Journal of Pacific Rim Psychology

www.cambridge.org/prp

A computerized adaptive testing advancing the measurement of subjective well-being

Yifang Wu, Yan Cai and Dongbo Tu

School of Psychology, Jiangxi Normal University, Nanchang, China

Original Article

Cite this article: Wu Y., Cai Y., and Tu D. (2019). A computerized adaptive testing advancing the measurement of subjective well-being. *Journal of Pacific Rim Psychology*, Volume 13, e6. https://doi.org/10.1017/prp.2019.6

Received: 11 July 2018
Revised: 17 January 2019
Accepted: 19 January 2019

Keywords:
subjective well-being; computerized adaptive testing; item bank

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

Abstract

This article aimed at developing an adaptive version of the subjective well-being (SWB) scale to measure a comprehensive concept of SWB among Chinese university students. Item response theory was employed to formulate the item bank of the SWB scale and computerized adaptive testing (CAT) for SWB (CAT-SWB), based on several commonly used SWB scales, after unidimensionality testing, model selection, local dependence testing, parameter estimation, item fit test and differential item functioning (DIF) analysis were performed. Finally, two CAT simulations using simulated-data and real-data were carried out to verify and evaluate the CAT-SWB. Results indicated that the proposed CAT-SWB had an excellent performance in that it largely reduces the number of test items and the length of test time without losing measurement precision.

The definition of happiness or well-being can be divided into three kinds (Diener, 1984). The first category emphasizes that well-being has been defined by external criteria such as virtue or holiness, and the criterion of happiness is not the subjective judgment of the actor, but the observer's value framework. However, the second category of happiness relies on the respondent's judgment of a good life, and Diener (1984) also pointed out that happiness was almost equal with life satisfaction. The third kind of definition of happiness emphasizes a pleasant emotional experience. Since more researchers have considered the latter two categories (i.e. life satisfaction and the positive and negative affect of subjective well-being; Diener, Oishi, & Lucas, 2003; Lucas, Diener, & Suh, 1996), subjective well-being in this research was defined as a combination of both life satisfaction and pleasure of experience.

Subjective well-being (SWB) is an important indicator of the quality of life, and it can make important predictions of future life outcomes (Diener, 2012). For example, Lyubomirsky, King, and Diener (2005) concluded that it may cause desirable characteristics, resources and successes correlated with happiness. Also, Diener (2012) put forward that SWB is linked with good citizenship behavior in the community. The research in adolescence also found that high SWB is of great significance to adolescence (Orkibi, Ronen, & Assoulin, 2014).

Over the last four decades, many scales have been developed to measure SWB, such as the Satisfaction with Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985), Affect Balance Scale (ABS; Bradburn, 1969), General Happiness Scale (GHS; Lyubomirsky & Lepper, 1999) and the Short Happiness and Affect Research Protocol (SHARP; Stones et al., 1996). However, it is difficult for a single scale to reveal the whole picture of SWB. For example, the SWLS only measures the cognitive component of SWB (Emerson, Guhn, & Gadermann, 2017), while the ABS only measures its affect composition. It seems that the goal of covering all aspects of SWB needs to be achieved at the expense of increasing items, but this would enlarge the test burden and decrease test motivation (Forkmann et al., 2009). No doubt, increasing the items would also prolong test time. So, it is necessary to explore how to measure SWB as a whole without increasing items and time, and computerized adaptive testing (CAT) was designed to meet this desire.

In fact, CAT measures the whole picture of trait ability, not through increasing the length of the test, but through constructing an item bank and setting some selection rules in the program to make sure the test-taker can meet the targeted items. The adaptive testing is based on the traditional paper and pencil test, which has shortcomings, such as the excessive length of the test. In an adaptive test, the additional items are selected from an item bank that is built in advance and whose item parameters were known (Weiss, 1985). The general principles of CAT were first applied in intelligence tests and developed in the early 1900s (Weiss, 1985). Such an approach can reduce the test-taker's time and tedium related to extended testing (Forbey & Benporath, 2007) while gaining an optimal amount of information needed (Žitný, 2011). The effectiveness of CAT has also been supported in psychological measurement; for instance, the individual abilities may be estimated more precisely and efficiently (Embretson, 1992). CAT was developed on the basis of adaptive testing, combined with computer and item response theory (IRT), which can assess the severity order of each item that is specified by one measure (Hagman et al., 2009) and

CAMBRIDGE UNIVERSITY PRESS

some assumptions (Anderson, Kahn, & Tindal, 2017), such as unidimensionality and local independence assumption.

To date, CAT has been widely used in the field of ability measurement and psychology, such as the Graduate Record Examination (GRE) and Test of English as a Foreign Language (TOEFL), and depression (Forkmannd et al., 2009) and personality (Forbey & Benporath, 2007). After investigating all the techniques used to measure SWB, no study applying CAT methods to SWB in university students was found. This study aimed to develop a CAT of SWB (CAT-SWB) to measure the whole picture of SWB while reducing test time and the number of test items without attenuating precision and accuracy. In all, the proposed CAT-SWB can provide a greater flexibility to the algorithm (Sunderland et al., 2017) while targeted items were selected to evaluate the whole picture of SWB using fewer items without losing measurement accuracy.

Methods

Samples

About 1,000 participants were recruited from 39 universities in 20 provinces in China. These universities included six types (according to the category of subject setting): comprehensive university, normal university, polytechnic university, medical university, national university and military university. The 20 provinces were divided into two categories (the eastern region and other region) in that Wu (2017) pointed out that there remain significant regional differences on SWB between the eastern region and other region. The participants volunteered to complete this test after being informed that their personal information would be kept secret and the test would occupy them for about half an hour. After excluding some invalid data (large missing responses data, consecutive responses data and inconsistent responses data), 724 participants remained, comprising 303 females and 421 males, with a mean age of 19.12 (SD = 1.11). Of the sample, 48% were freshmen, 23% were sophomores, 24% were juniors, and 5% were seniors; 38% of participants were from urban areas and 62% were from rural areas.

Measures

After referencing previous studies and literatures, seven commonly used SWB scales were chosen to construct the initial item bank of the CAT-SWB. The SWB scales used here included the Satisfaction with Life Scale (SWLS; Diener et al., 1985), General Well-Being Scale (GWB; Fazio, 1977), Well-Being Index Scale (WBIS; Campbell, Converse, & Rodgers, 1976), Global Happiness Scale (GHS; Lyubomirsky & Lepper, 1999), Scale of Happiness of the Memorial University of Newfounland (MUNSH; Kozma & Stones, 1980), Affect Balance Scale (ABS; Bradburn, 1969) and Subjective Well-Being Scale for Chinese Citizens (SWBS-CC; Xing, 2003).

Except for the GHS, the other six scales have a Chinese version. These are the GWB, WBIS, MUNSH and ABS, obtained from Wang, Wang, and Ma (1999), the SWLS (Yang, Tian, Wang, & Cui, 2015), and the SWBS-CC (Xing, 2003). Therefore, the GHS was translated into Chinese. The confirmatory factor analysis (CFA) showed the Chinese version of GHS has the same unidimensional structure with the original GHS (normed fit index [NFI] = 1, confirmatory fit index [CFI] = 1, root mean square error of approximation [RMSEA] = 0.052, standardized root mean square residual [SRMR] = 0.006).The reliability analysis showed the Chinese version of the GHS has an acceptable reliability (Cronbach's α = .719, split-half reliability coefficient = .747), and has a close association with life satisfaction (r = .387, p < .01), which indicated that the Chinese version of the GHS has a high convergent validity.

Item pool construction of the CAT-SWB

The construction of the CAT-SWB item bank included five steps.

Step 1: *Item selection for the initial item bank of the SWB*. All the items to construct the initial item bank are from the seven commonly used SWB scales.

Step 2: *The unidimensionality test of the initial item bank*. Any item which does not fit the unidimensionality assumption or has low loading on the main factor will be excluded from the initial item bank to guarantee the unidimensionality of the item bank.

Step 3: *IRT model comparison and selection*. In this step, a more suitable IRT model will be chosen to conduct the analysis of IRT.

Step 4: *Item analysis with IRT*. Item analysis will include the local independence test, parameter estimation, item fit and differential item functioning. We will delete items with local dependence, low discrimination, poor item fit or having differential item functioning (DIF).

Step 5: *The construction of the final item bank for SWB*. According to the above four steps, all items that meet the measurement requirements will be included to build the final item bank for SWB.

Item selection

All of the seven chosen scales are self-report scales. The SWLS contains five items and belongs to a 7-point, Likert-type scale. The GWB contains 24 items, and there are 3–11 choices for each item. Items 15–18 use a 0–10 rating scale that is anchored by adjectives (e.g. *very depressed* to *very happy*). Items 2, 5, 6, 7, 19 use a 5-point rating scale (e.g. *quite troubled* to *not troubled at all*), items 1, 3, 4, 8, 9, 10, 11, 12, 13, 14 use a 6-point rating scale (e.g. *all the time* to *never*), item 24 uses a 7-point rating scale (e.g. *yes, very helpful* to *have no problem*), and the remaining items use a 3-point rating scale (e.g. *yes, in the past year* to *no*); all these choices represent either intensity or frequency. The WBIS contains nine items and has a 7-point, Likert-type scale. The GHS contains four items, with a 7-point, Likert-type scale. The MUNSH contains 24 items, and each item has three levels (yes, unclear and no). The ABS contains 10 items and each item has two levels (yes and no). The SWBS-CC (short version) contains 20 items, with a 7-point, Likert-type scale.

Unidimensionality

According to Flens, Smits, Carlier, Van, and De (2016), when the ration of total variance explained by the first factor is above the Reckase criterion of 20% and the value of the first eigenvalue divided by the second eigenvalue is higher than Reeve criterion of 4, the scale is deemed to be unidimensional.

An exploratory factor analysis (EFA) was employed to investigate the above criterions. Items with the first factor load less than 0.3 (Nunally, 1978) were first removed to ensure that the remaining items met the assumption of unidimensionality of IRT. The EFA was conducted till the remaining items were unidimensionality.

IRT model comparison and selection

The fit of a parametric IRT model is very important when implementing IRT (Liang & Wells, 2009). Under the IRT framework, an

https://doi.org/10.1017/prp.2019.6 Published online by Cambridge University Press

IRT model can be divided into two main categories: the difference models (or cumulative logits models) and the divided-by-total models (or adjacent logits models; Tu, Zheng, Cai, Gao, & Wang, 2017). A representative model of difference models is the Graded Response Model (GRM; Samejima, 1969) while a typical model in divided-by-total models is the Generalized Partial Credit Model (GPCM; Muraki, 1992), and the Nominal Response Model (NRM; Bock, 1972) is the extreme form of the divide-by-total group, which allows truly nominal responses (Chen, 2017).

GPCM is an extension of the Partial Credit Model (PCM) proposed by Masters (1982) by adding the discrimination parameter. The NRM, developed by Bock (1972), may be applied to items with alternatives that cannot be ordered to represent varying degrees of the trait measured and attempts to increase the precision of the theta estimates (Dodd, Ayala, & Koch, 1995). The GRM, introduced by Samejima (1969), has the same number of item parameters with GPCM and belongs to the class of models for which the responses are measured on an ordinal scale. After investigating a sea of literature, the above three models were all commonly used polytomously scoring models in IRT, and were also commonly applied to CAT (e.g. Dodd et al., 1995; Paap, Kroeze, Terwee, Palen, & Veldkamp, 2017; Zhou & Reckas, 2014). Therefore, this article employed these models and investigated which one fitted the data best. In this research, the Akaike's information criterion (AIC), Bayesian information criterion (BIC) and -2 log-likelihood were used to investigate which model fitted the data best, since selecting a model with a smaller value of AIC, BIC and -2 log-likelihood is widely accepted in model selection. The smaller the AIC/BIC/-2 log-likelihood, the better the fit of the model (Posada, Crandall, & Simon, 2001).

Local independence

A Q_3 statistic was proposed by Yen (1984) to detect local independence of IRT. If the Q_3 value of an item was larger than an arbitrary cut value, it meant the item had local dependence (Finch & Jeffers, 2016). As Flens et al. (2016) pointed out that the value of Q_3 above 0.36 represents a moderate deviation or dependence, therefore in this article, items with a Q_3 value larger than 0.36 were removed from the item bank to ensure all remaining items met the IRT assumption of local independence.

Item discrimination parameter

The discrimination parameter in IRT is an important indicator to evaluate the quality of an item. Chang and Ying (1996) pointed out that discrimination with a value between 0.5 and 2.5 seems to be an acceptable item parameter. In this article, items with a discrimination less than 0.5 were removed from the item pool to ensure a high-quality item pool for the SWB.

Item fit

Testing goodness of item fit proves to be an important step when analyzing IRT-based analysis (Köhler & Hartig, 2017). Here the $S\text{-}X^2$ (Kang & Chen, 2008) statistic was used to evaluate item-fit. Items with a p value of $S\text{-}X^2$ less than .05 were deemed a misfit and removed from the item pool.

Differential item functioning (DIF)

DIF was analyzed to identify item bias for a wide range of variables, such as gender (male/female), region 1 (rural/urban) and region 2 (eastern region/other region), to build a non-biased item bank. The logistic regression (LR; Choi, Gibbons, & Crane, 2011) method and lordif package of Rstudio were used to detect DIF. The change of McFadden's (McFadden, 1974) pseudo R^2 with $\triangle R^2 > .02$ and $p > .05$ was used as the criterion of detecting DIF. That is to say, items with $\triangle R^2 > .02$ and $p > .05$ for the change of McFadden's pseudo R^2 were deemed as having DIF and were excluded from the item pool.

The construction of the final item bank

According to the above steps, all items that met the measurement requirements were included to build the final item bank for SWB.

Simulation of the CAT-SWB

The simulation of the CAT-SWB was to achieve the algorithm's checking of it and evaluate the item bank. First, 71 thetas were simulated whose values ranged from -3.5 to 3.5 with intervals of 0.1; each theta was repeated 100 times to simulate 7,100 new thetas or examinees. Based on the existing references (Bock & Mislevy, 1982; Chang & Ansley, 2003; Magis, 2015), the 71 thetas were from -3.5 to +3.5 with the intervals of 0.1. The theta values between -3.5 and +3.5 covered almost all the ability values (99.96%), and 71 thetas with 100 replications for each theta value can ensure that as many as possible of the representative ability points of different ability participant groups are covered. This simulation differs from the method of randomly extracting the theta from the standard normal distribution, since the latter may have uneven extraction problems. For example, the method of randomly extracting cannot ensure that there are enough examinees (e.g. $n = 100$) with very high theta values or very low theta values.

Second, the item parameters of the last valid item bank should be imported into the CAT program. Moreover, the item responses in the CAT-SWB were simulated with four stopping rules. The advantage of simulating new thetas was that it can reach full ranges of the participants with different theta values. According to Magis and Raiche (2012), the CAT process can be divided into four steps, which are initial step, test step, stopping step and final step. In the initial step of this CAT process, the first item was selected randomly from the item bank and then the participant's response was simulated according to the true value of the ability simulated in advance and the randomly selected initial item. Moreover, the theta value was estimated by the expected a posteriori method (EAP; Bock & Mislevy, 1982) based on the item response and item parameters. In the *test step*, all of the Fisher information values for each remnant item in the item pool were calculated. Then, at the provisional estimate of the new theta, according to the maximum Fisher information criterion (Linden, 1998), one of the popular item selection criterions in adaptive testing, the next item was selected. In the *stopping step*, the CAT program stopped when the standard error (SE) theta (θ) of measurement reached 0.500, 0.447, 0.387 or 0.316, which represented the measurement reliabilities of 0.75, 0.80, 0.85 and 0.90 respectively, according to the formula of reliability $(\theta) = 1\text{-}\ SE\ (\theta)^2$ in IRT. The *final step* yielded the final estimation of theta value, the numbers of response item and standard error of measurement. The number of items in the CAT program was explored and MATLAB 2016 was used to plot the test information function and standard error of measurement for checking its measurement precision using test information function (TIF) and SE (θ). All CAT simulations were run using catR package for Rstudio (Magis & Raiche, 2011). The code for CAT simulation and item pool construction was presented in the supplementary material.

https://doi.org/10.1017/prp.2019.6 Published online by Cambridge University Press

Table 1. Brief description of the scales in this study

Scale	Number of items	Cronbach's alpha	Dimension	Number of dimensions
Subjective Well-Being Scale for Chinese Citizens	20	.85	Experience of Health	3
			Experience of Satisfaction	
			Experience of Development	
Satisfaction with Life Scale	5	.83	**Life Satisfaction**	1
General Well-Being Scale	24	.82	Health Concerns	6
			Energy	
			Satisfaction and Interest in Life	
			A Melancholy or Cheerful State of Mind	
			Control of Emotions and Behavior	
			Relaxation and Tension (Anxiety)	
Well-Being Index Scale	9	.87	General Well-Being Index	2
			Life Satisfaction	
Global Happiness Scale	4	.72	Measuring happiness at a global level	1
Scale of Happiness of the Memorial University of Newfounland	24	.89	**Positive Affect**	4
			Negative Affect	
			General Positive Experience	
			General Negative Experience	
Affect Balance Scale	10	.63	**Positive Affect**	2
			Negative Affect	

Note: The 8th item of ABS scale and the 7th item of MUNSH scale are the same ("Depressed or very unhappy"). The first dimension of SWLS scale and the second dimension of WBIS scale are the same (Life Satisfaction shown in bold type). The first dimension of MUNSH scale and the first dimension of the ABS scale are the same (Positive Affect shown in bold type). The second dimension of MUNSH scale and the second dimension of ABS scale are the same (Negative Affect shown in bold type).

Also, a real-data simulation of the CAT-SWB was carried out to sufficiently investigate the accuracy of algorithm and the quality of item bank. The difference between the real-data simulation of the CAT-SWB and the simulated-data simulation of the CAT-SWB was that the former used the real 724 participants' responses collected in advance while the latter used the simulated examinees' responses. Item parameters and the real responses of the 724 participants were imported while using the real-data simulation of the CAT-SWB. The CAT program also stopped when the *SE* (θ) of measurement reached 0.500, 0.447, 0.387 or 0.316. The usage of items in the CAT program was considered. For a more intuitive presentation, MATLAB 2016 was used to plot the number of administered items across the latent trait.

Intending to investigate how the consistency between the estimated theta by full-item bank and the estimated theta by the adaptive CAT, the consistency was investigated by the Pearson's correlation (*r*). Also, Cohen's *d* was considered to evaluate whether the CAT scores sufficiently similar to the full-item bank scores (Flens et al., 2016).

Results

Item pool construction of CAT for SWB

Item selection

The initial item bank included 95 items from seven commonly used SWB scales. The index Cronbach's alpha was also calculated to check the reliability of each scale, which is displayed in Table 1. The reliabilities of the ABS (Cronbach's $\alpha = .63$) and the internal consistency of the GHS (Cronbach's $\alpha = .72$) were acceptable, and the reliabilities of other five scales were good (Cronbach's $\alpha > .80$). The dimension and the number of dimensions for each scale are also displayed in Table 1. Because there were several dimensions that were the same, the initial item bank eventually contained 16 different dimensions, which covered all the main domains of SWB. The correlation (Pearson correlation) analysis showed there were significant correlations among 16 dimensions (except the Health Concerns subscales) with the value from .213 to .713 (all $ps < .01$), which indicated that these dimensions measured a common component (i.e. SWB).That is to say, they can be regarded as unidimensionality.

Moreover, as the related research showed, the Well-Being Index Scale (Campbell et al., 1976), the Scale of Happiness of the Memorial University of Newfoundland (Diaz, Moraga, & Soromaa, 2011) and the Global Happiness Scale (Parackal, 2016) have significant correlations with life satisfaction while Xing (2002) regarded the Satisfaction with Life Scale, the Subject Well-Being Scale for Chinese Citizens and the Affect Balance Scale as scales for measuring SWB, and Duan (1996) also considered the General Well-Being Scale as a scale to measure SWB. All these indicated that these seven commonly used scales can refine a main common measurement factor or component (i.e. SWB). That is to say, all these seven commonly used scales measure SWB.

Unidimensionality

After 16 items were removed due to their first factor load less than 0.3, with the remaining 79 items, the first factor explained 26.41%

https://doi.org/10.1017/prp.2019.6 Published online by Cambridge University Press

Table 2. Indexes of model-fit based on test level

Model	AIC	BIC	−2 log-likelihood
GRM	141354.00	143229.20	140536.00
GPCM	142325.60	144200.80	141507.60
NRM	141793.10	144819.00	140473.10

Note: GRM = Graded Response Model, GPCM = Generalized Partial Credit Model, NRM = Nominal Response Model (NRM; Bock, 1972), AIC = Akaike's information criterion, BIC = Bayesian information criterion.

of the total variance, which was above 20%. And the ratio of the first eigenvalue to the second eigenvalue was 5.24, which was higher than 4. According to Reckase and Reeve criterion (Flens et al., 2016), it can be concluded that the remaining 79 items measured one main factor (i.e. SWB) and met the assumption of unidimensionality of IRT.

Model comparison and selection

Table 2 documented the model-data fit indicators. As shown in Table 2, the GRM had the smallest value of AIC, BIC and -2log-likelihood, which indicated that the GRM fit the data best. Therefore, the GRM was selected as the IRT model that was used to the subsequent IRT analysis.

Local independence

Using the 0.36 as the arbitrary cut-value (Flens et al., 2016), 15 items were removed in that its absolute Q_3 values were higher than 0.36; then the remaining items met the local independence well.

Item parameter

All remaining item discrimination parameters were higher than 0.5 (see Table 3) with a mean of 1.2 ($SD = 0.37$), which was regarded as a not bad value, and no item was removed from the current item bank.

Item fit

All the p values of $S\text{-}X^2$ for all remaining items were higher than .05 (see Table 3), which indicated that all the remaining items fitted well to the GRM.

Differential item functioning (DIF)

DIF results showed that all the items' ΔR^2 values were less than .02 and the corresponding p values were less than .05, which indicated there were no items with DIF in gender and the region (region 1 and region 2) of the participants.

Table 4 displayed the number of items used to measure each dimension in the final item pool. Fortunately, in the final item pool, all the 16 dimensions have been measured, and the number of items measured per dimension ranges from 1 to 8. Although only one item measured the dimension of health concern and two items measured the dimension of satisfaction and interest in life, each dimension used no less than three items, especially the dimension of negative affect, which used eight items to measure. To sum up, it can be concluded that the item bank performed well in content validity because it covered all 16 main domains of the SWB and provided a good guarantee for measuring the whole picture of the SWB.

Table 3. Some estimation values of the final item pool with 64 items

	Item parameter estimates	Item-fit estimates		
Item	a[a]	S-X^2	df[b]	p
1	0.66	143.93	156	.75
2	0.94	220.35	216	.41
3	0.90	227.27	228	.50
4	0.91	136.50	196	1.00
5	1.04	248.22	216	.07
6	1.21	226.58	222	.40
7	0.88	199.65	199	.47
8	1.09	207.20	216	.65
9	0.76	178.28	171	.34
10	1.10	156.37	169	.75
11	1.31	164.41	208	.99
12	0.84	246.42	241	.39
13	1.02	248.53	241	.36
14	0.82	227.63	267	.96
15	0.63	244.81	252	.62
16	1.31	216.76	223	.61
17	0.72	258.74	263	.56
18	0.75	226.86	218	.33
19	1.22	157.85	182	.90
20	1.64	215.66	242	.89
21	0.94	271.21	269	.45
22	0.89	273.42	281	.62
23	1.10	83.27	86	.56
24	1.71	117.34	116	.45
25	1.08	125.15	122	.40
26	1.62	118.45	120	.52
27	1.64	65.63	83	.92
28	2.02	73.72	88	.86
29	1.10	120.49	144	.92
30	1.66	75.24	100	.97
31	1.08	114.20	129	.82
32	1.03	85.35	87	.53
33	1.34	122.74	127	.59
34	1.40	96.38	106	.74
35	1.08	144.77	141	.40
36	1.25	75.02	91	.89
37	1.23	126.77	136	.70
38	1.07	119.61	125	.62
39	1.10	107.30	129	.92
40	1.58	119.19	115	.38
41	0.91	132.47	138	.62
42	0.91	96.01	90	.31

(*Continued*)

https://doi.org/10.1017/prp.2019.6 Published online by Cambridge University Press

Table 3. (*Continued*)

Item	Item parameter estimates	Item-fit estimates		
	a[a]	S-X^2	df[b]	p
43	2.01	65.26	60	.30
44	1.55	56.98	51	.26
45	2.32	63.75	66	.56
46	1.20	82.68	85	.55
47	1.84	122.73	124	.52
48	1.28	140.49	138	.42
48	0.86	147.36	155	.66
50	1.29	146.01	162	.81
51	1.04	145.35	141	.38
52	1.81	120.37	109	.21
53	0.91	161.58	162	.49
54	1.68	108.57	113	.60
55	0.88	200.38	176	.10
56	0.83	185.08	166	.15
57	1.29	168.31	161	.33
58	1.91	142.50	129	.20
59	1.12	173.58	165	.31
60	1.36	147.42	142	.36
61	1.70	284.59	277	.36
62	1.72	231.24	265	.93
63	1.02	151.13	143	.30
64	1.10	82.84	102	.92

Note: [a]discrimination parameter.
[b]degree of freedom.

The construction of the final item bank

The final item bank contained 64 items (see Table 3) after the above statistical analysis under the framework of IRT. The item bank of SWB met the IRT assumptions of unidimensionality and local dependence, fitted the GRM well, had high discrimination parameters, no DIF existed and had an acceptable content validity. All these results showed that the proposed item bank of SWB was acceptable.

Simulation of the CAT-SWB

Results based on the simulated data of the CAT-SWB

Table 5 displays the results of simulated data statistics for the CAT-SWB under four stopping rules of SE (θ) = 0.500/0.447/0.387/0.316; these were called Rule 1, Rule 2, Rule 3 and Rule 4 respectively. For the four stopping rules, the mean numbers of administered items ranged from 5.77 to 17.78, or 9–28% of the full-item bank. This result was satisfactory, especially with the stopping rule of SE (θ) = 0.500; the CAT-SWB used only about six items while achieving the full-item efficiency; mean of SE (θ) = 0.47. Obviously, the CAT-SWB saved the number of items to a large extent. Table 5 also shows the Pearson's correlations (r) between the CAT theta estimates and full-item bank theta estimates under each stopping rule, which were all higher than .9. That is to say, the theta value between CAT and the full-item bank was very similar. In addition, the mean value of theta estimated under Rule 1–Rule 4 was very close to the average of theta estimated using the full-item bank, and there was no statistically significant difference ($p > .05$). Moreover, the Cohen's d under each stopping rule was close to zero and indicated that there was no structure difference between theta estimates using CAT and theta estimates using the full-item bank.

Table 4. Number of items used to measure each dimension in the final item pool

Dimension	Number of items
Experience of Health	5
Experience of Satisfaction	4
Experience of Development	3
Life Satisfaction	4
Health Concerns	1
Energy	3
Satisfaction and Interest in Life	2
A Melancholy or Cheerful State of Mind	4
Control of Emotions and Behavior	3
Relaxation and Tension (Anxiety)	5
General Well-Being Index	3
Measuring happiness at a global level	3
Positive Affect	5
Negative Affect	8
General Positive Experience	5
General Negative Experience	6

Note: These dimensions are from the seven commonly used scales of SWB.

Figure 1 displays an intuitive result of test information function (TIF) and standard error of measurement (SEM) across the latent trait under the simulated-data simulation of CAT. The higher the theta value, the higher the SWB. Figure 1 documents the information about how precisely a test can measure the latent trait. It is easy to see almost all the SEM values were under 0.33; these values are regarded as acceptable since Michel et al. (2017) pointed out that a SEM between 0.33 and 0.55 was defined as acceptable and the smaller the SEM, the more accurate the CAT-SWB. Test information can be expressed as the sum of all item information at any relevant theta level. The larger the TIF at each theta level, the smaller the SEM. After considering these, we could conclude that the CAT-SWB reached sufficient information and acceptable standard error. In other words, the algorithm of the CAT-SWB performed well.

Results based on real-data of the CAT-SWB

Table 6 documents the results of real-data simulation statistic for the CAT-SWB under four stopping rules with SE (θ) = 0.500/0.447/0.387/0.316. The results were similar to the above results of simulated-data simulation for the CAT-SWB. Even better was that the number of items in real-data simulation was less than those in above simulated-data simulation. For example, under Rule 4, the mean of SE (θ) reached 0.316, and the average usage of an item was 13.80, which was 22% of the full-item bank, while

https://doi.org/10.1017/prp.2019.6 Published online by Cambridge University Press

Table 5. Simulated-data simulation statistic for the CAT-SWB under four stopping rules

	Number of items						θ estimates			
Stop rules	Min	Max	Mean	*SD*	% *all*[a]	Mean of *SE* (θ)	Mean	*SD*	Cohen's *d*	r^f
Full-item bank	64	64	64	0.00	100	0.22	0.00	2.05	–	1
Rule 1[b]	4	14	5.77	0.99	9	0.47	−0.01	1.68	−0.01	.97
Rule 2[c]	5	22	7.15	1.57	11	0.44	−0.01	1.74	−0.01	.98
Rule 3[d]	7	34	9.96	2.97	16	0.37	0.00	1.81	0.00	.98
Rule 4[e]	11	64	17.78	8.72	28	0.32	0.00	1.89	0.00	.99

Note:[a] The percentage of the mean numbers of administered items in the full-item bank. [b]Stopping rule *SE* (θ) = 0.500. [c]Stopping rule *SE* (θ) = 0.447. [d]Stopping rule *SE* (θ) = 0.387. [e]Stopping rule *SE* (θ) = 0.316. [f]Pearson's correlations between the CAT θ estimates and full-item bank θ estimates.

Table 6. Real-data simulation statistic for the CAT-SWB scale under four stopping rules

	Number of items						θ estimates			
Stop rules	Min	Max	Mean	*SD*	% *all*[a]	Mean of SE(θ)	Mean	*SD*	Cohen's *d*	r^f
Full-item bank	64	64	64	0.00	100	0.20	0.00	0.98	–	1
Rule 1[b]	4	15	5.36	0.76	8	0.47	0.00	0.93	0.00	.92
Rule 2[c]	5	21	6.50	0.97	10	0.42	−0.01	0.95	−0.01	.92
Rule 3[d]	7	29	8.60	1.35	13	0.37	−0.01	0.96	−0.01	.93
Rule 4[e]	11	64	13.80	3.00	22	0.32	0.00	0.97	0.00	.96

Note: [a]The percentage of the mean numbers of administered items in the full-item bank. [b]Stopping rule SE (θ) = 0.500. [c]Stopping rule SE (θ) = 0.447. [d]Stopping rule SE (θ) = 0.387. [e]Stopping rule SE (θ) = 0.316. [f]Pearson's correlations between the CAT θ estimates and full-item bank θ estimates.

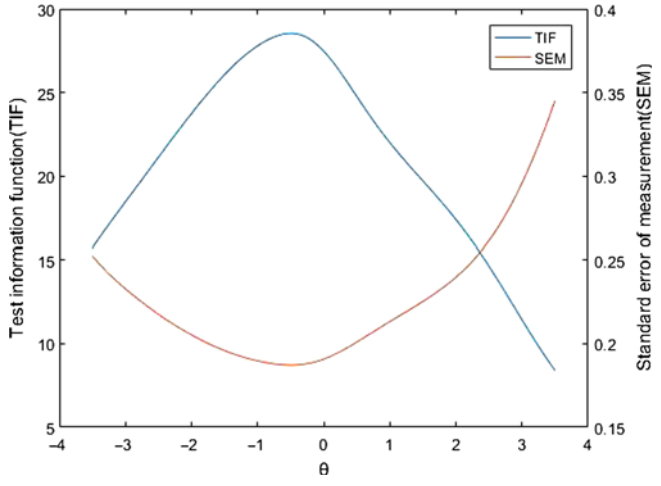

Figure 1. A bell-shaped test information function (TIF) of all 64 items of the bank (blue dotted line), and also plotted the standard error of measurement (SEM; red solid line).

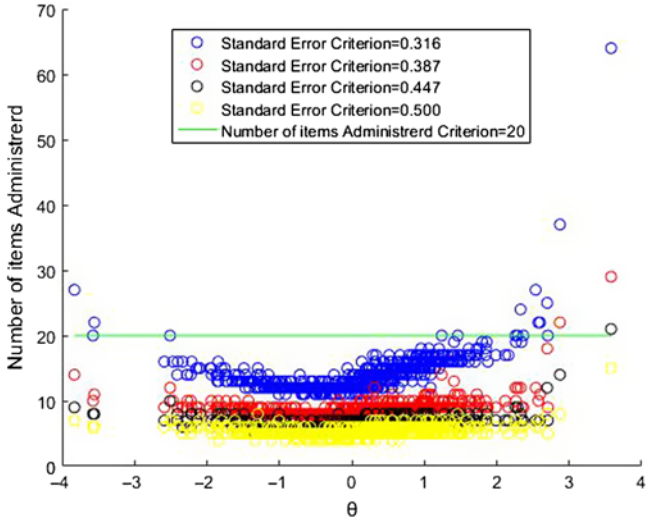

Figure 2. Number of administered items under four stopping rules.

in the simulated-data simulation, this was 28%. We drew an additional diagram (see Figure 2) that displays the number of items administered intuitively across the latent trait under the four stopping rules. Figure 2 indicates that almost all the items administered were under the horizontal line labeled 20 items, which was 31% of full-item bank. All of these results indicate that the CAT-SWB performed well in a real-data simulation.

Overall, it can be concluded that whether it is the simulated-data simulation or the real-data simulation, the algorithm of the CAT-SWB performed well, and it provides accuracy and low burden for the assessment of the Chinese university students' SWB.

Discussion

Until now, there are many examples that support the opinion that CAT can save the number of items to a certain extent without the loss of measurement precision. For example, when the CAT was applied to the Minnesota Multiphasic Personality Inventory

https://doi.org/10.1017/prp.2019.6 Published online by Cambridge University Press

(MMPI; Anderson et al., 2017) it showed that it can be reduced even by 119 items in the best case. For gross motor skills CAT (GM-CAT; Huang et al., 2018), the averages of items administered were from 7 to 11 while the item bank contained 44 items. There are many researchers who believe that CAT can improve the accuracy of the evaluation of trait level (θ) of the examinees (Barrada, Olea, Ponsod, & Abad, 2009).

This article proposed a CAT version of SWB (CAT-SWB) to measure the whole picture of SWB. After sequential analyses of unidimensionality, local dependence, item discrimination, item fit and DIF under the framework of IRT, the final item bank contained 64 items that were from seven commonly used SWB scales. The final item bank of the CAT-SWB covered 16 main domains of SWB and met the goal of measuring the whole picture of SWB. Diverging from other researches, this study investigated both simulated-data simulation and real-data simulation, which took full account of the efficiency of CAT. Both the simulated-data simulation and real-data simulation showed that the CAT-SWB had satisfactory accuracy and a low burden for measuring SWB. The averages of items administered were from 5.36 to 17.78, which were 8–28% of the full-item bank, and almost all the SEM were smaller than 0.33. In addition, the theta estimates of the CAT and the full-item bank were very similar and had very high correlations of more than .9. All these results indicate that the proposed CAT-SWB not only has a high measurement precision but also can greatly shorten test length. Moreover, there existed no DIF in gender and the regions of participants, and this enhances our confidence in promoting this CAT version.

Despite the promising results, there were also some limitations about this study. First, an extra finding shown in Figure 1 was that the CAT-SWB provided little information for those whose latent trait theta was higher than 2.5, which indicates that the CAT-SWB may not be appropriate for these participants. Second, the participants in this study were mainly recruited from two provinces in China. Therefore, more samples should be recruited from a wider range of provinces in the future. Third, as described in Table 4, some of the 16 dimensions were measured using only 1–2 items. This may be a shortcoming for measuring the whole picture of SWB. Therefore, future research, could consider adding more items to measure these dimensions, and this would ensure that the test measures the whole picture of SWB.

Fourth, although the item bank was consistent with the structure of unidimensionality under a series of criteria in accordance with the requirements of previous studies, this item bank also contained 16 domains. This inspired us to verify whether fitting results of the item bank and the multidimensional IRT (MIRT) model may be better. First, since the initial item bank contained 95 items from 16 domains, a CFA was used to investigate the fit of the item bank with the 16-dimensional structure. The results showed that the model with 16 dimensions cannot be identified. Then we tried to explore the structure of the item bank. Twenty-three factors were extracted with a principal component analysis that employed the variance maximum rotation method; the eigenvalue was greater than one criterion (EVG1) and the factor load higher than 0.3 criterion (Nunally, 1978). Second, given that the obtained pattern of eigenvalues is ambiguous, a parallel analysis (Horn, 1969; Humphreys & Ilgen, 1969) was employed to extract eight factors. Third, because the first 8 factors show that the proportion of the variation is small (42.973%), 20 factors can then be extracted to explain over 60% of the total variance. The results again showed that the model cannot be identified.

From this series of factor analysis, we can see that no matter whether it is 16 dimensions or 20 dimensions, the data is of a high dimension and a multidimensional CAT (MCAT) can be addressed in future research.

Finally, there are some reasons why it is not necessary to consider too many factors that may have an impact on SWB. For example, the main purpose of our research was not to explore factors that may affect SWB, but to construct a CAT-SWB to measure SWB for different participants. Moreover, the great advantage of IRT over classical test theory is that it only needs to cover all the respective theta values of the participants, which avoids the limitation of sample dependence when estimating item parameters (Hagman et al., 2009). That is to say, we just needed to guarantee that the sample size was sufficient and ensure there was a wide representation in theta values of SWB. However, there also remained some variables, such as culture and personality, that have an impact on SWB (Diener et al., 2003). It is preferable that those factors that might affect SWB are taken into account in future research to explore whether they affect SWB in mainland China.

Conclusions

The proposed item bank of SWB had acceptable psychometric properties under the framework of IRT and measured a comprehensive concept of SWB in that 16 main domains of SWB were covered. The proposed CAT-SWB had an excellent performance in saving the number of response items without attenuating measurement precisions. Above all, this CAT-SWB can advance the efficiency of measuring a comprehensive concept of SWB.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/prp.2019.6

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 individual participants who involved in this study. This research was funded by the National Natural Science Foundation of China (31760288, 31660278).

Conflict of interest. None.

Ethical approval. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

Anderson D., Kahn J.D. and Tindal G. (2017). Exploring the robustness of a unidimensional item response theory model with empirically multidimensional data. *Applied Measurement in Education*, **30**, 163–177.

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 & Social Sciences*, **5**, 7–17.

Bock R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, **37**, 29–51.

Bock R.D., and Mislevy R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, **6**, 431–444.

Bradburn N.M. (1969). *The Structure of Psychological Well-Being*. Chicago: Aldine.

Campbell A., Converse P.E. and Rodgers W.L. (1976). *The Quality of American Life: Perceptions, Evaluations, and Satisfactions*. New York, NY: Russell Sage Foundation.

Chang S. and Ansley T. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, **40**, 71–103.

Chang H.H. and Ying Z.L. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, **20**, 213–229.

Chen J. (2017). Advancing the bayesian approach for multidimensional polytomous and nominal irt models: Model formulations and fit measures. *Applied Psychological Measurement*, **41**, 3–16.

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.

Diaz E.M., Moraga E.F. and Soromaa H. (2011). Reliability and construct validity of MUNSH test to measure happiness in elderly MUNSH Chilean population. *Universitas Psychologica*, **10**, 567–580.

Diener E. (1984). Subjective well-being. *Psychological Bulletin*, **95**, 542–575.

Diener E. (2012). New findings and future directions for subjective well-being research. *American Psychologist*. **67**, 590–597.

Diener E.D., Emmons R.A., Larsen R.J. and Griffin S. (1985). The satisfaction with life scale. *Journal of personality assessment*, **49**, 71–75.

Diener E., Oishi S. and Lucas R.E. (2003). Personality, culture, and subjective well-being: Emotional and cognitive evaluations of life. *Annual Review of Psychology*, **54**, 403–425.

Dodd B.G., Ayala R.J.D. and Koch W.R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, **19**, 5–22.

Duan J.H. (1996). The trial results and analysis of the general well-being scale in Chinese college students. *Chinese Journal of Clinical Psychology*, **17**, 56–57.

Embretson S.E. (1992). Computerized adaptive testing: Its potential substantive contributions to psychological research and assessment. *Current Directions in Psychological Science*, **1**, 129–131.

Emerson S.D., Guhn M. and Gadermann A.M. (2017). Measurement invariance of the satisfaction with life scale: Reviewing three decades of research. *Quality of Life Research*, **26**, 2251–2264.

Fazio A.F. (1977). A concurrent validational study of the NCHS General Well-Being Schedule. *Vital and Health Statistics*, **73**, 1–53.

Finch W.H. and Jeffers H. (2016). A Q3-based permutation test for assessing local independence. *Applied Psychological Measurement*, **40**, 157–160.

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.

Forbey J.D. and Benporath Y.S. (2007). Computerized adaptive personality testing: A review and illustration with the MMPI-2 computerized adaptive version. *Psychological Assessment*, **19**, 14–24.

Forkmann T., Boecker M., Norra C., Eberle N., Kircher T., Schauerte P.... **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.

Hagman B.T., Kuerbis A.N., Morgenstern J., Bux D.A., Parsons J.T. and Heidinger B.E. (2009). An item response theory (IRT) analysis of the short inventory of problems-alcohol and drugs (SIP-AD) among non-treatment seeking men-who-have-sex-with-men: Evidence for a shortened 10-item SIP-AD. *Addictive Behaviors*, **34**, 948–954.

Horn J.L. (1969). On the internal consistency reliability of factors. *Multivariate Behavioral Research*, **4**, 115–125.

Huang C.Y., Tung L.C., Chou Y.T., Wu H.M., Chen K.L. and Hsieh C.L. (2018). Development of a computerized adaptive testing of children's gross motor skills. *Archives of Physical Medicine & Rehabilitation*, **99**, 512–520.

Humphreys L.G. and Ilgen D.R. (1969). Note on a criterion for the number of common factors. *Educational and Psychological Measurement*, **29**, 571–578.

Kang T. and Chen T.T. (2008). Performance of the generalized S-X2 item fit Index for polytomous IRT models. *Journal of Educational Measurement*, **45**, 391–406.

Köhler C. and Hartig J. (2017). Practical significance of item misfit in educational assessments. *Applied Psychological Measurement*, **41**, 388–400.

Kozma A. and Stones M.J. (1980). The measurement of happiness: Development of the Memorial University of Newfoundland Scale of Happiness (MUNSH). *Journal of Gerontology*, **35**, 906–912.

Liang T. and Wells C.S. (2009). A model fit statistic for generalized partial credit model. *Educational & Psychological Measurement*, **69**, 913–928.

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

Lucas R.E., Diener E. and Suh E. (1996). Discriminant validity of well-being measures. *Journal of Personality & Social Psychology*, **71**, 616–628.

Lyubomirsky S., King L. and Diener E. (2005). The benefits of frequent positive affect: Does happiness lead to success? *Psychological Bulletin*, **131**, 803–855.

Lyubomirsky S. and Lepper H.S. (1999). A measure of subjective happiness: Preliminary reliability and construct validation. *Social Indicators Research*, **46**, 137–155.

Magis D. (2015). Efficient standard error formulas of ability estimators with dichotomous item response models. *Psychometrika*, **81**, 184–200.

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

Magis D., and Raiche G. (2012). Random generation of response patterns under computerized adaptive testing with the R package cat*R*. *Journal of Statistical Software*, **48**, 1–31.

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

McFadden D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 105–142). New York, NY: Wiley.

Michel P., Baumstarck K., Lancon C., Ghattas B., Loundou A., Auquier P. and Boyer L. (2017). Modernizing quality of life assessment: Development of a multidimensional computerized adaptive questionnaire for patients with schizophrenia. *Quality of Life Research*, **27**, 1–14.

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

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

Orkibi H., Ronen T. and Assoulin N. (2014). The subjective well-being of Israeli adolescents attending specialized school classes. *Journal of Educational Psychology*, **106**, 515–526.

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.

Parackal M. (2016). A global happiness scale for measuring wellbeing: A test of immunity against hedonism. *Journal of Happiness Studies*, **17**, 1529–1545.

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

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

Stones M.J., Kozma A., Hirdes J., Gold D., Arbuckle T. and Kolopack P. (1996). Short Happiness and Affect Research Protocol (SHARP). *Social Indicators Research*, **37**, 75–91.

Sunderland M., Slade T., Krueger R.F., Markon K.E., Patrick C.J. and Kramer M.D. (2017). Efficiently measuring dimensions of the externalizing spectrum model: Development of the Externalizing Spectrum Inventory-Computerized Adaptive Test (ESI-CAT). *Psychological Assessment*, **29**, 868–880.

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.

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

Weiss D.J. (1985). Adaptive testing by computer. *Journal of Consulting & Clinical Psychology*, **53**, 774–789.

https://doi.org/10.1017/prp.2019.6 Published online by Cambridge University Press

Wu R.X. (2017). Urbanization, basic public service supply and subjective well-being: An empirical study on 56 cities in GGSS (2010). *Population & Development*, **2**, 37–48.

Xing Z.J. (2002). Report on several common self -reported subjective well-being scales used to citizen in China. *Health Psychology Journal*, **10**, 325–326.

Xing Z.J. (2003). Developing the brief subjective well-being scale for Chinese citizen. *Chinese Journal of Behavioral Medical Science*, **12**, 703–705.

Yang Q.F., Tian H.Z., Wang Q. and Cui H. (2015). Study on the intervention of relaxation training on the psychological stress of senile coronary heart disease patients received the interventional therapy during the stage of operation. *Progress in Modern Biomedicine*, **15**, 1474–1478.

Yen W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, **8**, 125–45.

Zhou X.C. and Reckase M.D. (2014). Optimal item pool design for computerized adaptive tests with polytomous items using GPCM. *Psychological Test and Assessment Modeling*, **56**, 255–274.

Žitný P. (2011). Computerized adaptive testing: Precision, validity and efficiency. *Ceskoslovenska Psychologie*, **55**, 167–179.

https://doi.org/10.1017/prp.2019.6 Published online by Cambridge University Press