tvst

Article

Computerized Adaptive Tests: Efficient and Precise Assessment of the Patient-Centered Impact of Diabetic Retinopathy

Eva K. Fenwick[1–3], John Barnard[4,11], Alfred Gan[1], Bao Sheng Loe[5], Jyoti Khadka[6–8], Konrad Pesudovs[9,10], Ryan Man[1,2], Shu Yen Lee[1], Gavin Tan[1], Tien Y. Wong[1,2], and Ecosse L. Lamoureux[1–3]

[1] Singapore Eye Research Institute, Singapore National Eye Centre, Singapore
[2] Duke–NUS Medical School, Singapore
[3] Centre for Eye Research Australia, University of Melbourne, Melbourne, Australia
[4] Excel Psychological & Educational Consultancy, Melbourne, Australia
[5] The Psychometrics Centre, University of Cambridge, Cambridge, UK
[6] Institute for Choice, University of South Australia, Adelaide, Australia
[7] Registry of Older South Australians, South Australian Health and Medical Research Institute, Adelaide, Australia
[8] Health and Social Care Economics Group, College of Nursing and Health Sciences, Flinders University, Adelaide, Australia
[9] University of New South Wales, Sydney, Australia
[10] Anglia Ruskin University, Cambridge, UK
[11] School of Medical Sciences, University of Sydney, Sydney, Australia

Correspondence: Ecosse L. Lamoureux, Singapore Eye Research Institute, 20 College Rd, Level 6, 169856, Singapore. e-mail: ecosse.lamoureux@seri.com.sg

Received: June 26, 2019
Accepted: February 21, 2020
Published: June 3, 2020

Keywords: diabetic retinopathy; quality of life; item banks; computerized adaptive testing; vision impairment

Citation: Fenwick EK, Barnard J, Gan A, Loe BS, Khadka J, Pesudovs K, Man R, Lee SY, Tan G, Wong TY, Lamoureux EL. Computerized adaptive tests: efficient and precise assessment of the patient-centered impact of diabetic retinopathy. Trans Vis Sci Tech. 2020;9(7):3, https://doi.org/10.1167/tvst.9.7.3

Purpose: Evaluate efficiency, precision, and validity of RetCAT, which comprises ten diabetic retinopathy (DR) quality of life (QoL) computerized adaptive tests (CATs).

Methods: In this cross-sectional clinical study, 183 English and/or Mandarin-speaking participants with DR (mean age ± standard deviation [SD] 56.4 ± 11.9 years; 38% proliferative DR [worse eye]) were recruited from retinal clinics in Singapore. Participants answered the RetCAT tests (Symptoms, Activity Limitation, Mobility, Emotional, Health Concerns, Social, Convenience, Economic, Driving, and Lighting), which were capped at seven items each, and other questionnaires, and underwent eye tests. Our primary evaluation focused on RetCAT efficiency (i.e. standard error of measurement [SEM] ± SD achieved and time needed to complete each CAT). Secondary evaluations included an assessment of RetCAT's test precision and validity.

Results: Mean SEM across all RetCAT tests was 0.351, ranging from 0.272 ± 0.130 for Economic to 0.484 ± 0.130 for Emotional. Four tests (Mobility, Social, Convenience, and Driving) had a high level of measurement error. The median time to take each RetCAT test was 1.79 minutes, ranging from 1.12 (IQR [interquartile range] 1.63) for Driving to 3.28 (IQR 2.52) for Activity Limitation. Test precision was highest for participants at the most impaired end of the spectrum. Most RetCAT tests displayed expected correlations with other scales (convergent/divergent validity) and were sensitive to DR and/or vision impairment severity levels (criterion validity).

Conclusions: RetCAT can provide efficient, precise, and valid measurement of DR-related QoL impact. Future application of RetCAT will employ a stopping rule based on SE rather than number of items to ensure that all tests can detect meaningful differences in person abilities. Responsiveness of RetCAT to treatment interventions must also be determined.

Translational Relevance: RetCAT may be useful for measuring the patient-centered impact of DR severity and disease progression and evaluating the effectiveness of new therapies.

Introduction

Diabetic retinopathy (DR) is a potentially sight-threatening microvascular complication of diabetes[1] that can have a detrimental impact on patients' visual functioning and socioemotional well-being.[2–4] Measuring the impact of disease and treatment effectiveness from the patient's perspective using patient-reported outcome measures (PROMs) is now mandated by decision-makers such as the Food and Drug Administration.[5] However, there are currently no DR-specific PROMs that measure the impact of the disease across the spectrum of quality of life (QoL).[6] Moreover, currently available PROMs in ophthalmology are paper- and pencil-based, which means they are inflexible (the number and order of items are fixed) and burdensome to administer (many questionnaires comprise >20 items, and patients have to answer every question).[7]

These limitations can be overcome by the use of item banking and computerized adaptive testing (CAT) systems.[8] An item bank is a pool of items (questions) measuring a latent construct, such as "Activity Limitation," that is (usually) calibrated using item response theory (IRT).[9] The items are administered from the bank using CAT algorithms, which customize the test for each test-taker by offering items that are most informative for the respondent at that point in the test.[10] The CAT selects each item according to the test-taker's previous responses and stops administering items when the stopping criterion (e.g., precision level or maximum number of items) is reached. Because items are targeted to the test-taker's level of the construct, test length can be minimized without loss of precision, making CAT tests more efficient than paper–pencil questionnaires. Moreover, with automated scoring and real-time feedback, CATs are ideal for use in clinical and research settings.[11]

In previously published work, we developed and psychometrically tested item banks to measure the impact of DR across ten domains of QoL,[12–14] and based on these promising findings, we subsequently developed ten final CATs. The aim of the current study is to evaluate the performance of our ten DR-QoL CATs—"RetCAT"—in a clinical sample of patients across the severity spectrum of DR, following the approach outlined in previous similar studies in other health fields.[15–19] Our primary evaluation includes a practical assessment of test efficiency (i.e., standard error of measurement [SEM] achieved and time needed to complete each CAT). Secondary evaluations include (1) a psychometric evaluation, including content range coverage, item exposure rate (IER), and test precision; and (2) a validity assessment of the score estimates derived by each CAT using classical test theory (CTT) methods.

Methods

Study Design and Participants

Participants in our cross-sectional study were consecutively recruited from retinal clinics at the Singapore National Eye Centre (SNEC) between December 2016 and June 2018. English- and/or Mandarin-speaking participants aged ≥21 years of Chinese, Malay, or Indian ethnicity with a primary diagnosis of DR and type 2 diabetes were included in the study. Those with significant hearing or cognitive impairment (measured by the 6-item Cognitive Impairment Test [6-CIT]),[20] physical disability excluding them from participating in the study protocol, and/or other ocular comorbidity affecting visual functioning (e.g., age-related macular degeneration, glaucoma, or late-stage cataract) were ineligible. For our convenience sample, we implemented a purposive recruitment strategy whereby we aimed to recruit approximately 60% Chinese (English- or Mandarin-speaking), 20% Malay, and 20% Indian participants, reflecting the ethnic split within Singapore. We also aimed to recruit patients across the spectrum of DR severity, according to the following allocations: 20% each mild and moderate nonproliferative DR (NPDR) and 60% severe NPDR and proliferative DR (PDR).

Participants underwent a standardized testing protocol conducted in either English (n = 131, 71.6%) or Mandarin (n = 52, 28.4%), including collection of clinical, sociodemographic, and other questionnaire data, at the Singapore Eye Research Institute clinic in SNEC. The study had ethical approval from the Singapore Eye Research Institutional Review Board (#2016/2763) and all participants provided written informed consent. The study was conducted in accordance with the Declaration of Helsinki.

DR QoL Item Banks

The development and psychometric assessment of our DR-QoL CATs have been described in detail previously. [12–14] In brief, domains and items were developed from extant vision-related questionnaires, published qualitative literature, focus groups, and semi-structured interviews with clinical experts and 57 patients with DR.[12] Domains and items were subsequently revised using a process of winnowing and binning, after which there were 314 items spread across nine QoL

domains.[13] Following in-depth psychometric testing using Rasch analysis with Winsteps software, version 3.91.2 (Winsteps, Chicago, IL),[21] the final number of items was 252 spread across eight QoL domains: Visual Symptoms (n = 18), Activity Limitation (n = 92), Mobility (n = 17), Emotional (n = 45), Health Concerns (n = 35), Convenience (n = 20), Driving (n = 15), and Lighting (n = 10).[14] Three domains—Ocular Surface Symptoms (n = 10), Social (n = 21), and Economic (n = 12)—failed to reach optimal fit to the Rasch model and were temporarily set aside. Subsequent work to optimize the psychometric properties of these problematic scales resulted in two of the three domains—Social (n = 20) and Economic (n = 15)—reaching adequate fit to the Rasch model. Therefore, in this study, we report CAT evaluation results for ten QoL domains, comprising a total of 287 items.

Linguistic and Cultural Adaptation of the Item Banks

Before development of the CATs, items were linguistically and culturally adapted into local parlance via consultation with an expert panel (Supplementary Table S1). Following an iterative process, a total of 75 items (30%) underwent some level of modification Most were minor (e.g., SocialQ16 "Meeting a partner" changed to "Looking for a partner"), while some were more substantial (e.g., Visual SymptomsQ17 "Difficulty distinguishing contrast" changed to "Difficulty telling the difference between similar tones and shades"). Following cultural adaptation, the item banks were professionally translated and back-translated into Mandarin. As appropriate translations were not possible for three items in the Emotional item bank, these items were excluded from the Mandarin Emotional CAT, leaving a total of 42 items available for administration.

Development of CAT

CATs for each domain were developed by Excel Psychological & Educational Consultancy. Using the known Rasch difficulty estimates of each category within each question, Monte Carlo simulations were used to generate abilities for cohorts of 1000 hypothetical test-takers.[22] No constraints were placed on exposure or content, as it was assumed that each domain was unidimensional with locally independent items. To minimize idiosyncrasies in the simulations, different random seeding was used in a number of replications of the same and different requirements. For each domain an initial simulation was based on normal (N ~0,1) distributions with abilities in the interval (–3,3) logits. No restrictions on the number of items were initially set and the precision in terms of the standard error (SE) of each ability estimate was stepwise reduced as SE $\leq$0.50, SE $\leq$0.40, SE $\leq$0.35, SE $\leq$0.30, and SE $\leq$0.25. Positively and negatively skewed ability estimate distributions were then explored. As simulations suggested that most domains achieved SE $\leq$0.35 with seven items, RetCAT was capped to administer seven questions from each domain; that is, 70 questions overall.

Assessment of DR and Visual Acuity and Related Definitions

Digital retinal photographs of two fields (macula and optic disc) were obtained in both eyes. DR was graded according to the modified Airlie House classification system for the Early Treatment Diabetic Retinopathy Study[23] as level 10 ("no DR"), levels 14 and 15 ("questionable DR," hemorrhage present, without any definite microaneurysm [MA]), level 20 ("minimal DR," MA only, with no other retinopathy lesions present), level 35 ("mild NPDR," MA and one or more hemorrhage or MA standard photograph 2A, hard exudates, venous loops, questionable cotton wool spot [CWS], intraretinal microvascular abnormality [IRMA], or venous beading), levels 43–47 ("moderate NPDR," MA and one or more CWS, IRMA standard photograph 8A), level 53 ("severe NPDR," MA and one or more venous beading, hemorrhage or MA 2A, IRMA 8A), levels 61–64 ("mild PDR," scatter laser photocoagulation scars, with retinopathy levels of 31–51), level 65 ("moderate PDR," PDR less than high-risk characteristics, as defined in the Diabetic Retinopathy Study), level 71 ("severe PDR," PDR with high-risk characteristics), levels 81 and 85 ("advanced PDR," fundus partially obscured or retina detached, total vitreous hemorrhage), or level 90 ("inactive PDR," laser scars and/or fibrous proliferation present but new vessels absent).

Presenting distance visual acuity (PDVA) was measured in the left, right, and both eyes using a logarithm of the minimum angle of resolution (LogMAR) number chart (Lighthouse International, New York) at a distance of 4 m with habitual correction (if any). If no numbers could be read at 4 m, the participant was moved to 3, 2, or 1 m or assessed as counting fingers, hand movements, perception of light, or no light perception, as required. If PDVA was >0.30 log units (<6/12 Snellen), pinhole was performed.

Other Measures

Sociodemographic, medical and ocular history, and other questionnaire data were collected by trained interviewers during face-to-face interviews. Questionnaires included the Impact of Vision Impairment (IVI) profile,[24,25] the Quality of Vision (QoV) questionnaire,[26] and the Generalized Self-Efficacy Scale (GSES).[27] The 28-item IVI is a vision-related QoL scale comprised of three independently scored scales; namely, Reading and Accessing Information ("Reading"), Mobility and Independence ("Mobility"), and Emotional Well-Being ("Emotional"). Higher scores indicate better VRQoL outcomes. The 30-item QoV questionnaire[26] assesses ten symptoms (e.g., glare, blurred vision, distortion), rated on a 4-point scale for frequency, severity, and degree of annoyance. The frequency scale was used in this study. Higher scores represent greater frequency of visual symptoms; scores were reversed during Rasch analysis. The 10-item GSES is designed to assess optimistic self-beliefs to cope with difficult demands in life. Higher scores indicate better self-efficacy. The IVI, QoV, and GSES were analyzed using Rasch analysis with Winsteps software, version 4.2.0 (Winsteps),[28] and the Andrich rating scale model.[21]

Data Analyses

Sociodemographic and clinical characteristics of the study population were examined using proportions, means, medians, percentiles, and standard deviation (SD) and computed using Stata version 14 (StataCorp, College Station, TX). Our primary goal was to assess the efficiency of the ten CATs, defined as mean SEM and time taken (in minutes) to complete each CAT. As Emotional scores were significantly lower for those who answered in Mandarin compared with English (β –1.10 [confidence Interval] CI –1.30 to –0.89, $P < 0.001$), independent of age, gender, DR severity, and visual impairment (VI), we report results for the Emotional test separately by language.

As a secondary evaluation, we explored test precision and IER. We used the test information function (TIF) to examine test precision. TIF is calculated by summing the information provided by all individual items in the bank and identifies where the test has the highest/lowest measurement precision. The TIF curve peak indicates the range of the trait best measured by that instrument. Therefore, TIF values are related to the calculation of the SE of the person ability estimates by the formula SE $1/\sqrt{\text{TIF}}$.[29] The average SE of estimates for people was calculated at four different score ranges (by centering the person measures to have a mean of 3.0) to determine the precision of each CAT score at different participant levels of each construct. CIs of the estimates were generated by multiplying the SE by a z score corresponding to certain CIs. The IER identifies which items are administered most often in each CAT test and is influenced by item difficulty, the distribution of patients' levels of each construct, and whether there are similar items in the item bank.[30] We assessed the proportion of items administered overall and ≥50% of the time.

Finally, we assessed the validity of RetCAT using CTT methods. For convergent validity, we correlated CAT scores with scores from the QoV questionnaire and the Reading, Mobility, and Emotional IVI scales using Pearson's correlation coefficient. Correlations were chosen based on a hypothesized moderate ($0.3 < r \leq 0.70$)[31] relationship between scores (e.g., Emotional CAT was correlated with Emotional IVI). For divergent validity, we correlated all ten CATs with the GSES, as we expected little to no relationship ($r < 0.3$). For criterion validity, we compared Student's t-test CAT scores across minimal to mild NPDR, moderate to severe NPDR, and PDR as well as three levels of binocular VI: none (LogMAR <0.3), mild (≥0.3 to LogMAR ≤0.60), and moderate to severe (LogMAR >0.60). P trend was calculated using a Wald test of the beta coefficient after performing a linear regression of each CAT score against DR severity and binocular VI as continuous variables. As 105 and 104 did not answer Economic and Driving, respectively, we did not assess criterion validity for these scales.

Results

Sociodemographic and Clinical Characteristics

A total of 183 participants (mean age ± SD, 56.4 ± 11.9 years; 61% male; 66% Chinese) answered RetCAT (Table 1). Mean ± SD duration of diabetes was 17.2 ± 15.0 years and 80 (44%) participants were on insulin. Of the 183 participants, 45 (26.8%), 58 (34.5%), and 64 (38%) had minimal to mild NPDR, moderate to severe NPDR, and PDR in the worse eye, respectively. Participants' mean ± SD binocular presenting distance visual acuity was 0.21 ± 0.20 LogMAR (Table 1).

Evaluation of RetCAT

CAT Efficiency

The mean SEM for RetCAT was 0.351, with values ranging from 0.272 ± 0.130 for Economic to

Table 1. Sociodemographic and Clinical Characteristics of Participants (N = 183)[a]

Variable	n (%)
Insulin use (yes)	80 (43.7)
Male	112 (61.2)
Ethnicity	
Chinese	120 (65.6)
Indian	28 (15.3)
Malay	35 (19.1)
Language of assessment	
English	131 (71.6)
Mandarin	52 (28.4)
Marital status	
Never married	28 (15.3)
Married/de facto	131 (71.6)
Divorced/separated/widowed	24 (13.1)
Education level	
None or primary level	35 (19.3)
Secondary level	79 (43.7)
A level, polytechnic diploma, vocational training	45 (24.9)
Undergraduate or postgraduate university degree	22 (12.2)
Main language spoken at home	
English	62 (33.9)
Mandarin	49 (26.8)
Malay	27 (14.8)
Tamil	12 (6.6)
Other[b]	33 (18.0)
Employment status	
Working	78 (42.6)
Not working	105 (57.4)
Monthly household Income (SGD)	
<$1000	28 (15.4)
$1000 to <$2000	27 (14.8)
$2000 to <$5000	41 (22.5)
≥$5000	39 (21.4)
Housing type	
HDB 1–2 rooms	15 (8.2)
HDB 3 rooms	40 (21.9)
HDB 4 rooms	69 (37.7)
HDB 5 rooms, executive flat	42 (23.0)
Condo, private apartment, landed property	17 (9.3)
How many people in dwelling	
1	46 (25.1)
2	94 (51.4)
>2	43 (23.5)

Table 1. Continued

Variable	n (%)
Other diabetes complications (self-reported)[c]	
None	121 (66.1)
1	39 (21.3)
>1	23 (12.6)
Comorbidities (self-reported)[d]	
None	29 (15.9)
1	39 (21.3)
>1	115 (62.8)
Self-rated health	
Very good to excellent	20 (10.9)
Good	67 (36.6)
Fair	72 (39.3)
Poor	24 (13.1)
History of other eye diseases (self-reported)	
Glaucoma	1 (0.6)
Age-related macular degeneration	1 (0.6)
Cataract	120 (65.6)
Other	22 (12.0)
Severity of DR (worse eye)	
Minimal NPDR (level 20)	9 (5.2)
Mild NPDR (level 35)	36 (20.8)
Moderate NPDR (level 43)	38 (22.0)
Moderately severe NPDR (level 47)	16 (9.3)
Severe NPDR (level 53)	4 (2.3)
Mild PDR (level 61)	4 (2.3)
Moderate PDR (level 65)	6 (3.5)
Severe PDR (level 71)	15 (8.7)
Advanced PDR (levels 81 and 85)	2 (1.2)
Inactive PDR (level 90)	37 (21.4)
Severity of DME (worse eye)	
No DME	86 (49.7)
DME (not CSME)	24 (13.9)
CSME	63 (36.4)
Severity of binocular vision impairment	
None	142 (77.6)
Mild	27 (14.8)
Moderate to severe	14 (7.7)
Continuous variables	Mean (SD); median (IQR)
Age (years)	56.36 (11.92); 58 (17)
Duration of diabetes (years)	17.23 (10.65); 17 (15)
Presenting distance binocular visual acuity (LogMAR)	0.21 (0.20); 0.18 (0.22)

CSME, clinically significant macular edema; DME, diabetic macular edema; HDB, housing development board; IQR, interquartile range; LogMAR, logarithm of the minimum angle of resolution; NPDR, nonproliferative diabetic retinopathy; PDR, proliferative diabetic retinopathy; SGD, Singapore dollar.

[a]Percentages for some variables may not equal 100% because of missing data.

[b]Includes Chinese dialects, such as Hokkein, Teochew, Hakka, Hainanese, and Cantonese.

[c]Includes diabetic coma, severe hypoglycemia, kidney disease, nerve damage, oral health problems, gangrene, foot ulcers, and impotence.

[d]Includes hypertension, angina or heart attack, irregular heartbeat, stroke, dyslipidemia, asthma, anemia, migraine, arthritis, and osteoporosis.

Table 2. CAT Results for Ten Diabetic Retinopathy CATs

Domain	Items Available for CAT	Mean (SD) Score	Mean (SD) SEM	Minimum Score	Maximum Score	Time Taken (Mins), Median (IQR)
Visual Symptoms	18	1.50 (0.55)	0.460 (0.13)	−0.94	1.91	1.48 (1.01)
Activity Limitation	84	1.07 (0.77)	0.315 (0.120)	−0.85	2.08	3.28 (2.52)
Mobility	17	2.30 (0.23)	0.360 (0.060)	0.73	2.59	1.22 (0.99)
Emotional	45	1.75 (0.49)[a]	0.390 (0.79)[a]	−0.76[a]	−1.29[a]	1.23 (1.22)[a]
		0.65 (0.83)[b]	0.484 (0.13)[b]	2.15[b]	2.15[b]	1.31 (0.89)[b]
Health Concerns	35	0.71 (0.69)	0.290 (0.110)	−1.43	1.60	1.78 (1.29)
Social	20	0.90 (0.21)	0.380 (0.100)	0.33	1.06	1.57 (1.13)
Convenience	20	1.13 (0.25)	0.351 (0.100)	0.40	1.43	2.12 (1.83)
Economic[c]	15	0.59 (0.73)	0.272 (0.130)	0.73	1.69	2.11 (1.63)
Driving[d]	15	1.60 (0.33)	0.350 (0.090)	0.65	1.86	1.12 (1.63)
Lighting	10	0.17 (0.36)	0.310 (0.130)	−0.78	0.76	1.95 (1.53)
Total	*252*		*0.351*			*1.79 mins*

[a]English speakers.
[b]Mandarin speakers.
[c]n = 105 did not answer this domain because they did not work for reasons other than their DR or vision.
[d]n = 104 did not answer this domain because they did not drive for reasons other than their DR or vision.

0.484 ± 0.130 for Emotional-Mandarin (Table 2). Mean SEM was lower for Emotional in English speakers compared with Mandarin speakers (0.390 vs. 0.484, respectively). For some CATs—namely, Mobility, Social, Convenience, and Driving—the average SE exceeded the observed SD (Table 2), suggesting that intrinsic measurement error was high for these CATs.[32] The median time to answer each RetCAT test was 1 minute 47 seconds (range, 1 minute 7 seconds for Driving to 3 minutes 17 seconds for Activity Limitation).

Psychometric Evaluation

Test Precision. Test precision (represented by TIF) of RetCAT was excellent (Supplementary Table S2), especially for the larger item banks, such as Activity Limitation (TIF = 48.02) and Emotional (TIF = 28.37). Smaller item banks, such as Lighting, had comparatively lower test precision (TIF = 6.01). As seen in the Figure, the TIF curve for the entire Activity Limitation test pool (n = 92 items) peaked at 0.02 logits on the ability scale, at which point the SE was lowest (and the test most precise) for participants. Test information decreased substantially and SE increased at the extreme ends of the spectrum (–4,4 logits). A similar pattern was observed for the other nine RetCAT tests (Supplementary Figure S1). When we categorized participants' scores into different bins across the ability spectrum, scores in the lowest two bins (<2.0 and 2.0 to <3.0) were unequivocally the most precisely estimated (Table 3). As scores moved into the highest two bins (3.0 to <3.5 and ≥3.5), precision levels decreased. For example, for Activity Limitation, the most and least precisely estimated score ranges were <2.0 (0.185 ± 0.007) and ≥3.5 (0.443 ± 0.020), respectively.

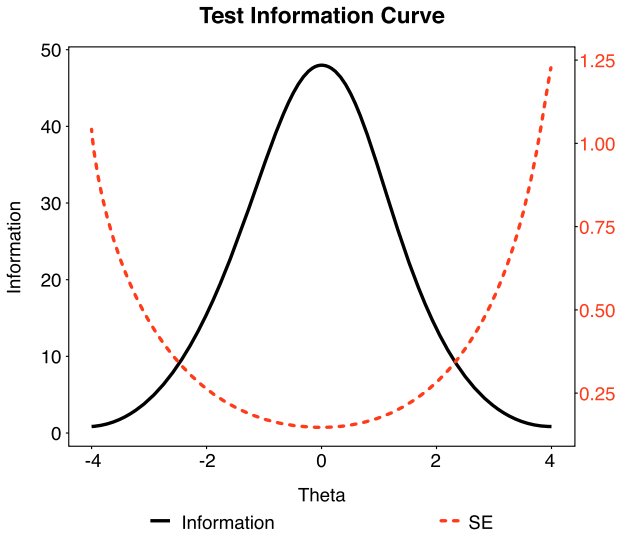

Figure. TIF curve of the Activity Limitation CAT. A higher level of information indicates greater measurement precision at that point along the scale. For the Activity Limitation CAT, the TIF curve peaked around zero on the ability scale (exact value = 0.02).

Item Exposure Rate. The IER varied across RetCAT (Table 4). For Visual Symptoms, Health Concerns, and Economic, all available items were administered (100% IER), while less than half the items available

Table 3. Average SE and 95% CI at Different Impairment Score Ranges for Diabetic Retinopathy Item Banks

	Score Range[a]				
QoL Domain	<2.0	2.0 to <3.0	3.0 to <3.5	≥3.5	Median Ability
Visual Symptoms	0.232 ± 0.008 n = 19	0.343 ± 0.017 n = 26	0.513 ± 0.017 n = 138	–	3.2
Activity Limitation	0.185 ± 0.007 n = 25	0.215 ± 0.008 n = 46	0.299 ± 0.013 n = 45	0.443 ± 0.020 n = 67	3.2
Mobility	0.428 ± 0.005 n = 2	0.316 ± 0.008 n = 35	0.373 ± 0.009 n = 146	–	3.0
Emotional	0.309 ± 0.068 n = 13	0.482 ± 0.038 n = 51	0.368 ± 0.003 n = 20	0.410 ± 0.011 n = 96	3.5
Health Concerns	0.194 ± 0.013 n = 18	0.193 ± 0.006 n = 43	0.271 ± 0.006 n = 77	0.449 ± 0.027 n = 45	3.2
Social	–	0.259 ± 0.010 n = 70	0.462 ± 0.003 n = 113	–	3.2
Convenience	–	0.269 ± 0.007 n = 96	0.443 ± 0.012 n = 87	–	3.0
Economic	0.202 ± 0.005 n = 13	0.189 ± 0.010 n = 27	0.232 ± 0.008 n = 36	0.439 ± 0.052 n = 27	3.2
Driving	–	0.236 ± 0.017 n = 21	0.387 ± 0.015 n = 59	–	3.1
Lighting	–	0.217 ± 0.003 n = 101	0.364 ± 0.026 n = 55	0.533 ± 0.000 n = 27	3.0

Shaded cells represent the score range with the lowest SE (i.e. the most precise measurement).
[a]Rasch scores for each domain were centered to have a mean of 3.0.

Table 4. Item Exposure Rates for Ten Diabetic Retinopathy CATs

Domain	Items Available for CAT	Items Administered (%)	Exposure Rate (>50%)
Visual Symptoms	18	18 (100.0)	7 (38.9)
Activity Limitation	84	61 (72.7)	11 (13.0)
Mobility	17	10 (58.8)	7 (41.2)
Emotional	45	28 (62.2)	7 (15.6)
Health Concerns	35	35 (100.0)	3 (8.6)
Social	20	13 (65.0)	7 (35.0)
Convenience	20	7 (35.0)	7 (35.0)
Economic	15	15 (100.0)	5 (33.3)
Driving	15	7 (46.7)	7 (46.7)
Lighting	10	8 (80.0)	7 (70.0)

in the Convenience (35%) and Driving (46.7%) item banks were administered. Similarly, some tests had a high proportion of items administered >50% of the time (e.g., Lighting, 70%), while some tests had only a small proportion of frequently administered items (e.g., Health Concerns, 8.6%). For most tests, 30%–40% of items were administered >50% of the time.

Validity

Convergent and Divergent Validity. Most RetCAT tests demonstrated expected moderate correlations with related scales (e.g., Mobility and IVI Mobility, $r = 0.461$; Supplementary Table S3). Although correlations between Convenience and Driving CATs and respective scales were statistically significant, they were slightly weaker than expected (<0.3), and

Visual Symptoms was not correlated at all with QoV ($r = 0.082$), although it was moderately correlated with IVI Mobility and IVI Reading. Activity Limitation and Lighting showed slightly stronger correlations than expected (>0.49). All RetCAT tests showed good divergent validity (Supplementary Table S3), with low correlations with GSES scores.

Criterion Validity. Four RetCAT tests (Activity Limitation, Health Concerns, Lighting, and Visual Symptoms) demonstrated reductions in test scores as DR severity increased (Supplementary Table S4). For example, Lighting scores were 0.28 (0.18–0.38), 0.16 (0.07–0.25), and 0.08 (–0.00 to 0.17) for minimal to mild NPDR, moderate to severe NPDR, and PDR, respectively (P trend = 0.004). The trend was not evident in the remaining RetCAT tests. For binocular VI, RetCAT scores consistently decreased as the severity of VI worsened for all domains except Convenience (Supplementary Table S5). For example, Activity Limitation scores were 1.21 (1.09–1.33), 0.77 (0.49–1.04), and 0.22 (–0.16 to 0.59) for no VI, mild VI, and moderate to severe VI, respectively (P trend < 0.001).

Discussion

Overall, RetCAT provides efficient, precise, and valid measurement of the impact of DR on QoL. While some CATs functioned well using only seven items and taking less than two minutes to administer per test, others would have benefited from more items to provide reliable measurement. To overcome this issue, future application of RetCAT will employ a stopping rule based on SE rather than number of items. Test precision was good overall, particularly for the larger item banks (>30 items). Measurement precision was highest for participants at the lower ends of the ability spectrum (i.e., most impaired) but comparatively lower for those at the higher ends (i.e., least impaired). As such, the tests are recommended for use in populations with vision-threatening DR, as measurement precision may be suboptimal in populations with early-stage disease. The IER varied across RetCAT; however, for most tests, 30%–40% of items were administered >50% of the time. Overall, RetCAT demonstrated excellent convergent and divergent validity and moderate criterion validity findings. With the potential to reduce respondent burden without sacrificing measurement precision, RetCAT may appeal to clinicians who wish to improve the patient experience of completing PROMs, pharmaceutical companies that wish to report the patient-centered impact of novel treatment interventions, health care organizations that wish to optimize care quality, and policy planners who wish to inform guidelines and resource allocation. RetCAT is available for use by contacting the corresponding author of the study.

The average SEM for RetCAT (0.351) was good, with certain CATs, such as Health Concerns and Economic, having very high measurement precision (SEM 0.290 and 0.272, respectively). However, others, such as Visual Symptoms (SEM 0.460) and Emotional-Mandarin (SEM 0.484), had comparatively lower precision. Moreover, four domains—Mobility, Social, Convenience, and Driving—had a high level of intrinsic measurement error impacting their ability to provide meaningful results,[32] which is likely due to the number of items being capped at seven. Administration of more items from the item banks of these domains would have improved their standard errors and increased the reliability of scores. To overcome this issue, future application of RetCAT will employ SE as the stopping rule rather than a maximum number of items. Although this may increase the time needed to complete the tests, it will greatly enhance the ability of these CATs to detect meaningful differences in person abilities.

Overall, RetCAT had excellent TIFs, suggesting that items within each bank carried a high level of relevant information. Generally, a TIF of 10 is considered excellent.[9] Six RetCAT tests achieved this, with some, like Activity Limitation, reaching a TIF of nearly 50. However, it is important to note that the maximum TIF values apply to one specific person measure, and for many CATs these maxima occurred outside the range of person measures observed in the study. As such, the TIF values reported in our study reflect the theoretical rather than actual information levels in our study sample. Smaller item banks (n = 10–18 items) had TIFs between 5 and 10, suggesting that having <20 items in a bank may not be optimal for outcomes measurement. However, specific QoL constructs, such as "economic" or "mobility," may only be defined by a small set of relevant items and, as such, may struggle to achieve high TIFs. In such cases, the importance of measuring these less commonly reported constructs may outweigh their lower TIF values.

RetCAT demonstrated the most precise measurement for patients at the lower end of the ability spectrum and was comparatively less precise for less impaired individuals. These results suggest that harder items are needed to improve measurement precision for those more able patients and to reduce ceiling effects. However, given that clinical focus is usually on patients with the most QoL impairment, having less precise measurement for those with few QoL issues may not

be problematic. Nonetheless, as part of the continuing process of item bank development and refinement, we aim to further improve the targeting and precision of RetCAT through the addition of more high-quality and sensitive items. One advantage of item banking and CAT is the ability to replenish and recalibrate item banks when content becomes outdated or gaps in measurement are observed.[33]

Overall, RetCAT displayed good convergent and divergent validity. However, the Visual Symptoms test showed almost no correlation with the QoV scores, which was unexpected since they shared similar content (e.g., "blurred vision," "fluctuating vision"). However, Visual Symptoms did correlate with IVI Reading and Mobility, providing sufficient evidence of convergent validity.

Six RetCAT tests (Activity Limitation, Visual Symptoms, Health Concerns, Lighting, Emotional, and Mobility) displayed evidence of criterion validity, being sensitive to DR and/or VI severity levels. Contrary to expectations, Convenience and Social demonstrated little relationship to either DR or VI. While these two QoL domains may lack relevance to people with DR, it is also possible that the study was not optimally powered to detect a statistically significant association between Convenience and Social and DR and VI severity. Despite oversampling patients with late-stage DR, we had only 27 patients with active PDR and 14 patients with moderate to severe binocular VI in our sample. More work is needed to explore the sensitivity of RetCAT across the spectrum of DR and VI in a larger sample as well as to determine which aspects of the visual function system (e.g., visual acuity, contrast sensitivity, depth perception, color vision) explain the most variance in QoL outcomes.

With its time-efficient administration and automated scoring, RetCAT will be a novel addition to ophthalmic research and clinical care. Results may be promptly integrated into patients' electronic medical records and immediately used to inform feedback and treatment,[34,35] which aligns well with the current global initiative to incorporate PROM data in clinical care and the push toward value-based medicine.[36–39] For example, RetCAT data could be synthesized with patients' corresponding clinical data and used to generate an at-a-glance report to treat poor vision–related mental health and monitor change over time or pre-/post-treatment therapies.[5] As recent advancements in treatments for eye diseases gain momentum, our comprehensive RetCAT instrument will be invaluable for use in clinical trials to compare the impact of novel treatment therapies from the patient's perspective. Similarly, RetCAT will allow researchers and policy planners to design and evaluate rehabilitation or educational programs for DR-related vision loss.

Strengths of our study include the robust practical and psychometric assessments of RetCAT and standardized eye tests, including fundus photographs and DR grading. Moreover, our results may be generalizable to Asian populations outside of Singapore, particularly English-speaking people with diabetic eye disease in China, Malaysia, and India as well as Mandarin speakers in China. However, more work may be required to replicate our results in Caucasian populations. Limitations include the relatively small sample size, particularly in those with severe disease, and the fact that test-retest and responsiveness data were not collected. While we endeavored to culturally and linguistically adapt the item banks, it is possible that differences in relation to perceptions of illness and responses to impairment may have persisted. Indeed, measurement precision was quite low in the Emotional-Mandarin CAT (SEM 0.484), suggesting that some items may have had a high degree of associated noise. Future work is required to better understand the cultural and linguistic issues associated with the EM-Mandarin CAT and to determine how to optimize its psychometric properties.

In summary, RetCAT is an efficient and psychometrically robust instrument to measure the impact of DR on QoL, particularly in people with greater levels of impairment. Future work will focus on improving the precision and targeting of some of the domains through the addition of high-quality items and recalibration of the item banks and employing a stopping rule based on SE rather than number of items. RetCAT may be useful for clinicians who wish to monitor patient DR risk and progress, pharmaceutical companies that wish to evaluate the patient-centered impact of new therapies, and eye clinics that wish to carry out value-based evaluations of patient care.

Acknowledgments

This project was funded by the National Health and Medical Research Council (NHMRC) Centre for Clinical Research Excellence (CCRE) (#529923); Translational Clinical Research in Major Eye Diseases, CCRE Diabetes, Novartis Pharmaceuticals Australia (#CRFB002DAU09T); SingHealth ACP Talent Development Grant (#R1383/69/2016); Duke-NUS Medical School Seed Funding; Royal Victorian Eye and Ear Hospital; and Lions Ride for Sight. Eva Fenwick was funded by an Australian NHMRC Early Career Fellowship (#1072987). Ecosse Lamoureux is

salary-supported by the Singapore National Medical Research Council Clinician Scientist Award. The Centre for Eye Research Australia receives operational infrastructure support from the Victorian Government. Sponsors and funding organizations had no role in the design or conduct of this research.

Disclosure: **E.K. Fenwick**, None; **J. Barnard**, None; **A. Gan**, None; **B.S. Loe**, None; **J. Khadka**, None; **K. Pesudovs**, None; **R. Man**, None; **S.Y. Lee**, None; **G. Tan**, None; **T.Y. Wong**, None; **E.L. Lamoureux**, None

References

1. Cheung N, Mitchell P, Wong T. Diabetic retinopathy. *Lancet*. 2010;376:124–136.
2. Fenwick E, Pesudovs K, Rees G, et al. The impact of diabetic retinopathy: understanding the patient's perspective. *Br J Ophthalmol*. 2010;95:774–782.
3. Fenwick E, Rees G, Pesudovs K, et al. Social and emotional impact of diabetic retinopathy: a review. *Clin Exp Ophthalmol*. 2012;40:27–38.
4. Khoo K, Man R, Rees G, et al. The relationship between diabetic retinopathy and psychosocial functioning: a systematic review. *Qual Life Res*. 2019;28:2017–2039.
5. Snyder CF, Jensen RE, Segal JB, et al. Patient-reported outcomes (PROs): putting the patient perspective in patient-centered outcomes research. *Med Care*. 2013;51(Suppl 3):73–79.
6. WHOQOL Group. *Measuring Quality of Life*. Geneva: The World Health Organisation; 1997.
7. Cella D, Gershon R, Lai JS, et al. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. *Qual Life Res*. 2007;16(Suppl 1):133–141.
8. Wainer H, Dorans N, Flaugher R, et al. *Computerized Adaptive Testing: A Primer*. 2nd ed. London & New York: Routledge; 2000.
9. Embretson S, Reise SP. *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaun Associates; 2000.
10. Bjorner J, Chang C-H, Thissen D, et al. Developing tailored instruments: item banking and computerized adaptive assessment. *Qual Life Res*. 2007;16:95–108.
11. Gershon RC. Computer adaptive testing. *J Appl Meas*. 2005;6:109–127.
12. Fenwick E, Pesudovs K, Khadka J, et al. The impact of diabetic retinopathy on quality of life: qualitative findings from an item bank development project. *Qual Life Res*. 2012;21:1771–1782.
13. Fenwick E, Pesudovs K, Khadka J, et al. Evaluation of item candidates for a diabetic retinopathy quality of life item bank. *Qual Life Res*. 2013;22:1851–1858.
14. Fenwick E, Khadka J, Pesudovs K, et al. Diabetic retinopathy and macular edema quality-of-life item banks: development and initial evaluation using computerized adaptive testing. *Invest Ophthalmol Vis Sci*. 2017;58:6379–6387.
15. Jette A, Haley S, Tao W, et al. Prospective evaluation of the AM-PAC-CAT in outpatient rehabilitation settings. *Phys Ther*. 2007;87:385–398.
16. Becker J, Fliege H, Kocalevent RD, et al. Functioning and validity of a computerized adaptive test to measure anxiety (A-CAT). *Depress Anxiety*. 2008;25:E182–194.
17. Abberger B, Haschke A, Wirtz M, et al. Development and evaluation of a computer adaptive test to assess anxiety in cardiovascular rehabilitation patients. *Arch Phys Med Rehabil*. 2013;94:2433–2439.
18. Barthel D, Otto C, Nolte S, et al. The validation of a computer-adaptive test (CAT) for assessing health-related quality of life in children and adolescents in a clinical sample: study design, methods and first results of the Kids-CAT study. *Qual Life Res*. 2017;26:1105–1117.
19. Marfeo EE, Ni P, Haley SM, et al. Scale refinement and initial evaluation of a behavioral health function measurement tool for work disability evaluation. *Arch Phys Med Rehabil*. 2013;94:1679–1686.
20. Brooke P, Bullock R. Validation of a 6 item cognitive impairment test with a view to primary care usage. *Int J Geriatr Psychiatry*. 1999;14:936–940.
21. Andrich D. A rating scale formulation for ordered response categories. *Psychometrika*. 1978;43:561–573.
22. Barnard J. From simulation to implementation: two CAT case studies. *Pract Assess Res Eval*. 2018;23:1–7.
23. Early Treatment Diabetic Retinopathy Study Research Group. Grading diabetic retinopathy from stereoscopic color fundus photographs—an extension of the modified Airlie House classification. ETDRS report number 10. *Ophthalmology*. 1991;98(Suppl):786–806.
24. Lamoureux E, Pallant JF, Pesudovs K, et al. The Impact of Vision Impairment questionnaire: an evaluation of its measurement properties using Rasch analysis. *Invest Ophthalmol Vis Sci*. 2006;47:4732–4741.

25. Lamoureux E, Pallant JF, Pesudovs K, et al. The Impact of Vision Impairment questionnaire: an assessment of its domain structure using confirmatory factor analysis and Rasch analysis. *Invest Ophthalmol Vis Sci*. 2007;48:1001–1006.
26. McAlinden C, Pesudovs K, Moore JE. The development of an instrument to measure quality of vision: the Quality of Vision (QoV) questionnaire. *Invest Ophthalmol Vis Sci*. 2010;51:5537–5545.
27. Luszczynska A, Scholz U, Schwarzer R. The General Self-Efficacy Scale: multicultural validation studies. *J Psychol*. 2005;139:439–457.
28. Linacre JM. *Winsteps Rasch measurement computer program User's Guide*. Beaverton, Oregon: Winsteps.com; 2020.
29. Bond TG, Fox CM. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. London: Lawrence Erlbaum Associates; 2001.
30. Revuelta J, Ponsoda V. A comparison of item exposure control methods in computerized adaptive testing. *J Educ Meas*. 1998;35:311–327.
31. Ratner B. The correlation coefficient: its values range between +1/−1, or do they? *J Target Meas Anal Market*. 2009;17:139–142.
32. Massof, R. Understanding Rasch and item response theory models: applications to the estimation and validation of interval latent trait measures from responses to rating scale questionnaires. *Ophthalmic Epidemiol*. 2011;18:1–19.
33. Haley SM, Ni P, Jette AM, et al. Replenishing a computerized adaptive test of patient-reported daily activity functioning. *Qual Life Res*. 2009;18:461–471.
34. Lai JS, Cella D, Chang CH, et al. Item banking to improve, shorten and computerize self-reported fatigue: an illustration of steps to create a core item bank from the FACIT-Fatigue Scale. *Qual Life Res*. 2003;12:485–501.
35. Forkmann T, Boecker M, Norra C, et al. Development of an item bank for the assessment of depression in persons with mental illnesses and physical diseases using Rasch analysis. *Rehabil Psychol*. 2009;54:186–197.
36. Basch E. Patient-reported outcomes—harnessing patients' voices to improve clinical care. *N Engl J Med*. 2017;376:105–108.
37. Bradley S, Rumsfeld J, Ho P. Incorporating health status in routine care to improve health care value: the VA Patient Reported Health Status Assessment (PROST) system. *JAMA*. 2016;316:487–488.
38. Baumhauer J, Bozic, K. Value-based healthcare: patient-reported outcomes in clinical decision making. *Clin Orthop Relat Res*. 2016;474:1375–1378.
39. Rotenstein L, Huckman R, Wagle N. Making patients and doctors happier—the potential of patient-reported outcomes. *N Engl J Med*. 2017;377:1309–1312.