# Development of the Presbyopia Impact and Coping Questionnaire

Nathan Johnson · Elaheh Shirneshan · Cheryl D. Coon ·

Jonathan Stokes · Ted Wells · J. Jason Lundy · David A. Andrae ·

Christopher J. Evans · Joanna Campbell

## ABSTRACT

*Introduction*: Presbyopia is a progressive, age-related visual condition that is characterized by reduced ability to focus on near/close objects, causing impacts on individuals' daily function and health-related quality of life. The Presbyopia Impact and Coping Questionnaire (PICQ) is a new patient-reported outcome (PRO) instrument that assesses presbyopia impact and use of coping behaviors among presbyopic individuals.

*Methods*: To document the impacts of presbyopia and associated coping behaviors, concept elicitation (CE) interviews were conducted with 20 presbyopic participants. Results from the CE interviews were used to develop draft items for additional testing. Following item generation, the draft PICQ was cognitively debriefed with 20 participants. Data from a phase 2 controlled clinical trial were used for psychometric analyses of the PICQ. The PICQ was administered at site visits throughout a 28-day treatment period. Confirmatory factor analysis (CFA) methods were used to guide the development of the scoring algorithm. The reliability (internal consistency, test–retest), construct validity (convergent and discriminant validity, known-groups methods), and responsiveness (Guyatt's responsiveness statistic [GRS]) of the PICQ scores were evaluated. Finally, anchor-based and distribution-based methods were used to inform thresholds for interpreting meaningful within-patient change.

*Results*: CE interviews identified the important and relevant presbyopia-related impacts and coping behaviors and 22 items were drafted and cognitively debriefed. Following minor revisions and item addition/deletion, a version of the PICQ including 23 items was subjected to psychometric testing. The analysis sample included 151 participants. The CFA established two PICQ domain scores, Coping and Impact, on 0-to-4 scales that demonstrate good model fit (root mean square error of approximation = 0.06, comparative fit index = 0.98, Tucker-Lewis index = 0.98, standardized root mean square = 0.07). Cronbach's alphas for the Coping and Impact scores were 0.89 and 0.84, respectively. Test–retest intraclass correlation coefficients were 0.77 for Coping and 0.67 for Impact. The pattern of results assessing construct validity was acceptable for the PICQ Coping and Impact scores, with the magnitude

N. Johnson · T. Wells · D. A. Andrae · C. J. Evans
Endpoint Outcomes, Boston, MA, USA

E. Shirneshan (✉) · J. Campbell
Allergan, an AbbVie company, 2525 Dupont Drive,
T2-2P, Irvine, CA 92629, USA
e-mail: Elaheh.Shirneshan@abbvie.com

C. D. Coon · J. J. Lundy
Outcometrix, St. Petersburg, FL, USA

J. Stokes
Allergan, an AbbVie company, Madison, NJ, USA

of correlations and effect sizes generally meeting a priori expectations. The corresponding GRS effect sizes for the PICQ Coping scores were −1.23 (i.e., large) for Patient Global Impression of Change (PGIC) and −0.72 (i.e., medium) for uncorrected near visual acuity (UNVA). The GRS effect sizes for the PICQ Impact scores were −0.60 (i.e., medium) for PGIC and −0.35 (i.e., small) for UNVA. Across three sets of anchor-based analyses for interpreting individual-level change, a responder threshold of −1.00 was identified for both PICQ Coping and PICQ Impact scores.

*Conclusions*: The totality of evidence from the qualitative and quantitative research establishes that the PICQ scores produced are valid and reliable measures of presbyopia impacts and coping behaviors that are important and relevant for assessing presbyopia treatment outcomes.

## Key Summary Points

Presbyopia is a progressive visual condition that has been found to significantly reduce quality of life.

To assess the impact of presbyopia and the coping strategies used to manage the condition, the Presbyopia Impact and Coping Questionnaire (PICQ) was developed in accordance with the development standards described in the Food and Drug Administration's Patient-Reported Outcomes guidance.

While there were some differences in the patient population between rounds of qualitative interviews, the Impact and Coping strategies experienced were similar across groups.

Although the psychometric evaluation of the PICQ Coping and Impact scores was based on a sample of modest size, the results were confirmed in a separate sample, thus increasing the confidence in the stability of the instrument's measurement properties.

The totality of the evidence from the qualitative research and the psychometric evaluation of the PICQ establishes a content-valid, parsimonious, and comprehensive instrument that produces domain scores with robust measurement properties.

# INTRODUCTION

Presbyopia is a vision condition in which the crystalline lens of the eye hardens, making it difficult to focus on close objects. Presbyopia sight reduction typically begins around age 40, with near-distance vision continuing to degrade as late as 65 years of age [1]. It is estimated that more than 1.8 billion people have presbyopia-related near-vision loss worldwide (including 128 million individuals in the United States) and that more than 50% of adults older than 50 years of age do not have adequate near-vision correction [2–4].

Presbyopia impacts task performance and productivity, such as reading text on paper at a close distance, compensatory behavior to change one's vision when reading information on an electronic device, and emotional impacts [3–5]. To assess the impact of presbyopia, as well as the coping strategies used to manage the condition, a new patient-reported outcome (PRO) instrument, the Presbyopia Impact and Coping Questionnaire (PICQ) was developed in accordance with the standards described in the US Food and Drug Administration (FDA's) PRO Guidance [6].

# METHODS

As part of the preliminary research to inform the assessment of presbyopia impacts and coping strategies, a targeted literature review was conducted to identify relevant measurement concepts and existing PRO instruments. Several instruments were identified and evaluated, including the National Eye Institute Visual Functioning Questionnaire-25 (NEI VFQ-25) [7, 8], National Eye Institute Refractive Error Quality of Life Instrument-42 (NEI RQL-42) [8, 9], and Near Activity Visual Questionnaire (NAVQ) [10]. Several impacts of presbyopia were identified in the context of these instruments, such as difficulty with near-distance reading ability; however, none of the instruments met the standards in the FDA's PRO Guidance due to inadequate documentation of content validity, poor psychometric measurement properties, and/or a lack of item content relevance to presbyopia symptoms and impacts [6, 11, 12]. Hence, it was determined that development of a de novo PRO instrument assessing presbyopia impacts and coping strategies was required. This study was conducted in accordance with the tenets of the Declaration of Helsinki, and all applicable local laws/regulations. Before study start, the Copernicus Group Independent Review Board (Ref. Number EOS1-14-332) reviewed and approved the qualitative study protocols, Quorum Independent Review Board (Ref. Number 00003226) reviewed and approved the clinical study protocols, and all patients provided written informed consent to participate in the studies.

To identify proximal impacts and coping mechanisms associated with presbyopia, concept elicitation (CE) interviews were conducted with individuals with clinically confirmed diagnosis of presbyopia, defined as uncorrected near visual acuity (UNVA) of 20/40 or worse at screening. Twenty individuals were chosen for in-person interviews as research has established that saturation of concepts would likely be achieved at this level [13]. Semi-structured 60-min interviews were conducted to identify relevant and bothersome presbyopia impacts and coping behaviors using inclusion and

exclusion criteria similar to the sample expected to be enrolled in presbyopia clinical trials. To achieve adequate participant representation, various recruitment targets for refractive error types (i.e., myopic, emmetropic, hyperopic, astigmatism with best-corrected visual acuity of 20/20) and sex were used. The qualitative sample was recruited from three US-based sites (Bakersfield, CA; St. Louis, MO; and Newport Beach, CA), and trained qualitative researchers conducted the interviews using a semi-structured interview guide. Audio recordings of the interviews were transcribed verbatim and anonymized by removing personal identifying information.

The interview data were coded using Atlas.ti qualitative software and then analyzed using grounded theory methods and traditional content analysis to identify themes and concepts relevant to presbyopia impacts and coping behaviors [14]. The qualitative data were assessed for conceptual saturation, and upon completion of the interview analysis, an item generation meeting was held. At the item generation meeting, items representing relevant presbyopia impact and coping concepts were discussed and reviewed for their clinical and content validity. The resulting draft instrument was comprised of a set of items covering a range of presbyopia impacts and coping behaviors.

To test the comprehension and comprehensiveness of the draft PICQ, cognitive debriefing (CD) interviews were conducted with a new sample of 20 participants with clinically confirmed presbyopia. The recruitment criteria were updated to better reflect the upcoming clinical trial population (i.e., natural or refractive surgery-corrected emmetropes at distance, with presbyopia); however, the interview methodology and data analysis for the CD interviews were similar to those used for the CE interviews.

After the CD interviews, the draft PICQ was administered during a phase 2, multicenter, double-masked, randomized, vehicle-controlled, parallel-group study (ClinicalTrials.gov identifier NCT02780115), with pilocarpine 1.25% in an optimized formulation administered once daily in the morning for 28 days in individuals with presbyopia. After the screening

visit, site visits occurred on Day 1 (Visit 1), Day 2 (Visit 2), Day 14 ± 2 (Visit 3), Day 21 ± 2 (Visit 4), and Day 28 ± 3 (Visit 5), with pilocarpine 1.25% administered in clinic at Hour 0. During the treatment-period site visits, Hour 1 was considered the peak efficacy assessment time point, whereas Hour 8 was considered outside of the peak efficacy period. All participants remained in the study for a 14-day follow-up period, during which site visits occurred on Day 1 (Visit 6), Day 7 ± 2 (Visit 7), and Day 14 ± 2 (Visit 8). The study included concurrent PRO measures; specifically, the NEI VFQ-25 and the Patient Global Impression of Change (PGIC) [7]. The modified intent-to-treat population, defined as all randomized participants with a baseline and at least one post-baseline assessment of mesopic, high-contrast UNVA, included 151 participants. The overall mean (range) age was 48.6 (40–55) years. The majority of participants were female (69.5%; 105/151), White (79.5%; 120/151), and non-Hispanic (64.9%; 98/151). To participate in the study, all subjects had to be natural or corneal laser refractive surgery-corrected emmetropes at distance with presbyopia in each eye and report subjective complaints of poor near vision that impacted activities of daily living. In terms of near visual acuity, subjects were required to have mesopic, high-contrast UNVA of 20/40 (J3) to 20/200 (J17) in each eye at the screening and baseline visits to participate in the study.

Psychometric testing began with an item-level evaluation based on response frequencies and item response theory using SAS Version 9.4. Confirmatory factor analysis (CFA) was conducted in the *lavaan* package in R Version 3.6.1 [15] to evaluate and refine the hypothesized PICQ conceptual framework consisting of a Coping domain and an Impact domain, developed based on the literature review and qualitative interviews. The CFA used weighted least squares estimation of polychoric correlations between the PICQ items at Day 28 Hour 1 based on pairwise deletion, with "not applicable" responses treated as missing to ensure a linear ordering to the response categories [16]. Factor loadings were reviewed to identify any items that did not appear to be adequate measures of the domain, and such items were considered for removal from the associated domain score. Modification indices were reviewed to identify additional model revisions that would improve overall model fit, such as the construction of testlets to account for common item content or the remapping of an item onto a different domain. Before making any revisions to the model, the item content and the qualitative evidence were reviewed to ensure that content validity was maintained. Where item content and qualitative evidence supported revisions, then alternative models were sequentially tested, allowing for each change to be evaluated before proceeding with another. The sequential model revisions concluded when factor loadings were strong (i.e., $\lambda \geq 0.40$), overall model fit was acceptable (i.e., nonsignificant chi-square statistic, root mean square error of approximation [RMSEA] < 0.10, comparative fit index [CFI] $\geq 0.95$, Tucker–Lewis index [TLI] $\geq 0.95$, standardized root mean square residual [SRMR] $\leq 0.08$), and there were no elevated modification indices (< 10.00) [17–19].

Next, the measurement properties of the PICQ domain scores were assessed. Specifically, the reliability, validity, and responsiveness of the PICQ scores were evaluated, and thresholds for interpreting meaningful within-patient changes were established. Cronbach's alpha, item-to-total correlations, and test–retest reliability were calculated to assess the internal consistency and stability of the PICQ scores over time [20, 21]. Cronbach's alpha and item-to-total correlations for the set of items comprising each domain score was computed using data from Day 28 Hour 1. Cronbach's alpha values exceeding 0.70 and item-to-total correlations of 0.40 or greater [22] were considered supportive of combining item responses into a domain score [20]. Test–retest reliability was computed using the intraclass correlation coefficient (ICC) for the PICQ scores between Day 21 Hour 1 and Day 28 Hour 1 using data from participants who were the same or slightly better or slightly worse on the PGIC at the retest time point [23]. The ICC [23] was computed using a 2-way mixed-effects regression model based on absolute agreement [24]. Excellent test–retest reliability is indicated with an ICC > 0.90, good reliability is indicated by

0.75 < ICC ≤ 0.9, moderate reliability is indicated by 0.5 < ICC ≤ 0.75, and values below 0.50 indicate poor reliability [25].

Two methods for assessing construct validity were applied to the PICQ scores. First, convergent and discriminant validity was evaluated by assessing the magnitude of the correlations between the PICQ scores and the NEI VFQ-25. Second, the construct validity of the PICQ scores was assessed using known-groups methods. For the known-groups analysis, three groups were defined based on the mesopic high-contrast UNVA: 20/125 or worse; 20/80 and 20/100; and 20/63 or better. An eta-squared effect size was computed as the between-groups sum of squares divided by the total sum of squares. Values of 0.01 to < 0.06 are considered small, 0.06 to < 0.14 are considered medium, and 0.14 or larger are considered large [26].

Responsiveness methods were used to examine score changes to assess that the PICQ scores were able to detect changes over time. Participants were classified based on concurrent PRO measures at the same time points according to the following definitions:

- PGIC

  o Improved = Complete improvement, far better, or moderately better
  o Not improved = Slightly better, no change, slightly worse, moderately worse, or far worse

- Mesopic high-contrast UNVA:

  o Improved = 3-line improvement or greater [27]
  o Not improved = Worsening, no change, or less than a 3-line improvement.

Guyatt's responsiveness statistic (GRS) was reported as an effect size comparing the improved group to the not-improved group [28]. The GRS is computed as the mean change for the target group (i.e., improved) minus the mean of the change for the comparison group (i.e., not improved) divided by the standard deviation (SD) of the comparison group (e.g., not improved). GRS is interpreted as small (0.20 to < 0.50), medium (0.50 to < 0.80), and large

(≥ 0.80) according to standards defined by Cohen [26].

Finally, anchor-based methods and distribution-based methods were used to establish thresholds for interpreting clinically meaningful within-person change on the PICQ scores [6, 29]. The primary anchor was the PGIC, for which meaningful improvement was defined as complete improvement, far better, or moderately better at Day 28 Hour 0. Mesopic high-contrast UNVA was used as a supportive anchor, with meaningful improvement defined as a 3-line improvement or greater from Day 1 Hour 0 to Day 28 Hour 1. Polyserial correlations between change on the PICQ domain scores and the anchors were computed to confirm that the anchors were related enough to the PICQ scores to be considered appropriate for use (i.e., $r \geq 0.30$) [30]. Anchor-based estimates were produced using empirical cumulative distribution functions (eCDFs), classification statistics (i.e., sensitivity, specificity, positive predictive value, negative predictive value), and discriminant analysis. Distribution-based estimates were produced using the one-half SD method and the standard error of measurement, both based on data at Day 1 Hour 0. Results were compared across all anchors and all methods to triangulate on one plausible threshold estimate for each PICQ domain score, and anchors with stronger correlations were given greater priority during the triangulation process.

# RESULTS

## Concept Elicitation Results

The institutional review board approval for the study was granted on July 29, 2014. The CE sample was mostly female ($n = 14$; 70.0%) with an average age of 50.1 years (range: 41–57 years). White was the most common ($n = 9$; 45.0%) race, followed by Black or African American ($n = 4$; 20.0%), Native Hawaiian or Pacific Islander ($n = 2$; 10.0%), and Asian ($n = 1$; 5.0%); four participants (20.0%) provided an "other" response for race. The majority of participants were non-Hispanic/Latino ($n = 15$; 75.0%). The median and mode for near visual

acuity in the right eye (OD), left eye (OS), and both eyes (OU) were 20/50.

A total of 15 coping activities were reported by participants to help with decreasing near-vision blurriness and/or improving their ability to read text up close. The most frequently reported coping activities that were identified as potential PRO items included changing font or screen size on electronic reading devices ($n = 19$; 95.0%); holding materials farther out to read ($n = 18$; 90.0%); squinting ($n = 14$; 70.0%); wearing glasses ($n = 12$; 60.0%); adjusting screen brightness/adjusting lighting conditions ($n = 10$; 50.0%); and resting eyes ($n = 10$; 50.0%). A total of 58 impacts of presbyopia were reported, with feeling angry or frustrated ($n = 17$; 85.0%), forgetting glasses ($n = 17$; 85.0%), and relying on others to read materials ($n = 14$; 70.0%) being reported most frequently. Impacts with the highest average bothersome ratings, rated by at least five (25.0%) participants on a 0–10 scale (0 being "not bothersome at all" and 10 being "extremely bothersome"), were reading books/newspapers/magazines, feeling older, and reading in low/dim light ($\overline{x}$= 7.5, 7.2, and 6.9, respectively). See Table 1 for a list of the most frequently reported coping activities and impacts reported by participants.

Saturation was assessed via a structured reporting process, as concepts emerging from the interviews were analyzed for saturation in sets in the order the data were collected (i.e., four rounds of five interviews each). Based on this analysis, it was determined that conceptual saturation was achieved, and the study sample size was sufficient to achieve the research objectives.

## PICQ Content Development

Frequently reported impacts and coping strategies identified in the CE interviews were included in the PICQ. The initial version of the PICQ consisted of 22 items assessing the impacts of near-vision problems and coping strategies using a 7-day recall period. Each item within this instrument measured a common activity or situation impacted by presbyopia or a coping strategy used to overcome the impact of presbyopia. For two concepts, "feeling annoyed reading materials at close distance" and "difficulty exercising or playing sports," two pairs of items, one assessing frequency and one assessing severity, were developed for further testing in CD interviews with the intent that only one item for each concept would be retained in the final version of the instrument. The majority of items assessed the frequency with which certain impacts and coping behaviors occur and are reported on a 0–to–4 verbal response scale ranging from "never" to "all of the time" or "not at all difficult" to "very difficult" or "not at all inconvenient" to "extremely inconvenient."

## Cognitive Debriefing Results

Twenty individuals with clinically confirmed presbyopia participated in CD interviews. More than half of the sample were female ($n = 11$; 55.0%) with an average age of 48.8 years (range: 41–59 years). Non-Hispanic/Latino represented the largest ethnicity group of participants ($n = 16$; 80.0%) and the majority of the participants were White ($n = 14$; 70.0%). The medians for UNVA in the OD, OS, and OU were 20/50, 20/45, and 20/40, respectively. Overall, the draft PICQ items were interpreted as intended by most participants, and a majority of participants ($n = 15$; 75.0%) reported that the recall period was acceptable and that they were able to recall impacts and coping behaviors over a 7-day period. In addition, minor item revisions were made to improve the clarity of the items and reduce the length of the item stems. Overall, for the draft items tested in the CD interviews, one item was deleted, one item was added, and one item was split into two new items. The resulting version of the PICQ that was taken into psychometric testing included a total of 23 items measuring the concepts of Coping and Impact.

## Psychometric Analysis Results

PICQ item evaluations at Day 1 Hour 0 identified one coping item (interact with tablet) that was suspected of instability based on an elevated item response theory (IRT) slope and two

**Table 1** Most frequently reported presbyopia coping activities and impacts

| Domain | Concept | $n$ | % | Average bothersome rating |
|--------|---------|-----|---|---------------------------|
| Coping | Changing font or screen size on electronic reading devices | 19 | 95.0 | N/A[a] |
| | Holding materials farther out to read | 18 | 90.0 | N/A |
| | Squinting | 14 | 70.0 | N/A |
| | Wearing glasses | 12 | 60.0 | N/A |
| | Adjusting screen brightness/adjust lighting conditions | 10 | 50.0 | N/A |
| | Resting eyes | 10 | 50.0 | N/A |
| | Holding materials closer to read | 5 | 25.0 | N/A |
| | Taking off distance glasses to read up close | 4 | 20.0 | N/A |
| Impact | Difficulty reading a restaurant menu | 20 | 100.0 | 7.5 |
| | Difficulty reading text on computer/laptop | 20 | 100.0 | 6.1 |
| | Difficulty reading text on a cell phone/caller ID | 19 | 95.0 | 5.9 |
| | Difficulty reading nutrition/recipe labels | 19 | 95.0 | 6.5 |
| | Difficulty reading in low/dim light | 19 | 95.0 | 6.9 |
| | Feeling angry/frustrated | 17 | 85.0 | 6.0 |
| | Forgetting glasses | 17 | 85.0 | N/A |
| | Difficulty reading a book | 15 | 75.0 | 7.1 |
| | Relying on others to read materials | 14 | 70.0 | 6.4 |
| | Difficulty working with small objects (e.g., sewing/crafting) | 13 | 65.0 | 7.3 |
| | Feeling older | 11 | 55.0 | 7.2 |
| | Feeling self-conscious | 8 | 40.0 | 6.3 |
| | Taking longer to complete a task | 7 | 35.0 | N/A |
| | Feeling less confident | 6 | 30.0 | 2.0 |
| | Inconvenience of switching between looking at something up close and then far away (e.g., looking at dashboard while driving) | 6 | 30.0 | N/A |

Participants were asked to provide bothersome ratings on a 0–10 scale (0, not bothersome at all; 10, extremely bothersome) for impacts

[a] Participants were not asked to provide bothersome ratings for coping activities. Not all participants provided impact bothersome ratings due to interview time constraints

impact items (difficulty with daily chores; difficulty with walking or balance) that demonstrated pronounced floor effects. Upon further review of the qualitative evidence, it was determined that these concepts were less frequently reported and not as relevant to participants as other, more salient concepts; therefore, these three items were removed from the instrument.
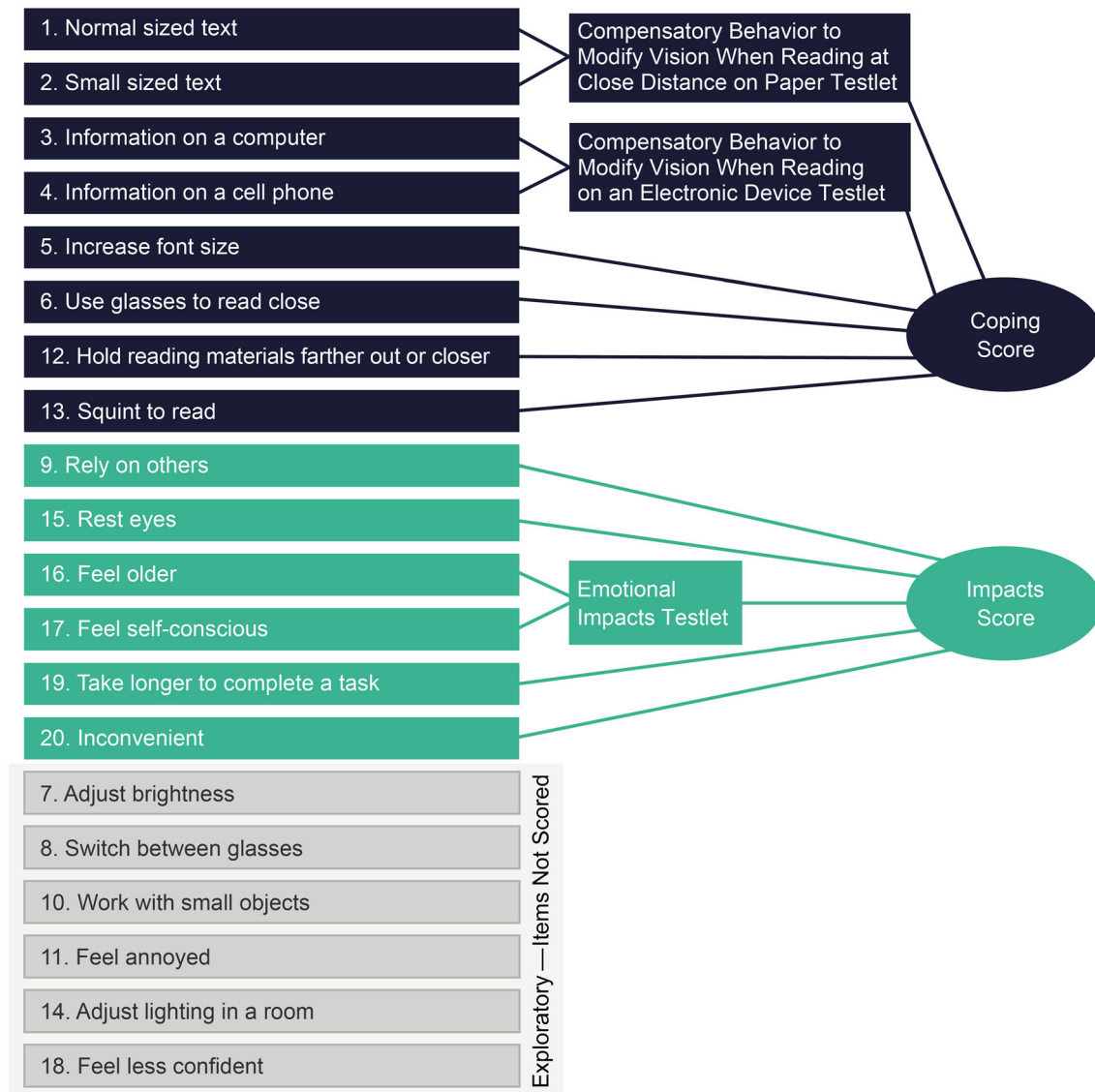
**Fig. 1** Final PICQ conceptual framework. *PICQ* Presbyopia Impact and Coping Questionnaire.

The initial 2-factor CFA of the remaining 20 items produced poor model fit and identified a number of very high inter-item correlations, indicative of item redundancy, as well as instances of item cross-loadings between factors. A series of CFA models refined the hypothesized relationships between items and factors and between item pairs as informed by modification indices and qualitative evidence for each item (Table 2). This stepwise approach resulted in the removal of six items from scoring that demonstrated poor fit to the conceptual framework and more limited qualitative evidence (i.e., adjust brightness, switch between glasses, work with small objects, feel annoyed, adjust lighting in a room, feel less confident), the remapping of one item to a different factor (i.e., rely on others from Impact to Coping), and the construction of testlets for items with related content (i.e., reading text at a close distance on paper; compensatory behavior to change one's vision when reading information on an

**Table 2** Summary of CFA model results and revisions

| Model | Factor loadings and covariances | Fit statistics[a] | Modification indices > 10.00 |
|---|---|---|---|
| Base case | Coping: 0.66–0.94<br>Impact: 0.73–0.99<br>Covariance = 0.88 | $P(\chi^2) < 0.001$<br>**RMSEA = 0.09**<br>**CFI = 0.98**<br>**TLI = 0.98**<br>**SRMR = 0.11** | Coping with i10, i11, i17, and i18<br>Impact with i2 and i14<br>i1 with i2<br>i4 with i3 and i6<br>i5 with i7<br>i11 with i14<br>i17 with i16 and i18<br>i19 with i20 |
| Revision #1: Free error covariances between items with similar content (i1 + i2, i3 + i4, i5 + i7) based on large modification indices and delete i18 based on limited content validity | Coping: 0.64–0.91<br>Impact: 0.70–0.98<br>Covariance = 0.92 | $P(\chi^2) < 0.001$<br>**RMSEA = 0.07**<br>**CFI = 0.99**<br>**TLI = 0.99**<br>SRMR = 0.09 | Coping with i11, i16, and i17<br>Impact with i14<br>i1 with i3<br>i4 with i6<br>i7 with i12<br>i11 with i14<br>i16 with i17<br>i19 with i20 |
| Revision #2: Delete i11 based on large modification index on Coping domain | Coping: 0.65–0.91<br>Impact: 0.73–0.92<br>Covariance = 0.87 | $P(\chi^2) < 0.001$<br>**RMSEA = 0.07**<br>**CFI = 0.99**<br>**TLI = 0.99**<br>SRMR = 0.09 | Coping with i10 and i17<br>Impact with i9<br>i1 with i3<br>i4 with i6<br>i7 with i12<br>i16 with i17 |
| Revision #3: Free error covariances between i16 + i17 based on large modification index | Coping: 0.65–0.91<br>Impact: 0.67–0.91<br>Covariance = 0.89 | $P(\chi^2) < 0.001$<br>**RMSEA = 0.07**<br>**CFI = 0.99**<br>**TLI = 0.99**<br>SRMR = 0.08 | Coping with i10<br>Impact with i9<br>i1 with i3<br>i4 with i6<br>i7 with i12 |

**Table 2** continued

| Model | Factor loadings and covariances | Fit statistics[a] | Modification indices > 10.00 |
|---|---|---|---|
| Revision #4: Delete i10 based on large modification index on Coping domain | Coping: 0.63–0.91<br>Impact: 0.70–0.85<br>Covariance = 0.84 | $P(\chi^2) < 0.001$<br>***RMSEA = 0.07***<br>***CFI = 0.99***<br>***TLI = 0.99***<br>SRMR = 0.08 | Impact with i9<br>i1 with i3<br>i4 with i6<br>i7 with i12 |
| Revision #5: Move i9 to Impact based on large modification index on Impact domain | Coping: 0.63 to 0.91<br>Impact: 0.69 to 0.85<br>Covariance = 0.83 | $P(\chi^2) < 0.001$<br>***RMSEA = 0.06***<br>***CFI = 0.99***<br>***TLI = 0.99***<br>***SRMR = 0.08*** | Impact with i5 and i14<br>i1 with i3<br>i4 with i6<br>i7 with i12 |
| Revision #6: Delete i14 based on large modification index on Impact domain | Coping: 0.62–0.92<br>Impact: 0.70–0.85<br>Covariance = 0.81 | $P(\chi^2) < 0.001$<br>***RMSEA = 0.06***<br>***CFI = 0.99***<br>***TLI = 0.99***<br>***SRMR = 0.08*** | Impact with i5<br>i7 with i12 |
| Revision #7: Delete i7 based on large modification index with i12 | Coping: 0.65–0.92<br>Impact: 0.71–0.86<br>Covariance = 0.82 | $P(\chi^2) < 0.001$<br>***RMSEA = 0.05***<br>***CFI = 1.00***<br>***TLI = 1.00***<br>***SRMR = 0.07*** | Impact with i5<br>i5 with i15 |
| Revision #8: Delete i8 based on poor content validity | Coping: 0.65–0.93<br>Impact: 0.71–0.86<br>Covariance = 0.81 | $P(\chi^2) < 0.001$<br>***RMSEA = 0.06***<br>***CFI = 1.00***<br>***TLI = 1.00***<br>***SRMR = 0.07*** | Impact with i5<br>i5 with i15 |
| Revision #9: Turn error covariances into testlet variables to evaluate final scoring model | Coping: 0.67–0.98<br>Impact: 0.66–0.85<br>Covariance = 0.88 | $P(\chi^2) < 0.001$<br>***RMSEA = 0.06***<br>***CFI = 0.98***<br>***TLI = 0.98***<br>***SRMR = 0.07*** | None |

*CFI,* comparative fit index, *RMSEA* root mean square error of approximation, *SRMR* standardized root mean square residual, *TLI* Tucker-Lewis index

[a] Fit statistics that met pre-specified values for indicating good model fit are bolded and italicized

**Table 3** PICQ construct validity as assessed using known-groups defined by mesopic high-contrast UNVA

| PICQ Score | Time | Groups | N | Mean | SD | Effect size[a] | p-value[b] |
|---|---|---|---|---|---|---|---|
| Coping | Day 1 Hour 0 | 20/125 or worse | 20 | 3.21 | 0.52 | 0.19 | < 0.001 |
| | | 20/80 and 20/100 | 78 | 2.69 | 0.67 | | |
| | | 20/63 or better | 53 | 2.15 | 0.86 | | |
| | Day 21 Hour 1 | 20/125 or worse | 2 | 3.18 | 0.02 | 0.15 | < 0.001 |
| | | 20/80 and 20/100 | 30 | 2.69 | 0.79 | | |
| | | 20/63 or better | 115 | 1.84 | 0.93 | | |
| Impact | Day 1 Hour 0 | 20/125 or worse | 20 | 1.94 | 0.89 | 0.05 | 0.021 |
| | | 20/80 and 20/100 | 78 | 1.61 | 0.87 | | |
| | | 20/63 or better | 53 | 1.30 | 0.99 | | |
| | Day 21 Hour 1 | 20/125 or worse | 2 | 1.00 | 0.71 | 0.04 | 0.056 |
| | | 20/80 and 20/100 | 30 | 1.28 | 0.92 | | |
| | | 20/63 or better | 115 | 0.89 | 0.76 | | |

*ANOVA* analysis of variance, *PICQ* Presbyopia Impact and Coping Questionnaire, *UNVA* uncorrected near visual acuity
[a] Effect size is eta-squared, computed as the between-groups sum of squares divided by the total sum of squares
[b] p-value is based on an *F*-test from an ANOVA comparing all levels of the group

electronic device; emotional impacts). This produced an 8-item Coping domain score and a 6-item Impact domain score that balance adequate model fit with retention of concepts that are important to patients, while accounting for item dependencies through the construction of testlets for pairs of items with similar content. Within each testlet, item pairs are averaged, and the testlet values are averaged with the other items within each domain to produce the domain score (Fig. 1).

The Cronbach's alpha value for both PICQ domain scores using item responses and testlet values at Day 28 Hour 1 exceeded the recommended threshold of 0.70, with alpha of 0.89 for the Coping domain score and 0.84 for the Impact domain score. This suggests that it is appropriate to combine the values from items/testlets to form Coping and Impact scores. Further, all of the item-to-total correlations exceeded 0.40, which also supports the retention and combination of items in the Coping and Impact domains. The ICC between Day 21 Hour 1 and Day 28 Hour 1 for stable participants was 0.77

for the Coping domain score and 0.67 for the Impact domain score. Thus, the PICQ scores were considered to be reliable over time.

The assessment of validity considered correlations between the PICQ domain scores and the NEI VFQ-25 domain scores at screening/Day 1 Hour 0 and at Day 21 Hour 1. Across all NEI VFQ-25 domain scores, the Near-vision Activities domain had the strongest correlations with the PICQ Coping score at Day 21 Hour 0 ($r = -0.45$), whereas the Role Difficulties domain had the strongest correlations with the PICQ Impact score at Day 21 Hour 0 ($r = -0.65$). The correlation with Near-vision Activities was expected to be strong ($r \geq 0.05$) for both PICQ domain scores, but the correlations were only low to moderate in strength (range: $-0.27$ to 0.45 for Coping and $-0.25$ to $-0.40$ for Impact). This slightly lower strength of correlations than expected may be due to the NEI VFQ-25 Near-vision Activities domain measuring difficulty when performing specific near-vision activities, while the PICQ measures compensatory behavior to make activities less difficult or the impact

**Table 4** Responsiveness statistics for the PICQ score changes from Day 1 Hour 0 to Day 28 Hour 1

| PICQ Score | Group | Levels[a] | N | Mean | SD | Guyatt's responsiveness statistic[b] |
|---|---|---|---|---|---|---|
| Coping | PGIC | Improved | 48 | −1.32 | 0.83 | −1.23 |
| | | Not improved | 95 | −0.48 | 0.68 | |
| | Mesopic high-contrast UNVA | Improved | 46 | −1.13 | 0.83 | −0.72 |
| | | Not improved | 100 | −0.57 | 0.78 | |
| Impact | PGIC | Improved | 48 | −1.01 | 0.89 | −0.60 |
| | | Not improved | 95 | −0.54 | 0.79 | |
| | Mesopic high-contrast UNVA | Improved | 46 | −0.88 | 0.84 | −0.35 |
| | | Not improved | 100 | −0.59 | 0.84 | |

Change is computed as Day 28 Hour 1 minus Day 1 Hour 0. Negative scores indicate improvement

*PGIC* Patient Global Impression of Change, *PICQ* Presbyopia Impact and Coping Questionnaire, *UNVA* uncorrected near visual acuity

[a] PGIC improved = complete improvement, far better, or moderately better at Day 28 Hour 1. PGIC not improved = slightly better, no change, slightly worse, moderately worse, or far worse, at Day 28 Hour 1. Mesopic high-contrast UNVA improved = 3-line improvement or greater from Day 1 Hour 0 to Day 28 Hour 1. Mesopic high-contrast UNVA not improved = worsening, no change, or less than a 3-line improvement from Day 1 Hour 0 to Day 28 Hour 1

[b] Guyatt's responsiveness statistic is computed as the mean change for the improved group minus the mean change for the not-improved group, all divided by the standard deviation of change scores for the not-improved group

of the difficulties. Correlations with the functional impact domains (i.e., Social Functioning, Mental Health, Role Difficulties, Dependency) were expected to be moderate ($r \geq 0.30$), particularly with the PICQ Coping domain, and they reached this threshold between PICQ Coping and NEI VFQ-25 Mental Health and Role Difficulties and between PICQ Impact and all four NEI VFQ-25 functional impact domains. Both PICQ Coping and Impact were poorly correlated with General Health, Ocular Pain, and Color Vision, ($r < 0.20$), which supports the discriminant validity of the PICQ based on its poor relationship with dissimilar concepts. This pattern of correlations is desirable, as more presbyopia-related impacts on the PICQ are associated with greater functional impacts of various types on the NEI VFQ-25 post-treatment.

PICQ domain scores were evaluated for three groups that were known to differ based on the clinical outcome of mesopic high-contrast UNVA at Day 1 Hour 0 and Day 28 Hour 1 (Table 3). PICQ Coping scores at both time points were able to significantly distinguish between clinically relevant levels of mesopic high-contrast UNVA ($p < 0.001$), while PICQ Impact scores reached significance at Day 1 Hour 0 ($p = 0.02$), but not at Day 21 Hour 1 ($p = 0.06$). The best mesopic high-contrast UNVA values (i.e., 20/63 or better) had the best scores on the PICQ domains, while the poorest mesopic high-contrast UNVA values (i.e., 20/125 or worse) had the poorest scores on the PICQ domains at Day 1 Hour 0, with mean of 3.21 for Coping and 1.94 for Impact (the sample size at Day 21 Hour 1 was too small to interpret for this group). The effect sizes for PICQ Coping ranged from 0.15 to 0.19, which are considered large, while the effect sizes for PICQ Impact ranged from 0.04 to 0.05, which are considered small.

To assess responsiveness, PICQ domain change scores were evaluated for groups defined as improved and not improved based on change in the patient outcome of PGIC and change in the clinical outcome of mesopic high-contrast UNVA from Day 1 Hour 0 to Day 28 Hour 1
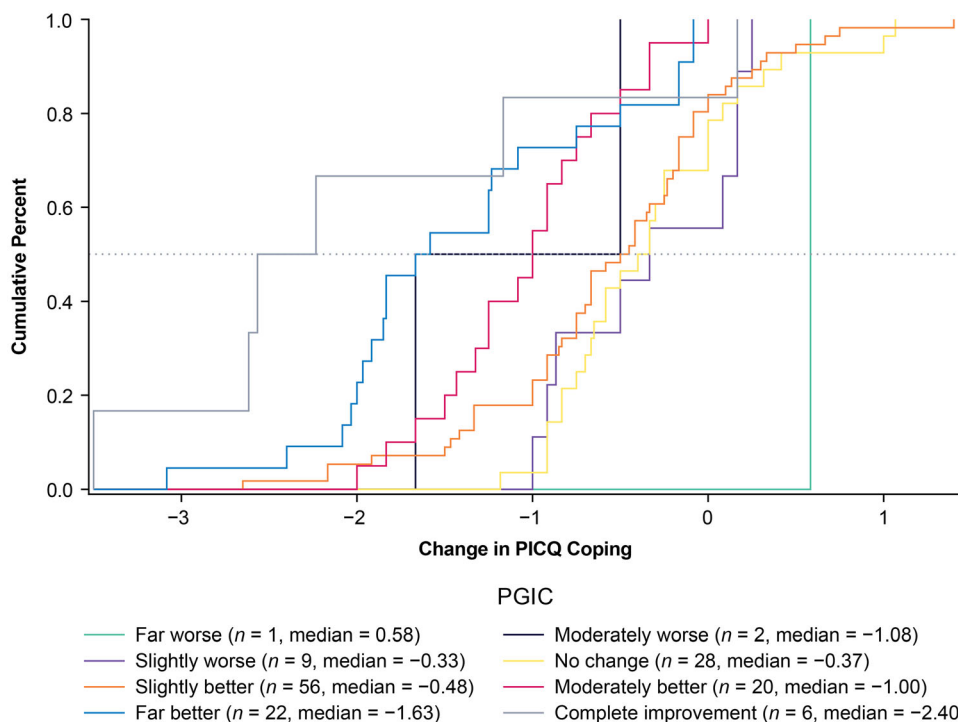
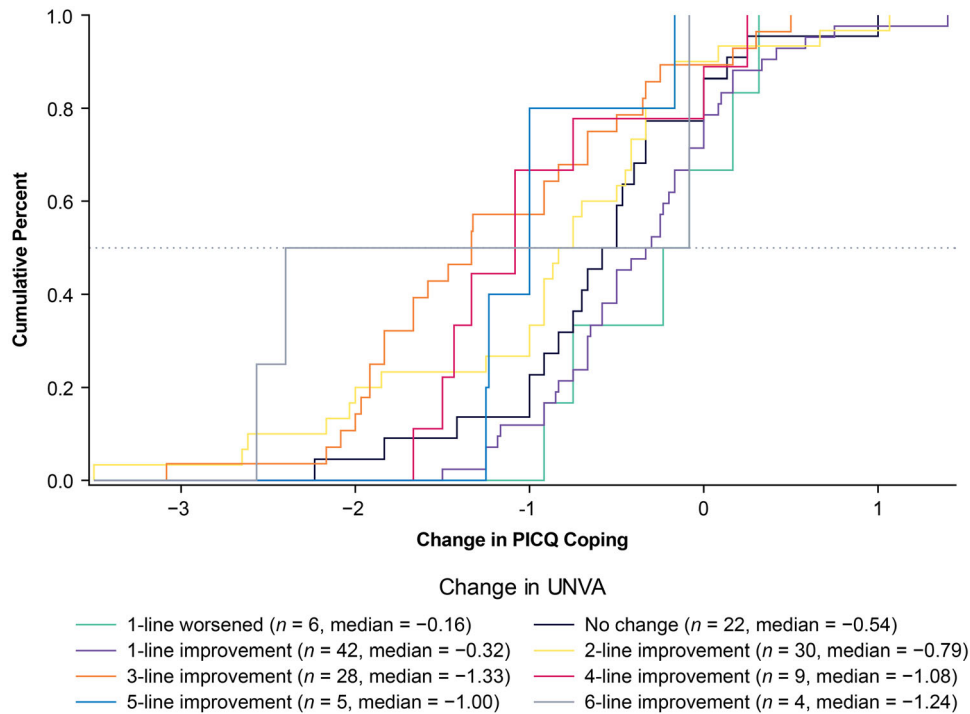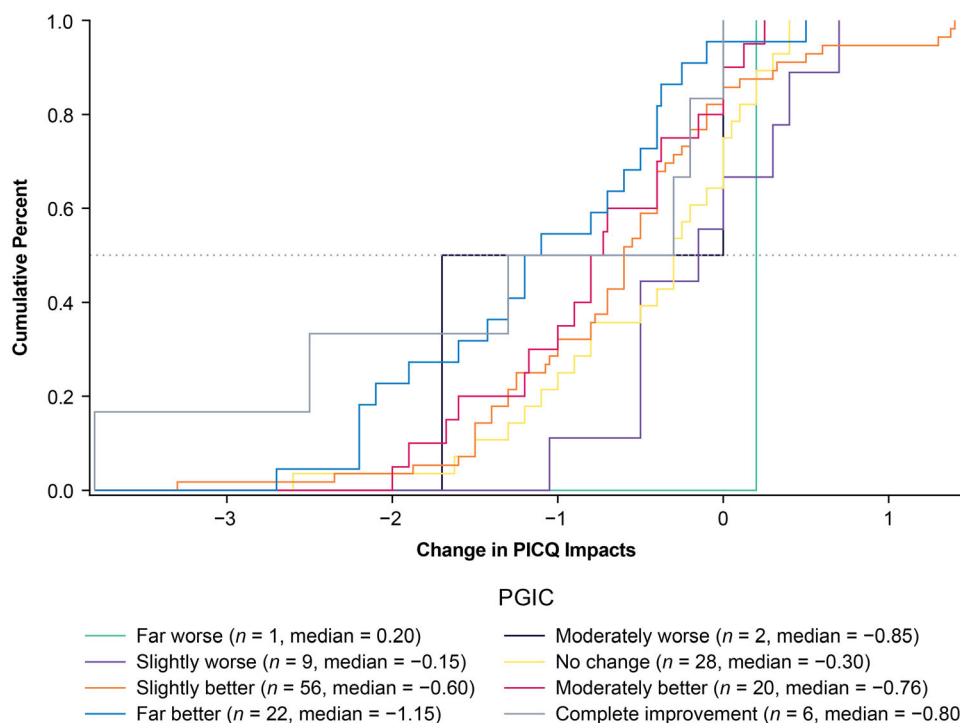**Fig. 2** eCDF of PICQ Coping by PGIC using change scores from Day 1 Hour 0 to Day 28 Hour 1. Note. Change is computed as Day 28 Hour 1 minus Day 1 Hour 0; negative scores indicate improvement. *eCDF,* empirical cumulative distribution functions, *PGIC* Patient Global Impression of Change, *PICQ* Presbyopia Impact and Coping Questionnaire

(Table 4). The corresponding GRS effect sizes comparing PICQ Coping scores by improvement groups were −1.23 (i.e., large) for PGIC and −0.72 (i.e., medium) for UNVA. The corresponding GRS effect sizes comparing PICQ Impact scores by improvement groups were −0.60 (i.e., medium) for PGIC and −0.35 (i.e., small) for UNVA.

Correlations were computed between the anchors chosen for meaningful change setting analyses and the PICQ domain scores. The correlation with PGIC was −0.52 for Coping and −0.33 for Impact, whereas the correlation with UNVA was −0.33 for Coping and −0.21 for Impact. Due to the insufficient correlation, analyses for Impact based on UNVA were reviewed but discounted during the triangulation process, which involved considering results across different methods and different anchors to identify a plausible responder threshold.

ECDFs, classification statistics, and discriminant analysis were used to produce responder threshold estimates. Across the two anchors and three types of anchor-based methods, the possible responder thresholds for PICQ Coping scores ranged from −0.50 to −2.90. The PICQ Coping score responder threshold proposed based on these results is −1.00. This value of −1.00 was highly informed by the eCDF estimate from the PGIC anchor (Fig. 2), as this anchor had a strong correlation with PICQ Coping ($r = -0.52$) and had a median value of −1.00 for the moderately better category. The eCDF estimate from the UNVA anchor (Fig. 3) was −1.33, but there was a large separation between the category corresponding to meaningful improvement (i.e., 3-line improvement) and the category corresponding to less than meaningful improvement (i.e., 2-line improvement) at −1.00, thus supporting the threshold of −1.00. This threshold exceeds the distribution-based estimates of 0.40 for one-half SD and 0.38 for standard error of measurement, suggesting that the proposed threshold is

**Fig. 3** eCDF of PICQ Coping by change in UNVA using change scores from Day 1 Hour 0 to Day 28 Hour 1. Note. Change is computed as Day 28 Hour 1 minus Day 1 Hour 0; negative scores indicate improvement. *eCDF* empirical cumulative distribution functions, *PICQ* Presbyopia Impact and Coping Questionnaire, *UNVA* uncorrected near visual acuity

sufficiently large to be able to be detected by the instrument.

Using the same anchors and anchor-based methods, the possible responder thresholds for PICQ Impact scores ranged from −0.30 to −1.90. The PICQ Impact score responder threshold proposed based on these results was also −1.00. This value of −1.00 was also informed by the eCDF estimate from the PGIC anchor (Fig. 4), which had a median value of −0.76 for the moderately better category, although the proposed threshold was rounded up to −1.00 to facilitate ease of interpretation. Despite its low correlation, the UNVA anchor eCDF (Fig. 5) also showed a large separation between 3-line improvement and 2-line improvement at −1.00, further supporting the threshold of −1.00. This threshold exceeds the distribution-based estimates of 0.47 for one-half SD and 0.54 for standard error of measurement, indicating that the proposed threshold is adequately large for measurement by the instrument.

## DISCUSSION

Presbyopia is a progressive visual condition that is caused by the inability of the eye to focus on near objects. Uncorrected presbyopia has been found to significantly reduce quality of life, as patients experience a reduction in near visual acuity. While there are existing PRO instruments that assess reduced quality of life due to presbyopia [8, 10–12], the team's landscape review found that these instruments were unsuitable for use in clinical studies for one or more of the following reasons: inadequate documentation of content validity or other requirements as per the FDA's PRO Guidance [6], poor psychometric measurement properties, or the focus was on the impact of interventions rather than impaired function. Additionally, based on the literature review, key concepts pertaining to presbyopia impacts and coping mechanisms were mapped into a preliminary
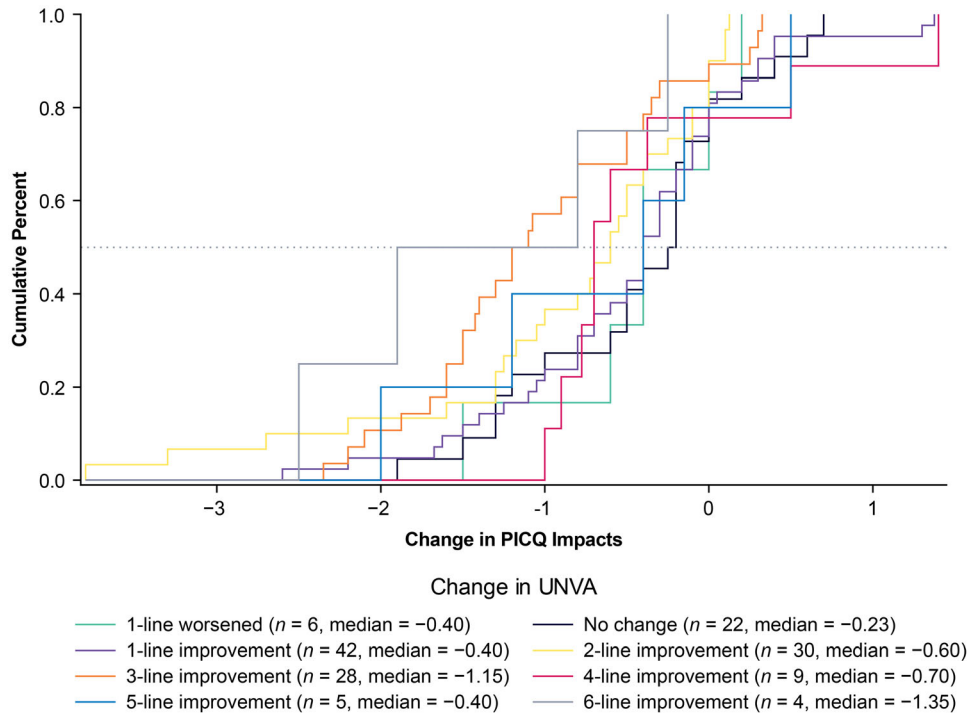
**Fig. 4** eCDF of PICQ Impact by PGIC using change scores from Day 1 Hour 0 to Day 28 Hour 1. Note. Change is computed as Day 28 Hour 1 minus Day 1 Hour 0; negative scores indicate improvement. *eCDF* empirical cumulative distribution functions, *PGIC* Patient Global Impression of Change, *PICQ* Presbyopia Impact and Coping Questionnaire, *UNVA* uncorrected near visual acuity

conceptual framework that was used to further aid in the subsequent development of the PICQ.

To evaluate the benefits of treatment for presbyopic individuals, a new PRO instrument, the PICQ, was developed following the standards described in the FDA's PRO Guidance [6]. As a foundational first step, qualitative research was conducted to identify important and relevant presbyopia impacts and coping behaviors used to manage the condition. Based on the CE research, the draft instrument was developed and then cognitively debriefed. Following the qualitative phase of development, the refined PICQ was included in a phase 2 clinical trial for psychometric evaluation. One limitation of the qualitative research is that the inclusion criteria for presbyopia participants differed between the CE and CD phases. While the CE phase included a broader range of participants (i.e., presbyopia participants that were emmetropic, myopic, hyperopic, and/or with astigmatism), the CD

phase only included natural or surgery-corrected emmetropes at distance. Regardless, most participants in the CD phase still found the PICQ item content to be relevant to their experience.

As part of the psychometric evaluation, scoring analyses were implemented to evaluate the domain structure and determine a scoring algorithm, followed by an assessment of the reliability, validity, and responsiveness of the PICQ scores and calculation of meaningful within-patient change thresholds for PICQ scores.

The PICQ Coping and Impact scores were found to have high levels of reliability and construct validity, and good responsiveness to change. It is notable that the responsiveness effect sizes for groups defined by the PGIC were stronger than the effect sizes for groups defined by UNVA, which is an understandable and desirable pattern because the PGIC is a direct

**Fig. 5** eCDF of PICQ Impact by change in UNVA using change scores from Day 1 Hour 0 to Day 28 Hour 1. Note. Change is computed as Day 28 Hour 1 minus Day 1 Hour 0; negative scores indicate improvement. *eCDF* empirical cumulative distribution functions, *PICQ* Presbyopia Impact and Coping Questionnaire, *UNVA* uncorrected near visual acuity

patient report of change that may be a more accurate expression of the patient experience than UNVA, a clinical outcome that does not involve direct patient input or interpretation.

In addition, assessment of the meaningful change thresholds using multiple methods and anchors produced estimates of −1.00 for both of the PICQ domain scores. The PICQ Coping and Impact thresholds exceeded the distribution-based estimates, and they can be considered sufficiently large to be reliably measured by the scale. Thus, there is support across the range of methods and studies that −1.00 is an appropriate responder threshold for both the PICQ Coping and Impact scores.

While the psychometric evaluation of the PICQ Coping and Impact scores identified good measurement properties, one limitation of this research is the modest sample size available from the phase 2 clinical trial. While a larger sample would have been desirable, this sample

of 151 meets COSMIN's (COnsensus-based Standards for the selection of health Measurement INstruments) criteria for being "very good" for reliability and validity (e.g., $N \geq 100$) and for being "adequate" for factor analysis (i.e., 5 times the number of items and $N \geq 100$) and responsiveness (i.e., 30–50 patients in the smallest group) [31]. To increase the confidence in the results of this psychometric study, the measurement properties of the PICQ domain scores were confirmed in a sample of 323 participants from an independent, blinded, phase 3 clinical trial (NCT03804268/GEMINI 1), which included a broad range of emmetrope and non-emmetrope participants.

## CONCLUSION

This PRO instrument development research has resulted in a content-valid and psychometrically

sound instrument designed to evaluate presbyopia-related impacts and coping behaviors that are important and relevant to patients. The totality of the evidence from the qualitative research and the phase 2 psychometric evaluation of the PICQ establishes a content-valid, parsimonious, and comprehensive instrument that produces domain scores with robust measurement properties. The psychometric properties of the PICQ domain scores were confirmed in an independent, blinded phase 3 clinical trial sample, finding the scores to be reliable, valid, and responsive, further supporting the use of the PICQ in constructing clinical trial endpoints.

# ACKNOWLEDGEMENTS

# REFERENCES

1. Wolffsohn JS, Davies LN. Presbyopia: effectiveness of correction strategies. Prog Retin Eye Res. 2019;68: 124–43.

2. Bourne RRA, Flaxman SR, Braithwaite T, et al. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. Lancet Glob Health. 2017;5(9): e888–97.

3. Fricke TR, Tahhan N, Resnikoff S, et al. Global prevalence of presbyopia and vision impairment from uncorrected presbyopia: systematic review, meta-analysis, and modelling. Ophthalmology. 2018;125(10):1492–9.

4. Frick KD, Joy SM, Wilson DA, Naidoo KS, Holden BA. The global burden of potential productivity loss from uncorrected presbyopia. Ophthalmology. 2015;122(8):1706–10.

5. Wolffsohn JS, Leteneux-Pantais C, Chiva-Razavi S, et al. Social media listening to understand the lived experience of presbyopia: systematic search and content analysis study. J Med Internet Res. 2020;22(9):e18306–e18306.

6. US Food and Drug Administration. Guidance for industry on patient-reported outcome measures: use in medical product development to support labeling claims. Fed Regist. 2009;74(235):65132–3.

7. Mangione CM. Development of the 25-list-item National Eye Institute Visual Function Questionnaire. Arch Ophthalmol. 2001;119(7):1050.

8. Berry S, Mangione CM, Lindblad AS, McDonnell PJ. Development of the National Eye Institute refractive error correction quality of life questionnaire. Ophthalmology. 2003;110(12):2285–91.

9. Hays RD, Mangione CM, Ellwein L, Lindblad AS, Spritzer KL, McDonnell PJ. Psychometric properties of the National Eye Institute-Refractive Error Quality of Life instrument. Ophthalmology. 2003;110 (12):2292–301.

10. Buckhurst PJ, Wolffsohn JS, Gupta N, Naroo SA, Davies LN, Shah S. Development of a questionnaire to assess the relative subjective benefits of presbyopia correction. J Cataract Refract Surg. 2012;38(1): 74–9.

11. Pesudovs K, Gothwal VK, Wright T, Lamoureux EL. Remediating serious flaws in the National Eye Institute Visual Function Questionnaire. J Cataract Refract Surg. 2010;36(5):718–32.

12. McAlinden C, Skiadaresi E, Moore J, Pesudovs K. Subscale assessment of the NEI-RQL-42 Questionnaire with Rasch analysis. Invest Ophthalmol Vis Sci. 2011;52(8):5685.

13. Turner-Bowker DM, Lamoureux RE, Stokes J, et al. Informing a priori sample size estimation in qualitative concept elicitation interview studies for clinical outcome assessment instrument development. Value Health. 2018;21(7):839–42.

14. Glaser BG, Strauss AL. The discovery of grounded theory: strategies for qualitative research. Mill Valley: Sociology Press; 1967.

15. Rosseel Y. lavaan: an R package for structural equation modeling. J Stat Softw. 2012;48(2):1–36.

16. Katsikatsou M, Moustaki I, Yang-Wallentin F, Jöreskog KG. Pairwise likelihood estimation for factor analysis models with ordinal data. Comput Stat Data Anal. 2012;56(12):4243–58.

17. Stevens JP. Applied multivariate statistics for the social sciences. New York: Routledge; 2012.

18. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Struct Equ Model. 1999;6(1):1–55.

19. Schumacker RE, Lomax RG. A beginner's guide to structural equation modeling. 3rd ed. New York: Taylor & Francis Group; 2004.

20. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16(3):297–334.

21. Nunnally J, Bernstein IH. Psychometric theory. New York: McGraw-Hill; 1994.

22. Fayers PM, Machin D. Quality of life: the assessment, analysis and interpretation of patient-reported outcomes. West Sussex: Wiley; 2007.

23. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86(2):420–8.

24. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–63.

25. Portney LG, Watkins MP. Foundations of clinical research applications to practice. Upper Saddle River: Prentice Hall Health; 2000.

26. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. New York: Routledge; 1988.

27. Csaky KG, Richman EA, Ferris FL. Report from the NEI/FDA ophthalmic clinical trial design and endpoints symposium. Invest Ophthalmol Vis Sci. 2008;49(2):479.

28. Guyatt G, Walter S, Norman G. Measuring change over time: Assessing the usefulness of evaluative instruments. J Chronic Dis. 1987;40(2):171–8.

29. Coon CD, Cook KF. Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. Qual Life Res. 2017;27(1):33–40.

30. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol. 2008;61(2):102–9.

31. Mokkink LB, Prinsen CAC, Patrick DL, et al. COSMIN Study Design checklist for patient-reported outcome measurement instruments. 2019. https://www.cosmin.nl/wp-content/uploads/COSMIN-study-designing-checklist_final.pdf.