tvst

Article

Psychometric Evaluation of Glaucoma Quality of Life Item Banks (GlauCAT) and Initial Assessment Using Computerized Adaptive Testing

Ryan Eyn Kidd Man[1,2], Eva K. Fenwick[1,2], Jyoti Khadka[3–5], ZhiChao Wu[6,7], Simon Skalicky[6–8], Konrad Pesudovs[9], and Ecosse L. Lamoureux[1,2,6,7]

[1] Singapore Eye Research Institute and Singapore National Eye Centre, Singapore
[2] Duke–NUS Medical School, National University of Singapore, Singapore
[3] Health and Social Care Economics Group, College of Nursing and Health Sciences, Flinders University, Adelaide, South Australia, Australia
[4] Registry of Senior Australians, South Australian Health and Medical Research Institute, Adelaide, South Australia, Australia
[5] Business School, University of South Australia, Adelaide, South Australia, Australia
[6] Centre for Eye Research Australia, Royal Victorian Eye and Ear Hospital, East Melbourne, Victoria, Australia
[7] Ophthalmology, Department of Surgery, The University of Melbourne, Melbourne, Victoria, Australia
[8] Department of Medicine, The University of Melbourne, Melbourne, Victoria, Australia
[9] School of Optometry and Vision Science, University of New South Wales, Sydney, New South Wales, Australia

Correspondence: Ecosse L. Lamoureux, Singapore Eye Research Institute, The Academia, 20 College Road, Level 6, Discovery Tower, Singapore 169856, Singapore.
e-mail: ecosse.lamoureux@duke-nus.edu.sg

Received: March 8, 2022
Accepted: May 18, 2022
Published: June 9, 2022

Keywords: glaucoma; quality of life; item banks; Rasch analysis; computerized adaptive testing

Citation: Man REK, Fenwick EK, Khadka J, Wu Z, Skalicky S, Pesudovs K, Lamoureux EL. Psychometric evaluation of glaucoma quality of life item banks (GlauCAT) and initial assessment using computerized adaptive testing. Transl Vis Sci Technol. 2022;11(6):9, https://doi.org/10.1167/tvst.11.6.9

Purpose: To evaluate the psychometric properties of glaucoma-specific quality of life (QoL) item banks (GlauCAT) and assess their performance using computerized adaptive testing (CAT) simulations.

Methods: In this cross-sectional study, 293 participants with glaucoma (mean age ± SD, 70.7 ± 13.2 years; 45% female) answered 342 items in 12 QoL item banks (IBs): Activity Limitation (AL); Driving (DV); Convenience (CV); Economic (EC); Emotional (EM); General Symptoms (GS); Health Concerns (HC); Lighting (LT); Mobility (MB); Ocular Surface Symptoms (OS); Social (SC); and Visual Symptoms (VS). These IBs were assessed using Rasch analysis, and CAT simulations with 1000 simulated respondents were utilized to determine the average number of items to be administered to achieve moderate and high precision levels.

Results: The AL, DV, EM, HC, LT, MB, EC, OS, SC, and VS IBs required relatively minor amendments to achieve satisfactory psychometric fit. To resolve multidimensionality, we split CV into Treatment Convenience (TCV) and General Convenience (GCV). Due to poor measurement precision, the GS IB was not pursued further. This resulted in 12 total IBs. In CAT simulations, an average of 3.7 and 7.3 items per IB were required to attain measurement at moderate and high precision, respectively.

Conclusions: Following rigorous psychometric assessment, we developed 12 valid glaucoma-specific QoL domains that can obtain highly precise person measure estimates using a small number of items.

Translational Relevance: GlauCAT will enable researchers and clinicians to quickly and comprehensively assess the impact of glaucoma and its associated interventions across a range of QoL domains.

Introduction

Glaucoma, the commonest cause of irreversible blindness globally, is projected to increase by over 70% by the year 2040.[1] Not only does this blinding disease impact substantially on a patient's quality of life (QoL),[2] but treatment options to slow down disease progression, including long-term topical medication use and surgery,[3,4] can also place a high burden on patients in terms of costs and side effects.[5–8]

With the transition to value-based care models, assessing the effectiveness of chronic disease interventions from the patient's perspective using

patient-reported outcome measures (PROMs) is crucial[9] and is mandated by several regulatory agencies.[10] Although several vision- and glaucoma-specific fixed-length PROMs are available, most have been developed and validated using classical test theory, which, in addition to having several psychometric shortcomings, also requires all items within the scale to be answered in order to arrive at an overall estimate of an individual's domain score.[11–13] Current paper-and-pencil PROMs therefore tend to focus on only one or two QoL domains (e.g., task difficulty) and have limited items within each domain so as to minimize participant burden,[14] resulting in suboptimal targeting of patients' impairment level across the spectrum of glaucoma severity.

These limitations can be addressed using modern psychometric techniques such as item banking and computerized adaptive testing (CAT). An item bank (IB) is a pool of items (questions), calibrated according to difficulty, that measure a defined latent construct such as visual functioning.[15] CAT is a method for administering items from a calibrated IB where, based on a person's previous responses, an algorithm selectively presents items that provide the greatest amount of information, until a predefined stopping criterion is reached.[16] CAT therefore requires fewer items (~7–10) than fixed-length PROMs yet maintains similar precision, allowing for a more comprehensive understanding of the QoL impact of the disease under assessment.[17,18]

We have previously reported on the development of content for a glaucoma-specific IB and CAT instrument to assess the specific QoL impact of glaucoma and its associated management strategies (GlauCAT),[19] as well as preliminary psychometric evaluation of the Activity Limitation IB.[20] Here, we present a thorough assessment of the psychometric properties of all IBs in GlauCAT in a clinical sample of patients with glaucoma using Rasch analysis and evaluate the efficiency of the final calibrated IBs by simulating a CAT application.

Methods

Sample Population

We recruited 293 glaucoma patients from two sources: (1) ophthalmic clinics at the Royal Victorian Eye and Ear Hospital (RVEEH), and (2) participants from the Glaucoma Initial Treatment Study (GITS), a multicentered, cluster-randomized, controlled clinical trial comparing selective laser trabeculoplasty and topical medication (eye drops) for patients with primary open-angle glaucoma (POAG) and exfoliative glaucoma (XFG).[21] All participants were recruited between 2013 and 2014; had confirmed diagnoses of glaucoma (POAG, XFG, angle closure, secondary and normal tension); were ≥35 years of age; had no significant hearing or cognitive impairment; and had no other late-stage eye diseases. Face-to-face interviews to administer all questionnaire materials were conducted by trained research assistants after obtaining each participant's written informed consent. Telephone interviews were offered if participants could not, or found it inconvenient, to make a scheduled visit to the study site, as several studies have found very little difference in data quality between face-to-face and telephone interviews.[22–24] This study was approved by the Human Research Ethics Committee of the RVEEH (#11/1024H and #11/995H) and was conducted according to the tenets of the Declaration of Helsinki.

Development of the Glaucoma IBs

The development of the domains and items for the GlauCAT IBs have been described elsewhere.[19] At the end of the content development phase, the GlauCAT instrument was comprised of 342 items under 10 QoL domains: Visual Symptoms (VS; $n = 19$); Ocular Surface Symptoms (OS; $n = 22$); General Symptoms (GS; $n = 15$), Activity Limitation (AL; $n = 66$); Mobility (MB; $n = 20$); Emotional (EM; $n = 49$); Health Concerns (HC; $n = 45$); Social (SC; $n = 23$); Convenience (CV; $n = 39$); and Economic (EC; $n = 22$). Domains were all rated on four- or five-category scales (see Supplementary Material).[19] During preliminary psychometric assessment,[20] the AL domain was found to be multidimensional, and the Driving (DV; $n = 13$) and Lighting (LT; $n = 9$) domains were subsequently split from AL and formed into separate domains, resulting in 12 total IBs for the current analyses.

Assessment of Glaucoma, Visual Fields and Visual Acuity

Humphrey visual field (VF) testing, using the 24-2 Swedish Interactive Thresholding Algorithm (SITA)-Standard or 30-2 SITA-Standard test, was performed to assess the extent of visual field loss, and the severity was defined from the mean deviation (MD) thresholds using the modified Hodapp–Parrish–Anderson criteria as early (MD > −6 dB), moderate (MD = −6 to −12 dB), and advanced (MD < −12 dB) glaucoma.[25] Refraction and best-corrected visual acuity (BCVA) measurements were performed by a trained orthoptist/optometrist. BCVA was measured for each eye and reported based on the logarithm of the minimum angle

of resolution (logMAR) visual acuity (VA) chart at 4 m. Participants' vision impairment was categorized as none (better-eye VA, logMAR $\leq$ 0.3), mild–moderate (0.3 < logMAR <1.0), or severe (logMAR $\geq$ 1.0).

Psychometric Evaluation of the IBs

We performed Rasch analyses separately on each of the 12 IBs using Winsteps 4.50 software (Winsteps, Beaverton, OR) using the Andrich single-rating scale model.[26] 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 invariant interval-level scale.[27] This allows for the transformation of ordinal categorical data into estimates of interval-level data, expressed in log of the odds units, or logits. A detailed description of the fit statistics and associated item modification processes has been published in previous item banking work by our group[17,18] but is outlined in brief below. Additionally, any psychometric amendments involving item deletion were reviewed and approved by the research team, comprised of members with content development and psychometric expertise (REKM, EF, JK, KP, EL) and/or clinical expertise (REKM, ZCW, SS, KP).

Category Probability Curves

A disordering of category probability curves (Supplementary Figs. S1–S12) may indicate an underutilization of certain response categories that can be resolved by collapsing adjacent categories as long as it is possible to combine the response options and other Rasch metrics are improved.[20]

Precision

Values of >2.0 and >0.8 are the minimum recommended values for person separation index (PSI) and person reliability (PR), respectively.[28] In our current analyses, participants with extreme scores (i.e., minimum or maximum) were removed; responses removed ranged from six (Emotional) to 169 (Social) (Supplementary Tables S1–S13). Extreme scores do not provide any information that can help with understanding how accurately an instrument is able to measure the QoL trait under assessment.[29]

Unidimensionality

We utilized principal component analysis (PCA) to assess the dimensionality of the IBs.[11] If multidimensionality was evident (eigenvalue $\geq$ 3), we checked the standardized residual loadings of the first contrast to determine if a cluster of loading items (>0.4) formed a conceptually relevant second dimension. We also checked the disattenuated correlations of each of these clusters; a disattenuated correlation value $\geq$ 0.8 with the primary domain indicates a high likelihood that the cluster is actually measuring a different strand of the same latent trait.[30]

Item Fit Statistics

We focused mainly on infit mean square (acceptable range, 0.5–1.5), as it is a greater threat to measurement than outfit.[31,32] To avoid unnecessary deletion of items, we explored the z-residuals of individual participant responses, with a score > |4| (i.e., >4 and <−4), indicating a high likelihood of an erroneous/unpredictable response. These were given a weightage of zero, with the process carried out iteratively until satisfactory item fit statistics were achieved. Item deletion was only considered if this process did not resolve item misfit and the expert panel did not deem the item important for face validity.

Local Item Dependency

A high correlation coefficient (>0.3) between item residuals of any two items is suggestive of local item dependency (LID).[33] To minimize the effects of LID on threshold calibrations, we first generated and then anchored LID-free person measures to all other person measures within a specific IB. This forces all item difficulties and rating-scale structures within the IB to conform with the LID-free person measures and prevents LID from impacting item difficulties.[17]

Targeting

Poor targeting occurs when there is a mismatch between respondents' ability levels and item difficulty levels or when items are clustered at particular difficulty levels leaving large gaps.[11] Targeting can be examined through visual inspection of the person–item map (Supplementary Figs. S13–S24) and calculated as the difference between the mean item difficulty and person ability, with a difference of >1.0 logits indicating notable mistargeting.[11]

Measurement Range

We determined the measurement range of each IB by calculating the difference in logits between highest and lowest item locations. The larger the measurement range, the more information about the measured construct provided by the items.

Differential Item Functioning

We utilized differential item functioning (DIF) to determine if item bias was present for sex and age group (median breakdown of <73 vs. ≥73 years). We took a DIF contrast of >1.0 logits with a corresponding $P < 0.05$ to indicate significant DIF; these items were subsequently deleted from the relevant IBs.[17]

Level of Dependence Between Different IBs

To assess the level of dependence between the final IBs and to ensure that they measured independent QoL constructs, we evaluated the Pearson correlation coefficient between each IB using individual person measures.[17,18]

CAT Simulations

A simulation of how a CAT algorithm performs is an important step in CAT applications as it allows developers to explore item selection and stopping rules and to determine whether the CAT algorithm is efficient before live testing.[34] We conducted CAT simulations in 1000 simulated respondents using Firestar-D software with the expected a posteriori (EAP) estimator and the maximum posterior weighted information item selection criteria.[35] In these simulations, we determined the average number of items required to achieve standard errors of measurement (SEMs) of 0.387 and 0.521 (approximating to a reliability of 0.85 and 0.72, respectively).[17] We also assessed the correlations between the IBs and CAT simulated person measure estimates for both levels of precision.

Results

Sociodemographic and Clinical Characteristics

Of the 293 participants recruited (mean age ± SD, 70.7 ± 13.2 years; 45% female), 161 had POAG (55%), 42 had angle closure glaucoma (14.3%), 40 had secondary glaucoma (13.7%), 24 had XFG (8.2%), and 12 had normal tension glaucoma (4.1%) (Table 1). Of the 247 participants with complete visual and sociodemographic data (84.3%), 165 had early disease (66.8%), 41 had moderate disease (16.6%), and 41 had advanced disease (16.6%), based on better-eye classifications. In terms of better-eye VA loss, 215 had no impairment (87.0%), 31 had mild–moderate loss (12.6%), and one had severe loss (0.4%).

Table 1. Participant Sociodemographic and Clinical Characteristics ($N = 293$)

Variable	Value
Age (y), mean (SD) ($N = 293$)	70.7 (13.2)
Female gender, *n* (%) ($N = 293$)	132 (45.0)
Glaucoma type ($N = 293$)	
POAG	161 (55.0)
PACG	42 (14.3)
Secondary	40 (13.7)
XFG	24 (8.2)
Normal tension	12 (4.1)
Marital status, *n* (%) ($N = 293$)	
Married	171 (58.3)
Single/divorced	122 (41.7)
Education, *n* (%) ($N = 293$)	
Primary	34 (11.6)
Secondary	165 (56.3)
Post-secondary	94 (32.1)
Employment status, *n* (%) ($N = 293$)	
Working	70 (23.9)
Unemployed/retired	223 (76.1)
Type of treatment, *n* (%) ($N = 293$)	
Medication	274 (93.6)
Laser	192 (65.5)
Surgery	167 (57.1)
Better-eye glaucoma severity,[a] *n* (%) ($N = 247$)	
Early	165 (66.8)
Moderate	41 (16.6)
Advanced	41 (16.6)
BCVA (logMAR), *n* (%) ($N = 247$)	0.06 (0.05)
Better-eye VA,[b] *n* (%) ($N = 247$)	
None	215 (87.0)
Mild–moderate	31 (12.6)
Severe	1 (0.4)

[a]Based on the Hodapp–Parrish–Anderson grading system.
[b]None, BCVA < 0.32; mild–moderate, 0.3 < BCVA ≤ 1.0; severe, BCVA > 1.0.

Psychometric Properties of the IBs

The initial psychometric properties, modifications made to fit Rasch model statistics, and final psychometric properties of the individual IBs are listed in Supplementary Tables S1 to S13 and summarized in the Figure. In brief, the DV, LT, MB, EM, and HC domains required relatively minor modifications to achieve satisfactory psychometric fit, most of which involved removal of those with extreme responses, giving unpredictable respondents a weightage of 0 (DV domain, $n = 2$ [Supplementary Table S4]; none needed

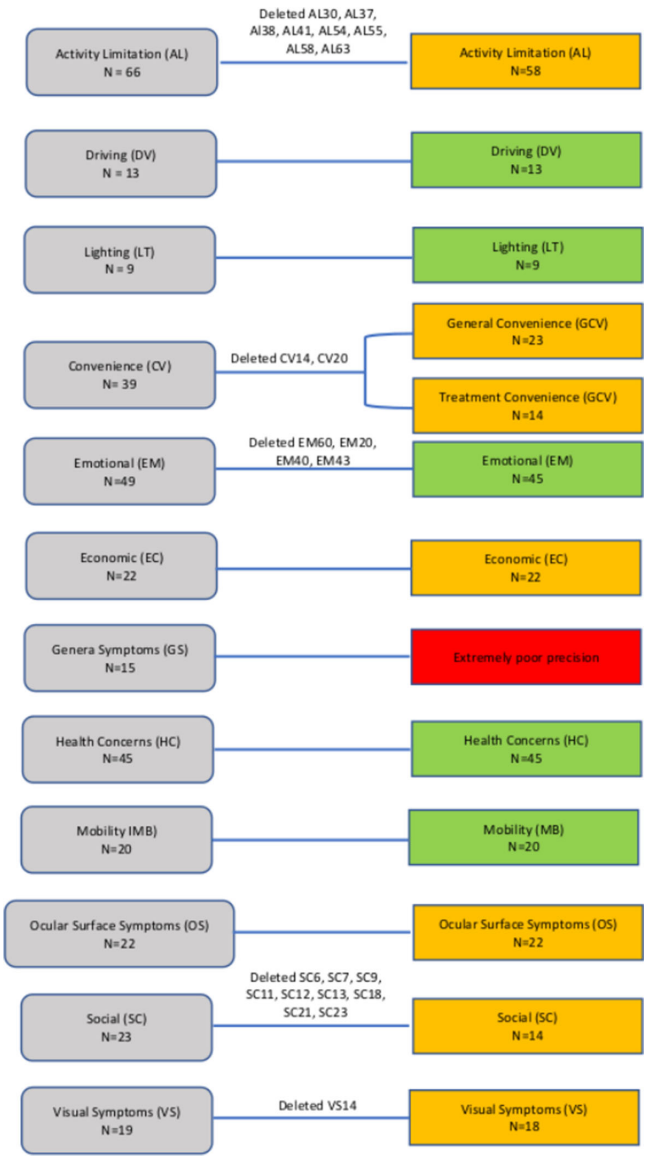

Figure. Summary of modifications made to the original 12 glaucoma item banks. *Green*, minor modifications (e.g., removal of extreme responses, misfitting items); *yellow*, more extensive modifications (i.e., minor modifications plus deletion of items with high missing data and discounting errant/unpredictable responses); *red*, unsolvable psychometric issues.

for the LT, MB, EM, and HC IBs) and removing very poorly performing items ($n = 4$ for the EM domain only) (Supplementary Table S7).

Initially, the SC, EC, and GS IBs demonstrated poor precision. Deletion of those with extreme responses, together with giving errant/unpredictable responders a weightage of zero ($n = 34$), resolved the poor precision for the EC IB (Supplementary Table S12). Precision for the SC IB improved after deletion of three items with high missing data (>30%) and six misfitting items, although the values remained somewhat suboptimal (PSI, 1.87; PR, 0.78) (Supplementary Table S11). Unfortunately, precision remained very poor for the GS domain even after extensive and iterative changes (PSI < 1.0; PR < 0.5) (Supplementary Table S13). As such, we did not to proceed with CAT simulations for this IB.

For the AL IB, adequate psychometric properties were obtained after deleting eight items with high missing data ($n > 50$) and giving a further 31 unpredictable respondents a weightage of zero (Supplementary Table S3). In contrast, the CV IB demonstrated multidimensionality, with 14 items related to treatment loading substantively (Supplementary Table S9). These items were split into a separate Treatment Convenience (TCV) IB. Two further items in the CV IB displayed gross and unsolvable fit statistics and were deleted, whereas two others with borderline infit values were retained. The modified 23-item CV IB was then renamed General Convenience (GCV) to better distinguish it from the TCV IB (Fig.).

The TCV IB displayed disordered thresholds and poor precision (Supplementary Table S10). The poor precision was resolved by assigning a weightage of zero to errant responders ($n = 23$). However, although collapsing underused response categories resolved disordering, it worsened other fit statistics; as such, we retained the five response categories.

Both the Visual Symptoms (VS) and Ocular Surface Symptoms (OS) domains showed poor precision, disordered thresholds, and poor item fit ($n = 2$ for VS [Supplementary Table S1]; $n = 3$ for OS [Supplementary Table S2]). These issues were remediated by removal of extreme responses, assigning a weightage of zero to errant respondents ($n = 3$ for VS; $n = 7$ for OS), removal of persons with high misfit (OS, $n = 14$), and removing one grossly misfitting item (VS). However, three items in OS still had infit values >1.5. After discussions with the research team and reviewing previous focus group discussion logs,[19] these three items were retained due to their perceived importance. Finally, although collapsing underused response categories for both IBs fixed the disordered thresholds, it worsened overall fit statistics, and, as such, the original response categories were retained.

We noted poor targeting of item difficulty to participant ability for all IBs, except for DV. In addition, higher than expected eigenvalues were noted for the final AL, EM, HC, and GCV IBs suggesting multidimensionality. However, an examination of the disattenuated correlations of the different item clusters within each bank revealed very high correlations (>0.9), indicating that they were most likely measuring strands of the same latent construct (Supplementary Tables S1–S12). Finally, correlation coefficients of the individual person measures between each IB were all <0.8,

Table 2. CAT Simulation Results for GlauCAT IBs

Domain	No. of Items Available for CAT	Average No. of Items Used by CAT	Correlation Between CAT and IB Person Measures	Mean SEM of CAT Administered Items
SEM = 0.387 (High Precision, for Use in Clinical Trials)				
VS	18	7.9	0.97	0.379
AL	58	5.5	0.94	0.368
MB	20	9.0	0.96	0.379
EM	45	6.7	0.94	0.368
HC	45	5.7	0.94	0.366
TCV	14	7.9	0.97	0.377
DV	13	7.1	0.97	0.375
LT	9	8.9	0.99	0.39
OS	22	7.9	0.96	0.376
GCV	23	6.6	0.95	0.374
SC	14	6.9	0.97	0.376
EC	22	7.8	0.96	0.374
Total	303	87.9[a]	0.96	0.375
SEM = 0.521 (Moderate Precision, for Use in Clinic Settings)				
VS	18	3.7	0.90	0.494
AL	58	3.0	0.88	0.475
MB	20	4.3	0.90	0.504
EM	45	3.3	0.89	0.495
HC	45	3.1	0.91	0.469
TCV	14	3.9	0.91	0.497
DV	13	3.4	0.92	0.493
LT	9	4.6	0.93	0.504
OS	22	3.9	0.90	0.495
GCV	23	3.3	0.88	0.486
SC	14	3.4	0.91	0.487
EC	22	4.1	0.89	0.479
Total	303	44[a]	0.90	0.490

[a]Total number of items required if all 12 item banks were administered using CAT.

Table 3. Correlation Coefficients Between the Final 12 Domains of the GlauCAT IBs

	AL	MB	EM	SC	VS	OS	DV	EC	LT	GCV	HC	TCV
AL	—	0.79	0.58	0.71	0.59	0.28	0.63	0.40	0.72	0.54	0.41	0.33
MB	—	—	0.54	0.62	0.48	0.15	0.63	0.35	0.64	0.52	0.39	0.34
EM	—	—	—	0.61	0.47	0.39	0.40	0.57	0.54	0.65	0.66	0.45
SC	—	—	—	—	0.36	0.19	0.51	0.40	0.61	0.63	0.52	0.39
VS	—	—	—	—	—	0.49	0.47	0.42	0.59	0.43	0.38	0.26
OS	—	—	—	—	—	—	0.15	0.47	0.38	0.41	0.50	0.44
DV	—	—	—	—	—	—	—	0.32	0.58	0.43	0.27	0.19
EC	—	—	—	—	—	—	—	—	0.40	0.62	0.61	0.56
LT	—	—	—	—	—	—	—	—	—	0.54	0.40	0.39
GCV	—	—	—	—	—	—	—	—	—	—	0.66	0.49
HC	—	—	—	—	—	—	—	—	—	—	—	0.44
TCV	—	—	—	—	—	—	—	—	—	—	—	—

supporting the independence of each of QoL domains assessed by the individual IBs (Table 2).

CAT Simulations

In CAT simulations (0.387 SEM, high precision), all IBs met this precision cutoff, except for LT, which was only able to achieve an average SEM of 0.39 with all nine items administered (Table 3). The average number of items administered per IB was 7.3 items, ranging between 5.5 items for AL and 9.0 items for MB and LT. Of the 303 total items within the 12 final IBs, 29.0% (87.9) were administered in order to reach the high precision benchmark. Correlations of the CAT simulated person measures with the IBs were high (0.94–0.99).

For moderate precision (0.521 SEM), all IBs met this precision cutoff. Forty-four of 303 items (14.5%) were administered to reach the stipulated precision, with an average item count of 3.7 per bank (range, 3.0 items for AL to 4.6 items for LT). Correlations between the CAT simulated person measures and the IBs were high (0.88–0.93).

Discussion

In this psychometric assessment of 12 initial glaucoma-specific QoL IBs (GlauCAT), extensive remediations, including deletion of poorly performing items, splitting scales to resolve multidimensionality, and dropping poorly performing domains, resulted in 12 final IBs, each measuring distinct QoL constructs. In CAT simulations, all IBs were able to achieve moderate precision with four or fewer items administered, and 11 of 12 were able to achieve high precision with eight or fewer items. These 12 final IBs will provide researchers and clinicians with a comprehensive understanding of the QoL impact of glaucoma, including relatively novel constructs such as DV, GCV, TCV, and HC. Moreover, the use of CAT to administer these IBs will provide considerable time-savings and reduction in participant burden. Importantly, as GlauCAT measures distinct and independent QoL constructs, researchers and clinicians can choose which domains are relevant to their research or clinical needs.

The CV IB displayed multidimensionality during psychometric testing that was resolved by further splitting the IB into individual subdomains assessing specific convenience traits: TCV for treatment-related convenience issues and GCV for other non–treatment-related convenience matters. This action resulted in two unidimensional IBs with good psychometric fit. The presence of multidimensionality in the original CV IB is not surprising, given that convenience issues related to treatment are very specific and distinct from convenience difficulties faced in other aspects of daily living. It is notable, though, that similar issues have not been observed with fixed-length questionnaires, and we suspect that this is because the high number of treatment-specific convenience items in our IB provides the magnification necessary to see how groups of items form into important traits relevant to people with disease undergoing treatment. As such, future IB work should consider separate banks to assess these treatment-related issues during initial conceptualization.

Despite iterative remedial steps, four IBs displayed suboptimal precision (VS, OS, SC, and GS), with GS demonstrating such poor performance that we opted not to carry out further analyses on this IB. The suboptimal performance of the GS domain may be related to the differential non-ocular symptomology among the different types of glaucomas (open-angle, angle-closure, and secondary glaucomas) present in our study population. For example, headaches are almost always associated with angle-closure rather than open-angle disease.[36] In contrast, because precision for the SC, VS, and OS IBs was close to the stipulated cut-off, we retained the scales due to perceived importance in the context of QoL changes consequent to glaucoma and its associated interventions based on a comprehensive content review of available glaucoma QoL questionnaires, qualitative participant responses,[20] and input from clinicians on our research team. The observed poor precision may be due to a lack of variance in our sample population, in particular the low number of persons with VA loss (mean VA, 0.06).[28]

In addition, we observed poor targeting of item difficulty to participant ability in all of our IBs except for DV. Specifically, the items appeared to be somewhat too "easy" relative to the average ability level of our participants, possibly related to the generally excellent VA of our clinical sample. This is expected as people with glaucoma suffer little loss of vision until the disease is quite advanced.[37,38] Poor targeting can be resolved by adding additional items of higher perceived difficulty level into the IBs, a task that is relatively simple to perform in item banking by estimating the calibration of new items relative to existing ones using Rasch analysis.[39]

Our promising simulation results demonstrate the viability of operationalizing IB administration using CAT, with this approach having distinct advantages over traditional paper-and-pencil questionnaire administration.[15,40] Only between six and nine items across 11 of the 12 IBs were required to achieve

high-precision QoL estimates; LT, comprised of only nine items, was the only bank not able to achieve the stipulated precision (0.387), although the final estimate was still relatively precise (0.39). Such brevity, together with automated scoring via CAT and the ability to pick and choose which IBs to administer, is highly valued in busy clinical settings, where clinicians might have little time to quantify patients' QoL scores. In view of the current push toward value-based clinical care,[9] we have developed an online CAT testing platform (PROMinsight) using Concerto open-source software that can be implemented on any smart device, enabling real-time scoring and reporting of data.[41] Indeed, a pilot study conducted in glaucoma clinics at Massachusetts Eye and Ear Institute, where 216 glaucoma patients completed six GlauCAT tests on a tablet device while waiting to see their treating doctor,[42,43] showed that patients took just 8 minutes and 5 seconds (median) to complete all six GlauCAT tests,[43] providing real-world evidence of the feasibility of implementing GlauCAT in routine clinical care.

Strengths of our study include a rigorous, comprehensive, and validated framework within which the IBs were developed and psychometrically assessed,[17,18] with the pragmatic efforts to rehome groups of items contributing to multidimensionality, rather than resorting to unnecessary deletion of items. We also purposively recruited participants across a range of glaucoma subtypes and severities in order to capture responses across the spectrum of participant ability levels. However, a few limitations should also be noted. Our cutoff for detecting LID was more lenient (0.3 rather than 0.2), and we may have missed noteworthy LID as a result, thus artificially inflating reliability and precision estimates. Similarly, we used a conservative cutoff for detecting DIF (>1.0) and therefore may have missed accounting for moderate to large DIF for some items. Moreover, some important clinical variables, such as VFD and VA, were missing from participant case notes, which then precluded us from evaluating potential DIF in these variables. Finally, our sample did not include many patients with primary angle-closure glaucoma (PACG), which may mean our calibrations are less robust in Asian populations, where the prevalence of PACG is higher (0.7%) compared to that in white populations (0.2%–0.4%).[44] Our group is currently developing and validating a glaucoma CAT system among multiethnic Singaporeans with glaucoma to address this issue.

In conclusion, our 12 GlauCAT IBs demonstrate adequate psychometric properties that enable a comprehensive and novel assessment of glaucoma-specific QoL. CAT simulations revealed the potential for highly precise QoL measurements with only a fraction of total items within each bank required. Several implementation trials are currently ongoing and will be informative in evaluating the performance of GlauCAT in real-world and clinical trial settings.

Acknowledgments

Supported by project grants from the National Health and Medical Research Council (NHMRC) (1031838 and 1009844 to Z.W.) and by a NHMRC fellowship (1104985 to Z.W.); by a National Medical Research Council Senior-Clinician Scientist Award (NMRC/CSASI/0009/2016 to E.L.L.); and by a NMRC Transition Award (MOH-TA19may-0002 to R.E.K.M.). The Centre for Eye Research Australia receives Operational Infrastructure Support from the Victorian Government. The funding bodies had no role in the design and conduct of this research.

Disclosure: **R.E.K. Man**, None; **E.K. Fenwick**, None; **J. Khadka**, None; **Z. Wu**, None; **S. Skalicky**, None; **K. Pesudovs**, None; **E.L. Lamoureux**, None

References

1. Tham YC, Li X, Wong TY, Quigley HA, Aung T, Cheng CY. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*. 2014;121(11):2081–2090.
2. Fenwick EK, Man RE, Aung T, Ramulu P, Lamoureux EL. Beyond intraocular pressure: optimizing patient-reported outcomes in glaucoma. *Prog Retin Eye Res*. 2020;76:100801.
3. Wormald R. Treatment of raised intraocular pressure and prevention of glaucoma. *BMJ*. 2003;326(7392):723–724.
4. Sihota R, Angmo D, Ramaswamy D, Dada T. Simplifying "target" intraocular pressure for different stages of primary open-angle glaucoma and primary angle-closure glaucoma. *Indian J Ophthalmol*. 2018;66(4):495–505.
5. Fiscella RG, Green A, Patuszynski DH, Wilensky J. Medical therapy cost considerations for glaucoma. *Am. J Ophthalmol*. 2003;136(1):18–25.
6. Cantor LB, Katz LJ, Cheng JW, Chen E, Tong KB, Peabody JW. Economic evaluation of medication, laser trabeculoplasty and filtering surgeries in treating patients with glaucoma in the US. *Curr Med Res Opin*. 2008;24(10):2905–2918.
7. Han JA, Frishman WH, Wu Sun S, Palmiero PM, Petrillo R. Cardiovascular and respiratory

considerations with pharmacotherapy of glaucoma and ocular hypertension. *Cardiol Rev*. 2008;16(2):95–108.
8. Servat JJ, Bernardino CR. Effects of common topical antiglaucoma medications on the ocular surface, eyelids and periorbital tissue. *Drugs Aging*. 2011;28(4):267–282.
9. Basch E. Patient-reported outcomes - harnessing patients' voices to improve clinical care. *N Engl J Med*. 2017;376(2):105–108.
10. Snyder CF, Aaronson NK, Choucair AK, et al. Implementing patient-reported outcomes assessment in clinical practice: a review of the options and considerations. *Qual Life Res*. 2012;21(8):1305–1314.
11. Pesudovs K, Burr JM, Harley C, Elliott DB. The development, assessment, and selection of questionnaires. *Optom Vis Sci*. 2007;84(8):663–674.
12. Massof R. Likert and Guttman scaling of visual function rating scale questionnaires. *Ophthalmic Epidemiol*. 2004;11(5):381–399.
13. Massof R. The measurement of vision disability. *Optom Vis Sci*. 2002;79(8):516–552.
14. Khadka J, McAlinden C, Pesudovs K. Quality assessment of ophthalmic questionnaires: review and recommendations. *Optom Vis Sci*. 2013;90(8):720–744.
15. Cella D, Gershon R, Lai JS, Choi S. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. *Qual Life Res*. 2007;16(suppl 1):133–141.
16. Gershon RC. Computer adaptive testing. *J Appl Meas*. 2005;6(1):109–127.
17. Fenwick EK, Khadka J, Pesudovs K, Rees G, Wong TY, Lamoureux EL. Diabetic retinopathy and macular edema quality-of-life item banks: development and initial evaluation using computerized adaptive testing. *Invest Ophthalmol Vis Sci*. 2017;58(14):6379.
18. Kandel H, Khadka J, Watson SL, Fenwick EK, Pesudovs K. Item banks for measurement of refractive error-specific quality of life. *Ophthalmic Physiol Opt*. 2021;41(3):591–602.
19. Khadka J, McAlinden C, Craig JE, Fenwick EK, Lamoureux EL, Pesudovs K. Identifying content for the glaucoma-specific item bank to measure quality-of-life parameters. *J Glaucoma*. 2015;24(1):12–19.
20. Khadka J, Fenwick EK, Lamoureux EL, Pesudovs K. Item banking enables stand-alone measurement of driving ability. *Optom Vis Sci*. 2016;93(12):1502–1512.
21. Ang GS, Fenwick EK, Constantinou M, et al. Selective laser trabeculoplasty versus topical medication as initial glaucoma treatment: the glaucoma initial treatment study randomised clinical trial. *Br J Ophthalmol*. 2020;104(6):813–821.
22. Vogl S. Telephone versus face-to-face interviews: mode effect on semistructured interviews with children. *Sociol Methodol*. 2013;43(1):133–177.
23. Novick G. Is there a bias against telephone interviews in qualitative research? *Res Nurs Health*. 2008;31(4):391–398.
24. Fenig S, Levav I, Kohn R, Yelin N. Telephone vs face-to-face interviewing in a community psychiatric survey. *Am J Public Health*. 1993;83(6):896–898.
25. Hodapp E, Parrish R, II, Anderson DR. *Clinical Decisions in Glaucoma*. St. Louis, MO: Mosby; 1993.
26. Linacre JM. *A User's Guide to Winsteps: Rasch–Model Computer Program.* San Diego, CA: MESA Press; 2002.
27. Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *Br J Clin Psychol*. 2007;46(1):1–18.
28. Bond TG, Fox CM. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Hove, UK: Psychology Press; 2013.
29. Boone WJ, Staver JR, Yale MS. *Rasch Analysis in the Human Sciences*. Dordrecht: Springer; 2014.
30. Schumacker R, Muchinsky P. Disattenuating correlation coefficients. *Rasch Measure Trans*. 1996;10(1):479.
31. Linacre JM. What do infit and outfit, mean-square and standardized mean. *Rasch Measure Trans*. 2002;16(2):878.
32. Pesudovs K. Item banking: a generational change in patient-reported outcome measurement. *Optom Vis Sci*. 2010;87(4):285–293.
33. Baghaei P. Local dependency and Rasch measures. *Rasch Measure Trans*. 2008;21(3):1105–1106.
34. Chen S-K, Cook KF. simpolycat: an SAS program for conducting CAT simulation based on polytomous IRT models. *Behav Res Methods*. 2009;41(2):499–506.
35. Choi SW. Firestar: computerized adaptive testing simulation program for polytomous item response theory models. *Appl Psychol Measure*. 2009;33(8):644–645.
36. Shindler KS, Sankar PS, Volpe NJ, Piltz-Seymour JR. Intermittent headaches as the presenting sign of subacute angle-closure glaucoma. *Neurology*. 2005;65(5):757–758.
37. Ang GS, Eke T. Lifetime visual prognosis for patients with primary open-angle glaucoma. *Eye (Lond)*. 2007;21(5):604–608.

38. Kwon YH, Kim CS, Zimmerman MB, Alward WL, Hayreh SS. Rate of visual field loss and long-term visual outcome in primary open-angle glaucoma. *Am J Ophthalmol*. 2001;132(1):47–56.
39. Haley SM, Ni P, Jette AM, et al. Replenishing a computerized adaptive test of patient-reported daily activity functioning. *Qual Life Res*. 2009;18(4):461–471.
40. Gibbons RD, Weiss DJ, Kupfer DJ, et al. Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatr Serv*. 2008;59(4):361–368.
41. Harrison C, Loe BS, Lis P, Sidey-Gibbons C. Maximizing the potential of patient-reported assessments by using the open-source concerto platform with computerized adaptive testing and machine learning. *J Med Internet Res*. 2020;22(10):e20950.
42. Halawa O, Roldan AM, Meshkin RS, et al. Factors associated with glaucoma-specific quality of life in a US glaucoma clinic using an online computerized adaptive test (GlauCAT) [published online ahead of print May 12, 2022]. *Br J Ophthalmol*, https://doi.org/10.1136/bjophthalmol-2022-321145.
43. Fenwick EK, Roldan AM, Halawa OA, et al. Implementation of an online glaucoma-specific quality of life computerized adaptive test system in a US glaucoma hospital. *Transl Vis Sci Technol*. 2022;11(2):24.
44. Zhang N, Wang J, Chen B, Li Y, Jiang B. Prevalence of primary angle closure glaucoma in the last 20 years: a meta-analysis and systematic review. *Front Med (Lausanne)*. 2021;7:624179.