Diabetic Retinopathy and Macular Edema Quality-of-Life Item Banks: Development and Initial Evaluation Using Computerized Adaptive Testing

Eva K. Fenwick,[1-3] Jyoti Khadka,[4] Konrad Pesudovs,[4] Gwyn Rees,[1] Tien Y. Wong,[2,3] and Ecosse L. Lamoureux[1-3]

[1]Centre for Eye Research Australia, The Royal Victorian Eye and Ear Hospital, University of Melbourne, Melbourne, Australia
[2]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore
[3]Duke–NUS, Singapore, National University of Singapore, Singapore
[4]Discipline of Optometry and Vision Science, Flinders University of South Australia, South Australia, Australia

Correspondence: Ecosse L. Lamoureux, Singapore Eye Research Institute (SERI), Population Health, The Academia, 20 College Road, Level 6, Singapore 169856;
ecosse.lamoureux@seri.com.sg.

Submitted: October 19, 2016
Accepted: October 11, 2017

Citation: 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;58:6379–6387. DOI:10.1167/iovs.16-20950

Purpose. The purpose of this study was to assess the psychometric properties of diabetic retinopathy (DR) and diabetic macular edema (DME) quality-of-life (QoL) item banks and determine the utility of the final calibrated item banks by simulating a computerized adaptive testing (CAT) application.

Methods. In this clinical, cross-sectional study, 514 participants with DR/DME (mean age ± SD, 60.4 ± 12.6 years; 64% male) answered 314 items grouped under nine QoL item pools: Visual Symptoms (SY); Ocular Comfort Symptoms (OS); Activity Limitation (AL); Mobility (MB); Emotional (EM); Health Concerns (HC); Social (SC); Convenience (CV); and Economic (EC). The psychometric properties of the item pools were assessed using Rasch analysis, and CAT simulations determined the average number of items administered at high and moderate precision levels.

Results. The SY, MB, EM, and HC item pools required minor amendments, mainly involving removal of six poorly worded, highly misfitting items. AL and CV required substantial modification to resolve multidimensionality, which resulted in two new item banks: Driving (DV) and Lighting (LT). Due to unresolvable psychometric issues, the OS, SC, and EC item pools were not pursued further. This iterative process resulted in eight operational item banks that underwent CAT simulations. Correlations between CAT and the full item banks were high (range, 0.88–0.99). On average, only 3.6 and 7.2 items were required to gain measurement at moderate and high precision levels, respectively.

Conclusions. Our eight psychometrically robust and efficient DR/DME item banks will enable researchers and clinicians to accurately assess the impact and effectiveness of treatment therapies for DR/DME in all areas of QoL.

Keywords: quality of life, item bank, computerized adaptive testing, Rasch analysis, psychometrics

Diabetic retinopathy (DR) is a common complication of diabetes,[1] which can result in substantial and sometimes irreversible vision loss in its proliferative stages. Diabetic macular edema (DME) can occur at any stage of DR and is responsible for severe loss of central vision.[2] The impact of DR on health-related quality of life (QoL) is substantial, especially at the vision-threatening stages.[3,4] Novel therapies such as anti-VEGF intravitreal injections have shown promising results for improving vision loss and QoL, particularly for DME.[5,6]

Patient reported outcome measures (PROs) are essential to guide service provision and improve care in clinical practice[7] and inform rehabilitation programs. They are required by regulatory authorities in clinical trials to assess the patient-centered effectiveness of novel treatments.[8] To date, however, most currently available QoL instruments are not specific to DR and DME, which means they may lack sensitivity in capturing issues specific to the conditions, such as the impact of laser treatment or intravitreal injections on QoL and the difficulty associated with managing diabetes (a chronic health condition that requires a high degree of self-management) with an eye condition that requires frequent monitoring and that is often associated with visual impairment. In addition, the QoL impact of DR/DME and treatment has only been assessed using paper-pencil PROs,[9,10] which have a finite number of items and often fail to optimally target participants' impairment level across the spectrum of disease severity even though participants must answer every item. This can increase respondent burden,[11,12] lower response rates, and reduce data quality.[13] Moreover, most QoL instruments predominantly focus on visual functioning, whereas QoL also encompasses vision-related symptoms, pain, concerns, inconvenience, social life, and work issues.[14] These limitations are addressed by modern psychometric techniques such as item banking and computerized adaptive testing (CAT).[12,15] The advantages of item response theory calibrated item banks are well recognized; for example, the National

Institutes of Health (NIH) Toolbox vision-targeted health-related quality of life measure.[16]

An item bank is a pool of calibrated items (questions) that measure a latent construct such as "health concerns."[17] CAT is a method for administering items from a calibrated item bank. It selectively chooses the questions asked based on the examinee's impairment level by presenting targeted items (i.e., those that will provide the greatest amount of information) to the respondent.[18] Subsequent items are selected based on the examinee's previous responses and selection proceeds until a predefined stopping criterion is reached. CAT requires fewer items than paper-pencil tests and may enhance measurement validity, precision, and accuracy.[18,19]

We have developed item banks to assess the specific impact of DR/DME on nine relevant aspects of patients' QoL.[20,21] Here, we assess the psychometric properties of these item banks and investigate the utility of the final calibrated item banks by simulating a CAT application.

Methods

Study Design and Participants

The 514 participants in our study were recruited from 1) a large cross-sectional study conducted in Melbourne, Australia (principle investigator, Ecosse Lamoureux), the Diabetes Management Project[11] (n = 200 participants); 2) ophthalmic clinics at the Royal Victorian Eye and Ear Hospital (RVEEH, n = 272); 3) the Royal Society of the Blind (RSB, n = 39); and 4) private clinics (n = 3). Participants had a primary diagnosis of DR and/or DME, were aged $\geq$18 years, had type 1 or 2 diabetes, had no significant hearing or cognitive impairment, and had no other late-stage eye diseases. Face-to-face or telephone interviews were conducted after obtaining participants' written informed consent. The study had ethical approval from the Royal Victorian Eye and Ear Hospital Human Research Ethics Committee (#09/888H/15) and was conducted in accordance with the Declaration of Helsinki.

Development of the DR/DME QoL Item Banks

Our hypothesized QoL domains were based on the World Health Organization Quality of Life (WHOQOL) conceptual framework of QoL[14] and a comprehensive review of the literature.[3,4] Items for the DR/DME item banks were extracted from 34 previously validated health-, vision-, and condition-specific QoL and vision-specific functioning questionnaires. New items were generated from 5 published qualitative papers reporting patient experiences with DR, 8 focus groups, and 18 semistructured interviews with 57 DR patients,[21] and 7 semistructured interviews with diabetes or ophthalmic experts. Items and domains were revised during three stages, namely binning (grouping) and winnowing (reduction), development of item stems and response options, and pretesting of final items via cognitive interviews with DR patients.[20] At the end of the content development phase, there were 314 items within nine QoL item banks (Supplementary Table S1): Visual Symptoms (SY, n = 18); Ocular Comfort Symptoms (OS, n = 10); Activity Limitation (AL, n = 120); Mobility (MB, n = 19); Emotional (EM, n = 48); Health Concerns (HC, n = 36); Social (SC, n = 21); Convenience (CV, n = 30); and Economic (EC, n = 12).

VS and OS have four-point summated frequency rating scales ranging from 4 (never) to 1 (very often). AL and MB are rated on a five-point difficulty scale ranging from 5 (none) to 1 (unable to do because of my vision). EM is rated on a five-point frequency scale from 5 (never) to 1 (all of the time). HC and EC are rated on a five-point "concern" scale ranging from 1 (not at all) to 5 (extremely). SC is rated on a five-point "problem" scale ranging from 1 (none) to 5 (unable to do because of my vision). CV is rated on a five-point "trouble" scale ranging from 1 (none) to 5 (extreme). A nonapplicable response option (e.g., don't do for other reasons) was also available for some items and was coded as missing.

Assessment of DR and Visual Acuity

For participants who completed a telephone interview, clinical data were extracted from medical files or obtained via fax from their treating ophthalmologist. In face-to-face interviews, we assessed DR using dilated fundus photography (macula and optic disc). DR grading adhered to the Modified Two-Standard Field Color Fundus Photography Procedure (Retinal Vascular Imaging Centre Grading Protocol 01; Assessment of DR); based on the Early Treatment Diabetic Retinopathy Study (ETDRS) and Multi-Ethnic Study of Atherosclerosis (MESA) Digital Grading Protocol.[11] We categorized DME severity using the American Academy of Ophthalmology classification.[22]

Visual acuity (VA) was assessed using a 3-m LogMAR (logarithm of the minimum angle of resolution) chart. Patients' presenting VA was first categorized into none (VA better eye LogMAR $\leq$ 0.3); mild/moderate (0.3 < LogMAR < 1.0); and severe vision impairment (LogMAR $\geq$ 1.0) and then into six groups of bilateral visual impairment: (1) normal vision in both eyes; (2) normal vision in one eye and mild or moderate visual impairment in the other; (3) normal vision in one eye and severe visual impairment in the other; (4) mild or moderate visual impairment in both eyes; (5) severe visual impairment in one eye, and mild or moderate visual impairment in the other; and (6) severe visual impairment in both eyes.[23] Sociodemographic and clinical statistics were computed using Stata version 14 (StataCorp, College Station, TX, USA).

Psychometric Evaluation of the Item Banks

We performed Rasch analysis on each of the nine item pools using Winsteps software (version 3.91.2; Chicago, IL, USA)[24] using the Andrich single rating scale model.[25] Rasch analysis is a one-parameter item response theory (IRT) model (Supplementary Fig. S1).[26,27] Unlike two- or three-parameter IRT models, the Rasch model assumes that guessing is part of person "impairment" and that all items have equivalent discriminations. As such, the Rasch model has the property of specific objectivity, meaning that the rank of the item difficulty is the same for all respondents independent of impairment, and that the rank of the person impairment is the same for items independently of difficulty.

During Rasch analysis, ordinal-level raw score data are converted into estimates of interval measures.[28] The Rasch model assumes that the probability of a given respondent affirming an item is a logistic function of the relative distance between the item's location (i.e., difficulty) and the respondent's location (i.e., impairment) on this linear scale. The resulting item locations and person measures are expressed in log of the odds units, or Logits. Rasch analysis also provides substantial insight into a scale's psychometric properties,[29] outlined below.

Correlations Between Person Measures

We calculated the correlations between person measure estimates to assess the level of independence of the nine QoL item banks. Although correlations between person measure estimates from the banks were relatively high (>0.5; Supplementary Table S2), we opted to treat each scale as independent of the other scales and use a unidimensional rather than multidimensional IRT model. This decision was based on our a priori hypotheses developed from the literature and our qualitative work that the nine areas of QoL were distinct constructs.

Category Probability Curves. The average measures for the person sample (i.e., average ability levels of the respondents who respond to the particular categories) should advance across categories in the expected order, such that, as person measures increase, each category in turn is more probable than any one of the others to be observed. Category ordering can be determined by the category probability curves (Supplementary Fig. S2). Disordering of categories may indicate that categories are underused, are poorly defined, that there are too many categories for respondents to sensibly distinguish, or that the scale is multidimensional due to different response categories being used for different groups of questions.[30] Category disordering can be resolved by collapsing adjacent categories so long as this improves the clarity of the response options and other Rasch metrics. The threshold boundary locations of each item on each response category are shown in Supplementary Tables S3 through S13.

Precision. Scale precision demonstrates the capacity of the item banks to discriminate between differing levels of the underlying construct (e.g., Activity Limitation). Precision is determined using person separation index (PSI) and person reliability (PR) coefficients, where values of <2.0 and <0.8, respectively, suggest that the instrument may not be sensitive enough to discriminate between high and low performers. Participants with extreme scores (i.e., minimum or maximum) were removed a priori from the analysis (number removed ranged from 16 [Activity Limitation] to 186 [Social], median n = 36; Supplementary Tables S14–S24) as extreme scores provide no information for estimating item measures and may lessen measurement precision.[26]

Unidimensionality, Item Fit, and Item Discrimination. Principal components analysis (PCA) of residuals was used to assess whether the item banks measured a single construct. During this process, the first component (dimension) is removed, and a PCA of residuals is conducted to look for patterns in the data that do not accord with the Rasch measures, suggesting that groups of items may be forming a secondary dimension.The raw variance explained for the first factor (i.e., primary dimension) should therefore exceed 50%, and the unexplained variance by first contrast (i.e., first component in the correlation matrix of the residuals) should be <3 eigenvalues.[31] If the first contrast has an eigenvalue of 3, this means that the secondary dimension has the strength of about 3 items, which is bigger than that expected by chance and is potentially substantial enough to suggest multidimensionality in the scale. If we found evidence of multidimensionality, the standardized residual loadings on the first contrast were assessed to determine if certain items load substantively (>0.4) and whether they formed a conceptually relevant second dimension.

Infit and outfit are fit statistics that indicate how well the data fit the Rasch model using χ^2 statistics. Item misfit may indicate that an item is measuring a different construct than the other items in the scale and is assessed through infit and outfit mean square (MnSq) statistics.[31] Infit is more sensitive to the pattern of responses to items targeted on the person, whereas outfit is more influenced by outliers (e.g., carelessness).[32] A relatively lenient range for infit and outfit statistics (0.50–1.50) has been adopted for our item banking work,[33] as this range is still conducive for productive measurement and it enables us to include as much content as possible given that many of our constructs are novel. As outfit is less threat to measurement,[32] we considered infit first followed by outfit when evaluating item fit.[31]

To understand the source of misfit and avoid unnecessary deletion of items, we reviewed item wording. Confusing or complicated wording provided evidence for deletion. We also explored z-residuals in Winsteps table 11.1 to determine whether odd or unexpected responses to an item by the respondents were responsible for item misfit, rather than other factors such as unclear item wording. Generally, respondents with higher z-residuals have high misfit to the Rasch model.[26] Therefore, respondents with responses with a z-residual >3 or >−4 were given a weighting of 0. This means that, although their measure and fit statistics are reported, they do not influence measures or fit statistics of other persons or items. Misfitting respondents were iteratively given a weighting of 0, each time followed by reassessment of the item fit statistics to assess if item fit had improved. The process was carried out until satisfactory item fit statistics were achieved (infit/outfit MnSq ≤1.5). However, if weighting "problematic" responders at 0 made no difference to the item fit statistics, we explored other reasons for the item misfit and considered the item as a candidate for deletion.

We also considered item discrimination when deciding whether to keep or remove items from the scales. Items with values >1.0 means that the item discriminates between high and low performers more than expected for an item of this difficulty. A value <1.0 means that the item discriminates between high and low performers less than expected for an item of this difficulty. From a Rasch perspective, overdiscriminating items tend to act like two distinct and opposing categories, classifying people as either highly impaired or highly functional. Underdiscriminating items, however, tend neither to stratify nor to measure. As underdiscrimination was more of a concern for us than overdiscrimination, we focused only on those items with values well under 1.0 as candidates for deletion. Item discrimination values are provided for all items in all domains in Supplementary Tables S3 through S13.

Local Item Dependency. Ideally, items should be independent, and there should be no correlation between two items after the effects of the underlying construct are accounted for.[34] High correlation (>0.3) among the residuals (those parts of the data not explained by the Rasch model) suggests local item dependency (LID). To remove the effects of LID on threshold calibrations for CAT, we generated person measures corresponding to only the LID-free items. We then anchored all person measures to those generated using the non-LID items. This forces all item difficulties and rating-scale structures to conform with LID-free person measures and prevents LID from impacting item difficulties.[24]

Targeting. Ideally, there should be a good spread of items across the full range of respondents' scores. Poor targeting occurs when persons generally have higher or lower impairment than the most or least item difficulty threshold, or when items are clustered at particular levels of difficulty leaving large gaps.[31] Targeting can be examined through visual inspection of the person-item map and calculated by determining the difference between the mean of item "difficulty" (defined as 0 logits) and the mean of person impairment; a difference of >1.0 logits indicates notable mistargeting.[31]

Measurement Range. Measurement range was determined by calculating the difference in logits between the highest and lowest item locations. The larger the measurement range, the more information about the measured construct is provided by the items. Item locations are provided in Supplementary Tables S3 through S13.

Differential Item Functioning. Differential item functioning (DIF) determines whether item bias exists for sample subgroups (e.g., sex). A DIF contrast of >1.0 logits and

Table 1. Sociodemographic and Clinical Characteristics of the 514 Participants*

Variable	No. (%)
Aged <60 y	239 (46.4)
Type 2 diabetes	412 (80.2)
Insulin use (yes)	357 (69.4)
Male	328 (63.6)
Country of birth	
Australia	278 (54.1)
Other	236 (45.9)
Main language spoken	
English	450 (87.5)
English plus another	64 (12.5)
Mode of interview	
Face to face	268 (0.52)
Phone	246 (0.48)
Marital status	
Never married	85 (16.5)
Married/de facto	306 (59.5)
Divorced/separated/widowed	123 (23.9)
Education level	
Primary	305 (59.3)
Secondary	190 (37.0)
TAFE/university degree	19 (3.7)
Employment status	
Working	145 (28.2)
Not working	369 (71.8)
Comorbidity (yes, self-reported)	436 (84.8)
Vision impairment (right eye)	
None ($\leq$0.3 LogMAR)	220 (44.0)
Mild (>0.3 LogMAR $\leq$0.48)	193 (37.5)
Moderate/severe (>0.48 LogMAR)	87 (17.4)
Vision impairment (left eye)	
None ($\leq$0.3 LogMAR)	228 (45.6)
Mild (>0.3 LogMAR $\leq$0.48)	170 (34.0)
Moderate/severe (>0.48 LogMAR)	102 (20.4)
Bilateral vision impairment	
Normal vision in both eyes ($\leq$0.3 LogMAR)	134 (26.8)
Normal vision in one eye and mild/moderate visual impairment (0.3 < LogMAR < 1.0) in the other	119 (23.8)
Normal vision in one eye and severe visual impairment ($\geq$1.0 LogMAR) in the other	58 (11.6)
Mild/moderate vision impairment in both eyes	91 (18.2)
Severe visual impairment in one eye, mild/moderate visual impairment in the other	60 (12.0)
Severe visual impairment in both eyes	38 (7.6)
Severity of DR (worse eye)	
Mild nonproliferative DR	42 (8.3)
Moderate nonproliferative DR	105 (20.8)
Severe nonproliferative DR	63 (12.5)
Proliferative DR	296 (58.5)
Severity of DME (worse eye)	
None	223 (48.7)
Mild diabetic macular edema	44 (9.6)
Moderate diabetic macular edema	71 (15.5)
Severe diabetic macular edema	120 (26.2)
DR treatment	
Laser therapy	424 (82.5)
Vitrectomy surgery	122 (23.7)
Intraocular injections	89 (17.3)
Other eye pathology (yes, self-report)	89 (8.4)

Continuous Variables	Mean (SD); Median (Range)
Age (y)	60.4 (12.6); 61.2 (22.3–91.7)
Duration of diabetes (y)	20.7 (11.0); 20.0 (1.0–68.0)

* Percentages for some variables may not equal 100% due to missing data. de facto, couples living in a marriage-like state for at least 2 years; TAFE, Technical and Further Education.

corresponding $P < 0.05$ for an item indicates very large DIF.[24] DIF was calculated using pairwise comparison (two-sided *t*-test) by presenting the effect size in logits (difference in measures of the items between two groups) and the probability of observing the difference in measures by chance when there is no systematic item bias present. DIF was assessed for sex, age group (<60 vs. $\geq$60 years [based on a median-split]), and better eye vision impairment (none versus at least mild vision impairment [<6/12]); DR severity (non–vision-threatening DR [VTDR] versus [VTDR]), and mode of interview (face to face versus phone).

CAT Simulation. A real-data simulation of CAT is an important step in CAT applications as it allows developers to evaluate important features of the CAT system, such as item selection and stopping rules, and determine whether CAT will work effectively before live testing.[35] CAT simulations were conducted on a sample size of 1000 using Firestar-D software with the EAP (expected a posteriori estimator) and the Max Posterior Weighted Info (MPWI) item selection criteria.[36] The first simulation calculated the average number of items required to achieve a standard error of measurement (SEM) of 0.387 (approximating to a reliability of 0.85, derived using the formula: $\alpha = 1 - \text{SEM}^2$), which represents the high precision that might be required in individual assessment. A second simulation was conducted to demonstrate the average number of items needed to achieve a SEM of 0.521 (approximating to a reliability of 0.72), representing moderate precision that might be required in large group studies.[37,38] Correlations between the full item bank and CAT simulation person measure estimates were calculated for both levels of precision. CAT simulations started with items of moderate difficulty (i.e., item location = 0.0). The next item to be selected was the one whose location most closely targeted the estimated person measure of the participant at that point in the test.

Results

Sociodemographic and Clinical Characteristics

A total of 514 participants (mean age ± SD, 60.4 ± 12.6 years; 64% male) answered the 314 items (Table 1). Median duration of diabetes was 20 years (range, 1–68 years). Of the 514 participants, 42 (8%), 105 (21%), 63 (13%), and 296 (59%) had mild nonproliferative DR (NPDR), moderate NPDR, severe NPDR, and proliferative DR (PDR), respectively, and 235 (46%) had DME. Around a quarter of participants (n = 134) had normal presenting vision in both eyes, whereas more than one third had some level of bilateral vision impairment (at least >0.3 LogMAR in the better eye).

Psychometric Properties of the Item Banks

The full psychometric properties of the item banks are presented in Supplementary Tables S14 through S24, and the modifications and final item bank structures are summarized in Figure 1. In brief, the Visual Symptoms (SY), Mobility (MB), Emotional (EM), and Health Concerns (HC) item banks required minor psychometric amendments to achieve satisfactory fit, mainly involving removal of items that were poorly worded, highly misfitting, had poor item discrimination, and/or displayed DIF ($n = 3$ for EM; $n = 2$ for MB, and $n = 1$ for HC).

However, the Activity Limitation (AL) item bank was multidimensional and standardized residual loadings for items revealed that the driving (AL108–120), lighting (AL81–90), and work (AL104–107) items grouped together (Supplementary Table S16). Therefore, we created separate 13-item driving (DV) and 10-item lighting (LT) item banks and added the 4 work items to the Economic (EC) item bank to explore whether they would fit under the EC construct (Fig. 1). The Convenience (CV) item bank was also multidimensional with the six driving items (CV25–30) loading together, and these were added to the new DV item bank for further consideration (Fig. 1). Four CV items relating to appointment/treatment inconvenience displayed misfit and/or poor item discrimination and were removed.

The psychometric properties of the two new item banks, Driving (DV) and Lighting (LT), were also tested. LT displayed acceptable fit to Rasch model parameters overall and required minimal modifications. After collapsing categories from five to four due to disordered thresholds and removal of five items due to multidimensionality, DV also achieved acceptable fit to Rasch model requirements.

However, three item banks showed suboptimal fit to the Rasch model and were therefore not further tested in CAT simulations. Ocular comfort symptoms (OS) demonstrated suboptimal precision and dimensionality metrics (Supplementary Table S15). Social (SC) demonstrated poor precision (PSI = 1.19) due to a substantial ceiling effect (36%). Although PSI increased to >2.0 once those with extreme scores ($n = 184$) were removed from the analysis one was item deleted due to misfit (Supplementary Table S20), loss of one third of the sample was deemed unacceptable. The Economic (EC) item bank was tested only in working individuals ($n = 202$). It initially contained 12 items but subsequently incorporated an extra five items from AL and MB (Fig. 1). Initially, the AL/MB items loaded together; however, on removal of two redundant items, multidimensionality was resolved. To resolve disordered thresholds (Supplementary Fig. S1), the feasibility of a three-category solution (12345→12223) was explored. However, the PSI dropped from 2.52 to 1.90, suggesting that the loss of information resulting from merging response categories was reducing measurement precision; moreover, when the three-category solution was tested in CAT simulation, all 15 items had to be administered to obtain an adequate level of measurement precision. Therefore, the EC item bank was not further considered.

CAT Simulation

CAT simulations were conducted for 1000 cases for the eight final item banks (Table 2). When the SEM was set at 0.387, an average of 7.23 items was required across the item banks (range: 5.27 for AL to 9.14 for MB). When the SEM was set at 0.521, 3.56 items on average were required (range: 3.06 AL to 4.45 MB). AL had the most substantial reduction in test length during CAT. From the 92 available items in the bank, only 5.3 and 3.1 items on average were required to achieve a SEM of 0.387 and 0.521, respectively. Correlations between the full AL item bank and CAT person measures were 0.94 and 0.88 for high and moderate precision levels, respectively, and results were very similar across all eight item banks. Item usage within the AL item bank was relatively evenly distributed except for the starting item AL18 “Reading the numbers on the front of a bus,” which was used 100% of the time (Fig. 2).

Discussion

The item banks resulting from this work will provide, for the first time, measurement of eight areas of QoL specific to people with DR and DME, in addition to the introduction of relatively novel constructs such as Mobility, Health Concerns, Convenience, Driving, and Lighting. Furthermore, with simulation testing indicating that less than 10 items are required to gain precise measurement of each QoL item bank, our CAT is likely to be a time-efficient modality for use in clinics and research settings. Overall, this work will enable researchers and clinicians to comprehensively explore the impact of DR/DME from the patient’s perspective for the first time. With the availability of eight item banks, researchers and clinicians can now choose the constructs relevant to their participants and patients, respectively.

Although the psychometric properties of most item banks were very good following amendments, AL, EM, and HC had higher than satisfactory eigenvalues, even after DV and LT were separated from AL. This may indicate multidimensionality, however, because the residual item loadings did not suggest meaningful secondary dimensions and because all items fit within their respective constructs, we did not split the scales. Although precision was good overall, targeting of item “difficulty” to participant “impairment” was suboptimal for the AL, MB, EM, HS, SC, CV, and DV item banks likely due to the relatively small number of participants with bilateral vision impairment (~36%). Targeting may improve by adding items relevant to those at the less impaired end of the spectrum. This is relatively simple in item banking, where new, uncalibrated items are added to the bank and their calibration, relative to the existing items, is determined using Rasch analysis.[39] However, poor targeting of item difficulty to person impairment is largely overcome by CAT as the test is tailored to the individual’s impairment level.[33]

Three item banks demonstrated suboptimal fit to key Rasch model criteria and were therefore not further considered in the current study. The OS item bank had poor measurement precision and dimensionality indictors, suggesting that ocular comfort symptoms may not be a relevant construct for people with DR/DME. The SC item bank only obtained adequate measurement precision after one third of participants with extreme scores (i.e., those who reported no problem to all items) were removed from the Rasch analysis, suggesting that restrictions in social life may only be relevant for those with substantial vision impairment from DR/DME. Finally, the rating scale of the EC domain displayed highly disordered thresholds. Although order was restored by collapsing categories from 5 to 3, the resulting loss of measurement precision meant that the efficiency of CAT for this item bank was compromised. Future work will involve crafting and pilot testing additional items to improve the psychometric properties of these item banks and ensure their suitability for CAT.

Our study demonstrates the potential for advancing QoL measurement in DR/DME using an item banking and CAT

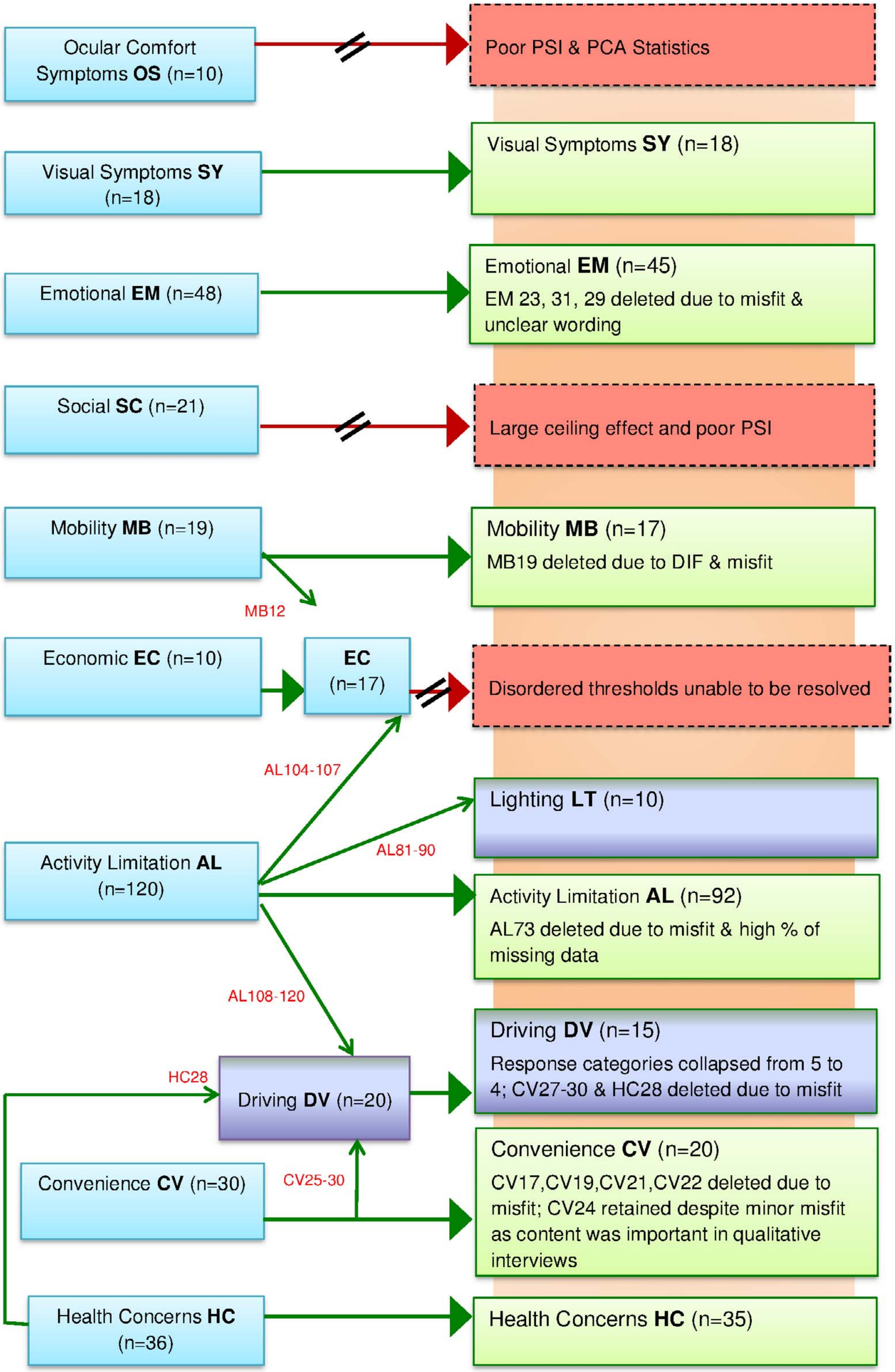
Ocular Comfort Symptoms OS (n=10)
Poor PSI & PCA Statistics
Visual Symptoms SY (n=18)
Visual Symptoms SY (n=18)
Emotional EM (n=48)
Emotional EM (n=45)
EM 23, 31, 29 deleted due to misfit & unclear wording
Social SC (n=21)
Large ceiling effect and poor PSI
Mobility MB (n=19)
Mobility MB (n=17)
MB19 deleted due to DIF & misfit
MB12
Economic EC (n=10)
EC (n=17)
Disordered thresholds unable to be resolved
AL104-107
Lighting LT (n=10)
AL81-90
Activity Limitation AL (n=120)
Activity Limitation AL (n=92)
AL73 deleted due to misfit & high % of missing data
AL108-120
Driving DV (n=15)
Response categories collapsed from 5 to 4; CV27-30 & HC28 deleted due to misfit
HC28
Driving DV (n=20)
Convenience CV (n=30)
CV25-30
Convenience CV (n=20)
CV17,CV19,CV21,CV22 deleted due to misfit; CV24 retained despite minor misfit as content was important in qualitative interviews
Health Concerns HC (n=36)
Health Concerns HC (n=35)

FIGURE 1. Flowchart explaining the iterative process of item bank modification following psychometric assessment. The flowchart shows that there were originally nine item banks (*left, blue boxes*), of which eight were retained (*right, green boxes*). Two new item banks were formed (*right, purple boxes*) from items rehomed from the original Activity Limitation, Convenience, Mobility, and Health Concerns item banks. Three item banks had suboptimal psychometric properties and were not further pursued in CAT testing.

approach, which addresses the shortcomings associated with short-form paper-pencil questionnaires.[12,17,19] For example, our CAT simulation tests indicated that only six to seven items were needed to gain measurement of the emotional impact of DR/DME with a high degree of precision. Such brevity may reduce test takers' burden and increase motivation because items are tailored to their individual situation.[12] Brief questionnaires are also highly valued in clinical settings where clinicians may have very little time to quantify patients' QoL using a PRO. Moreover, as CAT automates scoring, results can

TABLE 2. CAT Simulation Results for the DR Item Banks

Domain	No. of Items Available for CAT	Average No. of Items Used by CAT	Correlation Between CAT and Item Bank Person Measures	Mean SEM (sem.CAT)
SEM 0.387 (high precision, for use in clinical trials)				
Visual Symptoms	18	7.68	0.97	0.38
Activity Limitation	92	5.27	0.94	0.37
Mobility	17	9.14	0.97	0.38
Emotional	45	6.50	0.94	0.38
Health Concerns	35	5.82	0.95	0.37
Convenience	20	6.67	0.96	0.38
Driving	15	8.70	0.98	0.38
Lighting	10	8.05	0.99	0.38
Total	252*	57.8 (22.9%)†	0.96	0.38
SEM 0.521 (moderate precision, for use in the clinic setting)				
Visual Symptoms	18	3.28	0.88	0.50
Activity Limitation	92	3.06	0.88	0.47
Mobility	17	4.45	0.91	0.50
Emotional	45	3.26	0.89	0.50
Health Concerns	35	3.12	0.89	0.48
Convenience	20	3.30	0.90	0.49
Driving	15	3.87	0.91	0.50
Lighting	10	4.12	0.93	0.49
Total	252*	28.5 (11.3%)†	0.90	0.49

* The original item bank had 314 items; $n = 252$ excludes the Ocular Surface Symptoms 10-item scale, the Social 20-item scale, the Economic 15-item scale, and 17 other items from various scales that were deleted for gross misfit, differential item functioning, or extremely high missing data.

† Total no. of items needed on average if all eight item banks administered.

be integrated promptly into patient feedback and treatment,[40,41] which aligns well with the recent push to incorporate collection of PRO data in clinical care.[7]

As a result of these benefits, item banking and CAT are gaining momentum worldwide in health-related research, and item banks have been developed for cancer-related fatigue[40,42] arthritis[43], paediatrics,[44] spinal cord injury,[45] and low vision,[46] among others. Our rigorous methodology for the development and calibration of item banks is similar to that used by the PROMIS (Patient Reported Outcomes Measurement Information System) group, albeit with different IRT models for psychometric analysis and calibration (graded response model versus Rasch analysis). For example, Tulsky et al. recently provided a comprehensive description of the development of 14 unidimensional spinal cord injury QoL (SCI-QOL) item banks across physical, emotional, and social health domains, from qualitative content development to psychometric testing and item bank calibration and finally to CAT evaluation.[47] Item banks for vision-related activity limitation, symptoms and QoL have also recently been developed by Pesudovs et al. in cataract patients[33]; however, most of the items relate to activity limitation and the QoL bank requires additional content to become a comprehensive measure.[33] Moreover, because the item banks were formed by pooling items from 19 extant vision-related activity limitation questionnaires rather than developing content anew from qualitative work and were validated only in cataract patients, their applicability to DR/DME patients is likely to be limited.

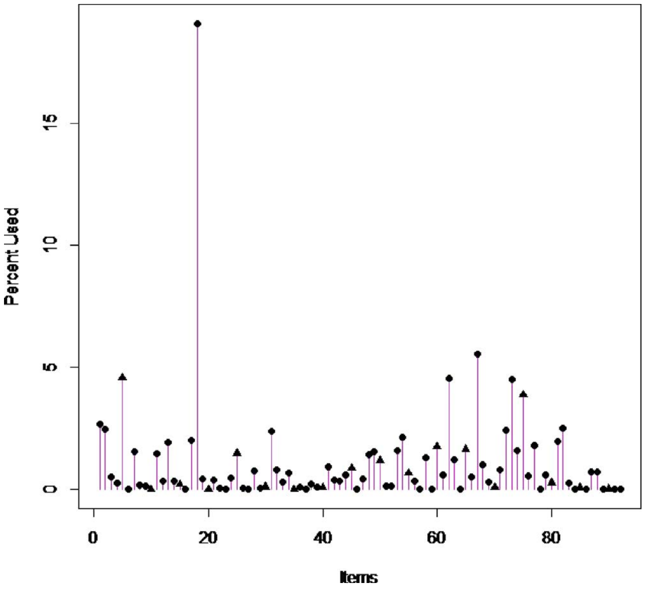

FIGURE 2. Item usage statistics for the 92-item Activity Limitation Computerized Adaptive Testing simulation when SEM is 0.387. The 18th item (AL18), "Reading the numbers on the front of a bus" was used 100% of the time as it was the starting item. Five other items were used approximately 5% of the time, with the remaining items used between 0% and 5% of the time.

One strength of our study is the large proportion of participants with severe DR/DME, which is often lacking in related studies. Another is the sophisticated psychometric techniques used to ensure that item banks were calibrated without LID and to address minor item misfit without having to delete numerous items unnecessarily. Similarly, efforts were made to rehome groups of items contributing to multidimensionality into related item banks. However, a few limitations should be noted. For example, nearly two-thirds of the sample was male, which may infer a sex bias in the results, although this may simply reflect the higher prevalence of diabetes and diabetic complications in men. The relatively high correlations between some of our measures may support a multidimensional latent structure underlying DR-specific QoL. However, our aim was to produce unidimensional measurement tools that provide users with the ability to administer selected scales for a given purpose. However, given that multidimensional IRT and bifactor models are available for use in item banking and CAT,[48–50] the potential to form a multidimensional DR/DME

QoL item bank should be considered. For practical reasons, both face-to-face and phone interviews were conducted. Ideally, data collection should be restricted to a single method since mode of administration may affect data quality.[51] However, when we stratified the sample by mode of administration ($n = 268$ face to face; $n = 246$ phone), we found very similar psychometric properties between the two groups and no DIF for mode of interview (Supplementary Tables S14–S24). In addition, the long interview duration may have reduced data quality. However, participants were given opportunities to rest and complete the questions over two sessions if desired. Finally, as our cutoff for detecting LID was 0.3 rather than the more commonly accepted value of 0.2, we may have missed noteworthy LID, thus artificially inflating reliability and precision estimates. Similarly, we used a conservative cutoff for detecting DIF (>1.0) and therefore may have missed detecting and accounting for moderate to large DIF for some items.

Our item banks will be validated in a future study using CAT by assessing completion time and average number of items administered; content range coverage and test precision; temporal reliability; and criterion, convergent, and divergent validity. We are currently developing an online testing platform which can be implemented through various platforms such as an iPad and can provide real-time scoring and recording of data. Given the rapidly increasing prevalence of diabetes and associated complications worldwide, the development of our item banks is timely. As recent advancements in treatments for DR and DME such as anti-VEGF therapy continue to gain momentum, a comprehensive PRO will be invaluable for use in clinical trials to compare the impact of novel treatment therapies from the patient's perspective. Similarly, the item banks will allow researchers and policy planners to effectively design and evaluate rehabilitation programs for DR/DME, and may also assist in identifying patients with specific QoL issues for timely referral for counselling or assistive services.

In summary, our eight item banks enable robust and comprehensive assessment of DR-specific QoL. CAT simulation results indicate that only a small number of items are required to obtain precise measurement of each QoL construct. Once validation using CAT is complete, our item banks offer clinicians and researchers the means to efficiently and accurately assess the impact of DR/DME and novel treatment therapies on eight aspects of QoL. In particular, relatively novel constructs such as Mobility, Health Concerns, Convenience, Driving, and Lighting can be explored in patients with DR/DME for the first time.

Acknowledgments

The authors thank Mike Linacre, Alan Tennant, and John Barnard for advice and support on the Rasch analysis and computerized adaptive testing simulations conducted in this study. The authors also thank the anonymous reviewers of this manuscript who provided comments and edits that enhanced the quality of this work.

Supported by National Health and Medical Research Council Centre for Clinical Research Excellence (CCRE) Grant 529923 (Translational Clinical Research in Major Eye Diseases); CCRE Diabetes; Novartis Pharmaceuticals Australia Grant CRFB002DAU09T; and the Royal Victorian Eye and Ear Hospital. EKF is funded by Australian National Health and Medical Research Council (NHMRC) Early Career Fellowship 1072987. GR is funded by NHMRC Career Development Award 1061801 The Centre for Eye Research Australia receives Operational Infrastructure Support from the Victorian Government.

The sponsor or funding organization had no role in the design or conduct of this research

Disclosure: **E.K. Fenwick**, None; **J. Khadka**, None; **K. Pesudovs**, None; **G. Rees**, 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. Wong T, Klein K. The epidemiology of eye diseases in diabetes. In: Ekoé J, Rewers M, Williams R, Zimmet P, eds. *The Epidemiology of Diabetes Mellitus.* 2nd ed. Oxford, UK: John Wiley and Sons; 2008:475–497.
3. 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.
4. Fenwick E, Rees G, Pesudovs K, et al. Social and emotional impact of diabetic retinopathy: a review. *Clin Experiment Ophthalmol*. 2012;40:27–38.
5. Penn JS, Madan A, Caldwell RB, Bartoli M, Caldwell RW, Hartnett ME. Vascular endothelial growth factor in eye disease. *Prog Retin Eye Res*. 2008;27:331–371.
6. Loftus JV, Sultan MB, Pleil AM. Changes in vision- and health-related quality of life in patients with diabetic macular edema treated with pegaptanib sodium or sham. *Invest Ophthalmol Vis Sci*. 2011;52:7498–7505.
7. Basch E. Patient-reported outcomes: harnessing patients' voices to improve clinical care. *N Eng J Med*. 2017;376: 105–108.
8. 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;21:1305–1314.
9. Mazhar K, Varma R, Choudhury F, McKean-Cowdin R, Shtir CJ, Azen SP. Severity of diabetic retinopathy and health-related quality of life: the Los Angeles Latino Eye Study. *Ophthalmology*. 2011;118:649–655.
10. Brose L, Bradley C. Psychometric development of the individualized retinopathy-dependent quality of life questionnaire (RetDQoL). *Value Health*. 2009;14:740–754.
11. Lamoureux EL, Fenwick E, Xie J, et al. Methodology and early findings of the Diabetes Management Project: a cohort study investigating the barriers to optimal diabetes care in diabetic patients with and without diabetic retinopathy. *Clin Exp Ophthalmol*. 2012;40:73–82.
12. Gibbons RD, Weiss DJ, Kupfer DJ, et al. Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatr Serv*. 2008;59:361–368.
13. Turner RR, Quittner AL, Parasuraman BM, Kallich JD, Cleeland CS. Patient-reported outcomes: instrument development and selection issues. *Value Health*. 2007;10(suppl 2):S86–S93.
14. WHOQOL Group. *Measuring Quality of Life*. Geneva: The World Health Organization; 1997.
15. Khadka J, McAlinden C, Pesudovs K. Quality assessment of ophthalmic questionnaires: review and recommendations. *Optom Vis Sci*. 2013;90:720–744.
16. Paz SH, Slotkin J, McKean-Cowdin R, et al. Development of a vision-targeted health-related quality of life item measure. *Qual Life Res*. 2013;22:2477–2487.
17. 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.
18. Gershon RC. Computer adaptive testing. *J Appl Measurement*. 2005;6:109–127.
19. Revicki DA, Cella DF. Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing. *Qual Life Res*. 1997;6:595–600.

20. Fenwick EK, Pesudovs K, Khadka J, Rees G, Wong TY, Lamoureux EL. Evaluation of item candidates for a diabetic retinopathy quality of life item bank. *Qual Life Res*. 2013;22: 1851–1858.

21. 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.

22. American Academy of Opthalmology. *International Clinical Classification of Diabetic Retinopathy Severity of Macular Edema*. San Francisco: American Academy of Ophthalmology; 2002.

23. Lamoureux EL, Chong EW, Thumboo J, et al. Vision impairment, ocular conditions, and vision-specific function: the Singapore Malay Eye Study. *Ophthalmology*. 2008;115: 1973–1981.

24. Linacre JM. *A User's Guide to Winsteps/Ministeps Rasch-Model Computer Programs. Program Manual 4.0.0*. Chicago, IL: MESA Press; 2017.

25. Andrich D. A rating scale formulation for ordered response categories. *Psychometrika*. 1978;43:561–573.

26. Boone W, Staver J, Yale M. *Rasch Analysis in the Human Sciences*. Dordrecht: Springer; 2014.

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(Pt 1):1–18.

28. Mallinson T. Why measurement matters for measuring patient vision outcomes. *Optom Vis Sci*. 2007;84:675–682.

29. Lamoureux E, Pesudovs K. Vision-specific quality-of-life research: a need to improve the quality. *Am J Ophthalmol*. 2011;151:195–197.

30. Khadka J, Fenwick E, Lamoureux E, Pesudovs K. Item banking enables stand-alone measurement of driving ability from an activity limitations item set. *Optom Vis Sci*. 2015;93:1502–1512.

31. Pesudovs K, Burr JM, Harley C, Elliott DB. The development, assessment, and selection of questionnaires. *Optom Vis Sci*. 2007;84:663–674.

32. Linacre JM. What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Trans*. 2002;16:878.

33. Pesudovs K. Item banking: a generational change in patient-reported outcome measurement. *Optom Vis Sci*. 2010;87:1–9.

34. Baghaei P. Local dependency and Rasch measures. *Rasch Measurement Trans*. 2008;21:1105–1106.

35. Chen SK, Cook KF. simpolycat: an SAS program for conducting CAT simulation based on polytomous IRT models. *Behav Res Methods*. 2009;41:499–506.

36. Choi SW. Firestar: computerized adaptive testing simulation program for polytomous item response theory models. *Appl Psychol Measurement*. 2009;33:644–645.

37. Latimer S, Meade T, Tennant A. Development of item bank to measure deliberate self-harm behaviours: facilitating tailored scales and computer adaptive testing for specific research and clinical purposes. *Psychiat Res*. 2014;217:240–247.

38. Forkmann T, Kroehne U, Wirtz M, et al. Adaptive screening for depression: recalibration of an item bank for the assessment of depression in persons with mental and somatic diseases and evaluation in a simulated computer-adaptive test environment. *J Psychosomatic Res*. 2013;75:437–443.

39. Haley SM, Ni P, Jette AM, et al. Replenishing a computerized adaptive test of patient-reported daily activity functioning. *Qual Life Res*. 2009;18:461–471.

40. Lai JS, Cella D, Chang CH, Bode RK, Heinemann AW. 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.

41. 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.

42. Lai JS, Cella D, Dineen K, et al. An item bank was created to improve the measurement of cancer-related fatigue. *J Clin Epidemiol*. 2005;58:190–197.

43. Kopec JA, Sayre EC, Davis AM, et al. Assessment of health-related quality of life in arthritis: conceptualization and development of five item banks using item response theory. *Health Quality Life Outcomes*. 2006;4:33.

44. Walsh TR, Irwin DE, Meier A, Varni JW, DeWalt DA. The use of focus groups in the development of the PROMIS pediatrics item bank. *Qual Life Res*. 2008;17:725–735.

45. Jette AM, Slavin MD, Ni P, et al. Development and initial evaluation of the SCI-FI/AT. *J Spinal Cord Med*. 2015;38:409–418.

46. Massof RW, Ahmadian L, Grover LL, et al. The Activity Inventory: an adaptive visual function questionnaire. *Optom Vis Sci*. 2007;84:763–774.

47. Tulsky DS, Kisala PA, Victorson D, et al. Methodology for the development and calibration of the SCI-QOL item banks. *J Spinal Cord Med*. 2015;38:270–287.

48. Stochl J, Bohnke JR, Pickett KE, Croudace TJ. Computerized adaptive testing of population psychological distress: simulation-based evaluation of GHQ-30. *Social Psychiatr Psychiatric Epidemiol*. 2016;51:895–906.

49. Croudace TJ, Bohnke JR. Item bank measurement of depression: will one dimension work? *J Clin Epidemiol*. 2014;67:4–6.

50. Reise SP, Morizot J, Hays RD. The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Qual Life Res*. 2007;16(suppl 1):19–31.

51. Bowling A. Mode of questionnaire administration can have serious effects on data quality. *J Public Health (Oxf)*. 2005;27: 281–291.