI revision sample of 326 college students, the alpha internal consistency coefficient was .94; it had a 45-day retest reliability of .91 for a sample of 31 college students, with a correlation of .96 with the RCBS. The MSS had a Cronbach's alpha of .90 and had a significant correlation of .01 with the RCBS (McCrosky & Richmond, 1982). The MSS and the SSI were translated into Chinese. The translation of the MSS and the SSI were performed by six researchers with extensive experience in translation of self-report measurement. Three of them performed a forward translation of the items, and the other researchers performed an independent review of these translations. Following this, if there were different opinions on translation, discussions and revisions were needed by the six researchers and a professor of psychology. Revisions and seminars were repeated until consistent results are obtained. The confirmatory factor analysis (CFA) showed that the Chinese version of the MSS had the same structure as the original MSS (Tucker-Lewis index [TLI] = 0.89, confirmatory fit index [CFI] = 0.91, root mean square error of approximation [RMSEA] = 0.07, standardized root mean square residual [SRMR] = 0.06). The alpha coefficient for the Chinese version of the MSS was .80, and it has a close association with the RCBS (r = .42, p < .01). Regarding the Chinese version of the SSI, after setting the error terms of item 10 with item 8, and item 2 with item 1 to be related due to their content being very similar, the SSI had the same structure with the original SSI, with TLI = 0.91, CFI = 0.93, RMSEA = 0.05 and SRMSR = 0.04. The alpha coefficient for the Chinese version of the SSI was .70, and the SSI has a close association with the RCBS (r = .73, p < .01). These indicated that the Chinese version of the MSS and SSI have acceptable reliability and validity.

To validate the proposed CAT-Shyness, the Shyness Questionnaire (Shy-Q; Bortnik et al., 2002) was chosen as the external criteria scale. It is commonly used to diagnose shyness symptoms in a clinical setting. It is considered that the average score of participants is more than 3.5 (Henderson, Gilbert, & Zimbardo, 2014). There are 35 items in the scale, which are divided into four dimensions: self-blame, seeking approval, fear of rejection, and self-restriction of expression. The scale is 5-point Likert-type scale, with 1 = *Not at all characteristic* and 5 = *Extremely characteristic*. In this study, the Cronbach's alpha was .88.

Construction of the CAT-Shyness Item Bank

For construction of the CAT-Shyness item bank, statistical analyses based on IRT were sequentially carried out, including the IRT analyses of unidimensionality, local independence, item fit, item discrimination, and DIF.

Unidimensionality

Within the framework of IRT, the unidimensionality assumption was checked first. Given the clear correlation shown between the different personality traits (Muñiz, Suárez-Álvarez, Pedrosa, Fonseca-Pedrero, & García-Cueto, 2014), a unidimensional hypothesis for the battery was established. Robust maximum likelihood estimation method was used in the exploratory factor analysis (EFA).

In EFA, the unidimensional hypothesis is established when the first factor explains at least 20% of the total variance (Reckase, 1979) and the explanatory variance ratio of the first factor to the second factor is more than 4 (Reeve et al., 2007).

To confirm acceptable unidimensionality of the dataset, we first ran an EFA and eliminated items with factor loadings below 0.30 (Nunnally, 1978) on the first factor, and then reran the EFA to investigate the unidimensionality of the item pool.

Parameter estimation

Based on the 1278 response data, item parameters were estimated by expectation-maximization (EM) algorithm via IRTPRO2.1.

Model selection

In IRT, choosing an appropriate model for data analysis is the premise to ensure the accuracy of data analysis results. In this study, the commonly used Akaike information criterion (AIC), Bayesian information criterion (BIC), and -2 log-likelihood (-2LL) were used to determine which model fit best. The smaller these test-fit indices are, the better the model fit (Posada & Crandall, 2001).

Under the IRT framework, IRT models can be divided into two main categories: the difference models (or cumulative logit models) and the divided-by-total models (or adjacent logit models; Tu, Zheng, Cai, Gao, & Wang, 2017). The graded response model (GRM; Samejima, 1969) is a typical model in difference models; in addition, the generalized partial credit model (GPCM; Muraki, 1992) is a representative model of divided-by-total models. The

https://doi.org/10.1017/prp.2020.15 Published online by Cambridge University Press

GPCM is an extension of the partial credit model (PCM; Masters, 1982) by adding the discrimination parameter. The GRM has the same number of item parameters as the GPCM and belongs to the class of models that measures the response in order. After investigating a large number of studies, the above two models were not only commonly used polytomously scoring models in IRT, but also commonly used in CAT (e.g. Paap, Kroeze, Terwee, Palen, & Veldkamp, 2017). Therefore, the model with the smaller test-fit indices between the GRM and the GPCM was selected for further analysis.

Local independence

Local independence is also a necessary assumption of IRT models. It means that when controlling for trait levels, the response to any item is unrelated to the response for any other item (Embretson & Reise, 2000). In other words, there are no other underlying factors explaining the response behavior. Yen's Q_3 statistic (Yen, 1993) was used to test local independence, where Q_3 values higher than 0.36 were represented as locally dependent (Flens, Smits, Carlier, van Hemert, & de Beurs, 2016). Therefore, items with a Q_3 larger than 0.36 were removed from the item pool.

Item fit

The item-fit test was used to determine whether the item fitted to the IRT model, and the item-fit test was performed using the S-χ^2 statistic (Orlando & Thissen, 2003). Items with p values of S-χ^2 less than .01 were eliminated from the original item bank (Flens et al., 2016).

Item discrimination parameters

In GRM and GPCM, which are both two-parameter models, the relation is determined by two parameters: the discrimination parameter (a), giving information about the discriminative ability of an item; and item threshold parameter (b), indicating the location or difficulty of an item. According to Fliege's criteria (Fliege, Becker, Walter, Bjorner, & Rose, 2005), we deleted items with low discrimination (<.7).

Differential item functioning

DIF was analyzed to identify item bias for a wide range of variables, such as gender or region, to build nonbiased item banks. DIF analyses were conducted using the polytomous logistic regression method (Swaminathan & Rogers, 1990) via the package lordif (Choi, Gibbons, & Crane, 2011). Change in McFadden's pseudo R^2 was used to evaluate effect size, and the hypothesis of no DIF was rejected when R^2 change $\geq .2$ (Flens et al., 2016), so these items were removed from the final analysis. We evaluated DIF for region (rural, city) and gender (male, female) groups.

The IRT analyses were all done in R package mirt (Version 1.24; Chalmers, 2012). The analyses of unidimensionality, local independence, item discrimination, item fit, and differential item functioning were repeated until all remaining items of CAT-Shyness sufficiently satisfied the above rules.

CAT-Shyness Simulated Study

After the final item bank was established, the CAT simulation was carried out. Based on the CAT-shyness real item bank parameters, the performance of the CAT-shyness in different shy levels was simulated to test its feasibility and rationality and its related algorithm. The social shyness trait levels of the subjects were simulated and ranged from −3.5 to 3.5 intervals of 0.25. Each point simulated 100 subjects, and a total of 2900 subjects were simulated. All analyses were done in R (Version 3.4.1) and catR package for R studio (Magis & Raiche, 2011).

Starting point, scoring algorithm, item selection algorithm, and stopping rule

The first step was to determine the starting point. In CAT simulation, item selection depends on the participants' responses to a given item. At first, however, the participant knows nothing about prior information. Therefore, a simple and effective method is to randomly select the first item from the final item bank (Magis & Barrada, 2017).

The second step used a scoring algorithm to estimate the score on the latent trait of the simulated subjects. The expected a posterior estimation (EAP) method was used to estimate the person parameters. First, this method can effectively utilize the information provided by the entire posterior distribution, and the EAP algorithm has high stability. Second, it does not need iteration and the calculation process is simpler. The simplicity and stability of the EAP makes it a widely used method for CAT simulations (e.g., Bulut & Kan, 2012; Chen, Hou, & Dodd, 1998). Third, the accuracy of EAP estimates are higher than the MLE (e.g., Sorrel, Barrada, de la Torre, & Abad, 2020).

The third step was to determine the item selection algorithm. Maximum Fisher information criterion (van der Linden, 1998) is the most widely used item selection algorithm in CAT programs. Its purpose is to improve the accuracy of measurement, but it is also likely to lead to uneven exposure of items in the item bank and reduce the safety of testing (Barrada, Olea, Ponsoda, & Abad, 2009). However, as Likert-type scales require participants to respond in the usual way, the response results without correct answers greatly reduces the safety of the test. Therefore, maximum Fisher information criterion was chosen as the item selection algorithm.

Finally, the stopping rules were based on the standard error (*SE*) of measurement. That is, the CAT will be stopped if participants' *SE* of measurement reaches the predefined *SE* of measurement, which is also called the variable length termination rule.

The relationship between the *SE* and the Fisher information can be defined as

$$\mathrm{SE}(\theta) = \frac{1}{\sqrt{\sum_{j=1}^{n} I_j(\theta)}}$$

where n is the number of items the participant has answered. In this study, several stopping rules with different *SE*s were performed, including $SE \leq .50$, $SE \leq .45$, $SE \leq .40$, $SE \leq .35$, $SE \leq .30$, $SE \leq .25$ and $SE \leq .20$.

Properties of the item pool

In order to explore the estimation results of simulated subjects under different stopping rules, bias, mean absolute deviation (MAD), root mean square error (RMSE), correlation coefficient between the subject's true shyness trait, and the estimated shyness trait by CAT-Shyness were all investigated to determine the effectiveness of the CAT-Shyness related algorithms.

The exposure rate (ER) index was used to measure the security of the item pool. $\mathrm{ER}_j = f_j/N$ ERj is the exposure rate of item j, and f_j is the number of times that j is selected. The smaller the ERj, the lower the exposure rate. The chi-squared statistic is used to reflect the overall exposure of the item bank as

$$\chi^2 = \sum_{j=1}^{\mathrm{M}} \frac{\left[ER_j - E\left(ER_j\right)\right]^2}{E\left(ER_j\right)}$$

where $\mathrm{E(ER_j)} = \mathrm{L/M}$ is the expected exposure rate of item j, L represents the test length, and M is number of items in the item pool

https://doi.org/10.1017/prp.2020.15 Published online by Cambridge University Press

Table 3. AUC indicator size description

Numerical interval of AUC	Predictive utilities
0.5	None
0.5–0.7	Small
0.7–0.9	Moderate
0.9–1	High

Note: AUC, area under curve.

(Chang & Ying, 1999). The chi-squared index reflects the difference between the observed item exposure rate and expected exposure rate. The smaller the chi-squared index, the safer the item pool.

The CAT-Shyness real study

In this part, we used real participants' data that had already been collected and used in development of the item pool. The CAT program stopping rules were also set to when the SE (θ) of measurement reached .50, .45, .40, .35, .30, .25 or .20. The parameter estimation method and item selection algorithm have been discussed above.

Characteristics of the CAT-Shyness

To investigate the characteristics of the CAT-Shyness, several statistics were calculated: number of items used (including the means and standard deviations), mean standard error of theta estimates, marginal reliability, Pearson's correlation between the estimated theta in the CAT-Shyness, and the estimated theta via the entire item bank. The marginal reliability is the mean reliability for all levels of theta (Smits, Cuijpers, & van Straten, 2011). The ER index is also calculated to measure the security of the item pool.

In addition, the number of selected items under several stopping rules was plotted as a function of the final theta estimation and test information curve. The test information shows the measurement precision of the CAT-Shyness: the larger the value, the smaller the error of the theta estimation.

Convergent-related validity of the CAT-Shyness

Convergent-related validity refers to how closely the new scale is related to other variables and other measures of the same construct (Paul, 2017). To further investigate the convergent-related validity of the CAT-Shyness, the Shy-Q (Bortnik et al., 2002), which is widely used in diagnosing shyness, was selected as the criterion scale. Pearson's correlation between the estimated theta in the CAT-Shyness and the score of the Shy-Q was calculated to address the convergent-related validity of CAT-Shyness.

Predictive utility (sensitivity and specificity) of the CAT-Shyness

The area under curve (AUC) under the receiver operating characteristic (ROC) curve index was used as an additional criterion to investigate the predictive utility (sensitivity and specificity; Smits et al., 2011) of the CAT-Shyness. A larger AUC index indicates a better diagnostic effect (Kraemer & Kupfer, 2006). We used the Shy-Q (Bortnik et al., 2002) as the classified variable for shyness. Moreover, the estimated theta in CAT-Shyness was used as a continuous variable to plot the ROC curve under each stopping rule. The meaning of the AUC sizes is shown in Table 3 (Forkmann et al., 2013).

Determination of the critical value was calculated by maximizing the Youden Index (YI = sensitivity + specificity – 1; Schisterman, Perkins, Liu, & Bondell, 2005). The sensitivity indicates the probability that a patient is accurately diagnosed with a disease, and specificity is the probability of patients without disease who test negative. The bigger the two values, the better the effect of the diagnosis.

Table 4. Test-level model-fit for three polytomously scored IRT models

Model	AIC	BIC	-2LL
GRM	229593.8	231387.0	228897.8
GPCM	231192.1	232985.4	230496.2

Note: GRM, graded response model; GPCM, generalized partial credit model; AIC, Akaike's information criterion; BIC, Bayesian information criterion; -2LL, -2 log-likelihood.

Results

Construction of the CAT-Shyness item bank

In the initial item bank development, 117 items were examined for unidimensionality, local independence and item characteristics (reliability curves, test information curves, differential item functioning, threshold, location parameters).

Unidimensionality

In order to determine whether the data were suitable for factor analysis, the Kaiser-Meyer-Olkin (KMO) test and the Bartlett spherical test were performed on the predicted data. The results showed that the KMO value was 0.96, which is higher than the minimum standard of .70, indicating that the data were suitable for factor analysis.

We performed an EFA on 117 items in the initial item bank and deleted the item with the lowest factor loading less than .30 on the first factor; the EFA model then re-estimated after each item removal. We estimated 43 EFA models. In the last EFA model, all items had a factor loading greater than .30, and 75 items remained from the initial item bank. The results show that the first eigenvalue of the factor was 17.34, the second eigenvalue was 3.08, the ratio of the two eigenvalues was 5.63, and the variance explained by the first eigenvalue was 23.12%. The results satisfy that the ratio of the two eigenvalues is greater than 4 (Reeve et al., 2007) and the variance explained on the first eigenvalue is greater than 20% (Reckase, 1979). Therefore, the initial item bank consisting of the remaining 75 items satisfies the IRT one-dimensional hypothesis.

Model selection

Table 4 shows the fitness of the GRM model and the GPCM model with the real data. It can be seen that the GRM model's -2LL, AIC, and BIC fitting index values were lower than the GPCM model, which indicated that the GRM fitted the data better than the GPCM. Therefore, the GRM was applied to the IRT analysis.

Local independence

Three items were deleted from the item bank because their Q3 values were all greater than .36. The remaining items met the local independence well.

Item fit

An item-fitting test was performed on the remaining 72 items, and it was found that the p values of $S\text{-}\chi^2$ of six items were less than the critical value of .01, so these items were removed from the item bank (Reeve et al., 2007). The p values of $S\text{-}\chi^2$ of the remaining 66 items were higher than .01 (see Table 5).

https://doi.org/10.1017/prp.2020.15 Published online by Cambridge University Press

Table 5. Item parameters for 66 item-bank with GRM

			Item parameters				Item-fit estimates			
Item	*M*	*SD*	a	b1	b2	b3	b4	*S-χ^2*	*df*	*p*	Subdomain
RCBS 1	1.82	1.15	1.24	−1.94	−0.09	0.48	3.01	205.68	233	.90	1
RCBS 4	1.80	1.07	1.61	−1.91	−0.09	0.63	2.70	200.13	207	.62	1
RCBS 5	1.79	1.05	1.33	−2.16	−0.13	0.84	2.90	227.80	220	.34	1
RCBS 7	1.90	1.09	1.35	−2.18	−0.25	0.52	2.76	230.35	223	.35	1
RCBS 8	2.34	1.07	1.01	−3.35	−1.23	−0.18	2.42	277.37	241	.05	1
RCBS 10	1.82	1.12	1.26	−2.04	−0.15	0.69	2.72	271.88	244	.11	1
RCBS 11	1.84	1.07	1.43	−2.12	−0.19	0.69	2.67	226.52	213	.25	1
RCBS 13	1.95	1.13	1.18	−2.27	−0.43	0.56	2.60	257.62	243	.25	1
RCBS 14	2.01	1.17	1.18	−2.11	−0.47	0.27	2.63	240.67	249	.64	1
SAD1	0.65	0.50	0.75	−0.87	–	–	–	96.40	110	.82	2
SAD4	0.36	0.48	0.85	0.75	–	–	–	103.11	104	.51	3
SAD7	0.31	0.46	0.78	1.16	–	–	–	123.41	105	.11	3
SAD8	0.37	0.48	0.72	0.80	–	–	–	129.19	110	.10	2
SAD9	0.55	0.50	1.02	−0.21	–	–	–	102.52	102	.47	3
SAD11	0.46	0.50	0.91	0.19	–	–	–	94.97	105	.75	3
SAD15	0.50	0.50	0.91	0.03	–	–	–	129.32	108	.08	2
SAD16	0.44	0.50	1.06	0.27	–	–	–	107.10	103	.37	3
SAD17	0.41	0.49	0.95	0.45	–	–	–	97.04	103	.65	2
SAD19	0.38	0.48	0.97	0.62	–	–	–	118.48	101	.11	2
SSI1	1.97	1.18	1.36	−1.95	−0.33	0.25	2.38	230.60	238	.62	7
SSI4	1.73	1.05	1.28	−2.18	0.04	0.95	3.06	224.90	221	.41	7
SSI5	1.30	1.06	0.98	−1.29	0.69	1.99	4.03	294.60	255	.04	4
SSI6	1.98	1.06	0.83	−3.27	−0.74	0.86	3.55	284.81	234	.01	4
SSI8	1.84	1.14	1.06	−2.18	−0.30	0.78	3.01	227.08	251	.86	4
SSI9	1.26	1.09	0.98	−1.12	0.75	1.89	4.14	252.25	257	.57	5
SSI10	2.11	1.12	1.22	−2.26	−0.68	0.11	2.78	252.17	237	.24	4
SSI11	2.09	1.11	1.23	−2.40	−0.61	0.27	2.56	261.34	233	.10	7
SSI13	2.02	1.08	1.56	−2.12	−0.45	0.36	2.39	200.62	200	.47	7
BFNS 1	2.08	1.13	1.05	−2.60	−0.70	0.34	2.63	285.31	253	.08	8
BFNS 3	1.93	1.04	1.03	−2.81	−0.52	0.72	3.43	200.47	220	.82	8
BFNS 5	1.93	1.06	1.11	−2.61	−0.50	0.71	3.07	274.01	223	.01	8
BFNS 6	1.79	1.05	1.07	−2.45	−0.21	0.97	3.47	216.44	235	.80	8
BFNS 8	2.06	1.05	1.07	−2.93	−0.76	0.49	2.98	216.63	217	.49	8
BFNS 11	2.08	1.13	1.07	−2.61	−0.72	0.42	2.49	254.40	258	.55	8
BFNS 12	2.19	1.08	1.20	−2.80	−0.82	0.16	2.43	254.51	228	.11	8
IAS1	1.42	1.00	1.74	−1.28	0.33	1.25	3.05	202.53	207	.57	9
IAS2	1.84	1.06	1.90	−1.73	−0.19	0.58	2.48	202.11	183	.16	9
IAS5	1.67	1.02	1.45	−1.83	−0.03	1.09	2.93	201.39	217	.77	9
IAS7	2.48	1.00	0.90	−4.05	−1.82	−0.42	2.51	261.57	247	.25	9
IAS10	2.11	1.08	1.39	−2.33	−0.65	0.26	2.41	199.60	211	.70	9
IAS11	1.90	1.08	1.31	−2.13	−0.38	0.63	2.74	233.56	231	.44	9

(*Continued*)

https://doi.org/10.1017/prp.2020.15 Published online by Cambridge University Press

Table 5. (*Continued*)

			Item parameters					Item-fit estimates			
Item	*M*	*SD*	a	b1	b2	b3	b4	*S-χ*^{2}	*df*	*p*	Subdomain
MSS3	1.88	0.99	1.01	−2.91	−0.57	1.08	3.59	215.44	217	.52	7
MSS6	1.87	1.07	1.05	−2.47	−0.41	0.84	3.27	248.59	236	.27	7
MSS9	2.03	0.94	0.80	−4.07	−1.20	0.93	4.54	199.27	197	.44	7
MSS10	1.95	0.99	0.80	−3.69	−0.85	1.08	4.24	217.85	220	.53	7
SS1	1.98	0.98	0.96	−3.22	−0.86	0.95	3.55	205.23	202	.42	7
SS2	2.19	0.96	0.94	−3.75	−1.30	0.39	3.37	230.60	193	.03	4
SS3	1.61	1.06	1.03	−2.06	0.08	1.43	3.46	258.30	250	.35	5
SS9	2.10	1.03	1.20	−2.89	−0.80	0.42	2.72	200.19	201	.50	5
SS10	2.22	1.00	1.37	−2.82	−1.03	0.26	2.28	188.09	206	.81	6
SS11	2.18	1.03	1.22	−2.91	−0.93	0.33	2.45	245.08	218	.10	7
SS14	2.19	0.98	0.90	−3.85	−1.42	0.49	3.23	239.42	217	.14	7
SS16	1.86	1.03	1.10	−2.66	−0.38	0.93	3.19	219.93	218	.45	5
SS18	1.83	0.98	1.21	−2.50	−0.42	1.03	3.41	185.57	206	.84	5
SS20	1.91	0.99	1.24	−2.64	−0.53	0.87	2.97	155.29	194	.98	7
SS21	1.82	1.00	1.38	−2.22	−0.38	1.02	2.79	219.56	210	.31	4
SS22	1.56	1.02	1.23	−1.83	0.13	1.36	3.37	259.46	236	.14	5
SS23	1.61	1.01	1.39	−1.83	0.06	1.24	3.15	220.24	224	.56	6
SS24	1.88	1.06	0.92	−2.67	−0.58	1.03	3.61	260.56	242	.20	7
SS25	2.20	1.01	0.82	−3.86	−1.47	0.37	3.52	231.35	224	.35	4
SS28	1.78	1.01	1.05	−2.65	−0.25	1.14	3.54	244.72	223	.15	4
SS29	1.58	1.02	1.12	−1.99	0.10	1.46	3.62	231.80	241	.65	5
SS30	1.57	1.01	1.21	−1.86	0.10	1.39	3.52	213.42	237	.86	6
SS31	1.91	1.03	1.41	−2.15	−0.48	0.74	2.70	237.05	209	.09	5
SS32	1.85	1.04	1.40	−2.17	−0.33	0.83	2.61	188.23	216	.91	6
SS36	2.08	1.01	1.24	−2.69	−0.90	0.55	2.60	228.54	204	.11	6

Note: The 10–19 items are scored with 0, 1, and the other items are scored with 0, 1, 2, 3, 4; subdomain: (1) shyness, (2) social avoidance, (3) social distress, (4) cognitive component of shyness, (5) somatic component of shyness, (6) emotional component of shyness, (7) behavioral component of shyness, (8) fear of negative evaluation, (9) interaction anxiousness. a, discrimination parameter; b, location parameters; *S-χ*^{2} = Orlando and Thissen's S-statistic.

Item discrimination parameters

After the item parameters were estimated, we excluded items with a slope below .70 and re-estimated the remaining items. The number of response options per item varied between 2 and 5, with 10 dichotomous items and 56 items with 5. The threshold parameters of the remaining 66 items in the final CAT varied between −4.07 and +4.54 (Table 5), which evenly covered a broad range of the shy symptoms. Slope parameters varied between .72 and 1.90 (Table 5).

Differential item functioning (DIF)

In order to ensure the fairness of the test, this study separately tested the functional differences of the subjects in terms of region (rural, city) and gender (male, female) groups, and found that the change of R^2 in all items was less than .02, that is, there was no DIF for the remaining 66 items, so no item was deleted.

Finally, the remaining 66 items were reanalyzed using the one-dimensionality test, local independence test, item fit test, discrimination test, and DIF test; all 66 items met the measurement requirements set out above. The average discrimination parameter of the 66 items was 1.14 ($SD = 0.24$), indicating that the whole item bank had a higher quality.

CAT-Shyness Simulated Study

Table 6 documents the ability estimation results of simulated subjects under different stopping rules. The results indicate that whatever stopping rules were used, the bias MAD was less than 0.5, which indicated that the estimation of the shyness of a participant in the CAT-Shyness had an ideal recovery. The RMSE was also less than 0.5 when the stopping rule was $SE \leq .20$, $SE \leq .25$, $SE \leq .30$, $SE \leq .35$, $SE \leq .40$. However, when the stopping rule was $SE \leq 0.45$ or $SE < 0.50$, the error was slightly larger (RMSE exceeds 0.5), which also indicated that the stopping rule with $SE \leq .45$ or $SE < .50$ may not be suitable for CAT-Shyness. Researchers or users can choose the other five stopping rules, which are $SE \leq .20$, $SE \leq .25$, $SE \leq .30$, $SE \leq .35$ or $SE \leq .40$. Table 6 also shows that under any stopping rule, the correlation coefficient between CAT-Shyness estimated shy traits and their true values was as high

Table 6. Ability estimation results of simulated subjects under different stopping rules

Stopping rule	Bias	MAD	RMSE	*r*	χ^2
SE (θ) ≤ .20	-0.008	0.183	0.229	.994***	0.162
SE (θ) ≤ .25	-0.005	0.213	0.268	.993***	20.682
SE (θ) ≤ .30	0.004	0.266	0.333	.989***	30.390
SE (θ) ≤ .35	0.003	0.309	0.387	.986***	35.914
SE (θ) ≤ .40	-0.004	0.373	0.468	.980***	38.411
SE (θ) ≤ .45	-0.015	0.412	0.512	.978***	37.024
SE (θ) ≤ .50	-0.005	0.469	0.577	.973***	38.323

Note: MAD, mean absolute deviation; RMSE, root mean square error; *r*, estimation of the relationship between the estimated value and the true value; ***$p < .001$; χ^2, the difference between the observed item exposure rate and expected exposure rate.

Table 7. Characteristics of CAT-Shyness under different stopping rules

	Number of items used					
Stopping rule	Mean	*SD*	Mean SE(θ)	Marginal reliability	*r*	χ^2
SE (θ) ≤ 0.20	62.489	2.992	0.202	0.959	0.999***	1.729
SE (θ) ≤ 0.25	31.782	4.562	0.249	0.938	0.986***	29.388
SE (θ) ≤ 0.30	20.153	3.688	0.297	0.912	0.967***	37.869
SE (θ) ≤ 0.35	13.979	2.144	0.345	0.881	0.950***	41.903
SE (θ) ≤ 0.40	10.322	1.552	0.391	0.847	0.927***	43.648
SE (θ) ≤ 0.45	7.925	1.237	0.438	0.808	0.903***	40.894
SE (θ) ≤ 0.50	6.321	0.870	0.482	0.768	0.883***	42.815

Note: *r*, the Pearson's correlation between the estimated theta in the CAT-Shyness and the estimated theta via the entire item bank; ***$p < .001$; χ^2, the difference between the observed item exposure rate and expected exposure rate.

as .97 or more ($p < .001$), and there was a highly significant positive correlation. It indicated that they had high consistency, and the correlation was stronger as the estimation accuracy increases. Table 6 shows the chi-squared index was smallest when the stopping rule was *SE* ≤ .20, which means the item pool is mostly safe. The chi-squared index was not significantly different when the stopping rule was *SE* ≤ .25, *SE* ≤ .30, *SE* ≤ .35, *SE* ≤.40, *SE* ≤ .45, and *SE* ≤ .50.

Figure 1 shows the average amount of items used by the simulated subjects in CAT-Shyness (*SE* ≤ 0.35). As the degree of shyness level increased, the average number of items used by the participants decreased. When a participant's shy trait value ranged from −1.5 to 1.5, the average test was less than 15 items (less than a quarter of the total item bank) and the marginal reliability of the test was greater than .88. This result shows that CAT-shyness is suitable for the measurement of shyness, and it can greatly reduce the test items without losing the accuracy of measurement.

Figure 2 displays the reliability and the whole test information for every simulated tester. For participants with a theta between −2 and 3, the marginal reliability of the test reached .95 or higher; and for participants with a theta between −3 and −2, the marginal reliability of the test was also around .94, which is higher than .9. All of these proved that the quality of this item bank is good.

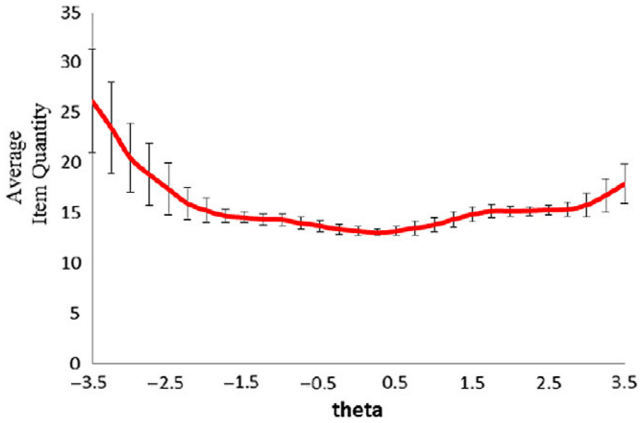

Figure 1. Item usage with *SE* ≤ .35 of competence.

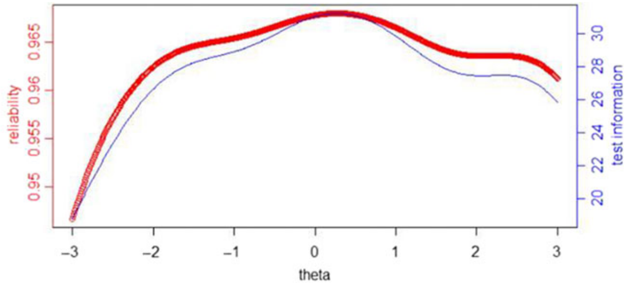

Figure 2. The reliability and the test information for every simulated tester completed for each item.

CAT-Shyness Real Study

Characteristics of CAT-Shyness

Table 7 shows the measurement characteristics based on the real test of CAT-Shyness, including the number of items used by the participants on CAT-Shyness, the average measurement error and the marginal reliability of the measurement. The last column of Table 7 shows the Pearson's correlation between the estimated theta in the CAT-Shyness and the estimated theta via the entire item bank.

Table 7 shows that the average length of the test used by the 1278 participants on the CAT-Shyness will increase as the measurement accuracy increases. When the stopping rule was *SE* ≤ .35, the correlation between the CAT-Shyness estimated trait value of the subject and the trait value estimated using the entire item bank was as high as .95 ($p < .001$). The length of the test was about 14 items for the stopping rule of *SE* ≤ .35, which only accounts for one fifth of the total number of items in the item bank and can save 1−1/5 = 80% of the items. Regardless of the stopping rule, the correlation coefficients between the estimated theta in the CAT-Shyness and the estimated theta via the entire item bank were all higher than .9 ($p < .001$), and the average measurement error was below .5, which met the measurement requirements under the framework of IRT.

Table 7 also shows that when *SE* ≤ .20, *SE* ≤ .25, *SE* ≤ .30 and *SE* ≤ .35, the measured marginal reliabilities of the CAT-Shyness were higher than .85; and when *SE* ≤ .40, *SE* ≤ .45 and *SE* ≤ .5, the measured marginal reliability of CAT-Shyness was between .75 and .85, which were all generally acceptable for individuals. This means that if the measurement reliability of CAT-Shyness in real measurement is to be above .85, the stopping rule strategy

https://doi.org/10.1017/prp.2020.15 Published online by Cambridge University Press

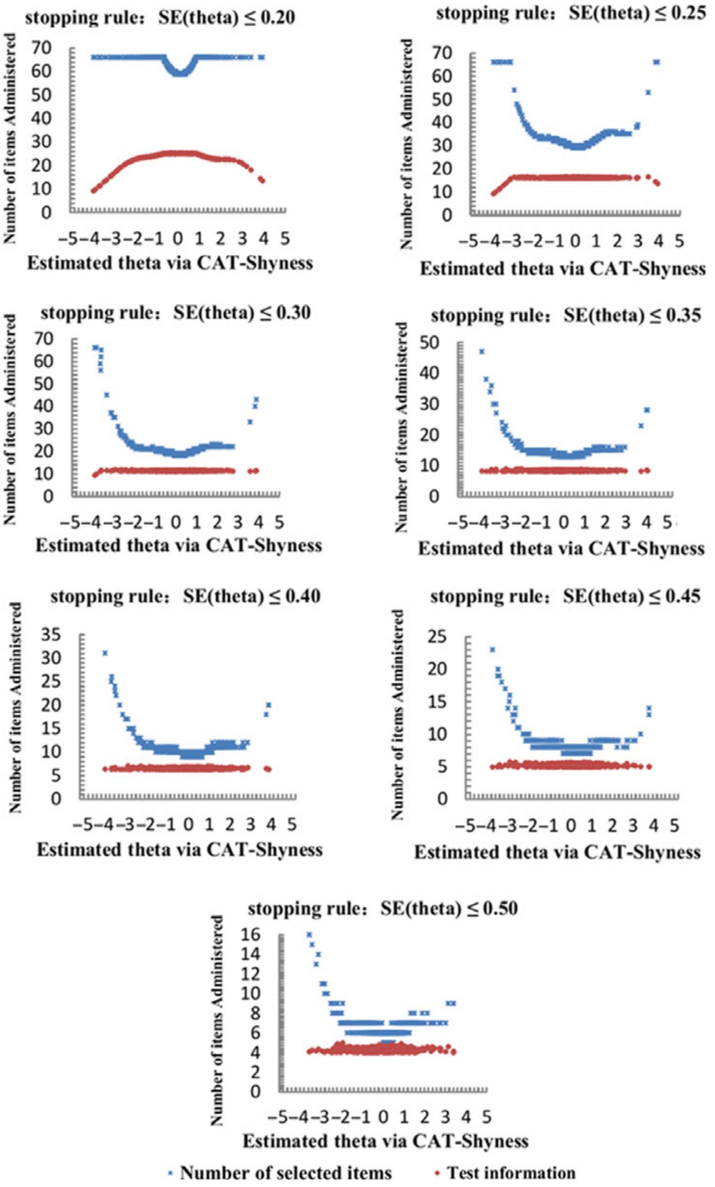

Figure 3. Number of selected items and test information curve under different stopping rules.

of CAT-Shyness can be set to $SE \le .35$ or less, but the test length used in consideration of $SE \le .3$ is more than 1.5 times than that of $SE \le .35$, so the termination rule of $SE \le .35$ can take into account both measurement efficiency and measurement accuracy.

Table 7 shows the chi-squared index was smallest when the stopping rule was $SE \le .20$, which means the item pool is mostly safe. The chi-squared index was not significantly different when the stopping rule was $SE \le .30$, $SE \le .35$, $SE \le .40$, $SE \le .45$, and $SE \le .50$.

Figure 3 displays the number of items administered intuitively across the latent trait under each stopping rule. As documented in Figure 3, despite a great quantity of items administered to individuals with lower trait/theta, the test information was still low. Individuals with middle or high trait/theta only needed to do fewer items and the test information was higher. For example, under the stopping rule SE (theta) ≤ 0.20, (1) the test information was less than 10 for those whose theta was less than −3 even if the entire item bank was administered to them; while (2) the test information was over 25 for those whose theta ranged from −1 to 1.

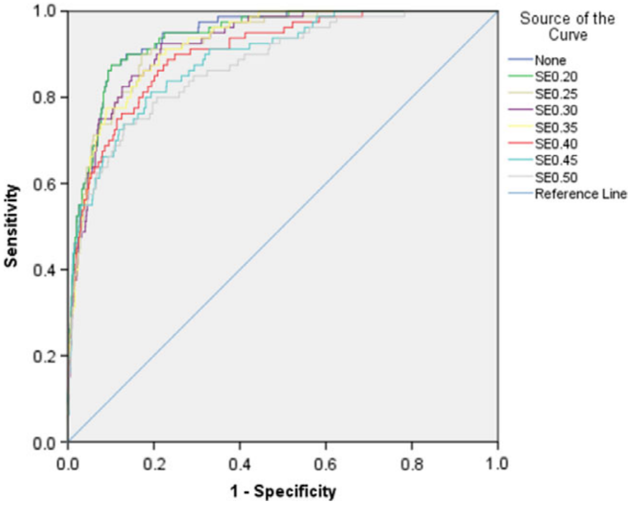

Figure 4. Receiver operating characteristic (ROC) curve for different stopping rules in the CAT-Shyness.

https://doi.org/10.1017/prp.2020.15 Published online by Cambridge University Press

Table 8. Criterion-related validity of CAT-Shyness with external criteria scales

Stopping rule	Shy-Q (95% CI)
SE (θ) ≤ .20	0.776 (0.741, 0.811)
SE (θ) ≤ .25	0.748 (0.712, 0.785)
SE (θ) ≤ .30	0.734 (0.697, 0.771)
SE (θ) ≤ .35	0.729 (0.692, 0.767)
SE (θ) ≤ .40	0.712 (0.674, 0.751)
SE (θ) ≤ .45	0.690 (0.651, 0.730)
SE (θ) ≤ .50	0.677 (0.637, 0.717)

Table 9. The predictive utility (sensitivity and specificity) of CAT-Shyness

Stopping rule	AUC (95% CI)	Sensitivity	Specificity	YI
SE (θ) ≤ .20	0.940 (0.918, 0.962)	0.875	0.899	0.774
SE (θ) ≤ .25	0.926 (0.900, 0.951)	0.900	0.829	0.729
SE (θ) ≤ .30	0.924 (0.899, 0.949)	0.925	0.783	0.708
SE (θ) ≤ .35	0.925 (0.900, 0.949)	0.900	0.788	0.688
SE (θ) ≤ .40	0.904 (0.871, 0.936)	0.888	0.769	0.656
SE (θ) ≤ .45	0.890 (0.854, 0.926)	0.800	0.819	0.619
SE (θ) ≤ .50	0.875 (0.834, 0.915)	0.738	0.868	0.606

Note: YI, Youden-Index = sensitivity + specificity – 1.

Convergent-related validity of CAT-Shyness

As shown in Table 8, the Pearson's correlations between CAT-Shyness score and the Shyness Questionnaire (Shy-Q) score ranged from .738 to .776 under different stopping rules. The results indicated that no matter which stopping rule was used, CAT-Shyness had an acceptable and reasonable external convergent-related validity.

Predictive utility (sensitivity and specificity) of CAT-Shyness

Table 9 presents the results of the CAT's diagnostic accuracy based on the Shyness Questionnaire. The AUC values under all stopping rules were all higher than the critical value of .7, which is widely used as the lower bound for moderate predictive utility. In addition, the sensitivity and specificity of CAT-Shyness were also both acceptable. These results suggested the predictive utility of CAT-Shyness was reasonable.

Discussion

In this study, CAT-Shyness was developed using the GRM with a large sample of college students and then the characteristics, marginal reliability, convergent-related validity and predictive utility (sensitivity and specificity) of CAT-Shyness were validated. In order to build a high-quality item bank, items were carefully selected from seven widely used shyness scales. Then, a strict one-dimensional test was carried out to test whether the assumptions of the IRT model were satisfied. In addition, two commonly used polytomous IRT models were compared, and the GRM model was selected to fit the real data due to it having better test level fit than the GPCM. Finally, after sequential analyses of IRT, a high-quality item bank was constructed that included 66 items, and these items had local independence, good item fit, high discrimination and no DIF. Besides, the mean IRT discrimination parameters of the item bank reached 1.14, which clearly showed that the final item bank of CAT-Shyness was of high quality.

Differing from other studies, this study took full account of the efficiency of CAT and investigated both simulated and real data. The simulation study indicated that the two stopping rules of *SE* (theta) ≤ .35 and *SE* (theta) ≤ .40 were the most effective and precise stopping rules. The results revealed that about 14 items on average were used to estimate shyness under stopping rule *SE* (theta) ≤ .35, while only about 6 items were needed with stopping rule *SE* (theta) ≤ .5. In addition, the theta estimates of the CAT and the full-item bank were very similar and the correlation was very high, exceeding .88 ($p < .001$). Moreover, CAT-Shyness had an acceptable marginal reliability with an average of .88, ranging from .77 to .96. These results show that the proposed CAT-Shyness not only has high measurement accuracy, but can also greatly shorten the test length (Smits et al., 2011).

A further investigation into the convergent-related validity and predictive utility (sensitivity and specificity) of CAT-Shyness was then carried out. The results revealed that: (1) CAT-Shyness had reasonable and acceptable convergent-related validities; (2) the sensitivity and specificity of the CAT-Shyness were both acceptable, and especially for stopping rule *SE* (theta) ≤ .35, the sensitivity and specificity of the CAT-Shyness were .90 and .79 respectively. So, the CAT-Shyness had good screening performance. The sensitivity (.738–.925) and specificity (.783–.899) of the CAT-Shyness were both acceptable. The minimum probability that a patient was accurately diagnosed with a disease, and that general people were accurately diagnosed with no illness were .738 and .783, respectively, which were higher than the random level (.5).

Despite the encouraging results, this study also had some limitations. First, the participants in this study mainly came from Jiangxi Province, China. Therefore, more samples should be recruited from a wider range of provinces in the future. Second, CAT-Shyness provided little information for those with a latent trait theta higher than 3 or lower than −3, suggesting that CAT-Shyness may not be suitable for these individuals. Future research could aim to develop a CAT appropriate for these people. Third, considering the problems of item exposure rate and item elimination, the existing item bank could be further expanded. Thus, future research needs to further supplement high-quality items and conduct in-depth research on exposure control and other related issues. Moreover, CAT-Shyness is a one-dimensional test and future research could further consider how to report nine domains under CAT-Shyness. Finally, the Shy-Q was selected to determine convergent validity of CAT-Shyness, but it could not show the relationship between CAT-Shyness and some external criteria. In future, other scales could be used to measure convergent-related validity that is a variable to determine the relationship between the CAT-Shyness and a criterion. Furthermore, we chose the maximum Fisher information item selection rule because of its popularity and availability (Magis & Barrada, 2017). However, this rule has a slightly worse performance in terms of accuracy, compared to global information item selection rules (Sorrel et al., 2020), and differences are greater for extreme latent trait levels (i.e., <−1.5, >1.5). Future studies could use different item selection rules, such as the global information item selection rule, to improve test accuracy; and the CAT could be probably shorter with a different item selection rule. As CAT-Sshyness has both dichotomous and polytomous items in the item pool, future research could add polytomous items to the item pool and delete dichotomous

https://doi.org/10.1017/prp.2020.15 Published online by Cambridge University Press

items to create a unified (polytomous) item pool, and to improve the item discrimination parameter of the item pool.

In addition, future research could use parallel analysis (Lim & Jahng, 2019), which is another method to determine the number of factors retained in exploratory factor analysis. Parallel analysis is a recommended procedure for deciding on the number of components involved in extracting eigenvalues from random datasets that parallel the actual dataset with regard to the number of cases and variables (O'Connor, 2000). Parallel analysis is more objective and rigorous in determining potential factors, and it is the gold standard for determining dimensionality (e.g., Lim & Jahng, 2019).

Conclusions

The proposed CAT-Shyness not only had acceptable psychometric properties, but also had a shorter yet efficient assessment of shyness, which can save significant test time and reduce the test burden for individuals, with less information loss.

Acknowledgments. The authors appreciated the anonymous reviewers who made very helpful comments on an earlier version of this article, and are very grateful to all the participants who were involved in this study. This research was funded by the National Natural Science Foundation of China (31960186, 31760288, 31660278).

References

Asendorpf J.B. (1990). Beyond social withdrawal: Shyness, unsociability, and peer avoidance. *Human Development*, **33**, 250–259.

Ban M.L. (2010). On college students' shyness and loneliness and their relationship. *Chinese Journal of Special Education*, **123**, 92–96.

Barrada J.R., Olea J., Ponsoda V. and Abad F.J. (2009). Item selection rules in computerized adaptive testing: Accuracy and security. *Methodology European Journal of Research Methods for the Behavioral and Social Sciences*, **5**, 7–17.

Bortnik K., Henderson L. and Zimbardo P.G. (2002, November). *The shy Q, A measure of chronic shyness: Associations with interpersonal motives and interpersonal values and self-conceptualizations.* Poster presented at the 36th Annual Conference of the Association for the Advancement of Behavior Therapy, Reno, NV.

Bulut O. and Kan A. (2012). Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Eurasian Journal of Educational Research*, **49**, 61–80.

Chalmers R.P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, **48**, 1–29.

Chang H.H. and Ying Z. (1999). A stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, **23**, 211–222.

Choi S.W., Gibbons L.E. and 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**, 1–30.

Cheek J.M. and Buss A.H. (1981). Shyness and sociability. *Journal of Personality and Social Psychology*, **41**, 330–339.

Chen S.K., Hou L. and Dodd B.G. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. *Educational and Psychological Measurement*, **58**, 569–595.

Crozier W.R. (1995). Shyness and self-esteem in middle childhood. *British Journal of Educational Psychology*, **65**, 85–95.

Embretson S.E. and Reise S.P. (2000). *Item response theory for psychologists.* London: Lawrence Erlbaum Associates.

Flens G., Smits N., Carlier I., van Hemert A.M. and de Beurs E. (2016). Simulating computer adaptive testing with the mood and anxiety symptom questionnaire. *Psychological Assessment*, **28**, 953–962.

Fliege H., Becker J., Walter O.B., Bjorner J.B. and Rose K.M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research*, **14**, 2277–2291.

Forkmann T., Boecker M., Norra C., Eberle N., Kircher T., Schauerte P. and Wirtz M. (2009). Development of an item bank for the assessment of depression in persons with mental illnesses and physical diseases using Rasch analysis. *Rehabilitation Psychology*, **54**, 186–197.

Forkmann T., Kroehne U., Wirtz M., Norra C., Baumeister H., Gauggel S., . . . Boecker, M. (2013). Adaptive screening for depression — Recalibration of an item bank for the assessment of depression in persons with mental and somatic diseases and evaluation in a simulated computer-adaptive test environment. *Journal of Psychosomatic Research*, **75**, 437–443.

Gibbons R.D., Weiss D.J., Kupfer D.J., Frank E., Fagiolini A., Grochocinski V.J. . . . Immekus, J.C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services*, **59**, 361–368.

Henderson L., Gilbert P., and Zimbardo P. (2014). Shyness, social anxiety, and social phobia. In S. Hofmann and P. DiBartolo (Eds.), *Social anxiety* (3rd ed., pp. 95–115). Cambridge, MA: Academic Press.

Kraemer H.C. and Kupfer D.J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry*, **59**, 990–996.

Leary M.R. (1983a). A brief version of the fear of negative evaluation scale. *Personality and Social Psychology Bulletin*, **9**, 371–375.

Leary M.R. (1983b). Social anxiousness: The construct and its measurement. *Journal of Personality Assessment*, **47**, 66–75.

Lim, S. and Jahng, S. (2019). Determining the number of factors using parallel analysis and its recent variants. *Psychological Methods*, **24**, 452–467.

McCroskey J.C. and Richmond V.P. (1982). Communication apprehension and shyness: conceptual and operational distinction. *Central States Speech Journal*, **33**, 458–468.

Magis D. and Barrada J.R. (2017). Computerized adaptive testing with R: Recent updates of the package catR. *Journal of Statistical Software*, **76**, 1–19.

Magis D. and Raiche G. (2011). CatR: an R package for computerized adaptive testing. *Applied Psychological Measurement*, **35**, 576–577.

Masters G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, **47**, 149–174.

Meijer R.R. and Nering M.L. (1999). Computerized adaptive testing: overview and introduction. *Applied Psychological Measurement*, **23**, 187–194.

Melchior L.A. and Cheek J.M. (1990). Shyness and anxious self-preoccupation during a social interaction. *Journal of Social Behavior and Personality*, **5**, 117–130.

Muñiz J., Suárez-Álvarez J., Pedrosa I., Fonseca-Pedrero E. and García-Cueto E. (2014). Enterprising personality profile in youth: *Components and assessment. Psicothema*, **26**, 545–553.

Muraki E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, **16**, 159–176.

Nunnally J.C. (1978). *Psychometric theory* (2nd ed.) New York, NY: McGraw-Hill.

O'Connor B.P. (2000). SPSS and BAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, and Computers*, **32**, 396–402.

Orlando M. and Thissen D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, **27**, 289–298.

Paap M.C.S., Kroeze K.A., Terwee C.B., Palen J.V.D. and Veldkamp B.P. (2017). Item usage in a multidimensional computerized adaptive test (MCAT) measuring health-related quality of life. *Quality of Life Research*, **26**, 2909–2918.

Paul F.M.K. (2017). *The measurement of health and health status: Concepts, methods and applications from a multidisciplinary perspective* (1st ed.). Groningen, The Netherlands: Academic Press.

Peng C.Z., Fan X.L. and Li L.C. (2003). The validity and reliability of Social Avoidance and Distress Scale in Chinese students. *Chinese Journal of Clinical Psychology*, **11**, 279–281.

Posada D. and Crandall K.A. (2001). Selecting the best-fit model of nucleotide substitution. *Systematic Biology*, **50**, 580–601.

Reckase M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, **4**, 207–230.

Reeve B.B., Hays R.D., Bjorner J.B., Cook K.F., Crane P.K., Teresi J.A., . . . PROMIS Cooperative Group. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-

https://doi.org/10.1017/prp.2020.15 Published online by Cambridge University Press

Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, **45**, S22–S31.

Samejima F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, **17**, 5–17.

Schisterman E.F., Perkins N.J., Liu, A. and Bondell H. (2005). Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology*, **16**, 73–81.

Smits N., Cuijpers, P., and van Straten A. (2011). Applying computerized adaptive testing to the CES-D scale: a simulation study. *Psychiatry Research*, **188**, 147–155.

Sorrel, M.A., Barrada, J.R., de la Torre, J. and Abad, F.J. (2020). Adapting cognitive diagnosis computerized adaptive testing item selection rules to traditional item response theory. *Plos One*, **15**, e0227196.

Su S.M. and Wu Y.Y. (2008). A study on the revise of the Shyness Scale and its validity. *Journal of Education and Psychology*, **31**, 53–82.

Swaminathan H. and Rogers H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, **27**, 361–370.

Tonidandel, S., Quinones, M.A. and Adams, A.A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology*, **87**, 320–332.

Tu D.B, Zheng C.J., Cai Y., Gao X.L. and Wang D.X. (2017). A polytomous model of cognitive diagnostic assessment for graded data. *International Journal of Testing*, **18**, 231–252.

van der Linden W.J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, **63**, 201–216.

Walter O., Becker J., Fliege H, Bjorner J.B., Kosinski M., and Klapp B.F. and Rose M. (2005). Developmental steps for a computer adaptive test for anxiety (A-CAT). *Diagnostica*, **51**, 88–100.

Wang X.D., Wang X.L. and Ma H. (1999). *Rating Scales of Mental Health* (rev. ed.). Beijing, China: Chinese Mental Health Journal Publisher.

Watson D. and Friend R. (1969). Measurement of social-evaluative anxiety. *Journal of Consulting and Clinical Psychology*, **33**, 448–457.

Xiang B.H., Ren L.J., Zhou Y. and Liu J.S. (2018). Psychometric properties of Cheek and Buss Shyness Scale in Chinese college students. *Chinese Journal of Clinical Psychology*, **26**, 268–271.

Yen W.M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement*, **30**, 187–213.

Zhao J.J., Kong, F. and Wang Y.H. (2012). Humor style as mediator on the relationship between shyness and loneliness in college students. *Chinese Journal of Clinical Psychology*, **20**, 102–104.

https://doi.org/10.1017/prp.2020.15 Published online by Cambridge University Press

Appendix: Final item bank of the CAT-Shyness

Item code	Item	Subdomain
RCBS 1	I feel tense when I'm with people I don't know well.	1
RCBS 4	I am often uncomfortable at parties and other social functions.	1
RCBS 5	When in a group of people, I have trouble thinking of the right things to talk about.	1
RCBS 7	It is hard for me to act natural when I am meeting new people.	1
RCBS 8	I feel nervous when speaking to someone in authority.	1
RCBS 10	I have trouble looking someone right in the eye.	1
RCBS 11	I feel inhibited in social situations.	1
RCBS 13	I am shyer with members of the opposite sex.	1
RCBS 14	During conversations with new acquaintances, I worry about saying something foolish.	1
SAD1	I try to avoid situations which force me to be very sociable.	2
SAD4	I often find social settings upsetting.	3
SAD7	I often feel nervous or tense in casual get-togethers in which both sexes are present.	3
SAD8	I often want to get away from people.	2
SAD9	I usually feel uncomfortable when I am in a group of people I don't know.	3
SAD11	Being introduced to people makes me tense and nervous.	3
SAD15	I don't mind talking to people at parties or social gatherings.	2
SAD16	I am seldom at ease in a large group of people.	3
SAD17	I often think up excuses in order to avoid social engagements.	2
SAD19	I try to avoid formal social occasions.	2
SSI1	During conversations with new acquaintances, I worry about saying something dumb.	7
SSI4	It is hard for me to act natural when I am meeting new people.	7
SSI5	I feel painfully self-conscious when I am around strangers.	4
SSI6	I am confident about my social skills.	4
SSI8	I often have doubts about whether other people like to be with me.	4
SSI9	Sometimes being introduced to new people makes me feel physically upset (for example, having an upset stomach, pounding heart, sweaty palms, or heat rash).	5
SSI10	I worry about how well I will get along with new acquaintances.	4
SSI11	I am shy when meeting someone of the opposite sex.	7
SSI13	I feel inhibited in social situations.	7
BFNS 1	I worry about what people will think of me even when I know it doesn't make any difference.	8
BFNS 3	I am frequently afraid of other people noticing my shortcomings.	8
BFNS 5	I am afraid that others will not approve of me.	8
BFNS 6	I am afraid that others will find fault with me.	8
BFNS 8	When I am talking to someone, I worry about what they may be thinking of me.	8
BFNS 11	Sometimes I am too concerned with what other people may think of me.	8
BFNS 12	I often worry that I will say or do the wrong things.	8
IAS1	I often feel nervous even in casual get-togethers.	9
IAS2	I usually feel uncomfortable when I'm in a group of people I don't know.	9
IAS5	Parties often make me feel anxious and uncomfortable.	9
IAS7	I would be nervous if I was being interviewed for a job.	9
IAS10	In general, I am a shy person.	9
IAS11	I often feel nervous when calling someone I don't know very well on the telephone.	9
MSS3	Other people think I am shy.	7

(*Continued*)

https://doi.org/10.1017/prp.2020.15 Published online by Cambridge University Press

Appendix: (*Continued*)

Item code	Item	Subdomain
MSS6	I don’t talk much.	7
MSS9	Most people talk more than I do.	7
MSS10	Other people think I am very quiet.	7
SS1	I’ll lower my voice when I talk to strangers.	7
SS2	I’ll pay attention to myself, especially my behavior, when I meet strangers.	4
SS3	I’ll blush when I ask the waiter to take back the incorrect meal.	5
SS9	When I speak to an authority, I’ll sweat and have a rapid heartbeat.	5
SS10	When I am the center of attention, I’ll be upset and anxious.	6
SS11	I’ll keep quiet when I’m in a crowded occasion.	7
SS14	In large communities, I’ll avoid to become the center of other’s attention.	7
SS16	When I introduce myself to someone, my heart will beat faster and my palms will sweat.	5
SS18	In unfamiliar social situation, my pulse will become rapid and I’ll quiver.	5
SS20	When I introduce myself, I’ll be incoherent.	7
SS21	When I speak to authority, I’d like to go away.	4
SS22	I often feel physical discomfort when I’m with a bunch of people basically unknown to me.	5
SS23	I often feel upset when I’m in a social occasion.	6
SS24	I’ll avoid sexual intimacy.	7
SS25	In a large group, I’ll worry about making a good impression on others.	4
SS28	When I ask for help, I’ll keep thinking about my clumsiness and my weakness.	4
SS29	I often feel physical discomfort when I’m in a large group.	5
SS30	I’ll be nervous when I ask the waiter to take back the incorrect meal.	6
SS31	When I become the center of attention, my heart beats faster, I’ll sweat and tremble.	5
SS32	I am afraid to ask questions in community for avoiding embarrassment.	6
SS36	If I knew someone was looking at me, I’ll get nervous.	6

Note: Subdomain: (1) shyness, (2) social avoidance, (3) social distress, (4) cognitive component of shyness, (5) somatic component of shyness, (6) emotional component of shyness, (7) behavioral component of shyness, (8) fear of negative evaluation, (9) interaction anxiousness.

https://doi.org/10.1017/prp.2020.15 Published online by Cambridge University Press