




Measuring glaucoma quality of life in an Asian population using item banking: psychometric evaluation and computerized adaptive testing simulations

Eva K. Fenwick^{1,2} · Belicia Lim¹ · Ryan E. K. Man^{1,2} · Mani Baskaran^{1,3} · Monisha Nongpiur¹ · Chelvin C. A. Sng^{1,5} · Jayant Venkatramani Iyer¹ · Rahat Husain¹ · Shamira Perera¹ · Tina Wong^{1,2} · Jin Rong Low¹ · Bao Sheng Loe⁵ · Olivia Shimin Huang^{1,2} · Katherine Lun⁴ · Tin Aung^{1,2,4,5} · Ecosse L. Lamoureux^{1,2,6,7} 

Accepted: 19 April 2023 / Published online: 28 April 2023
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2023

Abstract

Purpose To assess the psychometric properties of glaucoma-specific health-related quality of life (HRQoL) item banks (IBs), and explore their efficiency using computerized adaptive testing (CAT) simulations.

Methods In this cross-sectional, clinical study, 300 Asian glaucoma patients answered 221 items within seven IBs: Ocular Comfort Symptoms (OS); Activity Limitation (AL); Lighting (LT); Mobility (MB); Glaucoma Management (GM); Psycho-social (PSY); and Work (WK). Rasch analysis was conducted to assess each IB's psychometric properties (e.g., item “fit” to the construct; unidimensionality) and a set of analytic performance criteria guiding decision making relating to retaining or dropping domains and items was employed. CAT simulations determined the mean number of items for ‘high’ and ‘moderate’ measurement precision (stopping rule: SEM 0.3 and 0.387, respectively).

Results Participants' mean age was 67.2 ± 9.2 years (62% male; 87% Chinese). LT, MB, and GM displayed good psychometric properties overall. To optimize AL's psychometric properties, 16 items were deleted due to poor “fit”, high missing data, item bias, low discrimination and/or a low clinical/patient importance rating. To resolve multidimensionality in PSY, we rehomed 16 items into a “Concern (CN)” domain. PSY and CN required further amendment, including collapsing of response categories, and removal of poorly functioning items ($N = 7$). Due to poor measurement precision, low applicability and high ceiling effect, low test information indices, and low item separation index the WK IB was not considered further. In CAT simulations on the final seven IBs ($n = 182$ items total), an average of 12.1 and 15.7 items per IB were required for moderate and high precision measurement, respectively.

Conclusions After reengineering our seven IBs, they displayed robust psychometric properties and good efficiency in CAT simulations. Once finalized, GlauCAT™-Asian may enable comprehensive assessment of the HRQoL impact of glaucoma and associated treatments.

Keywords Glaucoma · Quality of life · Item banks · Rasch analysis · Computerized adaptive testing

✉ Ecosse L. Lamoureux
ecosse.lamoureux@duke-nus.edu.sg

¹ Singapore Eye Research Institute, Singapore National Eye Centre, The Academia, 20 College Road, Level 6, Singapore 169856, Singapore

² Duke–NUS Medical School, National University of Singapore, Singapore, Singapore

³ Medical and Vision Research Foundation, Sankara Nethralaya, Chennai, India

⁴ National University Health System, Singapore, Singapore

⁵ School of Psychology, University of Cambridge, Cambridge, UK

⁶ Department of Ophthalmology, National University of Singapore, Singapore, Singapore

⁷ The University of Melbourne, Melbourne, Australia

Glaucoma is the leading cause of permanent visual disability worldwide [1]. With the global prevalence estimated to increase from 64.3 million in 2013 to 111.8 million in 2040 [2], a populous continent like Asia with a rapidly ageing population will likely see a substantial rise in cases. Evidence suggests that glaucoma patients report reduced health-related quality of life (HRQoL) [3], including independence and mobility [4], increased risk of falls [5], and poor emotional well-being [6], and that HRQoL may decline over time [7]. Treatment-related side-effects and economic burden [8] can also negatively impact on patients' HRQoL, beyond the burden of the disease process itself [9]. Hence, it is important to understand the impact of glaucoma and the effectiveness of treatment therapies from the patients' perspective, especially as healthcare moves towards a value-based care model [10].

Measurement of patient-reported outcomes is often done using patient-reported outcome measures (PROMs). However, most PROMs in Ophthalmology are fixed-length meaning all items (questions) must be administered regardless of whether they are targeted to participants' underlying level of the construct (e.g., visual functioning), making them burdensome to administer. Moreover, despite most vision-related PROMs comprising more than 20 items, they generally only provide measurement of two or three HRQoL domains. Modern psychometric methods of instrument development, such as item banking and computerized adaptive testing (CAT), offer a solution to overcome these shortcomings [11]. In item banking, a collection of items that measure a latent construct is calibrated according to level of difficulty on an interval-level scale [11]. Items from the calibrated item bank (IB) are then administered by CAT using an algorithm to selectively present to participants items that provide the greatest amount of information until a stopping criterion (e.g., measurement precision, based on standard error of measurement [SEM]) is reached [12]. Capitalizing on these efficiency gains, CATs usually require only 7–10 items (dependent on the stopping rule) to obtain a score, which is substantially fewer items than required by most fixed-length PROMs. CATs therefore provide a fast, yet precise means of estimating patients' level of HRQoL [13, 14].

While an IB for glaucoma was developed by Matsuura and colleagues in 2017 [15], a CAT is not yet available. Our group has developed, validated and implemented a glaucoma IB and CAT system (GlauCAT™-Western) [16–18]; however, the content was developed in patients from Australia and the UK and, as such, may not be relevant to an Asian population where lifestyles, healthcare systems and access and illness perceptions may differ from the West.

Our team has recently developed domains and items for a glaucoma-specific instrument HRQoL item bank (IB) and CAT based on qualitative information from Asian patients with glaucoma (GlauCAT™-Asian), which differs from the

GlauCAT™-Western in terms of number of domains (7 vs. 12, respectively) and content (e.g., more items about glaucoma management in GlauCAT™-Asian) [19]. Indeed, in a head-to-head comparison of the GlauCAT-Western and GlauCAT-Asian instruments, we found that only 57% of content in GlauCAT-Western was similar to GlauCAT-Asian, and that 25% of the content in the new GlauCAT instrument was unique (Online Resource 1). The current study reports on the psychometric properties of the GlauCAT™-Asian IBs in a multi-ethnic, clinical sample glaucoma patients and explores the functionality of the final calibrated IBs using CAT simulations.

Methods

Sample population

Patients aged ≥ 40 years (English- or Mandarin-speaking) with a primary diagnosis of glaucoma (i.e., primary open angle glaucoma (POAG) or primary angle closure glaucoma (PACG)) in at least one eye were recruited from the Singapore National Eye Centre (SNEC) between December 2019 and January 2021. Patients were excluded if they had other ocular comorbidities including secondary glaucomas, severe cataract, neurological conditions affecting vision, and/or hearing or cognitive impairment (assessed using the 6-CIT questionnaire [20]). During recruitment, we utilised specific target quotas to ensure we included patients across the spectrum of ethnicity, gender, age, and glaucoma severity) in order to obtain a diverse sample and enhance the applicability of our calibrations.

The study protocol was approved by the SingHealth Centralised Institutional Review Board (CIRB #2018/2459) and conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from participants prior to study participation.

Assessment of glaucoma, visual fields, and visual acuity

Glaucoma subtype, Snellen visual acuity (VA) and visual field (VF) data (both eyes) were extracted from patients' files. We also conducted binocular Esterman tests using the Humphrey Visual Field Analyzer-3 (Carl Zeiss AG, Jena, Germany). Details on grading of glaucoma severity and definitions of vision impairment (VI) are provided in Online Resource 2.

Development of the glaucoma IBs

The content development process of the GlauCAT™-Asian IBs has been described in detail elsewhere [19] and more

information is provided in Online Resource 3. In brief, the final instrument comprised 221 items under seven QoL domains with 4 to 5 response options each rated on a Likert-type scale with a non-applicable option available when the task or issue was not relevant to the participant (Online Resource 4): Activity Limitation (AL; $n = 72$); Lighting (LT; $n = 15$); Mobility (MB; $n = 19$); Psychosocial (PSY; $n = 55$); Ocular Comfort Symptoms (OS; $n = 19$); Glaucoma Management (GM; $n = 28$); and Work (WK; $n = 13$).

Psychometric evaluation of the IBs

We performed Rasch analysis separately on each IB using Winsteps software (version 4.50; Chicago, IL, USA) using the Andrich single rating-scale model [21]. Rasch analysis is a probabilistic model that estimates the relative difficulty of items (item measures) and relative abilities of respondents (person measures) and aligns them on a common scale [22] enabling transformation of ordinal data into estimates of interval-level data, expressed in log of the odds units (logits). Rasch analysis was chosen over Item Response Theory (IRT) models such as the Graded Response Model as, unlike IRT models [23, 24], Rasch analysis enables stable item calibrations and person measure estimates in small sample sizes [25]. Rasch analysis allows a thorough assessment of a scale's psychometric properties, which are outlined in detail below. Similarly, a set of analytic performance criteria guiding decision making relating to retaining or dropping domains (i.e., adequate fit statistics, plus Applicability, Ceiling effect, Test Information Function (TIF), Item Separation Index (ISI), Measurement range, and Clinical importance) and items (misfit, Differential Item Functioning (DIF), Applicability, Discrimination, Clinical importance and Patient importance; Online Resource 2) was taken into account by the research team, comprising members with content development and psychometric (EF, RM, BL, EL); and/or clinical (MB, MN) expertise.

Rating-scale assessment

To evaluate rating-scale performance, the average observed measures for the person sample were investigated. The average observed measures should increase monotonically as the rating-scale increases meaning that, as person measures increase, each category in turn should be more probable than any one of the others to be observed, as shown by a distinct peak on the category probability curve graphs (Online Resources 12–20) [26]. Disorder may indicate that categories are underutilized, are poorly defined, or that there are too many categories for respondents to sensibly distinguish [27]. Category disordering can be resolved by collapsing adjacent categories if sensible to do so and this results in improvements to other Rasch metrics [26].

Precision

Person separation indicates if the scale can adequately distinguish between different person groups according to their level of performance. Values of > 2.0 and > 0.8 for person separation index (PSI) and person reliability (PR), respectively, suggest that the scale can distinguish at least three different person groups [22]. Extreme scores (i.e., minimum or maximum) were removed a priori from the analysis (range 29 [Lighting] to 126 [Mobility]), as these provide minimal information when estimating item measures and can reduce precision [22].

Unidimensionality

We utilized principal components analysis of residuals (PCA) to assess the dimensionality of the IBs [28]. An eigenvalue ≥ 3 and raw variance explained by measures $< 50\%$ was suggestive of multidimensionality. If PCA targets were not met, we then considered four other metrics [28]: 1. A high number of misfitting items, which may indicate poor fit by a group of items to the underlying construct rather than problems with the items per se; 2. The ratio of the raw variance explained by items (i.e., % 'Rasch' dimension) and the % unexplained variance in the first contrast. If the Rasch dimension was bigger than the secondary dimension by several-fold, this indicated the secondary dimension was noticeable but not problematic; 3. The standardized residual loadings of the first contrast to determine if loading items (> 0.4) formed a potential second dimension (i.e., they formed a conceptually relevant alternative construct; 4. The disattenuated correlations of each potential cluster to ascertain if they were measuring an independent construct (disattenuated correlation < 0.82), hence providing evidence for the scale to be split, or whether they could be assessing a different "strand" of the same construct (disattenuated correlation ≥ 0.82) [29].

Item fit statistics

Item fit was explored using infit and outfit MnSq statistics (acceptable range 0.5–1.5) [22]. Before deleting misfitting items, we explored the z -residuals of individual participant responses according to the process outlined by Boone and colleagues (p. 185) [22]. Scores $> |4|$ (i.e., > 4 and < -4) indicated a high likelihood of an erroneous/unpredictable response. Such responses were given a weightage of zero and item fit statistics reassessed. If misfit persisted, we iteratively checked for z -residuals $> |3|$ and then $> |2|$ if needed, until satisfactory item fit statistics were achieved [22]. Item deletion was only considered if this process did not resolve item misfit.

We also considered discrimination when considering item performance. Items with discrimination values > 1.0 and < 1.0 means that the item discriminates between high and low performers more and less than expected for an item of this difficulty, respectively. Over-discriminating items tend to classify people as either ‘highly impaired’ or ‘highly functional’. Under-discriminating items, however, are more of a concern as they tend neither to stratify nor to measure [30]; as such, we focused on items with values substantially under 1.0 as candidates for deletion.

Local item dependency

Rasch measurement requires items that approximate local independence. That is, there should not be any correlation between two items after the effect of the underlying construct is conditioned out [31]. Local item dependency (LID) between two items was deemed present if the correlation of residuals was > 0.2 . To ensure we had LID-free item calibrations, we generated and then anchored LID-free person measures to all other person measures within a specific IB. This process forces item difficulties and rating-scale structures within the item bank to conform with the LID-free person measures and prevents LID from impacting item difficulties [13].

Targeting

Targeting was inspected via the person-item map (Online Resources 12–20). Poor targeting is evident if there are gaps in item coverage leaving certain levels of patient ability unmeasured and/or when there is a difference of > 1.0 logits between the mean item difficulty and person ability [22].

Differential item functioning

Differential item functioning (DIF) indicates if item bias is present for certain participant characteristics [32]. We examined uniform DIF for gender, age group (< 65 vs. ≥ 65 years), glaucoma type (POAG vs. PACG/normal tension glaucoma (NTG)), and language of interview (English vs. Mandarin). A DIF contrast of > 1.0 logits with a corresponding Rasch–Welch probability of $P < 0.05$ indicated notable DIF.

Measurement range

Measurement range was calculated as the difference in logits between highest and lowest item locations. A larger measurement range indicates a greater spectrum of measurement of the latent construct.

Test information function

The test information function (TIF) represents the sum of the information provided by all items in the IB and identifies where the test has highest/lowest SE (Online Resources 12–20). A higher level of information indicates greater measurement precision (i.e., low SE) at that point along the scale [33]. Generally, a $TIF \geq 10$ is considered excellent [34].

Item Separation Index (ISI)

Item separation is used to verify the item hierarchy or construct validity of the instrument. Low item separation (< 3.00) can mean that the difficulty range of items is low and/or that the person sample is not large enough to confirm the item difficulty hierarchy [35].

Level of dependence between different IBs

We applied the Pearson correlation coefficient to individual person measures from each IB to determine the level of dependence between them, with $r < 0.8$ suggesting that they were measuring independent HRQoL constructs [13].

CAT simulations

We performed simulations to assess the efficiency of our Winsteps threshold calibrations (JMLE; Joint Maximum Likelihood Estimation method) and associated CAT algorithm [36] in 1000 simulated respondents using R Statistical Computing Environment [37]. Individual packages were loaded in R to conduct IRT including CAT simulations (“catR” [38]). Simulations were based on a standard normal distribution ($M = 0$, $SD = 1$) and used the Rating-Scale Model (RSM), the ML (maximum likelihood) estimator and the Maximum Fisher Information (MFI) item selection criteria [39]. We determined the average number of items required based on two different stopping rules: SEM of 0.30 representing “high precision” and 0.387 representing “moderate precision” (approximating to a reliability of 0.91 and 0.85, respectively) [13]. Model fit was assessed using the root mean square error (RSME) and level of bias between true and estimated ability levels (low values are desirable). We also calculated the Pearson correlation coefficient between the IBs and CAT simulated person measure estimates. We hypothesized high ($r \geq 0.85$) and moderate-high ($0.75 \geq r < 0.85$) correlations for simulations with the high and moderate precision stopping rules, respectively. Results are summarized both overall (Table 3), and in deciles (D1–D10 $n = 100$ each, where D1 and D10 includes simulees at the lowest and highest ‘ability’ levels, respectively) across the latent trait ranging from -4 to 4 (Online Resources 21–27).

Results

Sociodemographic and clinical characteristics

Of the 300 patients (mean \pm SD age 67.2 ± 9.2 years; 62.3% male, 87.3% Chinese), 60 (20.0%), 114 (38.0%), 72 (24.0%), 34 (11.3%), and 20 (6.7%) had no, mild, moderate, severe, advanced/end-stage glaucoma in the better eye, respectively (Table 1). Of the 300 patients, 208 (69.3%), 45 (15.0%), and 47 (15.7%) had POAG/NTG, PACG, or different glaucoma types in each eye, respectively. Most ($n=275$; 91.7%) patients had received topical treatment in at least one eye, with 121 (20.2%) and 188 (31.1%) receiving laser or surgery in at least one eye, respectively.

Psychometric properties of the IBs

A summary of the modifications made to the GlauCAT™-Asian IBs is provided in Fig. 1, while a detailed tabulation of the initial and final psychometric properties is found in Online Resources 5–11).

In brief, the LT, MB, and GM domains required few modifications to achieve acceptable psychometric properties. These involved collapsing categories to resolve disordered thresholds and giving unexpected respondents a weightage of 0 (LT: $N=11$ [Online Resource 7]; GM: $N=12$ [Online Resource 9]; none needed for MB).

Initial analysis of the OS IB showed suboptimal precision (PSI < 2.0), five misfitting items (infit/outfit MnSq > 1.5) and possible measurement “noise” (raw variance explained by measures < 50%; Online Resource 5). Upon iterative removal of three highly misfitting items, precision level and measurement “noise” improved. With ordered thresholds (Online Resource 12), no item misfit, strong evidence of unidimensionality, lack of DIF, high applicability (100% response rate), lack of ceiling effect (< 20% extreme high responders), high ISI (6.60), good measurement range (3.28), and high clinical importance rating, the OS IB was retained despite lower than desirable measurement precision (PSI = 1.75) and TIF (~ 8.0).

The AL domain initially had > 10 misfitting items, five items that displayed DIF for age, gender, glaucoma type or language, and > 10 items with a high amount of missing data (> 20%) due to use of the ‘non-applicable’ response (Online Resource 6). High item misfit for eight items was resolved by giving unpredictable responders (total $n=10$) a weighting of 0. However, misfit for several items could not be remedied by this process and these items were deleted after considering the other performance criteria outlined in Online Resource 2. For example, item AL33 “sewing” displayed substantial misfit (infit MnSq 1.78), DIF for gender,

had 50% missing data, poor discrimination (0.62) and was deemed to have low clinical importance; as such, this item was deleted. After these amendments, a high eigenvalue for the first contrast (3.90) was suggestive of possible multidimensionality; however, the variance explained by the 1st factor was > 50%, the Rasch dimension was bigger than the secondary dimension by 3.5-fold, and the high disattenuated correlations between primary/secondary dimensions suggested that any secondary strands within the AL domain were contributing to the same underlying construct.

Initial evaluation of the PSY domain revealed strong evidence of multidimensionality, with seven misfitting items, a high eigenvalue (5.35), items relating to “concern” aspects loading together, and borderline disattenuated correlation values (Online Resource 10), suggesting that the PSY domain comprised > 1 construct. As such, we separated the “concern” items into a new domain (“CN”; $n=18$) and conducted a new Rasch analysis on the CN domain and remaining items in PSY.

The new PSY domain initially still displayed evidence of multidimensionality; however, upon removal of five items with misfit, evidence overall suggested the scale was unidimensional notwithstanding slight measurement noise (raw variance explained < 50%). While one item (PS18 “Passing glaucoma onto your children”) had an infit MnSq value > 1.5, it was retained due to its high applicability (missing data < 20%), adequate discrimination (0.91), and high perceived importance by patients during qualitative sessions. Disordered thresholds were observed in both the new PSY and CN IBs (Online Resources 18–19), which were remedied by collapsing categories ‘Moderately’ and ‘A little bit’. After category collapse, DIF for gender for CN17 “Losing your driver’s license” and DIF for glaucoma type CN18 “The appearance of your eyes” emerged. CN17 had very high missing data (56%) and both items had poor discrimination (< 0.7); as such, both items were removed after which psychometric properties improved. While the eigenvalue for the first contrast was > 3, the variance explained by the 1st factor was > 50% and high disattenuated correlations suggested that the potential secondary clusters were not measuring separate constructs.

The Work domain initially displayed disordered thresholds, poor precision (PSI = 1.55), one misfitting item and four items with DIF for age/language (Online Resource 11). Although fit statistics improved after deletion of three items and measurement range was high (6.91), precision remained suboptimal (PSI = 1.92) and the IB had low applicability (41% response rate), low test information (< 6), a low item separation index (3.02), high ceiling effect (39.8%), and a low clinical importance rating. Taken together, evidence suggested that the WK IB was not valid in this population and testing for this domain ceased.

Table 1 Sociodemographic and clinical characteristics of the 300 participants

Variable	<i>N</i>	% ^b
<i>Gender</i>		
Male	187	62.3
<i>Age (years)</i>		
40–49	15	5.0
50–59	36	12.0
60–69	120	40.0
69–90	129	43.0
<i>Ethnicity</i>		
Chinese	262	87.3
Malay	17	5.7
Indian	21	7.0
<i>Duration of glaucoma (years)</i>		
0–2	55	18.3
3–5	109	36.3
6–10	100	33.3
11–15	17	5.7
> 15	19	6.3
<i>Allergic reactions/side-effects to medication</i>		
Yes	45	15.0
No	239	79.7
<i>Glaucoma type (worse eye)</i>		
POAG/NTG	208	69.3
PACG	45	15.0
Combination ^a	47	15.7
<i>Glaucoma laterality</i>		
Unilateral	46	15.3
Bilateral	254	84.7
<i>Glaucoma severity (better eye)</i>		
None	60	20.0
Mild	114	38.0
Moderate	72	24.0
Severe	34	11.3
Advanced/end-stage	20	6.7
<i>Glaucoma severity (worse eye)</i>		
Mild	23	7.7
Moderate	107	35.7
Severe	91	30.3
Advanced/end-stage	79	26.3
<i>Glaucoma treatments(per eye)</i>		
Laser	121	20.2
Surgery	188	31.3
<i>Glaucoma treatments (per person)</i>		
Topical medication	275	91.7
<i>Vision impairment (better eye)</i>		
None (≤ 0.3 LogMAR or $\leq 20/40$ Snellen)	265	88.3
Mild (> 0.3 LogMAR ≤ 0.48 or $> 20/40$ Snellen $\leq 20/60$)	19	6.3
Moderate/severe (> 0.48 LogMAR or $> 20/60$ Snellen)	16	5.3
<i>Vision impairment (worse eye)</i>		
None (≤ 0.3 LogMAR or $\leq 20/40$ Snellen)	199	66.3
Mild (> 0.3 LogMAR ≤ 0.48 or $> 20/40$ Snellen $\leq 20/60$)	36	12.0

Table 1 (continued)

Variable	<i>N</i>	% ^b
Moderate/severe (> 0.48 LogMAR or > 20/60 Snellen)	65	21.7
<i>Marital status</i>		
Single	34	11.3
Married	227	75.7
Divorced/separated/widowed	39	13.0
<i>Highest education level^b</i>		
Primary or lower	50	16.7
Secondary	114	38.0
A Level	23	7.7
Polytechnic/Diploma/Vocational <i>Training</i>	53	17.7
University or higher	60	20.0
<i>Employment status</i>		
Working	117	39.0
Not working	183	61.0
<i>Chronic health conditions^a</i>		
Hypertension	139	53.7
Dyslipidaemia	149	49.7
Diabetes	71	23.7
Heart attack	12	4.0
Stroke	16	5.3
<i>Visual symptoms (self-reported, yes)</i>		
Blurred vision	119	39.7
Fluctuating vision (eg: your vision change regularly)	77	25.7
Dimming of your vision	57	19.0
Blocking of your vision	53	17.7
<i>Use of assistive devices/low vision aids</i>		
None	42	12.8
Corrective aids	254	77.2
Mobility aids (white cane, walking stick)	1	0.3
Magnifiers (handheld, lenses, electronic)	24	7.3
Adjusted lighting	8	2.4
<i>Support for glaucoma</i>		
None	181	60.3
A little	25	8.3
A moderate amount	41	13.7
A great deal	53	17.7
Continuous variables		
	Mean/Median	SD/IQR
Age (years)	67.2/68.0	9.2/11.0
Presenting VA (better eye), LogMAR; Snellen	0.16; 20/25/0.1; 20/25	0.18; 20/25/0.16; 20/25
Presenting VA (worse eye), LogMAR; Snellen	0.32; 20/40/0.26; 20/40	0.21; 20/32/0.26; 20/40
Visual fields (better eye), mean deviation	− 7.2/− 5.2	6.9/7.1
Visual fields (worse eye), mean deviation	− 14.9/− 13.3	7.8/11.1
No. of topical treatments within past 6 months	1.7/2.0	1.0/1.0

LogMAR logarithm of the minimal angle of resolution, *NTG* normal tension glaucoma, *PACG* primary angle closure glaucoma, *POAG* primary open angle glaucoma, *SD* standard deviation, *VA* visual acuity

^aCombination patients included those with POAG in one eye/PACG in the other ($n=2$); POAG in one eye/Other glaucoma in the other ($n=8$); PACG in one eye/Other glaucoma in the other ($n=30$); NTG in one eye/Other glaucoma in the other ($n=7$)

^bPercentages for some variables may not equal 100% due to missing data or participants selecting > 1 category

Fig. 1 Flowchart explaining the iterative process of item bank modification following psychometric assessment. The left hand side shows there were seven original item banks, of which six were retained and one new item bank was created (right, green boxes). One item bank had suboptimal psychometric properties and was not further pursued in CAT testing (right, red box)

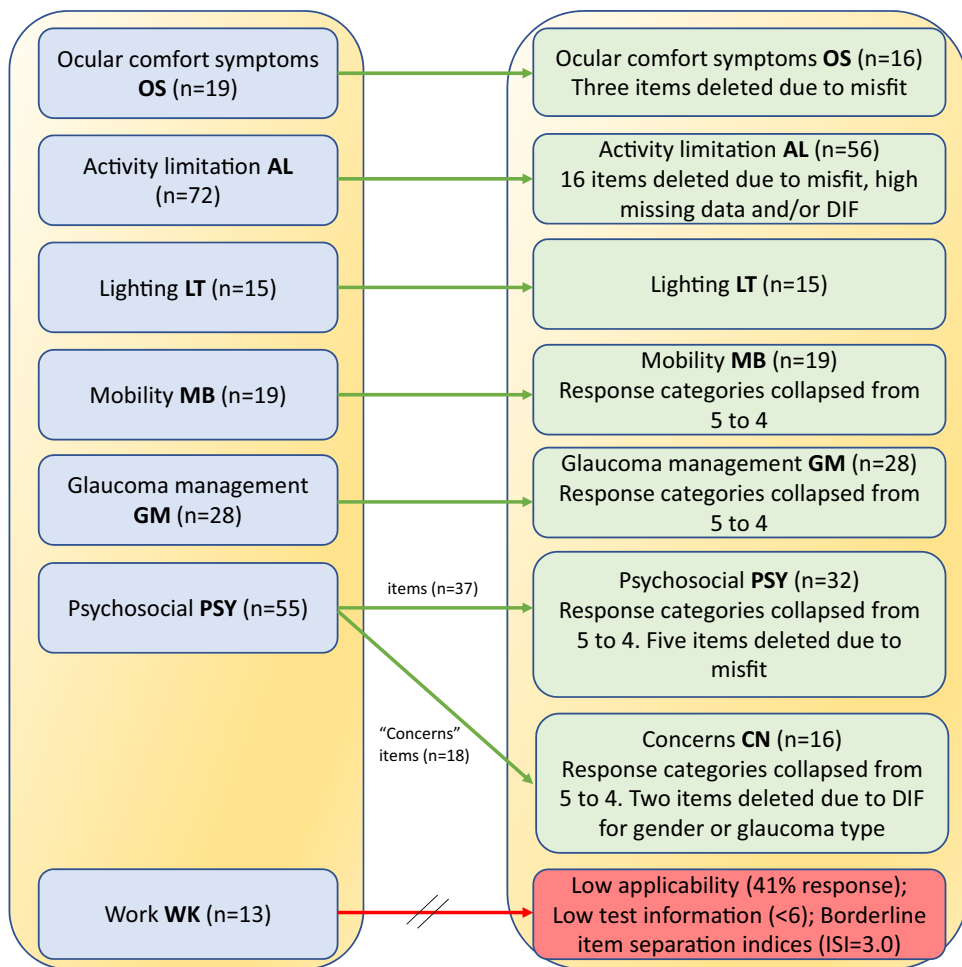


Table 2 Correlation coefficients between the final seven domains of the GlauCAT™-Asian item banks

	OS	AL	LT	MB	GM	CN	PSY
OS	–	0.42	0.44	0.27	0.47	0.41	0.37
AL		–	0.71	0.71	0.37	0.52	0.45
LT			–	0.73	0.39	0.51	0.45
MB				–	0.39	0.51	0.46
GM					–	0.55	0.61
CN						–	0.76
PSY							–

AL activity limitation, CN concerns, GM glaucoma management, LT lighting, MB mobility, OS ocular comfort symptoms, PSY psychosocial

Overall, targeting was suboptimal for all IBs (Online Resources 5–11), with items, on average, too easy to endorse for participants’ level of the QoL construct (Online Resources 12–20). Test information was good overall (excluding WK), ranging from 8.1 for OS to 44.2 for AL (Online Resources 12–20). Finally, correlation coefficients of the person measures for all final IBs were < 0.8 (range 0.27–0.76), supporting the independence of each QoL domain assessed by the individual IBs (Table 2).

Following psychometric analyses (Fig. 1), there were seven IBs comprising 182 items: AL (*n* = 56); LT (*n* = 15); MB (*n* = 19); PSY (*n* = 32); OS (*n* = 16); GM (*n* = 28); and CN (*n* = 16).

CAT simulations

Overall, good model fit (low RMSE and bias values) was achieved for all simulations (Table 3). With a 0.3

Table 3 CAT simulation results for the seven GlauCAT-Asian quality of life item banks

Domain, items	Mean no. items	Correlation with true theta	RMSE	Bias	% satisfied stopping rule
<i>SEM 0.3 (high precision)</i>					
OS, 16	16.00	0.80	0.79	0.02	0
AL, 56	12.55	0.82	0.72	0.10	98
LT, 15	15.00	0.81	0.77	-0.02	0
MB, 19	19.00	0.89	0.55	-0.01	0
GM, 28	16.44	0.85	0.72	-0.03	85
CN, 16	16.00	0.77	0.91	0.02	0
PSY, 32	15.40	0.82	0.76	-0.04	92
<i>SEM 0.387 (moderate precision)</i>					
OS, 16	14.50	0.80	0.80	0.01	66
AL, 56	7.87	0.81	0.72	0.08	99
LT, 15	13.02	0.80	0.82	< -0.001	85
MB, 19	15.0	0.87	0.57	0.01	88
GM, 28	10.55	0.85	0.73	-0.02	92
CN, 16	13.87	0.76	0.92	0.03	69
PSY, 32	9.86	0.80	0.80	-0.03	95

AL activity limitation, CAT computerized adaptive test, CN concerns, GM glaucoma management, LT lighting, MB mobility, OS ocular comfort symptoms, PSY psychosocial, RMSE root mean square error, SEM standard error of measurement

SEM stopping rule, the mean number of items administered ranged between 12.55 for AL and 19 for MB (overall mean 15.7 items; Table 3). For the smaller IBs (i.e., those with < 20 items: OS, LT, MB and CN), all items were administered and the stopping rule target was not reached. For AL, GM and PSY, the target stopping rule was met 100% of the time except for at the very ‘unable’ (D1) and very ‘able’ (D9-10) ends of the spectrum (Online Resources 21–27).

For ‘moderate’ precision (0.387 SEM), the mean number of items administered ranged between 7.87 for AL and 14.5 for OS (overall mean 12.1 items; Table 3). Except for the extreme deciles (D1 and D10), the proportion of simulees reaching the stopping rule target was > 75% (Online Resources 21–27).

Correlations between the CAT simulated person measures and the IBs were moderate to high (0.77–0.89 for SEM 0.3, and 0.76–0.87 for SEM 0.387).

In simulations in which test length was capped at 10 items, the SEM achieved ranged between 0.309 for AL and 0.466 for CN, corresponding to a reliability of ~0.91 and ~0.78, respectively (Table 4).

Discussion

Following a comprehensive assessment of seven initial glaucoma-specific HRQoL IBs designed specifically for an Asian population (GlauCAT™-Asian), LT, MB and GM required

Table 4 CAT simulation results using a combined stopping rule based on test length and precision

Domain	Mean SEM	Reliability ^a	Correlation with true theta	RMSE	Bias
OS, 16	0.463	0.79	0.77	0.87	0.04
AL, 56	0.309	0.91	0.88	0.54	-0.07
LT, 15	0.431	0.81	0.78	0.87	0.03
MB, 19	0.457	0.79	0.82	0.68	0.04
GM, 28	0.385	0.85	0.83	0.81	0.00
CN, 16	0.466	0.78	0.74	0.97	0.04
PSY, 32	0.357	0.87	0.80	0.83	-0.03

Test length = 10 items and precision = SEM 0.3

AL activity limitation, CAT computerized adaptive test, CN concerns, GM glaucoma management, LT lighting, MB mobility, OS ocular comfort symptoms, PSY psychosocial, RMSE root mean square error, SEM standard error of measurement

^aDerived from the formula $1 - SEM^2$

minimal optimization of their psychometric properties. AL and PSY required more extensive remediations, including category collapse, deletion of poorly performing items, and/or splitting scales to resolve multidimensionality resulting in a new IB, “CN”. Due to unresolvable psychometric issues, the WK IB was not pursued further. In CAT simulations of the seven final IBs (n = 182 items in total), two IBs achieved moderate precision with ≤ 10 items administered, and 5/7 achieved moderate precision with ≤ 15 items. Researchers

and clinicians can utilize GlauCAT™-Asian to obtain a comprehensive yet low burden understanding of the HRQoL impact of glaucoma and associated treatments.

Sixteen items in the AL domain were deleted due to high missing data (> 20%) or other associated psychometric issues (e.g., misfit). Items relating to social participation (e.g., Participating in community groups) were particularly susceptible to missing data. This may have been because these activities were not relevant to our elderly participants or because data collection occurred during the COVID-19 pandemic, during which Singapore spent ~ 100 days under strict lockdown [40] with many community activities suspended. As items that are likely to elicit the ‘non-applicable’ response are a threat to CAT efficiency, six ‘social participation’ items with high missing data were removed. While several ‘social participation’ items remain in the IB, users with a focus on measuring activity limitation in a social context may wish to include additional related questionnaires.

Despite iterative remedial steps, the OS IB displayed suboptimal precision (PSI 1.75), which may be due to a lack of variance in our sample population. Indeed, few participants had VI in the better eye (11.6%) resulting in a sample skewed towards the higher end of the ‘ability’ spectrum. However, because the PSI/PR values fall well within a useful range for a measurement tool [22], we retained the OS scale due to high perceived importance in our qualitative participant responses [19], and input from our team’s research clinicians that changes in OS consequent to glaucoma and its associated interventions are integral to clinical care outcomes.

The PSY domain needed to be split to resolve multidimensionality, which is not unexpected given that it contained items relating to ‘concerns’ (e.g., falling), social life (e.g., becoming socially isolated), and emotional reactions (e.g., feel frustrated). Indeed, sister instruments for diabetic retinopathy (RetCAT™ [13, 41]) and GlauCAT™-Western [17] both have > 1 domain measuring the psychosocial aspects of HRQoL. In a head-to-head comparison of the two GlauCAT instruments, the newly developed Asian version is more compact, comprising only seven domains compared to 12. While 5/7 GlauCAT™-Asian domains (namely OS, AL, MB, LT, CN) are like those in the original, the remaining two domains differ. The PSY domain in GlauCAT™-Asian covers content found in two domains of GlauCAT™-Western (Emotional and Social) and, while the GM domain in GlauCAT™-Asian has some similarities to the Convenience-Treatment domain in GlauCAT™-Western, it has twice the number of items and covers a broader range of treatment-related issues.

Despite iterative remedial steps, the WK IB displayed suboptimal psychometric properties including low test information and item separation. These issues may be due to the small sample size, with only 41% of participants providing responses for this domain and, of these, 40% recording an

‘extreme’ high score and thus removed a priori from analysis. This was unexpected as ‘work’ issues were important to our working focus group participants (25%) [19]. However, with only 17% of our current sample aged < 60 years (median age 68), the WK IB lacked applicability and was not pursued. Further work, including additional data collection in a younger glaucoma sample, are warranted to optimize the WK IB.

Overall, targeting was suboptimal for all our IBs, with items too “easy” relative to the average ability level of our participants. This affected CAT efficiency in the highest ability ‘bins’ (D9-10), with more items needed on average to provide a score estimate. Similar targeting issues were observed in the GlauCAT™-Western instrument, where 11/12 IBs lacked items targeted to very high performers [17]. The poor targeting may be due to the lack of VI in our sample, with most people with glaucoma having good VA until late-stage disease [42, 43]. Consultation with more early-stage patients to create ‘harder’ items is warranted. Addition of novel items to IBs is possible by estimating the calibration of new items relative to existing ones using Rasch analysis [44]. While users should be aware of the limitations of the GlauCAT™-Asian in providing stable ability estimates in patients at the very able end of the spectrum, poor targeting in the upper score range is not necessarily problematic, as such patients may not be the focus of clinician monitoring in healthcare or drug development in pharma companies.

Overall, our simulation results were promising, particularly for our moderate precision stopping rule target where CAT administration of our GlauCAT™-Asian IBs could reduce the number of items needed by 10–86% (depending on the domain) compared to the full IB. CAT efficiency was reduced with the high precision stopping rule, where the smaller IBs needed all items to be administered. Correlations between the CAT and full IB thetas (person measures) was somewhat lower than expected (~ 80% compared to the expected ≥ 0.85) and also lower than that obtained by Patient-Reported Outcomes Measurement Information System CATs in other health conditions (> 0.90) [45, 46], suggesting that the reduction in response burden offered by GlauCAT™-Asian may come with some potential loss of accuracy in scores. This may be particularly important for individual level comparisons, where measurements may require reliabilities ≥ 0.90 to be considered appropriate. Future work will conduct a formal comparison between CAT and full item bank theta with real data, where a subset of patients ($n \sim 50$) will answer both the CAT tests and associated full item banks. This will provide a better understanding of the effectiveness of GlauCAT in reducing response burden while replicating full IB scores.

Overall, our results demonstrate that IB and CAT approaches are advantageous compared to fixed-length questionnaire administration [11, 47]. Moreover, as the

GlauCAT™-Asian IBs function independently, users can select which to administer depending on their measurement goals, which is useful in clinical settings, where time for PROM administration is limited yet increasingly valued in value-based care [10]. Indeed, using our online CAT testing platform (PROMinsight), six QoL domains of GlauCAT™-Western (taking 8 min on average) were recently administered to > 200 glaucoma patients while waiting to see their treating doctor at Mass. Eye and Ear Institute, with the HRQoL scores discussed during the patients' consultations [18, 48]. Analysis of the group data found that VA was associated with greater activity limitation and worse mobility, while poorer VF mean deviation was associated with worse emotional well-being. These pilot results provide real-world evidence of the implementation feasibility of GlauCAT™ in routine clinical care enabling real-time discussion of results [49] and subsequent analysis of group data to better understand the factors associated with HRQoL outcomes in glaucoma patients [48].

Strengths of our study include our rigorous psychometric assessment and wealth of reliability and validity data provided. We also employed techniques to save items with minor misfit rather than deleting them outright, thus retaining as many issues important to glaucoma patients as possible. Our sample included patients with bilateral POAG ($n=202$) and PACG ($n=45$), ensuring that our calibrations are robust in Asian populations where the prevalence of PACG is high (0.7%) compared to Western populations (0.2–0.4%) [50]. Our recruitment strategy using clinical and demographic target quotas ensured we recruited participants across a range of glaucoma subtypes and severities enabling us to capture diverse responses. However, several limitations of our work must be acknowledged. First, we did not explore how IB score estimates differed across levels of glaucoma severity and associated VI, and different glaucoma subtypes. Future work using the final GlauCAT-Asian™ instrument will assess the HRQoL impact of glaucoma types and severity in another large sample of Asian patients with glaucoma. Second, while our new instrument has been developed in Asian patients, our sample comprised relatively few Indians or Malays meaning that we may have missed issues pertinent to these minority groups. Third, while Singapore comprises a multi-ethnic population of Chinese, Malays and Indians, who form some of the largest ethnic groups in Asia, further studies to test the cross-cultural validity of GlauCAT™-Asian beyond Singapore are warranted. For example, item calibrations between our Singaporean population and glaucoma patients from China, Malaysia and India could be assessed for agreement, and DIF for culture could be explored. If found, cultural DIF could be managed using the item anchoring method described by Gibbons and colleagues, where parameters of items with DIF are allowed to vary, while item calibrations for the other items

which displayed DIF remain constant across each country [51]. Fourth, some correlations between our measures were relatively high (e.g., PSY and CN) which may support a multidimensional latent structure underlying glaucoma-specific QoL and the application of multidimensional IRT and bifactor models [52–54]. However, our aim was to produce unidimensional measurement tools that provide users with the ability to administer selected scales for a given purpose. Finally, our use of the ML estimator in our CAT simulations may have resulted in inaccurate estimates for participants at the extreme ends of the ability spectrum due to lack of variability in responses, and other estimators such as EAP (expected a posteriori) or BM (Bayesian Method) could be used instead to overcome this issue. Similarly, we used the standard normal distribution ($M=0$, $SD=1$) to run our simulations. Given that our item bank calibrations were based on a relatively able participant sample with high mean person measures, simulating latent traits from a normal distribution with these parameters may have affected the accuracy of our results.

In conclusion, following reengineering using Rasch analysis to optimize their psychometric properties, our seven IBs (OS, AL, LT, MB, GM, PSY, and CN) allow comprehensive assessment of glaucoma-specific HRQoL relevant to glaucoma populations in Asia. CAT simulations suggest that relatively few items within each IB will be required, particularly in low-stakes situations. Future work aims to assess the ability of GlauCAT™-Asian to provide efficient, valid, and reliable assessment of glaucoma-specific HRQoL in a clinical sample of glaucoma patients across the range of glaucoma severity and types; and the feasibility and acceptability of its implementation into routine clinical care.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11136-023-03428-8>.

Acknowledgements The authors would like to thank Prof Mike Linacre for his assistance in Rasch analysis and CAT simulations.

Funding This study was funded by a National Medical Research Council Health Services Research Grant (PI Lamoureux, and Co-Is Fenwick, Man, Baskaran, and Aung, #0070/2016). The sponsor or funding organization had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data availability statement Data are not available to other researchers for replication purposes as the study group has not completed planned analyses for future publications; and due to strict data privacy laws in Singapore. Analytic methods or materials are available upon request.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

- Chan, E. W., Li, X., Tham, Y. C., Liao, J., Wong, T. Y., Aung, T., & Cheng, C. Y. (2016). Glaucoma in Asia: Regional prevalence variations and future projections. *British Journal of Ophthalmology*, *100*(1), 78–85.
- Tham, Y. C., Li, X., Wong, T. Y., Quigley, H. A., Aung, T., & Cheng, C. Y. (2014). Global prevalence of glaucoma and projections of glaucoma burden through 2040: A systematic review and meta-analysis. *Ophthalmology*, *121*(11), 2081–2090.
- Szegedi, S., Boltz, A., Scharinger, E. M., & Vécsei-Marlovits, P. V. (2022). Quality of life in patients with glaucoma assessed by 39-item National Eye Institute Visual Functioning Questionnaire (NEI VFQ-39). *Graefes Archive for Clinical and Experimental Ophthalmology*, *260*(5), 1623–1631.
- Fenwick, E., Ong, P., Man, R., Cheng, C.-Y., Sabanayagam, C., Wong, T., & Lamoureux, E. L. (2016). The impact of vision impairment and major eye diseases on mobility and independence in a Chinese population. *JAMA Ophthalmology*, *134*(10), 1087–1093.
- Lamoureux, E. L., Chong, E., Wang, J. J., Saw, S. M., Aung, T., Mitchell, P., & Wong, T. Y. (2008). Visual impairment, causes of vision loss, and falls: The Singapore Malay Eye Study. *Investigative Ophthalmology & Visual Science*, *49*(2), 528–533.
- Fenwick, E., Ong, P., Man, R., Cheng, C.-Y., Sabanayagam, C., Wong, T., & Lamoureux, E. L. (2016). Vision impairment and major eye diseases reduce vision-specific emotional well-being in a Chinese population. *British Journal of Ophthalmology*, *101*(5), 686–690.
- Koenig, S. F., & Hirneiss, C. W. (2022). Changes in quality of life of glaucoma patients over a period of 8 years. *Ophthalmologie*, *119*(6), 599–604.
- Rahman, M. Q., Beard, S. M., Discombe, R., Sharma, R., & Montgomery, D. M. (2013). Direct healthcare costs of glaucoma treatment. *British Journal of Ophthalmology*, *97*(6), 720–724.
- Stagg, B. C., Granger, A., Guetterman, T. C., Hess, R., & Lee, P. P. (2022). The burden of caring for and treating glaucoma: The patient perspective. *Ophthalmol Glaucoma*, *5*(1), 32–39.
- Basch, E. (2017). Patient-reported outcomes—Harnessing patients' voices to improve clinical care. *New England Journal of Medicine*, *376*(2), 105–108.
- Cella, D., Gershon, R., Lai, J. S., & Choi, S. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research*, *16*(Suppl 1), 133–141.
- Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement*, *6*(1), 109–127.
- Fenwick, E. K., Khadka, J., Pesudovs, K., Rees, G., Wong, T. Y., & Lamoureux, E. L. (2017). Diabetic retinopathy and macular edema quality-of-life item banks: Development and initial evaluation using computerized adaptive testing. *Investigative Ophthalmology & Visual Science*, *58*(14), 6379.
- Kandel, H., Khadka, J., Watson, S. L., Fenwick, E. K., & Pesudovs, K. (2021). Item banks for measurement of refractive error-specific quality of life. *Ophthalmic and Physiological Optics*, *41*(3), 591–602.
- Matsuura, M., Hirasawa, K., Hirasawa, H., Yanagisawa, M., Murata, H., Mayama, C., & Asaoka, R. (2017). Developing an item bank to measure quality of life in individuals with glaucoma, and the results of the interview with patients: The effect of visual function, visual field progression rate, medical, and surgical treatments on quality of life. *Journal of Glaucoma*, *26*(2), e64–e73.
- Khadka, J., McAlinden, C., Craig, J. E., Fenwick, E. K., Lamoureux, E. L., & Pesudovs, K. (2015). Identifying content for the glaucoma-specific item bank to measure quality-of-life parameters. *Journal of Glaucoma*, *24*(1), 12–19.
- Man, R. E. K., Fenwick, E. K., Khadka, J., Wu, Z., Skalicky, S., Pesudovs, K., & Lamoureux, E. L. (2022). Psychometric evaluation of Glaucoma Quality of Life Item Banks (GlauCAT) and initial assessment using computerized adaptive testing. *Translational Vision Science & Technology*, *11*(6), 9.
- Fenwick, E., Roldan, A., Halawa, O., Meshkin, R., Zebardast, N., Popov, V., Lis, P., Friedman, D., & Lamoureux, E. (2022). Implementation of an online glaucoma-specific quality of life computerized adaptive test (GlauCAT) system in a US glaucoma hospital. *Translational Vision Science & Technology*, *11*(2), 24.
- Fenwick, E., Lim, B., Man, R., Baskaran, M., Nongpiur, M., Sng, C. C. A., Iyer, J. V., Husain, R., Perera, S., Wong, T., Low, J., Shimin, O. H., Lun, K., Aung, T., & Lamoureux, E. (2022). The development of a glaucoma-specific health-related quality of life item bank supporting a novel computerized adaptive testing system in Asia. *Journal of Patient-Reported Outcomes*. <https://doi.org/10.1186/s41687-022-00513-3>
- Brooke, P., & Bullock, R. (1999). Validation of a 6 item cognitive impairment test with a view to primary care usage. *International Journal of Geriatric Psychiatry*, *14*(11), 936–940.
- Linacre, J. M. (2002). *A user's guide to Winsteps: Rasch-Model Computer program*. MESA Press.
- Boone, W., Staver, J., & Yale, M. (2014). *Rasch analysis in the human sciences*. Springer.
- Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in Psychology*, *7*, 109.
- Reeve, B., & Fayers, P. (2005). Applying item response theory modelling for evaluating questionnaire item and scale properties. In P. Fayers & R. Hays (Eds.), *Assessing quality of life in clinical trials: Methods and practice* (2nd ed., pp. 55–73). Oxford University Press.
- Linacre, J. (1994). Sample size and item calibration [or Person Measure] stability. *Rasch Measurement Transactions*, *7*(4), 328.
- Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*, 85–106.
- Khadka, J., Gothwal, V. K., McAlinden, C., Lamoureux, E. L., & Pesudovs, K. (2012). The importance of rating scales in measuring patient-reported outcomes. *Health and Quality of Life*, *10*, 80–93.
- Linacre, M. Dimensionality and Structural Validity investigation—An example. Retrieved August 16, 2022, from <https://www.winsteps.com/winman/multidimensionality.htm>
- Schumacker, R., & Muchinsky, P. (1996). Disattenuating correlation coefficients. *Rasch Measurement Transactions*, *10*(1), 479.
- Linacre, M. (1999). DISCRIMINATION= report item discrimination = No. <https://www.winsteps.com/winman/discrim.htm>
- Baghaei, P. (2008). Local dependency and Rasch measures. *Rasch Measurement Transactions*, *21*(3), 1105–1106.
- Linacre, M. Table 30.1 Differential item functioning DIF pairwise. Retrieved December 18, 2022, from https://winsteps.com/winman/table30_1.htm
- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Lawrence Erlbaum Associates.
- Embretson, S., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.
- Linacre, J. (1999). Reliability and separation of measures. <https://www.winsteps.com/winman/reliability.htm>
- Chen, S.-K., & Cook, K. F. (2009). SIMPOLYCAT: An SAS program for conducting CAT simulation based on polytomous IRT models. *Behavior Research Methods*, *41*(2), 499–506.

37. R Core Team R-project. (2016). R: A language and environment for statistical computing. <http://www.r-project.org>
38. Magis, D., & Raîche, G. (2011). CatR: An R package for computerized adaptive testing. *Applied Psychological Measurement*, 35(7), 576–577.
39. Choi, S. W. (2009). Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. *Applied Psychological Measurement*, 33(8), 644–645.
40. 2020–21 Singapore circuit breaker measures. 2022. Retrieved August 22, 2022, from https://en.wikipedia.org/wiki/2020%E2%80%9321_Singapore_circuit_breaker_measures
41. Fenwick, E., Barnard, J., Gan, A., Loe, B., Khadka, J., Pesudovs, K., Man, R., Lee, S., Tan, G., Wong, T., & Lamoureux, E. (2020). Computerised adaptive tests: An innovative, efficient and precise method to assess the patient-centred impact of diabetic retinopathy. *Translational Vision Science & Technology*, 9(7), 3.
42. Ang, G. S., & Eke, T. (2007). Lifetime visual prognosis for patients with primary open-angle glaucoma. *Eye (London, England)*, 21(5), 604–608.
43. Kwon, Y. H., Kim, C. S., Zimmerman, M. B., Alward, W. L., & Hayreh, S. S. (2001). Rate of visual field loss and long-term visual outcome in primary open-angle glaucoma. *American Journal of Ophthalmology*, 132(1), 47–56.
44. Haley, S. M., Ni, P., Jette, A. M., Tao, W., Moed, R., Meyers, D., & Ludlow, L. H. (2009). Replenishing a computerized adaptive test of patient-reported daily activity functioning. *Quality of Life Research*, 18(4), 461–471.
45. Rose, M., Bjorner, J. B., Gandek, B., Bruce, B., Fries, J. F., & Ware, J. E., Jr. (2014). The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. *Journal of Clinical Epidemiology*, 67(5), 516–526.
46. Stucky, B. D., Edelen, M. O., Tucker, J. S., Shadel, W. G., Cerully, J., Kuhfeld, M., Hansen, M., & Cai, L. (2014). Development of the PROMIS negative psychosocial expectancies of smoking item banks. *Nicotine & Tobacco Research*, 16(Suppl 3), S232–240.
47. Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., Bhaumik, D. K., Stover, A., Bock, R. D., & Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services (Washington, D. C.)*, 59(4), 361–368.
48. Halawa, O., Roldan-Vasquez, A., Meshkin, S., Zebardast, N., Fenwick, E., Lamoureux, E., & Friedman, D. (2022). Factors associated with glaucoma-specific quality of life in a US glaucoma clinic in a pilot implementation of an online computerised adaptive test (GlauCAT). *British Journal of Ophthalmology*. <https://doi.org/10.1136/bjophthalmol-2022-321145>
49. Harrison, C., Loe, B. S., Lis, P., & Sidey-Gibbons, C. (2020). Maximizing the potential of patient-reported assessments by using the open-source concerto platform with computerized adaptive testing and machine learning. *Journal of Medical Internet Research*, 22(10), e20950–e20950.
50. Zhang, N., Wang, J., Chen, B., Li, Y., & Jiang, B. (2021). Prevalence of primary angle closure glaucoma in the last 20 years: A meta-analysis and systematic review. *Frontiers in Medicine*, 7, 624179–624179.
51. Gibbons, C. J., & Skevington, S. M. (2018). Adjusting for cross-cultural differences in computer-adaptive tests of quality of life. *Quality of Life Research*, 27(4), 1027–1039.
52. Stochl, J., Bohnke, J. R., Pickett, K. E., & Croudace, T. J. (2016). Computerized adaptive testing of population psychological distress: Simulation-based evaluation of GHQ-30. *Social Psychiatry and Psychiatric Epidemiology*, 51(6), 895–906.
53. Croudace, T. J., & Bohnke, J. R. (2014). Item bank measurement of depression: Will one dimension work? *Journal of Clinical Epidemiology*, 67(1), 4–6.
54. Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, 16(Suppl 1), 19–31.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.