tvst

Article

Calibration of the Dutch EyeQ to Measure Vision Related Quality of Life in Patients With Exudative Retinal Diseases

T. Petra Rausch-Koster[1,2], Michiel A. J. Luijten[3,4], F. D. Verbraak[1], Ger H. M. B. van Rens[1], and Ruth M. A. van Nispen[1]

[1] Amsterdam UMC, Vrije Universiteit Amsterdam, Ophthalmology, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands
[2] Bergman Clinics, Department of Ophthalmology, The Netherlands
[3] Emma Children's Hospital, Amsterdam UMC, University of Amsterdam, Child and Adolescent Psychiatry & Psychosocial Care, Amsterdam Reproduction and Development, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands
[4] Department of Epidemiology and Data Science, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands

Correspondence: T. Petra Rausch-Koster, Amsterdam UMC, Location VUmc, Ophthalmology PK4X, PO Box 7700, 1000 SN Amsterdam, The Netherlands. e-mail: t.p.rauschkoster@amsterdamumc.nl

Received: September 10, 2021
Accepted: February 24, 2022
Published: April 5, 2022

Keywords: vision-related quality of life; item-bank; development; calibration; exudative retinal diseases

Citation: Rausch-Koster TP, Luijten MAJ, Verbraak FD, van Rens GHMB, van Nispen RMA. Calibration of the dutch EyeQ to measure vision related quality of life in patients with exudative retinal diseases. Transl Vis Sci Technol. 2022;11(4):5, https://doi.org/10.1167/tvst.11.4.5

Purpose: This study aims to develop an item-bank to measure vision-related quality of life (Vr-QoL) and subsequently calibrate this set of items.

Methods: Three Vr-QoL instruments were searched for suitable items to be added in the EyeQ. Patients who received antivascular endothelial growth factor treatment for various retinal diseases involving macular edema were included in the study and completed the 47-item EyeQ. Item response theory (IRT) was used to calibrate the EyeQ items, which was performed multiple times in subsets as a novel approach, containing 80% of the data. Differential item functioning (DIF) was evaluated for various variables.

Results: Responses of 704 patients were used in analysis. One item violated the local independence IRT-assumption and showed a high percentage of missing values, after which this item was deleted from the item-bank. The data of the five subsets fitted the graded response model adequately, and no DIF was detected for items between subsets, after which mean item parameters were calculated. Item fit statistics were found to be good. DIF was detected for gender, age, and administration mode by the patient (independently vs. with help), this involved three items, which all showed negligible impact on total scores.

Conclusions: Because of separate calibrations of the EyeQ in multiple subsets, a high robustness of item parameters is expected.

Translational Relevance: The calibrated EyeQ can now be used for the assessment of Vr-QoL in patients suffering from exudative retinal diseases and is promising for use as a computer adaptive test.

Introduction

The prevalence of nonrefractive vision impairment, according to the WHO's definition of having a best corrected visual acuity < 20/60, in European countries is approximately 1% to 2%.[1] The major causes of nonrefractive low vision worldwide are cataract, age-related macular degeneration, glaucoma, and diabetic retinopathy.[2] because of the pathophysiology of these eye diseases, individuals over age 50 years are more frequently affected. For patients suffering from macular edema caused by underlying retinal diseases, such as neovascular age-related macular degeneration (nAMD), retinal vein occlusion, or diabetic retinopathy (DR), intraocular injections with anti-vascular endothelial growth factor (anti-VEGF) are often beneficial: In most treatment-naïve patients anti-VEGF leads to a stabilization (about 50%) or improvement (≥10 letters gain (two lines on EDTRS chart)) (>30%) of their vision after three years' follow-up. However, for about 15% of

Copyright 2022 The Authors

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

these patients anti-VEGF is less effective, resulting in reduced vision (≥10-letter loss).[3] Eventually, the loss of vision will cause limitations in physical functioning, daily activities and might have impact on the quality of life.[4–7]

Evaluating patients' disabilities in daily activities and vision-related quality of life has become more important in ophthalmology. These outcomes are from the patient's own perspective and therefore of direct relevance to them. Furthermore, patient-reported outcome measures (PROMs) might help clinicians in their communication toward patients.[8] Additionally, the assessment of quality of life is increasingly introduced because of the interest of the government and health insurance companies to evaluate the quality of care.[9] Even though the added value and benefits of measuring and evaluating PROMs are clearly seen by patients, professionals and health care institutes, the implementation of measuring and evaluating PROMs in clinical practice and supporting the effort of the patient to periodically fill out the questionnaires is still a major challenge.

In the past, various PROMs were developed. To solve measuring and scoring problems regarding first-generation PROMs (in which equidistance between response categories and equal value of items is assumed) second-generation PROMs (in which Item Response Theory [IRT] is used to calibrate items and respondents on the same scale to provide a better scoring mechanic that takes the psychometric properties of items into account) were developed frequently.[10] Recently, as a proposed solution for the limitations of first- and second-generation PROMs in clinical practice (e.g., logistical, technical) and to reduce patient's burden filling out a PROM to a minimum, item-banks have been developed, which are collections of items across a disability spectrum.[11] These item-banks can be used to apply computerized adaptive testing (CAT) and is currently more frequently introduced in health care.[12–14] In contrast to long questionnaires including a broad range of items on the health continuum that all must be answered by the patient to accurately measure their ability, a CAT, or tailored test, selects the next question from the item-bank using an algorithm. The selection of the next item from the item-bank is based on the response option that the participant has chosen on the previous question: after each response the patient's summary score ("theta") is recalculated, and a next item is selected by the algorithm. The CAT will continue selecting items depending on which administration rules have been determined. Stopping criteria that can be considered are length, precision, classification, or information, and combinations of these criteria are also a possible solution to optimize the performance of the CAT.[15] A significant advantage of a CAT is that the level of the patient's ability can be estimated very precisely, requiring a considerably smaller number of items. This will considerably reduce time and effort, as well as frustration and careless responses[16] caused by administration fatigue. Two recent examples in ophthalmology are the Impact of Vision Impairment–Computer Adaptive Test (IVI-CAT) and the Diabetic Retinopathy and Macular Edema Quality-of-Life (DR/DME) QoL item-banks.[17–19] The first study focused on developing a CAT based on the 28-item Impact of Vision Impairment Profile (IVI), whereas the other, the DR/DME QoL item-bank, contains 287 items, categorized under 10 domains, each responsible for a separate item-bank. Based on the outcome of interest, a choice can be made regarding which domains are used. The average amount of items that was required to estimate a person level of disability was approximately seven items per item-bank.

In previous research we translated the IVI forward from English into Dutch and backward twice and evaluated its content validity by performing cognitive debriefing interviews in Dutch patients (The Netherlands) who receive intraocular injections with anti-VEGF for exudative retinal diseases.[20] This led to some adaptations and the necessity to expand the IVI with other relevant items for the development of the Dutch EyeQ item-bank. The purpose of the current study is to calibrate the EyeQ item-bank for measuring vision-related quality of life (Vr-QoL) as an important step in the development of a CAT.

Methods

The EyeQ

The EyeQ is based on the 28-item Dutch version IVI.[20] As a broad range of items on the construct continuum is preferred to develop a CAT that can provide precise measurement for the wide range of ability levels,[15] we investigated the content of the Dutch versions of the low-vision quality-of-life questionnaire and the National Eye Institute Visual Functioning Questionnaire 25.[21–23] Most items appeared to be similar to the items of the IVI, but we found 19 items that had relevant unique content, based on results of our previous research.[20] Before adding items, we reformulated the specific items to fit into identical response categories. In total we selected 47 items for the EyeQ. The EyeQ items are scored using a four-point Likert scale with the following response categories: never (1), sometimes (2), often (3), and always (4)

because in previous research no comprehensibility problems or other issues arose regarding these response categories.[20] The response category "not applicable" was supplemented with "I don't do this for reasons other than my eyesight" and was treated as a missing value. The order of the 47 items was randomized using a random number generator to avoid possible effects of careless behavior or fatigue toward the end of the test, to the detriment of the same items. This resulted in 10 different versions of the EyeQ.

Study Design and Participants

The study protocol was approved by the Medical Ethics Committee of Amsterdam University Medical Centers and conducted according to the Declaration of Helsinki. The Medical Ethics Committee declared that the protocol did not fall under the scope of the Medical Research Involving Human Subjects Act (Dutch law).

Adults aged over 18 years who are diagnosed with macular edema caused by nAMD, DR, and retinal vein occlusion and currently receiving anti-VEGF injections in the Bergman Clinics eye hospitals were invited by letter. We explained the aim, procedure, and duration of the study, and we asked whether they would agree to participate. To create a reliable representation of the clinical variety of patients receiving anti-VEGF treatment in ophthalmic clinical practice, no restrictions for participation were made based on visual acuity or the duration of treatment with intravitreal anti-VEGF. Patients had to have adequate knowledge and understanding of the Dutch language. All patients signed written informed consent and subsequently were included in the study. Participants were given the possibility to fill out the questionnaire via an online form, by a printed copy, sent to their address, or by telephone. In addition, participants were asked to complete various socio-demographic questions, questions regarding comorbidities and a generic health QoL questionnaire; EuroQol 5 Dimensions (EQ5D-3L). The EQ5D is a commonly used generic health status measurement, and it evaluates five dimensions of functional impairment including mobility, self-care, usual activities, pain/discomfort, and anxiety/depression with a three-level response option.[24,25] Clinical characteristics, regarding ocular comorbidities, visual acuity, treated site anti-VEGF, and diagnosis for which anti-VEGF treatment was received were manually searched in digital patient records.

Statistical Analysis

All statistical analyses were performed using SPSS (version 26.0)[26] and R using the ltm package.[27] Patient characteristics were analyzed using descriptive statistics. Before the calibration of the EyeQ using IRT (i.e. the graded response model [GRM]), we investigated the response percentages on the items. Because items with high missing rates are indicative for less reliable measurement properties, we removed items if the missing proportion was higher than 50%, whereas percentages between 30% and 50% were flagged for potential removal, these limits were arbitrarily chosen. Additionally, participants with missing responses above 25% were removed. After filtering the high proportioned missing values in items and participants, we checked the distribution of responses over response categories for possible floor- and ceiling effects and possible conjunction of categories in order to create a more equal distribution. Item-pairs with >0.75 inter-item collinearity were flagged, because this could be a sign of similarity and therefore be considered redundant.

Important assumptions that are required for IRT modeling were checked:

Unidimensionality of the Construct

Unidimensionality of the construct (i.e., all items representing a single latent trait) was examined by the output of a confirmatory bi-factor analysis and an explanatory factor analysis (EFA), the principal component analysis. A bi-factor analysis tests the item loadings on other factors in addition to the general factor. If omega hierarchical (ω_H) is >0.80, it is accepted to consider total scores as essentially unidimensional. An explained common variance attributed by the general factor >0.70 is indicative to assume that the factor loadings obtained from a unidimensional model might approximate well the factor loadings obtained from a bi-factor model.[28] An EFA tests the amount of explained variance by the first factor. EFA was performed for a one-factor and a two-factor model. Thereafter the ratio of explained variance by factor one and factor two was determined that should have a minimum of four to assume unidimensionality.[29] In a subjective approach we examined the item loadings on the first factor by evaluating the eigenvalues in a scree plot. Additionally, we used the acceleration factor as a nongraphical alternative, which determines the coordinate where the slope of the curve changes most abruptly.[30,31]

Local Independence

This states that every item on a measure, given a particular latent trait value (theta), is statistically independent of responses to all other items on that measure.[31–33] Values of item residuals above 0.25 are considered as items violating local independence.

Monotonicity

This assumption implies that as a respondent moves to a higher level of the latent trait (i.e., increased disability), the probability of endorsement of a successive threshold never decreases. We used Mokken scale analysis for the assessment of manifest monotonicity by examining graphs.[34,35] Additionally, Loevinger H coefficients of the items were calculated as a function of the Guttman errors between pairs of items to examine their scalability,[36] where values below 0.30 were considered as unsatisfactory.[35]

As a novel approach in ophthalmology, we used the full dataset and also created five random subsets, each consisting of 80% of the data. This way, we assessed to what extent the estimates varied across subsets due to a possible selection bias (e.g., for age or gender) and subsequently calculate mean estimates. A random number generator was used to create five reproducible datasets. The GRM estimates a discrimination (α) parameter and location (i.e. thresholds) (β) parameters for each response-category of the item. The discrimination parameter reflects how well the item can distinguish differences in patient's level of ability, where a higher discrimination parameter refers to a higher separative power. The item thresholds parameters locate the item response categories on the disability continuum. Item parameters were estimated using a marginal maximum likelihood approach as it easily handles perfect response patterns and is applicable in polytomous IRT models.[37] The assessment whether the data fits the GRM was performed by comparing the full GRM model fit to a constrained GRM model using the marginal maximum likelihood estimates with a Likelihood-Ratio test. The constrained model, which is similar to the Rasch model, does not allow the discrimination parameter to vary between items. This procedure was repeated for all five subsets and full data. Differential item functioning (DIF) was inspected using an iterative hybrid ordinal logistic regression analyses to assess differences in probabilities of selecting a certain item response between subsets and full data. The Likelihood-Ratio χ^2 test at a level of 0.01 was used as detection criterion for both uniform DIF (DIF that is proportional across levels of the underlying latent trait) and nonuniform DIF (DIF that is nonproportional across levels of the underlying latent trait).[33,38] In case of significant DIF, McFadden's pseudo R^2 was used to measure change in DIF magnitude, where a 2% change was considered as critical value.[39] Finally, mean GRM estimates were calculated out of five subsets to create the pooled dataset.

Subsequently, item goodness-of-fit was evaluated using the generalized S-X^2 index, which is used for polytomous items to compare observed and predicted response proportions,[40–42] and item and test information were assessed. Item information refers to the information content of an item in relation to the total test information and therefore, item information is a representative for measurement precision or reliability.[31] Items contributing <3/4 of ideal item information across the disability continuum (based on total test information) were (arbitrarily) considered for elimination. However, we acted reservedly in actual deletion of items that contributed little to the test information, as a balanced item-bank should contain items that cover the whole range of ability levels.[15] Little informative items were identified by evaluating item information curves and category response curves. The range of theta over which the item is most informative is visible in an item information curve.

DIF was inspected to assess whether participants with different characteristics, having the same level of disability, have equal probabilities of selecting a specific response category.[33,38] Again, the Likelihood-Ratio χ^2 test and McFadden's pseudo R^2 were used to detect DIF and to measure the change in DIF magnitude, respectively, using the same detection criteria as mentioned above. In addition, the impact of DIF on test scores was inspected by plotting test characteristic curves, which represent the relation between expected test scores on the y-axis and the thetas on the x-axis. Detection criteria for DIF were kept equal to DIF detection in subsets. DIF was evaluated for gender, nationality, visual acuity, age, diagnosis, civil status, EQ5D score, the number of nonocular comorbidities, administration mode (independently vs. with help), and completion method (paper vs. digital).

Results

Patient Characteristics

Patients (N = 3783) were invited, and 746 were willing to participate (response rate 19.7%), met the inclusion criteria, and gave their written informed consent. Seven hundred thirteen participants filled out the EyeQ. Nine patients with an excessive number of missing responses (>25%) were excluded from the analyses. Sociodemographic and clinical characteristics of the remaining 704 participants are summarized in Table 1.

Calibration of the EyeQ and Item Analyses

The confirmatory bi-factor analysis showed a ω_H and explained common variance of 0.85 and 0.78,

Table 1. Sociodemographic and Clinical Characteristics of Participants ($n = 704$)

Age in years, mean ± SD	76.2 ± 9.2
Sex, *n* (%)	
Male	342 (48.6)
Female	362 (51.4)
Nationality, *n* (%)	
Dutch	663 (94.2)
Other	41 (5.8)
Educational background in years, median (IQR)	11.0 (9.0–15.0)
Civil status, *n* (%)	
Single	235 (33.4)
Not single	469 (66.6)
EQ5D-3L scores, mean ± SD	0.86 ± 0.18
Administration mode questionnaires, *n* (%)	
Digital (online form)	454 (64.5)
Printed copy	231 (32.8)
Telephone	19 (2.7)
Nonocular comorbidity,[a] *n* (%)	
Lung disease[b]	64 (9.1)
Cardiovascular disease[c]	222 (31.5)
Diabetes Mellitus	138 (19.6)
Rheumatoid arthritis	37 (5.3)
Cancer	70 (9.9)
Psychiatric disorder	34 (4.8)
Other nonocular comorbidities	170 (24.1)
Number of nonocular comorbidities within person, *n* (%)	
Zero	254 (36.1)
One	261 (37.1)
Two or more	189 (26.8)
Ocular comorbidity,[d] *n* (%)	
Front segment[e]	198 (23.9)
Back segment[f]	266 (37.8)
Glaucoma	80 (11.4)
Amblyopia	10 (1.4)
Other	18 (2.6)
Number of ocular comorbidities within person,[d] *n* (%)	
Zero	262 (37.2)
One	339 (48.2)
Two	93 (13.2)
Three	10 (1.4)
Treated site anti-VEGF, *n* (%)	
Monocular	574 (81.5)
Binocular	128 (18.2)
Unknown	2 (0.3)
LogMAR visual acuity better eye, median (IQR)	0.10 (0.01 – 0.22)
Degree Vision Impairment,[g] *n* (%)	
No vision impairment: LogMAR ≤0.50 (≥20/60)	652 (92.6)
Low vision: LogMAR 0.51 – 1.29 (<20/60–>20/400)	45 (6.4)
Blind: ≥1.30 (≤20/400)	6 (0.8)
Diagnosis for which anti-VEGF treatment, *n* (%)	
nAMD	446 (63.4)
Diabetic macular edema	56 (8.0)
Cystoid macular edema secondary to retinal vein occlusion	175 (24.9)
Unknown	27 (3.8)

SD, standard deviation; IQR, interquartile range.

[a]Only if currently under treatment or monitored by a physician.

[b]Lung disease response categories were asthma, COPD, chronic bronchitis.

[c]Cardiovascular disease response categories were myocardial infarction or heart disease, stroke, vascular disease in vessels of the abdomen or lower extremity.

[d]Apart from eye diseases for which participant was included in study, present within one or both eyes.

[e]Front segment includes cornea and lens disorders.

[f]back segment includes retina and corpus vitreous disorders.

[g]In accordance with the ICD-11, where no to mild vision impairment (normal vision) is defined as visual acuity of the better eye equal to or better than 0.50 LogMAR (Snellen 6/18 or 20/60), moderate to severe vision impairment (low vision) is defined as visual acuity of the better eye worse than 0.50 LogMAR and equal to or better than 1.30 LogMAR (20/60 - 20/400), and blindness is defined as visual acuity of the better eye worse than 1.30 LogMAR (≤20/400).

respectively. Both values are supportive to assume unidimensionality.[28] The principal component analysis showed a variance of 49% that could be explained by the first factor, whereas the second factor contributed 4% of variance; thus the ratio explained by the first and second factor is 12.25, which is well above the required minimum of 4.[29] The scree plot and acceleration factor were also supportive for unidimensionality (Supplement 1). One item pair, CAT33 "Driving a car during the night" and CAT34 "Driving a car under difficult circumstances (bad weather, rush hour, etc.)," violated the local independence assumption with residuals above 0.25 and showed inter-item correlation >0.75. Mokken analysis showed that all items complied with monotonicity, and all Loevinger H coefficients were above the required 0.3, which indicated sufficient scalability. The internal consistency reliability coefficient (Cronbach's alpha) of the one-factor scale was 0.98. All response categories in all items were endorsed (Table 2); however, response categories "often" and "always" were chosen infrequently by the participants. To create a more equal distribution of the responses, these categories were collapsed. Finally, the already flagged CAT33 item was removed because of the high percentage of missing data (31.0%).

Calibration of the EyeQ Item-Bank

A Likelihood-Ratio test showed that the unconstrained GRM was preferred above the constrained model for the 46 items, which was tested for all five subsets (1 to 5) and the full dataset (1: LRT = 258.9, $P < 0.001$; 2: LRT = 263.2, $P < 0.001$; 3: LRT = 286.1, $P < 0.001$; 4: LRT = 253.0, $P < 0.001$; 5: 272.2, $P < 0.001$; Full data: LRT = 327.8, $P < 0.001$). The overall fit of the 46 items to the GRM model was adequate for all subsets (Table 3).

Differential Item Functioning Between Subsets With Full Data

No item was flagged for DIF in the subsets compared to the full dataset using the Likelihood-Ratio χ^2 test at a level of 0.01 (Table 4). Subsequently, item parameter means were calculated out of five subsets to get robust estimates. Item discrimination coefficients ranged from 1.17 to 2.86, with CAT3 "Going out, such as seeing cinema films, theater plays or sports events" showing highest item discrimination and CAT29 "Suffering from glare" the lowest. Item thresholds parameters ranged from −1.45 to 4.11. The total test information was 156.55 of which 97.45% fell within the range −4 to 4. The S-X^2 goodness-of-fit

Table 2. Distribution of Responses Over the Response Categories of the EyeQ Item-Bank

Item	Description of Item Content[a] The Items Were Formulated To Measure to What Extent the Eyesight Interfered With an Activity or Task, or to What Degree of Impact the Eyesight Had on Their Emotional Well-Being	Distribution of Responses (%) Over the Response Categories				
		Never (1)	Sometimes (2)	Often (3)	Always (4)	Missing or N/A
CAT1	Ability to see and enjoy TV-programs	51.7	33.6	8.7	6.0	1.5
CAT2	Recreational activities, such as biking, walking or other activities	70.5	18.9	5.5	5.1	8.7
CAT3	Going out, such as seeing cinema films, theater plays, or sports events	66.7	21.0	6.4	5.9	21.2
CAT4	Shopping	64.3	23.4	7.3	5.0	4.2
CAT5	Visiting friends and family	83.2	11.5	2.8	2.5	3.8
CAT6	Meeting people	78.5	16.1	3.7	1.7	1.7
CAT7	Recognizing people at the opposite side of the street	50.2	28.1	10.6	11.1	1.7
CAT8	Recognizing people at the opposite side of the room	78.3	12.8	4.1	4.8	1.3
CAT9	Looking after appearance	71.0	20.7	5.3	3,0	2,0
CAT10	Activities or hobbies that require good near vision	39.5	38.3	14.7	7.5	2.9
CAT11	Operating household appliances	72.1	21.8	4.2	2.0	2.0
CAT12	Needing (more) help from others because of eyesight	63.7	27.9	6.3	2.1	1.4
CAT13	Mobility outdoors	59.9	27.2	9.3	3.6	2.4
CAT14	Going carefully to avoid falling or tripping	53.3	29.7	9.6	7.5	3.1
CAT15	Travelling or using transport (bus and train)	75.4	13.6	4.5	6.5	16.1
CAT16	Going down steps, stairs or pavements in good lighting conditions	58.7	27.8	8.0	5.5	3.6
CAT17	Going down steps, stairs, or pavements in poor lighting conditions or at night	38.5	36.4	14.5	10.6	6.3
CAT18	Reading ordinary size print	46.3	30.7	12.6	10.3	2.2
CAT19	Reading large text	80.0	12.5	4.2	3.3	2.7
CAT20	Reading labels or instructions	30.1	37.5	17.9	14.5	2.2
CAT21	Reading colored text or text on a colored background	40.1	34.7	14.0	11.2	2.7
CAT22	Writing	63.1	21.1	8.5	7.3	3.8
CAT23	Mobility indoors under poor lighting conditions	59.1	30.3	5.7	4.9	1.8
CAT24	Mobility indoors under normal lighting conditions	76.2	19.1	3.3	1.4	1.5
CAT25	Depth perception and estimating the distance of objects	55.3	31.7	10.5	2.4	0.1
CAT26	Seeing street signs or other signs from a distance	3.0	40.7	16.4	10.8	0.1
CAT27	Seeing moving objects	63.0	27.9	7.3	1.8	0.3
CAT28	Suffering from tiredness of the eyes	42.9	42.1	12.3	2.7	0.0
CAT29	Suffering from glare	20.1	46.2	23.7	10.0	0.1
CAT30	Getting the right amount of light to see properly	29.4	44.0	20.1	6.6	0.1
CAT31	Seeing how people react to what you say	78.1	14.9	4.5	2.5	0.3
CAT32	Driving a car during the day in a well-known environment	78.2	11.3	2.3	8.3	25.2
CAT33	Driving a car during the night	34.4	32.7	11.8	21.1	31.0
CAT34	Driving a car under difficult circumstances (bad weather, rush hour, etc.)	37.9	39.0	8.9	14.2	27.8
CAT35	Feeling worried or concerned about your safety at home	85.6	12.8	1.0	0.7	0.0
CAT36	Feeling worried or concerned about spilling food/drinks, dropping/breaking things	81.0	13.8	3.9	1.3	0.4
CAT37	Feeling worried or concerned about safety outdoors	75.5	18.9	3.7	2.0	0.0
CAT38	Feeling worried about coping with everyday life	67.7	24.1	6.3	1.8	0.0
CAT39	Feeling worried about eyesight	25.8	54.1	15.7	4.4	0.1
CAT40	Caused eyesight stop doing the things you wanted to do	62.8	28.9	7.2	1.1	0.1
CAT41	Interfered eyesight life in general	53.9	33.6	8.7	3.8	0.3
CAT42	Feeling embarrassed	91.0	7.4	1.3	0.3	0.1
CAT43	Feeling frustrated or annoyed	56.7	29.7	10.4	3.2	0.3
CAT44	Feeling lonely or isolated	84.0	12.6	2.8	0.6	0.0
CAT45	Feeling sad	62.3	30.2	6.8	0.7	0.3
CAT46	Feeling like a burden	79.5	16.7	3.0	0.8	0.3
CAT47	Using a mobile phone or smartphone	64.2	23.7	8.2	3.9	5.5

[a]Item content description in this table is not representing a formal translation of the Dutch EyeQ item-bank (Dutch EyeQ item-bank is available upon request).

index of the items ranged from 14.25 to 60.1. No items showed a significant S-X^2 value. Item information ranged from 2.11 to 5.05. Four items (CAT19 "Reading large text," CAT28 "Suffering from tiredness of the eyes," CAT29 "Suffering from glare" and CAT45 "Feeling sad") contributed less than 2.55 (75% of the ideal item information [3.40] based on the total test information (156.55)). Nevertheless, we decided not to

Table 3. Overall Fit Indices of 46 Items to the GRM Model of Five Subsets

Dataset	RMSEA[a]	SRMR[b]	CFI[c]	TLI[d]
Subset 1	0.031	0.064	0.995	0.995
Subset 2	0.037	0.078	0.994	0.994
Subset 3	0.031	0.079	0.996	0.996
Subset 4	0.026	0.063	0.997	0.997
Subset 5	0.029	0.081	0.997	0.996
Full data	0.035	0.071	0.995	0.995

[a]Root mean square error of approximation (values ≤0.06 represent good fit) (unscaled).

[b]Standardized root mean square of residuals (values ≤0.08 represent good fit).

[c]Comparative fit index (values ≥0.95 represent good fit) (unscaled).

[d]Tucker-Lewis index (values ≥0.95 represent good fit) (unscaled).[43]

remove these items from the scale, given their locations of item difficulty on the disability continuum and their content. The person-item map shows that the items are distributed over almost the entire disability continuum. The scores of the respondents matched the difficulty of the items reasonably well, however, there are relatively few items at the ends of the continuum (Fig. 1).

Differential Item Functioning

The following variables were dichotomized: nationality (Dutch vs. non-Dutch), visual acuity (<0.30 LogMAR versus ≥0.30 LogMAR), age (<75 years versus ≥75 years), diagnosis (nRMD versus other), civil status (living single versus not single), EQ5D score (perfect score versus other scores). A three-factor variable was created for number of nonocular comorbidities (no comorbidities vs. one comorbidity vs. two or more comorbidities) to obtain enough cases in response categories. DIF analysis for nationality was performed for only 27 items as after response category conjunction and dichotomizing this variable the minimum of five cases in each category could not be reached.

Table 5 and Figures 2 and 3 show the results of the DIF analyses. The items CAT9 "Looking after appearance" showed uniform DIF for gender, and CAT42 "Feeling embarrassed" showed uniform DIF for age. CAT35 "Feeling worried or concerned about your safety at home" showed nonuniform DIF for independently versus with help (proxy). For the item "Looking after appearance," the threshold parameters for females were lower than the thresholds for males, indicating that females endorse higher response categories at the same level of Vr-QoL score. For the item "Feeling embarrassed" the threshold parameter was lower for ≤75 years of age than >75 years of age, indicating that younger patients endorse higher response categories at the same level of Vr-QoL score. Furthermore, item "Feeling worried or concerned about your safety at home" showed a lower threshold for independently completed questionnaires, indicating patients filling out the Vr-QoL questionnaire independently, endorse higher response categories at the same level of Vr-QoL score. This effect was nonproportional across the levels of the trait: the difference between independent versus with help (proxy) widened with higher levels of theta.

Discussion

This study describes the development of the EyeQ item-bank and its calibration. In addition DIF was investigated for several subgroups. The EyeQ is a PROM which aims to measure Vr-QoL in patients having exudative retinal diseases. The content of the EyeQ is based on three instruments measuring Vr-QoL to provide an extensive range of items covering the whole disability trait, which is preferable for a CAT. This new EyeQ item-bank now also covers domains that were reported as under-represented in previous qualitative research on the content validity of the Dutch-IVI.[20] Despite the fact that the EyeQ was developed mainly based on items originating from other instruments, we consider the content of this instrument to be new, as various questions have been rewritten to be applicable to the response categories.

A slightly high inter-item correlation was found between CAT33 "driving a car during the night" and CAT34 "driving a car under difficult circumstances" (0.77), with residuals of 0.33. Although this is not a severe threat this could influence the IRT parameter estimates and could pose a problem in the construction of the scale.[29] Even when the instrument is implemented as a CAT, the estimation of the level of disability will be inaccurate. In addition, the high percentage of missing values of CAT33 warranted removal of this item.

The final EyeQ contains 46 items with difficulties across the disability trait, however there are more items that are targeted for patients with a higher level of disability. In our study we included patients with a relatively low level of disability; however, the instrument is also likely to be suitable for patients with a higher level of disability as at the higher levels of theta, the EyeQ contains several items applicable

Table 4. Mean GRM Item Parameters of the EyeQ With Standard Deviation, Item Information, and Fit Statistics of the Full Dataset

Item	Description of Item Content the Items Were Formulated To Measure to What Extent the Eyesight Interfered With an Activity, or Task, or to What Degree of Impact the Eyesight Had on Their Emotional Well-Being	Results of Pooled Subsets			Results of Full Dataset		
		Threshold $\beta 1$[a]	Threshold $\beta 2$[a]	Discrimination α[b]	Item Information[c]	S-X^2 [d]	*P* value S-X^2 [e]
CAT1	Ability to see and enjoy TV programs	0.12 (0.06)	1.72 (0.06)	1.82 (0.06)	3.20	33.35	0.46
CAT2	Recreational activities, such as biking, walking, or other activities	0.77 (0.04)	1.89 (0.06)	2.42 (0.10)	4.18	32.72	0.21
CAT3	Going out, such as seeing cinema films, theater plays, or sports events	0.49 (0.03)	1.57 (0.05)	2.86 (0.10)	5.05	14.86	0.96
CAT4	Shopping	0.52 (0.03)	1.69 (0.05)	2.76 (0.08)	4.82	28.67	0.33
CAT5	Visiting friends and family	1.44 (0.03)	2.49 (0.06)	2.18 (0.08)	3.59	17.94	0.27
CAT6	Meeting people	1.23 (0.04)	2.48 (0.06)	2.30 (0.11)	4.11	19.06	0.39
CAT7	Recognizing people at the opposite side of the street	0.09 (0.03)	1.30 (0.06)	1.82 (0.06)	3.00	35.80	0.30
CAT8	Recognizing people at the opposite side of the room	1.27 (0.04)	2.21 (0.06)	1.83 (0.06)	2.78	22.47	0.38
CAT9	Looking after appearance	0.93 (0.02)	2.29 (0.08)	1.77 (0.05)	2.89	31.34	0.40
CAT10	Activities or hobbies that require good near vision	−0.32 (0.01)	1.15 (0.04)	2.20 (0.06)	3.93	42.77	0.07
CAT11	Operating household appliances	0.86 (0.03)	2.31 (0.07)	2.25 (0.10)	4.08	24.83	0.48
CAT12	Needing (more) help from others because of eyesight	0.55 (0.03)	2.12 (0.05)	2.39 (0.04)	4.50	32.22	0.23
CAT13	Mobility outdoors	0.37 (0.02)	1.68 (0.04)	2.59 (0.13)	4.78	37.22	0.15
CAT14	Going carefully to avoid falling or tripping	0.16 (0.05)	1.46 (0.04)	2.41 (0.05)	4.31	49.79	0.02
CAT15	Travelling or using transport (bus and train)	0.97 (0.03)	1.83 (0.05)	2.59 (0.06)	4.21	21.75	0.36
CAT16	Going down steps, stairs or pavements in good lighting conditions	0.36 (0.03)	1.73 (0.05)	2.02 (0.07)	3.62	26.29	0.76
CAT17	Going down steps, stairs or pavements in poor lighting conditions or at night	−0.37 (0.02)	1.00 (0.04)	2.25 (0.08)	4.01	33.94	0.38
CAT18	Reading ordinary size print	−0.1 (0.04)	1.2 (0.07)	1.88 (0.09)	3.05	55.52	0.06
CAT19	Reading large text	1.47 (0.07)	2.5 (0.07)	1.55 (0.06)	2.35	15.87	0.87
CAT20	Reading labels or instructions	−0.7 (0.03)	0.74 (0.04)	1.88 (0.08)	3.22	34.39	0.80
CAT21	Reading colored text or text on a colored background	−0.32 (0.02)	1.08 (0.07)	1.71 (0.08)	2.91	48.14	0.24
CAT22	Writing	0.53 (0.04)	1.57 (0.07)	2.07 (0.12)	3.31	25.49	0.75
CAT23	Mobility indoors under poor lighting conditions	0.42 (0.04)	2.08 (0.06)	1.7 (0.05)	2.96	31.00	0.67
CAT24	Mobility indoors under normal lighting conditions	1.26 (0.05)	2.89 (0.15)	1.62 (0.09)	2.79	24.46	0.55
CAT25	Depth perception and estimating the distance of objects	0.28 (0.03)	1.9 (0.07)	1.67 (0.07)	2.86	49.01	0.05
CAT26	Seeing street signs or other signs from a distance	−0.62 (0.04)	0.99 (0.07)	1.8 (0.06)	3.22	52.87	0.09
CAT27	Seeing moving objects	0.54 (0.01)	2.15 (0.06)	2.03 (0.08)	3.62	40.10	0.13
CAT28	Suffering from tiredness of the eyes	−0.21 (0.03)	1.89 (0.05)	1.44 (0.07)	2.43	32.10	0.66
CAT29	Suffering from glare	−1.45 (0.05)	0.9 (0.06)	1.17 (0.02)	2.11	53.58	0.31
CAT30	Getting the right amount of light to see properly	−0.78 (0.03)	1.1 (0.05)	1.63 (0.03)	2.82	41.65	0.45
CAT31	Seeing how people react to what you say	1.25 (0.04)	2.42 (0.03)	1.98 (0.04)	3.42	23.08	0.24
CAT32	Driving a car during the day in a well-known environment	0.89 (0.03)	1.61 (0.03)	2.47 (0.06)	3.91	27.30	0.21
CAT34	Driving a car under difficult circumstances (bad weather, rush hour, etc.)	−0.54 (0.04)	0.9 (0.05)	1.94 (0.07)	3.38	36.37	0.64
CAT35	Feeling worried or concerned about your safety at home	1.92 (0.05)	3.92 (0.14)	1.49 (0.03)	2.66	15.53	0.56
CAT36	Feeling worried or concerned about spilling food/drinks, dropping/breaking things	1.41 (0.02)	2.61 (0.06)	2.00 (0.09)	3.30	16.61	0.56
CAT37	Feeling worried or concerned about safety outdoors	1.21 (0.04)	2.79 (0.08)	1.58 (0.07)	2.62	37.11	0.12
CAT38	Feeling worried about coping with everyday life	0.76 (0.02)	2.29 (0.05)	1.85 (0.06)	3.33	33.16	0.32
CAT39	Feeling worried about eyesight	−0.95 (0.04)	1.46 (0.08)	1.58 (0.07)	2.95	35.10	0.29
CAT40	Caused eyesight stop doing the things you wanted to do	0.56 (0.03)	2.18 (0.06)	2.08 (0.09)	3.70	34.68	0.30
CAT41	Interfered eyesight life in general	0.2 (0.03)	1.82 (0.06)	2.15 (0.07)	4.06	31.99	0.47
CAT42	Feeling embarrassed	2.42 (0.1)	4.11 (0.29)	1.5 (0.09)	2.59	12.64	0.32
CAT43	Feeling frustrated or annoyed	0.31 (0.02)	1.79 (0.08)	2.02 (0.13)	3.56	32.39	0.45
CAT44	Feeling lonely or isolated	1.67 (0.06)	3.01 (0.14)	1.92 (0.13)	3.28	15.02	0.46
CAT45	Feeling sad	0.62 (0.05)	2.7 (0.13)	1.33 (0.08)	2.32	42.28	0.26
CAT46	Feeling like a burden	1.41 (0.07)	2.99 (0.12)	1.75 (0.06)	3.08	23.04	0.29
CAT47	Using a mobile phone or smartphone	0.55 (0.03)	1.8 (0.03)	2.19 (0.05)	3.69	24.53	0.75

[a] Location parameters of item response categories.
[b] Discrimination parameter (i.e. indicates the relationship of the item with the underlying construct of interest).
[c] Item information (i.e. refers to the information content of an item in relation to the total test information).
[d] S-X^2 index (which is used for polytomous items to compare observed and predicted response proportions).
[e] S-X^2 *P* values < 0.01 considered significant.

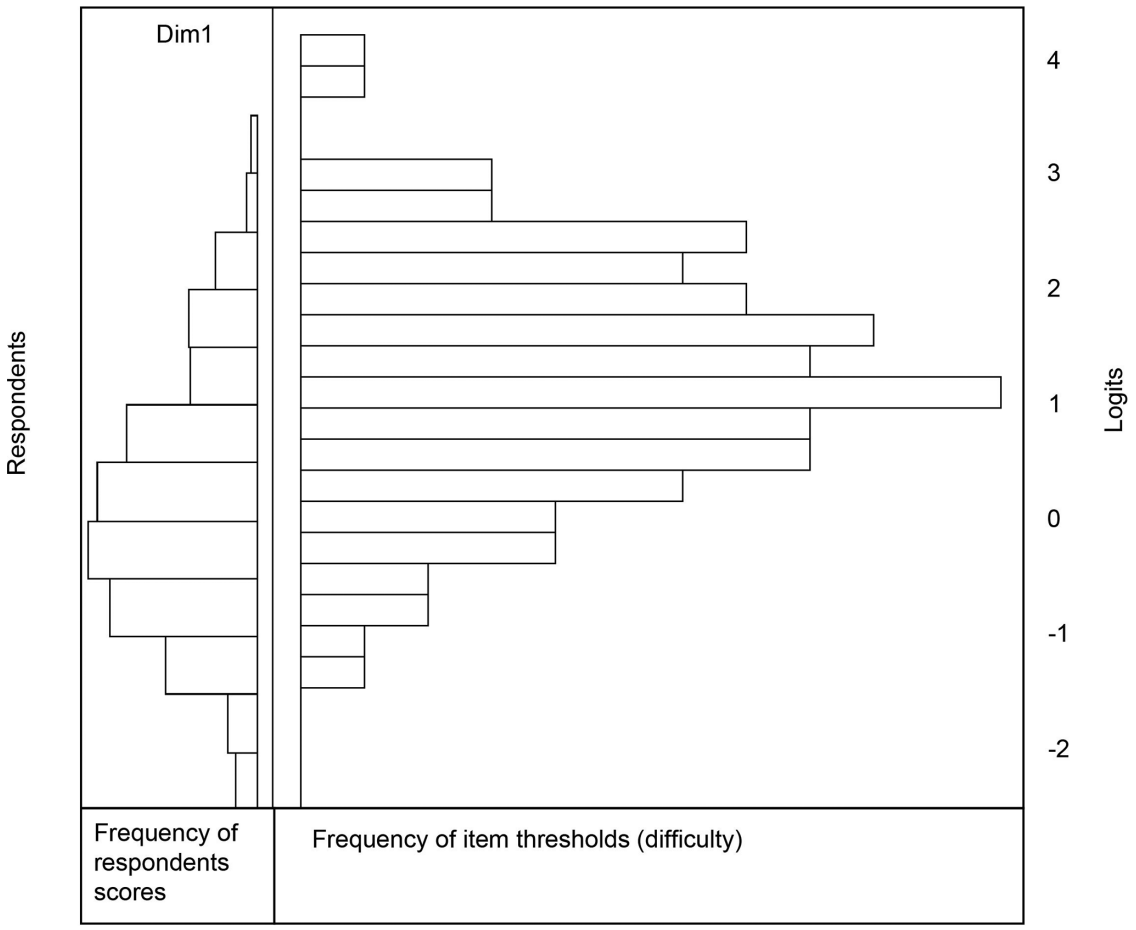

Figure 1. Person–item map of the EyeQ item-bank. Respondents and items are calibrated along the same scale (latent trait). The histogram on the left represents the respondents. The histogram on the right represents the item location on the latent trait continuum. The Y-axis represents the theta range of the latent trait continuum where a higher theta represents a higher level of disability.

without DIF for visual acuity. On the other hand, possibly because of the relatively low level of disability of participants, it was decided to collapse response categories "often" and "always," which could lead to a decrease of sensitivity to detect changes in Vr-QoL scores of patients having high levels of disability.

A strength of this research is the relatively large study sample that made it possible to divide the data in subsets and perform separate calibrations, a novel approach in ophthalmology leading to robust item estimates. After DIF analyses between these subsets and the full data showed no significant differences in item performance because of a possible selection bias, it was possible to calculate mean item estimates. This supports our expectation that the items included in the final EyeQ are robust and stable and therefore suitable for measuring Vr-QoL in patients with exudative retinal diseases.

A limitation of this study is the relatively small group of patients with diabetic macular edema (8%), which may limit the generalizability of the results for this particular group. Response rates of patients with diabetic macular edema, cystoid macular edema, and age-related macular edema were, respectively, 13%, 24%, and 22%. A possible explanation is the presence of other diabetes-related health concerns, which may carry a heavier burden than the visual impairment and may have discouraged participation in this study.

We performed DIF analyses for a series of variables; however, we only found three items that showed DIF, which all had a negligible impact on the total score. The impact of DIF when administering the EyeQ as a CAT could be higher, because the algorithm selects items from the item-bank until a specified level of precision is reached or a predetermined number of items is answered, which could possibly be the items displaying DIF. Future CAT simulations will show to what extent the items displaying DIF are actually administered in a CAT, in order to better estimate the impact of DIF

Table 5. McFadden's Pseudo R2 and IRT Parameters for Items Displaying DIF (Likelihood-Ratio χ^2 Test Criterion of 0.01)

Item With DIF Number and Content	DIF Type	McFadden's Pseudo R^2	Slope; and Threshold Parameters	
Items with DIF for gender: CAT9: Looking after appearance	Uniform[a]	R^2_{12}[b] = 0.031 R^2_{23}[c] = 0.002[*]	Male Female	2.39; 1.03, 2.04 1.98; 0.47, 1.77
Items with DIF for age: CAT42: Feeling embarrassed	Uniform	R^2_{12} = 0.025 R^2_{23} = 0.002[*]	<=75 >75	2.33; 1.43[d] 1.88; 2.11
Items with DIF for independently completed vs. with help: CAT 35: Feeling worried or concerned about your safety at home	Nonuniform[e]	R^2_{12} = 0.011[*] R^2_{23} = 0.023	Independent With help	2.37; 1.27[d] 0.99; 2.23

[a]DIF that is proportional across levels of the underlying latent trait.
[b]R^2_{12} represents nonuniform DIF.
[c]R^2_{23} represents uniform DIF.
[d]CAT42 and CAT35, both located at one end (high disability) of the disability continuum, and response category 'often/always' (collapsed) was hardly chosen, therefore one threshold parameter was calculated, representing threshold between categories 'never' and 'sometimes'.
[e]DIF that is nonproportional across levels of the underlying latent trait.
[*]DIF type that exceeds the test criterion of 0.01.

on the total scores. Until then, we recommend using the group-specific item parameter estimates for these items in the algorithm. In addition, our future research will involve post hoc CAT simulations to assess how well the EyeQ item-bank performs as a CAT under different administration conditions. In this step, the theta estimated by CAT is compared with the true theta estimated by the full set of items.

This calibration and assessment of DIF of the EyeQ is a step forward in the development of a new item-bank. However, in future research several psychometric properties need to be investigated (e.g., concurrent and discriminant validity, and test-retest reliability, as well as the responsiveness to detect changes over time).[44,45]

Even though the IVI-CAT and the DR/DME QoL item-banks have been developed recently as well,[17,18] we assume that the EyeQ is a valuable addition. First, because the IVI-CAT is based on the 28-item IVI, the EyeQ potentially is more comprehensive as we added items after searching other Vr-QoL instruments; in addition, we evaluated the relevance and comprehensibility of the IVI using three-step test-interviewing.[20] Second, the DR/DME QoL item-bank requires still items to fill in case all 10 item-banks (for several domains) are administered as CAT; however, the possibility exists not to evaluate all domains. The EyeQ could provide us the best of both worlds: an in-between version with an all-around performance; a large distribution of items across the disability spectrum, while still measuring an unidimensional construct. The results from this study are promising for the use of the EyeQ as a generic instrument of Vr-QoL in patients with several retinal diseases. However, future research will involve longitudinal measurement invariance of the EyeQ and post hoc CAT simulations. The results of these analyses are important for use of the EyeQ in clinical practice.

Conclusion

In conclusion, this study described the development and calibration of the EyeQ item-bank, which is a new instrument that can be used for the periodical and systematic assessment of Vr-QoL in patients suffering from exudative retinal diseases and receiving treatment in ophthalmic clinical practice. The model and item fit statistics of the EyeQ were found to be satisfactory, and robust item estimates could be estimated for 46 items because of separate calibrations. The calibration of the EyeQ allows use of this instrument in clinical practice. Future research should focus on the impact of DIF items on test scores while administered as a CAT

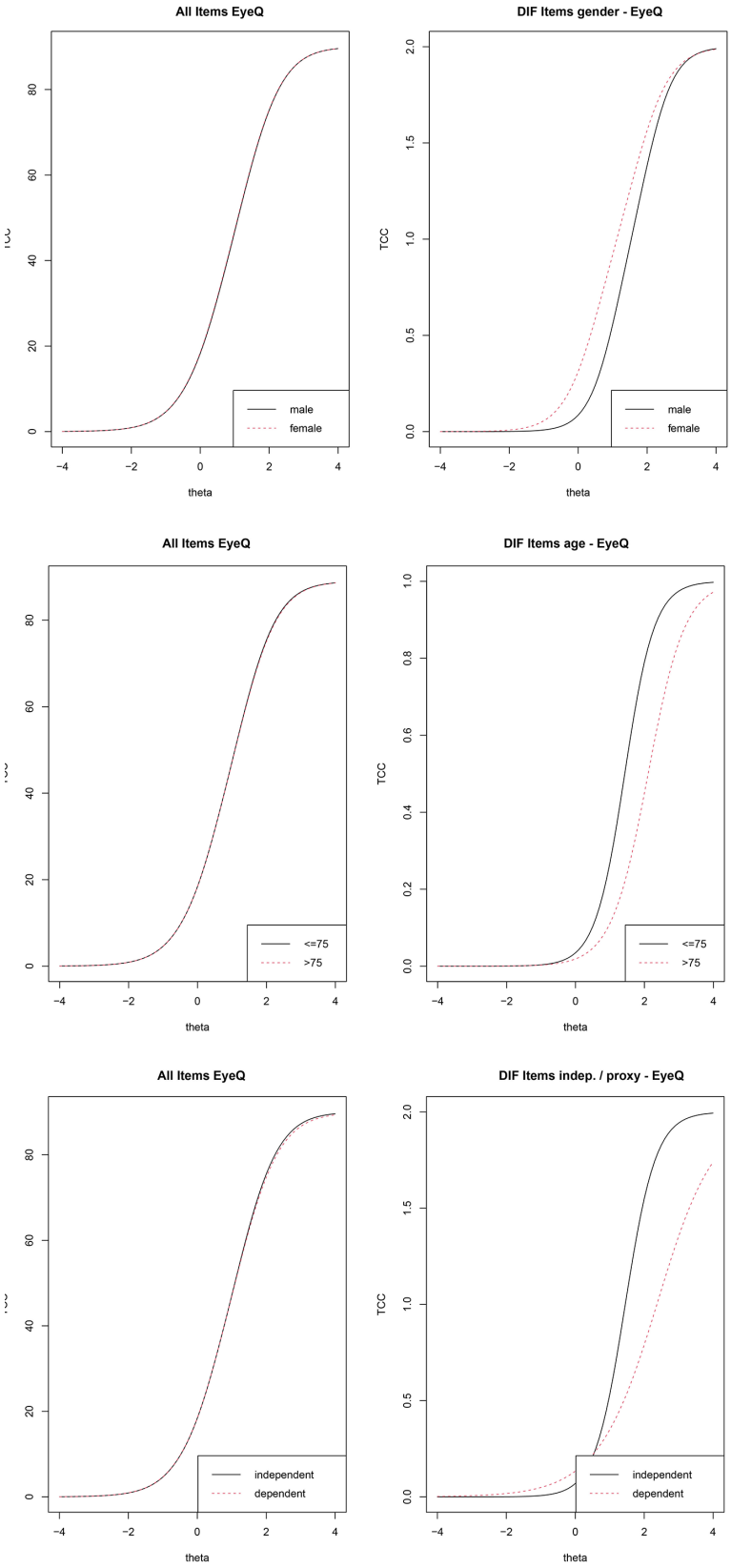

Figure 2. Impact of DIF on the test characteristic curve (TCC) for the EyeQ. The plots on the left show the impact of differential time functioning (DIF) considering all items, the plots on the right show the DIF impact considering only items displaying DIF.

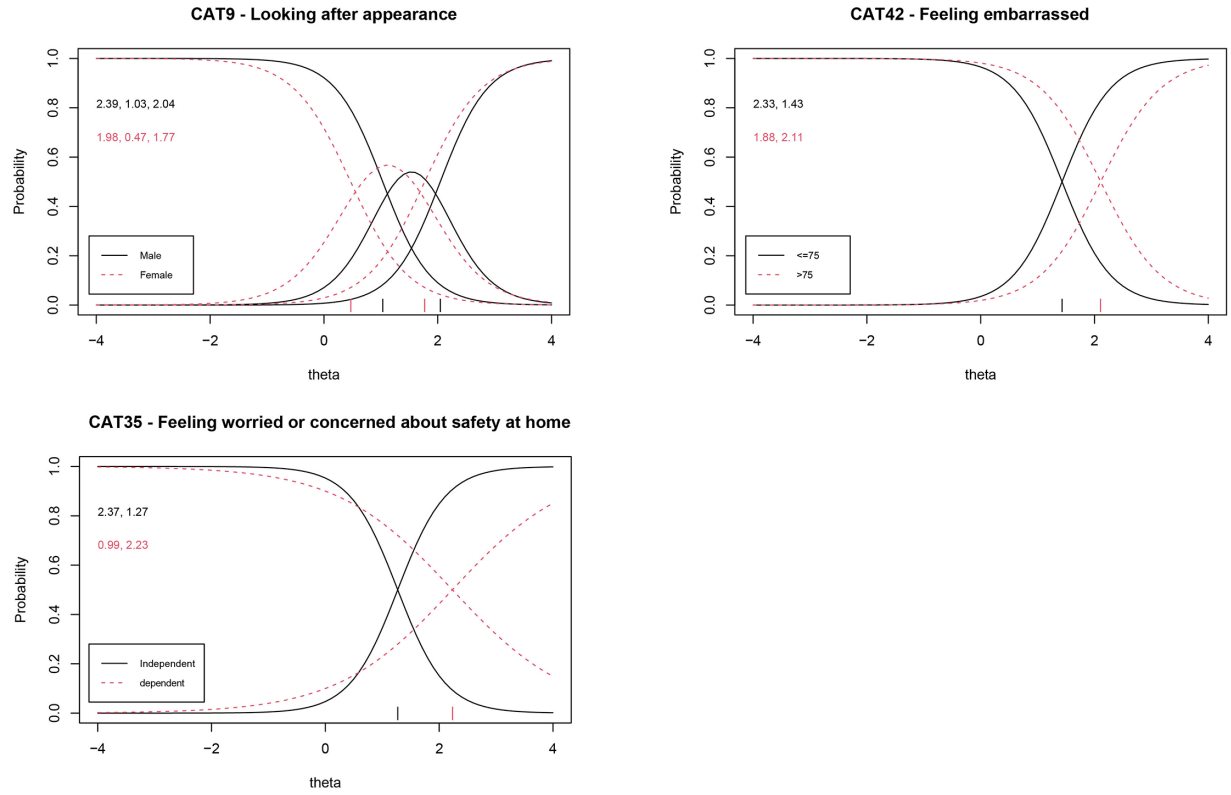

Figure 3. Category response curves for items displaying DIF.

and evaluate the longitudinal measurement invariance of the instrument.

Acknowledgments

Supported by Bayer Healthcare Mijdrecht (Fellowship name/number: IMP20112/VUmc PROM). The sponsor had no role in the design and conduct of the study, the data collection, data analysis, data interpretation, or writing of the report.

Disclosure: **T.P. Rausch-Koster**, None; **M.A.J. Luijten**, None; **F.D. Verbraak**, None; **G.H.M.B. van Rens**, None; **R.M.A. van Nispen**, None

References

1. Delcourt C, Le Goff M, Von Hanno T, et al. The decreasing prevalence of nonrefractive visual impairment in older Europeans: a meta-analysis of published and unpublished data. *Ophthalmology*. 2018;125:1149–1159.
2. GBD 2019 Blindness and Vision Impairment Collaborators; Vision Loss Expert Group of the Global Burden of Disease Study. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study. *Lancet Glob Health*. 2021;9(2):e144–e160.
3. Verbraak FD, Ponsioen DL, Tigchelaar-Besling OA, et al. Real-world treatment outcomes of neovascular age-related macular degeneration in the Netherlands. *Acta Ophthalmol*. 2021;99(6):e884–e892.
4. Brody BL, Gamst AC, Williams RA, et al. Depression, visual acuity, comorbidity, and disability associated with age-related macular degeneration. *Ophthalmology*. 2001;108:1893–1900; discussion 1900-1901.
5. Bookwala J, Lawson B. Poor vision, functioning, and depressive symptoms: a test of the activity restriction model. *Gerontologist*. 2011;51:798–808.
6. Hayman KJ, Kerse NM, La Grow SJ, Wouldes T, Robertson MC, Campbell AJ. Depression in older people: visual impairment and subjective ratings of health. *Optom Vis Sci*. 2007;84:1024–1030.

7. van Nispen RM, de Boer MR, Hoeijmakers JG, Ringens PJ, van Rens GH. Co-morbidity and visual acuity are risk factors for health-related quality of life decline: five-month follow-up EQ-5D data of visually impaired older patients. *Health Qual Life Outcomes*. 2009;7:18.
8. Greenhalgh J, Gooding K, Gibbons E, et al. How do patient reported outcome measures (PROMs) support clinician-patient communication and patient care? A realist synthesis. *J Patient Rep Outcomes*. 2018;2(1):1–28.
9. van Nispen RMA, Virgili G, Hoeben M, et al. Low vision rehabilitation for better quality of life in visually impaired adults. *Cochrane Database Syst Rev*. 2020;1(1):CD006543.
10. Pesudovs K. Item banking: a generational change in patient-reported outcome measurement. *Optom Vis Sci*. 2010;87:285–293.
11. Braithwaite T, Calvert M, Gray A, Pesudovs K, Denniston AK. The use of patient-reported outcome research in modern ophthalmology: impact on clinical trials and routine clinical practice. *Patient Relat Outcome Meas*. 2019;10:9–24.
12. Flens G, Smits N, Terwee CB, Dekker J, Huijbrechts I, de Beurs E. Development of a computer adaptive test for depression based on the Dutch-Flemish version of the PROMIS item bank. *Eval Health Prof*. 2017;40:79–105.
13. You DS, Cook KF, Domingue BW, et al. Customizing CAT administration of the PROMIS misuse of prescription pain medication item bank for patients with chronic pain. *Pain Med*. 2021;22:1669–1675.
14. Patel RN, Esparza VG, Lai JS, et al. Comparison of PROMIS computerized adaptive testing versus fixed short forms in juvenile myositis [published online ahead of print July 30, 2021]. *Arthritis Care Res*. 2021, https://doi.org/10.1002/acr.24760.
15. Magis D, Yan D, von Davier AA. *Computerized Adaptive and Multistage Testing with R*. Cham: Springer; 2017.
16. Wainer H, Dorans NJ, Flaugher R, Green BF, Mislevy RJ. *Computerized Adaptive Testing: A primer*. New York: Routledge; 2000.
17. Fenwick EK, Khadka J, Pesudovs K, Rees G, Wong TY, Lamoureux EL. Diabetic retinopathy and macular edema quality-of-life item banks: development and initial evaluation using computerized adaptive testing. *Invest Ophthalmol Vis Sci*. 2017;58:6379–6387.
18. Fenwick EK, Barnard J, Gan A, et al. Computerized adaptive tests: efficient and precise assessment of the patient-centered impact of diabetic retinopathy. *Transl Vis Sci Technol*. 2020;9(7):3.
19. Fenwick EK, Loe BS, Khadka J, Man RE, Rees G, Lamoureux EL. Optimizing measurement of vision-related quality of life: a computerized adaptive test for the impact of vision impairment questionnaire (IVI-CAT). *Qual Life Res*. 2020;29:765–774.
20. Rausch-Koster TP, van der Ham AJ, Terwee CB, Verbraak FD, van Rens GH, van Nispen RM. Translation and content validity of the Dutch Impact of Vision Impairment questionnaire assessed by Three-Step Test-Interviewing. *J Patient Rep Outcomes*. 2021;5(1):1.
21. Wolffsohn JS, Cochrane AL. Design of the low vision quality-of-life questionnaire (LVQOL) and measuring the outcome of low-vision rehabilitation. *Am J Ophthalmol*. 2000;130:793–802.
22. Mangione CM, Lee PP, Gutierrez PR, et al. Development of the 25-item National Eye Institute Visual Function Questionnaire. *Arch Ophthalmol*. 2001;119:1050–1058.
23. van Nispen RM, Knol DL, Langelaan M, van Rens GH. Re-evaluating a vision-related quality of life questionnaire with item response theory (IRT) and differential item functioning (DIF) analyses. *BMC Med Res Methodol*. 2011;11:125.
24. EuroQol Group. EuroQol—a new facility for the measurement of health-related quality of life. *Health Policy*. 1990;16:199–208.
25. Lamers LM, Stalmeier PF, McDonnell J, Krabbe PF, van Busschbach J. [Measuring the quality of life in economic evaluations: the Dutch EQ-5D tariff]. *Ned Tijdschr Geneeskd*. 2005;149:1574–1578.
26. *Corp. I. IBM SPSS Statistics for Windows 26.0.*
27. Rizopoulos D. ltm: An R package for latent variable modeling and item response theory analyses. *J Stat Software*. 2006;17.
28. Rodriguez A, Reise SP, Haviland MG. Applying bifactor statistical indices in the evaluation of psychological measures. *J Pers Assess*. 2016;98:223–237.
29. Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care*. 2007;45(5 Suppl 1):S22–S31.
30. Courtney M. Determining the number of factors to retain in EFA: Using the SPSS R-Menu v2.0 to make more judicious estimations. *Practical Assess Res Eval*. 2013;18:1–14.
31. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Qual Life Res*. 2007;16(Suppl 1):5–18.

32. Liu Y, Maydeu-Olivares A. Local dependence diagnostics in IRT modeling of binary data. *Educ Psychol Measure*. 2012;73:254–274.
33. Nguyen TH, Han HR, Kim MT, Chan KS. An introduction to item response theory for patient-reported outcome measurement. *Patient*. 2014;7(1):23–35.
34. Sijtsma K, Meijer RR, van der Ark LA. Mokken scale analysis as time goes by: An update for scaling practitioners. *Pers Individual Diff*. 2011;50(1):31–37.
35. Pilkonis PA, Choi SW, Reise SP, et al. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS): depression, anxiety, and anger. Assessment. 2011;18:263–283.
36. Loevinger J. The technic of homogeneous tests compared with some aspects of scale analysis and factor analysis. *Psychol Bull*. 1948;45:507–529.
37. Magis D, Yan D, von Davier AA. *Computerized Adaptive and Multistage Testing with R: using packages catR and mstR*. Cham: Springer International Publishing; 2017.
38. Jones RN. Differential item functioning and its relevance to epidemiology. *Curr Epidemiol Rep*. 2019;6:174–183.
39. Choi SW, Gibbons LE, Crane PK, lordif: an R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *J Stat Softw*. 2011;39(8):1–30.
40. Orlando M, Thissen D. Further Investigation of the Performance of S - X2: An item fit index for use with dichotomous item response theory models. *Appl Psychol Measure*. 2003;27:289–298.
41. Kang T, Chen TT. Performance of the Generalized S-X^2 Item Fit Index for Polytomous IRT Models. *J Educ Measure*. 2008;45(4):391–406.
42. Orlando M, Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models. *Appl Psychol Measure*. 2000;24:50–64.
43. Hu Lt, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling*. 1999;6:1–55.
44. Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care*. 2007;45(5):S22–S31.
45. Pellicciari L, Chiarotto A, Giusti E, Crins MH, Roorda LD, Terwee CB. Psychometric properties of the patient-reported outcomes measurement information system scale v1.2: global health (PROMIS-GH) in a Dutch general population. *Health Qual Life Outcomes*. 2021;19:226.