Optimizing Computer Adaptive Test Performance: A Hybrid Simulation Study to Customize the Administration Rules of the CAT-EyeQ in Macular Edema Patients

T. Petra Rausch-Koster¹⁻³, Michiel A. J. Luijten⁴⁻⁷, Frank D. Verbraak¹,², Ger H. M. B. van Rens¹,², and Ruth M. A. van Nispen¹,²

¹ Amsterdam UMC location Vrije Universiteit Amsterdam, Ophthalmology, De Boelelaan 1117, Amsterdam, the Netherlands
² Amsterdam Public Health, Quality of Care, Aging and Later Life, Amsterdam, the Netherlands
³ Bergman Clinics, Department of Ophthalmology, Naarden, the Netherlands
⁴ Emma Children’s Hospital, Amsterdam UMC location University of Amsterdam, Department of Child and Adolescent Psychiatry & Psychosocial Care, Amsterdam, the Netherlands
⁵ Amsterdam UMC location Vrije Universiteit Amsterdam, Department of Epidemiology & Data Science, Amsterdam, the Netherlands
⁶ Amsterdam Public Health, Mental Health & Methodology, Amsterdam the Netherlands
⁷ Amsterdam Reproduction & Development, Child Development, 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: July 11, 2022
Accepted: October 10, 2022
Published: November 17, 2022

Keywords: vision-related quality of life, macular edema, computer adaptive test (CAT), patient-reported outcome; patient perspective

Citation: Rausch-Koster TP, Luijten MAJ, Verbraak FD, van Rens GHMB, van Nispen RMA. Optimizing computer adaptive test performance: A hybrid simulation study to customize the administration rules of the CAT-EyeQ in macular edema patients. Transl Vis Sci Technol. 2022;11(11):14, https://doi.org/10.1167/tvst.11.11.14

Purpose: In previous research the EyeQ item bank, which measures vision-related quality of life (Vr-QoL), was calibrated for future use as a computer adaptive test (CAT). The aim of the current study was to define optimal administration rules.

Methods: CAT simulations were performed using real responses. Patients (N = 704; mean age, 76.2 years), having macular edema completed the EyeQ. Four CAT simulations were performed, which were set with different administration rules regarding length, accuracy level and the association with best health, which means the test was aborted after the first 4 responses of having no complaints.

Results: The CATDefault showed a mean test length of 6.9 and 15.1% unreliable estimations. Extending the test length to 15 items (CATAlt1) resulted in a mean test length of 7.3 and slightly decreased the percentage unreliable estimations (11.5%). Under CATAlt2, the percentage unreliable estimations was 15.1% and the mean test length was 9.7. Percentages of floor/ceiling effects for CATDefault, CATAlt1, and CATAlt2 were 3.1, 3.0, and 3.1, respectively. CATBestHealth reduced the mean test length to 5.9 and showed 18.2% unreliably estimated patients, of which 14.2% had floor/ceiling scores.

Conclusions: This study shows that the CATBestHealth provided reliably estimated ability scores, with a negligible increase in the number of unreliably estimated patients and ensures that patients having little or no vision-related quality of life problems are minimally burdened with completing items.

Translational Relevance: The computer adaptive test EyeQ, set with optimal administration rules, can now be used for the computer adaptive assessment of vision-related quality of life in patients suffering from exudative retinal diseases in ophthalmic clinical practice.

Introduction

Nonrefractive vision impairment in European countries has a prevalence of approximately 1% to 2%,¹ having cataract, age-related macular degeneration, glaucoma, and diabetic retinopathy as major causes.² Loss of vision will cause limitations in patient’s daily activities, physical functioning and might impact their emotional well-being and quality of life.³⁻⁶
Patient-reported outcome measures (PROMs) are proven to be suitable for measuring and evaluating patient’s disabilities in daily activities and are supportive for patient–doctor communication and shared decision-making. Nowadays, as a solution for the limitations of first generation PROMs with their related measuring and scoring problems, and second-generation PROMs that provide robust measurement instruments using techniques from item response theory to calibrate items, third-generation PROMs are increasingly introduced in health care. The latter are item banks, which refer to collections of items across a disability continuum, that can be used to apply computer adaptive test (CAT). When administering a CAT, an algorithm selects the items that are relevant to the patient as the selection is based on the response the patient gave on previous items. This practice allows a test to limit the number of items that need to be administered, which results in an unique sequence of items that are tailored to a patient’s individual level of ability. The process of selecting items from the item bank continues until a predefined level of precision of the patient’s theta estimation is reached. In addition, the item selection process depends on other administration criteria that have been set, such as maximum test length. In this way, administering PROMs and thereby evaluating the impact of the disease on a patient’s daily life can be done in significantly less time and effort and will decrease administration fatigue while obtaining acceptable reliability levels.

Several administration criteria can be considered when administering CATs, namely, maximum length, level of accuracy, classification, or information, and combinations of these criteria are also a possible solution to optimize the performance of the CAT. Additionally, in case the first few responses to items are associated to best health (i.e., patient indicated not to experience any difficulties), it could be considered to abort the CAT, contributing to minimize the burden on the patient.

In previous research we developed and calibrated the EyeQ-46 item bank, which measures vision-related quality of life (Vr-QoL), using a graded response model. The rationale for the development and content of this new item bank has been described. The overall fit of the EyeQ-46 item bank to the model was adequate. Differential item functioning (DIF) was performed differently in subgroups (gender, age, and administration mode), which had negligible impact on the total EyeQ-46 score. However, when administering the EyeQ as a CAT, the impact of DIF could be higher because only a few items are administered.

The aim of the current study was to define the optimal administration rules (i.e., stopping rules) of the CAT-EyeQ by performing hybrid CAT simulations using mainly real responses and to investigate to what extent items displaying DIF are administered in the CAT-EyeQ.

## Methods

### Ethical Statement

This study is part of a prospective cohort study, the Dutch EyeQ study, which aims to develop a new computer adaptive PROM to assess Vr-QoL in patients suffering from exudative retinal diseases. The study protocol was approved by the Medical Ethics Committee (approval reference number 2018.361) of Amsterdam University Medical Centers and conducted according to the Declaration of Helsinki.

### Participants and Study Design

In a previous study, the EyeQ-46 item bank has been calibrated in patients, aged more than 18 years, having macular edema caused by age-related macular degeneration, retinal vein occlusion and diabetic retinopathy who received anti-vascular endothelial growth factor injections at Bergman Clinics eye hospitals. All patients who participated in the calibration study were informed about the aim, procedure, and duration of the study and were asked whether they would participate and subsequently signed a written informed consent. The current study is a CAT simulation study using the real response patterns of patients who participated in the calibration study and imputed responses for patients with missing values: a hybrid simulation.

### Statistical Analyses

#### Missings Investigation and Data Imputation

A patient’s level of ability (theta [θ]) on the EyeQ-46 scale was estimated, using the expected a posteriori estimator, with item parameters obtained after fitting the graded response item response theory model (GRM). An excessive percentage of missing
responses in patients (>25%) led to exclusion from further analyses. Post hoc CAT simulations use real data and during the simulation process the person's ability level is estimated after each response. In case the CAT selects an item that does not have a response, the CAT will not continue or give an estimation for this person’s record. Therefore, we investigated the possibility to perform hybrid simulations. This type of simulation uses a combination of full response data and imputed responses for patients with missing values. Because the response category of not applicable is scored as missing value as well, first, the amount of missing values and the amount of patients having one or more missing responses was evaluated. To gain a better understanding of whether or not data were missing at random, locations of missing values and the presence of missing values related to other missing values were examined. Additionally, we investigated if sex, age, and vision impairment affected the presence of missing values. After this, we imputed plausible responses based on the a posteriori estimator estimated level of ability (θ), obtained from item parameters of the initial EyeQ-46 GRM model. After imputing plausible data again, we fitted the GRM model and obtained patient’s thetas based on complete data containing imputed plausible responses. We evaluated Pearson’s correlation and the mean difference between patient’s theta calculated after fitting the GRM model with incomplete response pattern and patient’s theta calculated after fitting the GRM model with imputed data, to ensure that imputation led to similar outcomes of theta.

**CAT Administration Rules**

Subsequently, three CAT simulations (CATDefault, CATAlt1, and CATAlt2) were performed under different conditions by customizing the administration rules. The CATs used the maximum posterior weighted information criterion for selecting items. The applied administration rules in the CAT simulations varied in the minimum and maximum test lengths: CATDefault followed the default CAT administration rules developed by the Patient-Reported Outcomes Measurement Information System with a minimum test length of 4 items, a maximum of 12 items and an accuracy level of 0.32. For CATAlt1 a minimum of 2 and a maximum of 15 items was set, with an accuracy level of 0.32 (corresponding with a reliability of 0.90) and CATAlt2 was set with a minimum of four and a maximum of 12 items, with an accuracy level of 0.25. In addition, it was investigated what the effect would be on the average test length, the mean standard error (SE), the percentages floor/ceiling effects (i.e., percentages of partici- pants with only the lowest [floor] response categories administered [best possible Vr-QoL], and with only the highest [ceiling] response categories administered), and unreliably estimated tests, if the CAT was aborted after the first four responses were associated with best health (CATBestHealth).

**CAT Evaluation**

The accuracy of the CAT simulations were evaluated by comparing the theta estimates obtained by the CAT with the true theta estimates obtained from the full-length EyeQ (46 items): we calculated the mean differences and Pearson’s correlation between these thetas. Additionally, we evaluated the mean test length, mean SEs and conditional SEs of each simulation. The latter provided insight into which part of the ability continuum the errors were the lowest and thus in which part of the ability continuum the CAT performs best. Additionally, percentages unreliably estimated scores (SE of >0.32, reliability of 0.90) were calculated for each simulation. Last, we evaluated to what extent the DIF items, which we found in our previous calibration study, were actually administered in the CATs. All statistical analyses were performed using R (version 3.6.1).

**Results**

Patients (N = 3783) were invited and 746 were willing to participate in the Dutch EyeQ study (response rate of 19.7%) and written informed consent was obtained. Seven hundred thirteen participants filled out the EyeQ. Nine patients, having an excessive number of missing responses (>25.0%), were excluded from the analyses. The sociodemographic and clinical characteristics of the remaining 704 respondents are reported in Table 1.

In total, 339 patients (48.2%) scored a missing response for at least one item, of which 287 (84.7%) had one to four missings (<10% of total item bank). In total, there were 989 (3.1%) missing values. More than one-half of the missings (504 [51.0%]) were found for the EyeQ3 [Going out, such as seeing cinema films, theater plays or sports events] 143 [20.3%], EyeQ32 [Driving a car during the day in a well-known environment] 172 [24.4%], and the EyeQ34 [Driving a car under difficult circumstances (bad weather, rush hour, ...] 189 [26.8%]). The missing values in the EyeQ32 and EyeQ34 were partly related to each other: 79 times a combination of missing values of the EyeQ32 (79 [46.9%]) and EyeQ34 (79 [41.8%]) occurred. In total,
Table 1. Sociodemographic and Clinical Characteristics of Participants (n = 704)

| Characteristic                                      | Value                        |
|----------------------------------------------------|------------------------------|
| Age in Years                                       | 76.2 ± 9.2                   |
| Sex                                                 |                              |
| Male                                               | 342 (48.6)                   |
| Female                                             | 362 (51.4)                   |
| Nationality                                         |                              |
| Dutch                                              | 663 (94.2)                   |
| Other                                               | 41 (5.8)                     |
| Educational background in years                     | 11.0 (9.0–15.0)              |
| Civil status                                        |                              |
| Single                                             | 235 (33.4)                   |
| Not single                                         | 469 (66.6)                   |
| Nonocular comorbiditya                              |                              |
| Lung diseaseb                                      | 64 (9.1)                     |
| Cardiovascular diseasec                            | 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     |                              |
| 0                                                  | 254 (36.1)                   |
| 1                                                  | 261 (37.1)                   |
| ≥2                                                 | 189 (26.8)                   |
| Ocular comorbidityd                                |                              |
| Anterior segmente                                  | 198 (23.9)                   |
| Posterior segmentf                                | 266 (37.8)                   |
| Glaucoma                                           | 80 (11.4)                    |
| Amblyopia                                          | 10 (1.4)                     |
| Other                                              | 18 (2.6)                     |
| Number of ocular comorbidities within persond       |                              |
| 0                                                  | 262 (37.2)                   |
| 1                                                  | 339 (48.2)                   |
| 2                                                  | 93 (13.2)                    |
| 3                                                  | 10 (1.4)                     |
| Treated site anti-VEGF                             |                              |
| Monocular                                          | 574 (81.5)                   |
| Binocular                                          | 128 (18.2)                   |
| Unknown                                            | 2 (0.3)                      |
| LogMAR visual acuity better eye                    | 0.10 (0.01–0.22)             |
| Degree of vision impairmentg                       |                              |
| 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            |                              |
| Neovascular age-related macular degeneration       | 446 (63.4)                   |
| Diabetic macular edema                             | 56 (8.0)                     |
| Cystoid macular edema secondary to retinal vein occlusion | 175 (24.9)             |
| Unknown                                            | 27 (3.8)                     |

logMAR, logarithm of the minimum angle of resolution.

Data are mean ± standard deviation, number (%) or median (interquartile range).

a Only if currently under treatment or monitored by a physician.
b Lung disease response categories were asthma, chronic obstructive pulmonary disease, 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 Anterior segment includes cornea and lens disorders.
f Posterior segment included retina and corpus vitreous disorders.
g In accordance with the International Classification of Diseases, 11th edition, 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).
Table 2. Defined Administration Rules and Performance of CAT Simulations

| CAT Simulation   | CAT_{Default} | CAT_{Alt1} | CAT_{Alt2} | CAT_{BestHealth} |
|------------------|---------------|------------|------------|-----------------|
| Defined administration rules |               |            |            |                 |
| Minimum length   | 4             | 2          | 4          | 4               |
| Maximum length   | 12            | 15         | 12         | 12              |
| Accuracy level   | 0.32          | 0.32       | 0.25       | 0.32            |
| Best health association | n/a          | n/a        | n/a        | 4 items         |
| Performance      |               |            |            |                 |
| Mean difference$^a$ (all tests) | 0.21         | 0.21       | 0.18       | 0.25            |
| Mean difference (excl. floor/ceiling tests) | 0.22         | 0.21       | 0.18       | 0.22            |
| Correlation$^a$  | 0.96          | 0.96       | 0.97       | 0.96            |
| Mean CAT length (n items) | 6.89         | 7.26       | 9.70       | 5.85            |
| Mean SE          | 0.32          | 0.32       | 0.28       | 0.34            |
| Unreliably estimated CAT $\theta$, n (%) | 106 (15.1)   | 81 (11.5)  | 106 (15.1) | 128 (18.2)     |
| Floor/ceiling effects, n (%) | 22 (3.1)     | 21 (3.0)   | 22 (3.1)   | 100 (14.2)     |

$^a$Comparisons between thetas estimated by CAT and thetas estimated by full-length item bank.

28 times a combination of missing values of the EyeQ3 (28 [19.6%]), EyeQ32 (28 [16.3%]) and the EyeQ34 (28 [14.8%]) occurred. The number of missing values in the EyeQ3, EyeQ32, and EyeQ34 increased with advancing age and the degree of vision impairment. Additionally, a higher amount of missing values of these items were found for females. After the exploration of missing values we considered the data to be missing conditionally at random. In addition, the amount of missing values within respondents appeared to be less than 10% for the majority of respondents (84.7%), having little impact on the estimation of a patient’s level of ability that the imputation was based on.

Subsequently, we performed a plausible data imputation for 339 patients. The Pearson correlation coefficient between a patient’s theta calculated after fitting the GRM model with incomplete response pattern and patient’s theta calculated after fitting the GRM model with imputed data was 0.99. The mean difference between thetas before and after imputation was 0.02.

The mean differences between the full length thetas and CAT thetas for CAT_{Default}, CAT_{Alt1}, and CAT_{Alt2} and CAT_{BestHealth} were 0.21, 0.21, 0.18, and 0.23, respectively. And the correlation coefficients were 0.96, 0.96, 0.97, and 0.95, respectively. The CAT_{Default} showed a mean test length of 6.89 and 15.1% unreliably estimated patients, of whom 3.1% had floor or ceiling scores. CAT_{Alt1} showed a mean test length of 6.89 and 15.1% unreliably estimated patients, of whom 3.1% had floor or ceiling scores. CAT_{Alt2} showed a mean test length of 9.70 and 15.1% unreliably estimated patients, of whom 3.1% had floor or ceiling scores. CAT_{BestHealth}, in which the fourth item (the most difficult one) in a series that had to be answered with having no problems was about reading small labels and instructions, showed a mean test length of 5.85 and 18.2% unreliably estimated patients, of whom 14.2% had floor or ceiling scores. The administration rules that were defined for CAT_{Default}, CAT_{Alt1}, CAT_{Alt2}, and CAT_{BestHealth}; the results of these four simulations are summarized in Table 2. The accuracy of the CAT simulations are shown in Figure 1.

Under all conditions, SEs were found to be highest at low thetas (i.e., lower level of disability). The results regarding conditional SEs of the four simulations are shown in Figure 2.

The items displaying DIF that we found in our previous calibration study were EyeQ9 looking after appearance (DIF male vs. female), EyeQ35 feeling worried or concerned about your safety at home (DIF independently vs. with help [proxy]), and EyeQ42 feeling embarrassed (DIF age ≤75 years vs. age >75 years), turned out to be administered to a negligible extent in the CAT simulations. The EyeQ9 was not administered at all and the EyeQ35 was administered in two tests (0.3%), which did not differ between different administration conditions. The EyeQ42 was administered in four tests (0.6%) under the CAT_{Default}, CAT_{Alt1}, and CAT_{BestHealth} conditions and seven times (1.0%) under the CAT_{Alt2} condition.
Figure 1. Accuracy of the CAT_Default, CAT_Alt1, CAT_Alt2 and CAT_BestHealth simulations. The x axis represents thetas estimated by the full-length EyeQ item bank (46 items) and the y axis represents the thetas estimated by CAT.

Figure 2. Conditional SEs of the CAT_Default, CAT_Alt1, CAT_Alt2 and CAT_BestHealth simulations. The x axis represents thetas estimated by the full-length EyeQ item bank (46 items) and the y axis represents the SE of the thetas estimated by CAT. The dashed line (red) represents the preferred minimal level of accuracy (0.32), corresponding with a reliability of 0.90.\textsuperscript{16}
**Discussion**

The aim of this study was to define the optimal administration rules for the use of the CAT-EyeQ in clinical practice. We used mainly real responses (96.9%) of patients having exudative retinal diseases in this study. The findings of this study suggest that a combination of the default administration rules and a best health criterion (i.e., the first four responses are associated with best health [having no complaints]) is useful.

**Missing Values and Data Imputation**

The exploration of missing values showed that missing values were partly related to other missing values in items and sex, age, and the degree of vision impairment affected the amount of missing values in patients. This could have influenced the estimation of patient’s theta. However as the missings were partly related to each other, and the within-person missings were for most patients (84.7%) fewer than five items, we considered the data to be missing (conditionally) at random. In addition, data imputation was performed with the use of patient’s estimated theta. Patients having more than 11 missings were excluded prematurely. A smaller number of within-person missing values was considered to not affect patients’ estimation. The high correlation and the small mean difference between the estimated patient theta before and after imputation, and the small percentage of missing values overall (3.1%) were considered as justification for data imputation.

**Strengths and Limitations**

A strength of this research is the use of mostly real response data for the simulations. This nature allows that the CATs are simulated more efficiently, and in this way the data used in this study most closely matches the population of patients in the ophthalmic clinical practice. The four simulations were evaluated for a variety of properties, such as the accuracy level, mean test length, mean SEs, and conditional SEs of the CATs. The latter gave insight in the performance of the CATs over the whole construct continuum, which provided information on what range of ability levels the CAT could provide reliable scores. This was also valuable information that supports the choice of applying the best health administration rule when administering the CAT-EyeQ for this target population. Although similar studies have been published about the optimization of CATs, to our knowledge no gold standard for the evaluation of the CAT simulations has been described. However, several studies describe comparing the CAT estimates with the full bank estimates regarding accuracy and efficiency as a function of the CAT test length, and the evaluation of the test length reduction or response burden.

The data on which this hybrid simulation study is based concerned the responses of patients who had a relatively good binocular visual acuity, which may be explained by the large proportion of patients who were treated monocularly with anti-vascular endothelial growth factor. In addition, patients could participate in the study if they received anti-vascular endothelial growth factor treatment for exudative retinal diseases: no minimum period of treatment was used as an inclusion criterion, because we aimed to achieve a study sample that is highly representative for the patient population in ophthalmological practice. Owing to the relatively healthy study population, a large proportion scored a low Vr-QoL score, representing a good Vr-QoL. Nevertheless, we chose to use the real response data supplemented with imputed data (hybrid method) for the simulations as this allows CATs to be simulated more effectively. Although Monte Carlo simulations, in which all response data is generated, could have led to a more equal distribution of simulees over the construct continuum, a disadvantage would be a less representative sample regarding clinical practice, which would have made it more difficult to define the optimal administration rules for the target population.

The results show that, at a lower level of theta, all four CAT simulations could not reach the required accuracy levels that were set. A possible explanation for this is that there are fewer items available at the ends of the EyeQ scale; this is the case especially for patients having fewer difficulties. This nature means that, for patients who are situated at the lower levels of the continuum, the algorithm continues to select items until another administration rule stops the CAT (e.g., the maximum number of items is reached or the best health condition is present). Although the application of the best health administration rule ensures that the burden of completing the CAT remains low, especially for patients who do not (yet) experience any complaints as a result of their vision, this does not lead to more reliable CAT scores of patients experiencing fewer difficulties in daily life owing to loss of vision. This factor might be considered as a limitation of the EyeQ item bank as a broad range of items covering the construct continuum is preferred. However, the EyeQ item bank was developed for measuring Vr-QoL with the intention to support the health care professionals to identify patients who experience difficulties in daily life owing to their eyesight and who might need additional support and enhance the referral to
rehabilitation services. From that perspective, the decreased accuracy at the lower ability levels cannot be seen as a major limitation.

In previous research the EyeQ-46 item bank was developed and calibrated, it seemed to be a unidimensional instrument measuring Vr-QoL. However, in case the unidimensional EyeQ-46 is administered as a CAT, it is very likely to assume that not all aspects of quality of life are measured within a patient. This can be seen as a limitation of using the CAT-EyeQ instead of the total EyeQ-46.

**DIF in a CAT**

Evaluating DIF and its impact on the estimation of person’s ability during the development of quality of life questionnaires is important. In our previous study, in which we calibrated the EyeQ item bank, three items in the EyeQ item bank showed DIF: for gender, age, and filling out the questionnaire with or without help. Their impact on total EyeQ scores was negligible. However, the impact of DIF on the estimation of patient’s theta obtained from a CAT could be higher because fewer items are administered and, therefore, it is possible that CAT scores are not comparable across groups. In theory, the deletion of DIF items from the item bank can be a solution, however, this is often not preferable as the affected items could assess an important aspect of the construct of interest. Another solution could be the use of group-specific item parameters. In this case, registration of patient’s characteristics in the CAT is necessary. This would mean that, for the EyeQ, where a mean test length of six items was reached using the CAT administration rules, the patient also needs to provide the additional information that is needed for group-specific testing, before filling out the PROM. However, extending the length of the questionnaire is not preferred; the aim of using CAT is mainly to minimize the burden for filling out the PROM and to decrease administration fatigue. In our study, however, because the DIF items hardly appeared in the CAT simulations, we believe that using group-specific item parameters for the CAT-EyeQ would be redundant.

**Conclusions**

In this study, we searched for the optimal administration rules that can be applied for administering the CAT-EyeQ: an instrument that is intended to assess Vr-QoL in patients who are diagnosed with exudative retinal diseases. Overall, all four simulations showed reliable CAT scores; however, conditional SEs were higher at lower levels of disability. To keep the burden to fill out the questionnaire as low as possible for patients, a combination of the default administration rules and the best health stopping rule seemed to be the best way to set the administration conditions for the use of this PROM in clinical practice. In this way, the mean test length remains low and additionally, patients who do not experience any difficulties in their daily life owing to their vision, only need to answer the minimum of four items. This study shows that measuring Vr-QoL in clinical practice using the CAT administration rules provide reliable estimated ability scores, with a negligible increase in the number of unerably estimated patients. This practice leads to a high measurement efficiency and ensures that patients having little or no Vr-QoL problems are minimally burdened with the completion of PROMs.

**Acknowledgments**

Financial support for this study was provided by Bayer B.V. 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(8):1149–1159, doi:10.1016/j.ophtha.2018.02.005.

2. 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, doi:10.1016/s2214-109x(20)30489-7.

3. Brody BL, Gamst AC, Williams RA, et al. Depression, visual acuity, comorbidity, and disability associated with age-related macular degeneration. *Ophthalmology*. 2001;108(10):1893–900; discussion 1900-1, doi:10.1016/s0161-6420(01)00754-0.
4. Bookwala J, Lawson B. Poor vision, functioning, and depressive symptoms: a test of the activity restriction model. *Gerontology*. 2011;51:798–808, doi:10.1093/geront/gnr051.

5. Hayman KJ, Kerse NM, La Grow SJ, Woulfes T, Robertson MC, Campbell AJ. Depression in older people: visual impairment and subjective ratings of health. *Optom Vis Sci*. 2007;84(11):1024–1030, doi:10.1097/OPX.0b013e318157a6b1.

6. van Nispen RMA, de Boer MR, Hoeijmakers GJ, Ringens PJ, van Rens GHMB. 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(1):18, doi:10.1186/1477-7525-7-18.

7. 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. Journal article. *J Patient Rep Outcomes*. 2018;2(1):42, doi:10.1186/s41687-018-0061-6.

8. Detmar SB. Use of HRQOL questionnaires to facilitate patient-physician communication. *Expert Rev Pharmacoecon Outcomes Res*. 2003;3(3):215–217, doi:10.1586/14737167.3.3.215.

9. Greenhalgh J, Pawson R, Wright J, et al. Functionality and feedback: a protocol for a realist synthesis of the collation, interpretation and utilisation of PROMs data to improve patient care. *BMJ Open*. 2014;4(7):e005601, doi:10.1136/bmjopen-2014-005601.

10. Pesudovs K. Item banking: a generational change in patient-reported outcome measurement. *Optom Vis Sci*. 2010;87(4):285–293, doi:10.1097/OPX.0b013e3181d408d7.

11. 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(1):79–105, doi:10.1177/0163278716684168.

12. 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* (Malden, Mass). 2021;22(7):1669–1675, doi:10.1093/pm/pnab159.

13. Patel RN, Esparza VG, Lai JS, et al. Comparison of PROMIS computerized adaptive testing versus fixed short forms in juvenile myositis. *Arthritis Care Res*. 2021 Jul 30. Online ahead of print, doi:10.1002acr.24760.

14. 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, doi:10.2147/prom.S162802.

15. Cella D, Gershon R, Lai JS, Choi S. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. *Qual Life Res*. 2007;16(Suppl 1):133–141, doi:10.1007/s11136-007-9204-6.

16. Wainer H, Dorans NJ, Eignor D, et al. Computerized adaptive testing: a primer. London: Routledge; 2000;

17. David Magis DY, von Davier AA. *Computerized adaptive and multistage testing with R*. New York: Springer; 2017:171.

18. Thompson NA, Weiss D. A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation*. 2011;16:Article 1.

19. Kallen MA, Cook KF, Amtmann D, Knowlton E, Gershon RC. Grooming a CAT: customizing CAT administration rules to increase response efficiency in specific research and clinical settings. *Quality Life Res*. 2018;27(9):2403–2413, doi:10.1007/s11136-018-1870-z.

20. Rausch-Koster TP, Luijten MAJ, Verbraak FD, van Rens G, 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, doi:10.1167/tvst.11.4.5.

21. Jones RN. Differential item functioning and its relevance to epidemiology. *Curr Epidemiol Rep*. 2019;6:174–183, doi:10.1007/s40471-019-00194-5.

22. David Magis GR, Barrada Juan Ramon. Package ‘catR’. Available at: https://cran.r-project.org/web/packages/catR/catR.pdf. Accessed September 6, 2022.

23. Chalmers RP. mirt: a multidimensional item response theory package for the R environment. *J Stat Softw*. 2012;48(6):1–29, doi:10.18637/jss.v048.i06.

24. Khorramdel LvDM, Gonzalez E, Yamamoto K. Plausible values: principles of item response theory and multiple imputations. In: Maehler D., Rammstedt B. (eds). *Large-scale Cognitive Assessment Methodology of Educational Measurement and Assessment*. New York: Springer; 2020.

25. Magis D, Barrada JR. Computerized adaptive testing with R: recent updates of the package catR. *J Stat Softw Code Snippets*. 2017;76(1):1–19, doi:10.18637/jss.v076.c01.
26. Choi SW, Swartz RJ. Comparison of CAT item selection criteria for polytomous items. *Appl Psychol Meas*. 2009;33(6):419–440, doi:10.1177/0146621608327801.

27. Abma IL, Butje BJD, Ten Klooster PM, van der Wees PJ. Measurement properties of the Dutch-Flemish patient-reported outcomes measurement information system (PROMIS) physical function item bank and instruments: a systematic review. *Health Qual Life Outcomes*. 2021;19(1):62, doi:10.1186/s12955-020-01647-y.

28. R Development Core Team R: a language and environment for statistical computing [Internet]. Available at: http://www.R-project.org. Accessed September 6, 2022.

29. Yu L, Buysse DJ, Germain A, et al. Development of short forms from the PROMIS sleep disturbance and Sleep-Related Impairment item banks. *Behav Sleep Med*. 2011;10(1):6–24, doi:10.1080/15402002.2012.636266.

30. Seo DG, Choi J. Post-hoc simulation study of computerized adaptive testing for the Korean Medical Licensing Examination. Journal of *Educational Evaluation for Health Professions*. 2018;15:14, doi:10.3352/jeehp.2018.15.14.

31. Haley SM, Coster WJ, Dumas HM, et al. Accuracy and precision of the Pediatric Evaluation of Disability Inventory computer-adaptive tests (PEDI-CAT). *Dev Med Child Neurol*. 2011;53(12):1100–1106, doi:10.1111/j.1469-8749.2011.04107.x.

32. Magis D, Raîche G. Random generation of response patterns under computerized adaptive testing with the R package catR. *J Stat Softw*. 2012;48(8):1–31, doi:10.18637/jss.v048.i08.

33. Borsboom D. When does measurement invariance matter? *Med Care*. 2006;44(11):S176–S181, doi:10.1097/01.mlr.0000245143.08679.cc.