CHAPTER 6

COMPARISON OF THE MEASUREMENT PROPERTIES OF THE EQ-5D-3L, ICECAP-O AND ASCOT IN FRAIL OLDER ADULTS

Karen M. van Leeuwen
Judith E. Bosmans
Aaltje P.D. Jansen
Emiel O. Hoogendijk
Maurits W. van Tulder
Henriette E. van der Horst
Raymond W. Ostelo

ABSTRACT

BACKGROUND The ICEPOP CAPability measure for Older people (ICECAP-O) and Adult Social Care Outcomes Toolkit (ASCOT) are preference-based measures for assessing quality of life (QoL) from a broader perspective than do traditional health-related QoL measures such as the EuroQol five-dimensional questionnaire (EQ-5D). Measurement properties of these instruments have not yet been directly compared.

OBJECTIVE The purpose of this study was to compare the test-retest reliability, construct validity and responsiveness of the three-level EQ-5D (EQ-5D-3L), ICECAP-O and ASCOT in frail older adults living at home.

METHODS Cross-sectional data and longitudinal data were used. Parameters for reliability (the intraclass correlation coefficient) and agreement (standard error of measurement) were used to assess test-retest reliability after 1 week. We formulated hypotheses about correlations with other measures and tested these to assess construct validity and responsiveness (longitudinal validity).

RESULTS The reliability parameters for all three scales were considered good (intraclass correlation coefficient values above 0.70). Standard error of measurement values were less than 10% of the scale. Hypotheses regarding construct validity were in general accepted; the EQ-5D-3L was more strongly associated with physical limitations than were the ICECAP-O and ASCOT and less strongly with instruments measuring aspects beyond health. Longitudinally, as hypothesized, mental health was most strongly associated with the ICECAP-O, and self-perceived QoL, mastery and client-centeredness of home care most strongly with the ASCOT.

CONCLUSIONS Our findings support the adoption of the ICECAP-O and ASCOT as outcome measures in economic evaluations of care interventions for older adults that have a broader aim than health-related QoL, because they are at least as reliable as the EQ-5D-3L and are associated with aspects of QoL broader than health.
INTRODUCTION

A common measure of outcomes of health interventions or care services is health-related quality of life. In economic evaluations, preference-based measures (PBMs) are used to value health-related quality of life lost or gained in study populations, with the EuroQol five-dimensional questionnaire (EQ-5D)\(^1,2\) as the most well-known and widely used instrument. It is increasingly recognized, however, that although health is an important determinant of quality of life, consequences and objectives of many care services include aspects of quality of life beyond health, particularly in areas such as public health, mental health and care for the elderly.\(^3\)–\(^12\) Recently, two new PBMs were developed for the purpose of evaluating care services for older adults/social care service users from a broader perspective, the ICEpop CAPability measure for Older people (ICECAP-O) and the Adult Social Care Outcomes Toolkit (ASCOT).\(^13,14\)

PBMs such as the EQ-5D are used to measure and indirectly value health-related quality of life to calculate quality adjusted life years (QALYs) for cost-utility analysis. An index score results from the categorization of a participant in health profiles valued by pre-existing weights based on preferences of the general population. The more preferable a health profile, described by several health-related quality of life domains, the larger the value assigned to a profile. The scale of the index score is traditionally anchored to two points: 1 (full health) and 0 (equal to being dead). Values below 0 are possible as well, for profiles that are considered to be less preferable than being dead.

The ICECAP-O is a PBM with a broader perspective than health-related quality of life and is conceptually based on the capability approach.\(^14\)–\(^17\) The capability approach defines well-being in terms of an individual’s ability to ‘do’ and ‘be’ the things that are important in life. The ICECAP-O was developed in the United Kingdom for use in the economic evaluation of health and social care interventions, covering the domains attachment (love and friendship), security (thinking about the future without concern), role (doing things that make you valued), enjoyment (enjoyment and pleasure), and control (independence).

The ASCOT was developed in the United Kingdom as well and includes a PBM for use in economic evaluations that measures social care-related quality of life.\(^13,18\) The ASCOT covers eight domains, varying from basic aspects such as personal cleanliness and feeling safe to higher order aspects such as social participation, occupation, dignity and control over daily life. The ASCOT distinguishes between capabilities and functionings in response levels.\(^13,19\)

Not much is known yet about the measurement properties of the ICECAP-O and ASCOT. One study indicated that the test-retest reliability of the ASCOT after 74 days was acceptable.\(^13\) Other studies have shown that both the ICECAP-O and ASCOT are moderately to strongly associated
with general quality of life scales, but also with the EQ-5D, ADL limitations, illness and depressive symptoms. There is some evidence that compared to the EQ-5D, the ICECAP-O and ASCOT are equally or stronger associated with the quality of care services and the intensity of care, as measured by hours delivered. Together with studies that reported improvements in ICECAP-O scores after hip or knee (replacement) surgery, this indicates that the ICECAP-O and ASCOT may be responsive to the impact of care services, one of the most important measurement properties of scales with an evaluative purpose. So far, there is no head-to-head comparison of measurement properties of the ICECAP-O, the ASCOT and the EQ-5D. We conducted a study in frail older adults with the aim of comparing the test-retest reliability, construct validity, and responsiveness of the EQ-5D-3L, ICECAP-O and ASCOT.

**METHODS**

**Design and Participants**

The study population consisted of a sample of participants of the ‘frail older Adults: Care in Transition’ (ACT) study in The Netherlands. This stepped wedge cluster randomized controlled trial was designed to evaluate a geriatric care model based on the Chronic Care Model for frail older adults living at home. The ACT study is described in more detail elsewhere. Eligible participants were identified in 35 general practices in two regions of The Netherlands, using the ‘Program on research for integrating services for the maintenance of autonomy’ tool (PRISMA-7), a brief seven-item questionnaire containing risk factors for functional decline. Signed informed consent was obtained from all study participants. The study received approval from the Medical Ethics Committee of the VU University Medical Center (10/003).

The subsample used for the current study consisted of 190 participants from one of the regions (Amsterdam) who completed the 12-month follow-up measurements (T12) after November 1, 2011 (Fig. 1). They were invited for a retest measurement 1 week after T12, and their 18 month follow-up measurements (T18) were used. All measurements were collected at participants’ homes by means of a computer-assisted personal interview between Nov 2011 and Oct 2012.

**Measures**

In this study, three preference-based quality of life measures were compared: the three level EQ-5D (EQ-5D-3L), the ICECAP-O and the ASCOT. Domain weights of the three measures are depicted in Figure 2. The measures were administered in the same order as they are described below.
Measurement properties of the EQ-5D, ICECAP-O & ASCOT

**EQ-5D-3L**
The EQ-5D-3L\textsuperscript{1,2} is a brief five-item instrument (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) with three response options (no problems, some problems, and extreme problems), resulting in 243 health states. Each health state can be converted to a preference-weighted index score, with values ranging between -0.594 (extreme problems on every domain) and 1 (full health), and a value of 0 being equivalent to the state of being dead. We used a Dutch version of the EQ-5D-3L and applied the York A1 population-based tariff.\textsuperscript{37}

**ICECAP-O**
The ICECAP-O\textsuperscript{14–16} is a five-item instrument (attachment, security, role, enjoyment, and control) with four level response options that are described as statements representing four levels of

---

**Figure 1.** Flow chart of the study
Figure 2. Population based preference-weights.

ASCOT dimensions: C: Control over daily life; PC: Personal cleanliness and comfort; Fd: Food and drink; Ps: Personal safety; Sp: Social participation and involvement; Oc: Occupation; Ac: Accommodation cleanliness and comfort; D: Dignity. The EQ-5D-3L index score is calculated by subtracting the weights from 1 and subtracting an additional 0.081 for any dysfunctional state and 0.269 for extreme problems in at least one domain. The ICECAP-O index score is calculated by the sum of the weights. ASCOT index score is calculated by multiplying the sum of the weights with 0.203 and subtracting 0.466 from the result.
capability: none, a little, a lot and all. This descriptive system results in 1024 different states. A total index score is obtained by summing preference-based weights corresponding to levels of each domain. The scale of this score ranges from 0 to 1, with 0 representing no capability and 1 representing full capability. Preference-based values were elicited in a sample of older adults from the UK general population.\textsuperscript{14} We used a Dutch version of the ICECAP-O, translated by Makai et al. following a forward and backward translation procedure.\textsuperscript{38} The index score was computed using the British tariff of the ICECAP-O with population-based values.\textsuperscript{14}

**ASCOT**

The four level self-completion version of the ASCOT includes eight domains (personal cleanliness, safety, meals and nutrition, activities/occupation, control over daily life, social participation, home cleanliness and comfort, and dignity) with four levels (ideal state, no needs, some needs, and high needs). The ideal state level was added to reflect the distinction between capabilities and functioning.\textsuperscript{13,19} The descriptive system of the ASCOT generates 65,536 different states, and the total preference-weighted index scale ranges from -0.171 (high needs on all domains) to 1, with ‘0’ equivalent to ‘being dead’ and ‘1’ being the ‘ideal’ social care-related quality of life state, in which all needs are met to the desired level. The preference-based values we applied to calculate the index score were elicited in a sample of the UK general population.\textsuperscript{13} We used a Dutch translation of the ASCOT that was recently produced by our team, following forward and backward translation procedures as described by Beaton et al.\textsuperscript{39}

**Test-retest reliability**

To assess the test-retest reliability of the PBM index scores, T12 respondents were asked to participate in an extra measurement within 7 to 14 days of T12. We assumed that during this time period the quality of life of the participants remained stable while at the same time recall bias was minimized.

We used two parameters for the test-retest reliability: the intraclass correlation coefficient (ICC\textsubscript{AGREEMENT})\textsuperscript{40,41} and the standard error of measurement (SEM).\textsuperscript{42} The ICC is relevant for discriminative purposes and the SEM for measurement of changes.\textsuperscript{43} Both parameters were calculated on the basis of the variance components obtained with the repeated-measures analysis of variance technique: the between-subject variance (‘true’ differences between older adults), the between-measures variance (systematic error), and the random error variance. The analysis of variance and all further analyses were performed on all available complete data using IBM SPSS Statistics 20.

The ICC\textsubscript{AGREEMENT} was calculated as the ratio of the between-subject variance and the total variance.\textsuperscript{41,42} For the SEM\textsubscript{AGREEMENT} we used the squared root of the within-subject variance (between
measures + random error variance).\textsuperscript{42}

We converted the $\text{SEM}_{\text{AGREEMENT}}$ to the smallest detectable change ($\text{SDC}_{\text{AGREEMENT}} = 1.96 \times \sqrt{2} \times \text{SEM}_{\text{AGREEMENT}}$).\textsuperscript{43} The SDC represents the smallest within-person change that is with $p<0.05$ ‘detectable’ by an instrument beyond measurement error in one individual.\textsuperscript{31} A within-person change smaller than the SDC may be due to measurement error.

**Construct validity**

To assess the construct validity of the PBMs, the PBM index scores were compared to other scales available from the ACT study T12 measurements. We a priori defined hypotheses about the difference in association of these measures with the EQ-5D-3L compared with the associations of these measures with the ICECAP-O or the ASCOT. The hypotheses were defined on the basis of the different scope of the PBMs and extensive discussion within our project team. All measures and general hypotheses are described below; specific hypotheses are presented in Table 4. We rejected the hypotheses when the difference in strength of Spearman correlation coefficients was less than 0.1, which has previously been used by De Boer et al.\textsuperscript{44} as smallest amount of expected differences between correlation coefficients.

First of all, we expected that the ICECAP-O and ASCOT were more strongly correlated to each other than to the EQ-5D-3L (hypotheses 1a and 1b), because they both intend to measure broader aspects of quality of life whereas the EQ-5D-3L is restricted to health-related quality of life.

**Self-perceived health (Health GRS)** We also expected that the ICECAP-O and ASCOT were less strongly correlated than the EQ-5D-3L to a single-item question about perceived health (hypotheses 2a and 2b). The Health Global Rating Scale (Health GRS) asked respondents to rate their current health status on a five-point scale from ‘poor’ to ‘excellent’.

**ADL limitations (Katz ADL index)** ADL limitations were assessed using the Katz Index of Activities of Daily Living including six questions about the need of assistance with basic functions.\textsuperscript{45} Responses were binary (yes=1 and no=0) and summed up to calculate the number of limitations. These limitations are health related and were expected to correlate less strongly with the ICECAP-O and ASCOT than with the EQ-5D-3L (hypotheses 3a and 3b).

**Impact of physical limitations (SF-12 PCS)** The Physical health Component Score of the 12-item Short-Form Health Survey (SF-12) reflects the impact of physical limitations on functioning.\textsuperscript{46} Summary scores were transformed using a standard scoring algorithm to have a mean of 50 and an SD of 10 in the US general population, with higher scores reflecting better functioning.\textsuperscript{47} Again, we expected that this health related scale was less strongly correlated with the ICECAP-O and ASCOT than with the EQ-5D-3L (hypotheses 4a and 4b).

**Impact of emotional influences (SF-12 MCS)** The Mental health Component Score of the
SF-12 (SF-12 MCS) reflects the extent of emotional influences on functioning.\textsuperscript{46,47} Because the ICECAP-O and ASCOT measure quality of life from a broad perspective including some emotional aspects, we expected that they were more strongly correlated than the EQ-5D-3L to the SF-12 MCS (hypotheses 5a and 5b).

**Self-perceived quality of life (QoL GRS)** Respondents were asked to rate their current quality of life on the single-item Quality of Life Global Rating Scale (QoL GRS), with five response options from ‘poor’ to ‘excellent’. Because the ICECAP-O and ASCOT measure quality of life from a broad perspective, we expected that they were more strongly correlated than the EQ-5D-3L with global quality of life as perceived by the respondents (hypotheses 6a and 6b).

**Mastery (Pearlin Mastery Scale)** Sense of mastery was measured by an abbreviated version (five items) of the Pearlin Mastery Scale.\textsuperscript{48,49} The sum score of this scale reflects the extent to which a person perceives himself or herself to be in control of events and ongoing situations, and ranges from 5 to 25, with a higher rating indicating higher sense of mastery. We expected that the ICECAP-O and ASCOT were more strongly correlated than the EQ-5D-3L to mastery because they both include an item about control over daily life (hypotheses 7a and 7b).

**Client-centeredness (CCCQ)** The Client-Centred Care Questionnaire (CCCQ) measures the client-centeredness of professional home nursing care from a client perspective.\textsuperscript{50} The CCCQ comprises 15 items about the extent to which respondents feel recognized and respected by nurses and to which they experience autonomy with respect to the way in which care is delivered. The ICECAP-O and ASCOT were developed to evaluate outcomes of social care, which includes home care. Therefore, we expected that the CCCQ was more strongly correlated to the ICECAP-O and ASCOT than to the EQ-5D-3L (hypotheses 8a and 8b).

**Responsiveness**

We used a similar construct validity approach to assess the responsiveness of the PBMs. We calculated correlation coefficients between change scores and defined hypotheses about the change scores equivalently to those described above. Change scores were calculated as the difference in scores between T18 and T12. The specific hypotheses about the responsiveness of the scores can be found in Table 4.
RESULTS

Sample
The characteristics of the 190 frail older adults at T12 can be found in Table 1. The flowchart of the study (Fig. 1) shows the study design, number of participants, and reasons for drop-out. The mean age of participants was 82.4 years (SD 7.7) and 71.6% were women. The largest amount of missing data was for the ASCOT index score (14.7%), mostly due to missing responses on the dignity domain (n=24). The CCCQ was only administered to the 97 participants who received home care at that time.

Test-retest reliability
For the retest of the PBMs, 147 participants agreed to participate in the additional measurements, which were conducted on average 9 days after T12 measurements. Reasons for not participating are presented in the flowchart (Fig. 1).

Table 2 presents the variance components of the PBMs and the test-retest reliability parameters. There were no systematic differences between the T12 and T12 retest measures. Compared to the ICECAP-O and ASCOT, there was more variance in EQ-5D-3L scores between the participants, but the error variance was larger as well. The estimated ICC$^\text{AGREEMENT}$ values for all three PBMs were higher than 0.70, which is the recommended minimum reliability,\textsuperscript{51} although the lower bound of the 95% confidence interval for the ICC of the ASCOT was below 0.7 (i.e., 0.60).

The SEM and SDC were slightly higher for the EQ-5D-3L than for the ICECAP-O and ASCOT, due to the larger error variance in EQ-5D-3L index scores.

Construct validity
Correlation coefficients between the PBM index scores and other measures at T12 are given in Table 3. The strength of the correlations was in general moderate according to Cohen’s convention for interpreting effect sizes,\textsuperscript{52} with only some correlations over 0.50. The strongest correlation of the EQ-5D-3L was with the SF-12 PCS (0.60). The strongest correlation of the ICECAP-O and ASCOT was with the SF-12 MCS (0.53 and 0.50, respectively). Although there is some overlap between the three measures, the strength of associations with other constructs differs. The evaluation of the hypotheses about differences in correlations between the EQ-5D-3L and the other measures compared with the ICECAP-O and ASCOT is presented in Table 4.

The ICECAP-O scores were less strongly correlated than the EQ-5D-3L to health related measures, and more strongly correlated than the EQ-5D-3L to mastery, client-centeredness and the impact of emotional influences. Only the hypotheses concerning the global rating scales were
Table 1. Sample characteristics at T12

<table>
<thead>
<tr>
<th></th>
<th>N complete</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age. mean (SD)</td>
<td>82.4 (7.7)</td>
</tr>
<tr>
<td>Sex. % women</td>
<td>71.6%</td>
</tr>
<tr>
<td>Partner status. % no partner</td>
<td>77.9%</td>
</tr>
<tr>
<td>Educational level</td>
<td>190</td>
</tr>
<tr>
<td>- Low. %</td>
<td>16.8%</td>
</tr>
<tr>
<td>- Middle. %</td>
<td>49.5%</td>
</tr>
<tr>
<td>- High. %</td>
<td>33.7%</td>
</tr>
<tr>
<td>EQ-5D-3L (-0.594-1). Mean (SD)</td>
<td>0.59 (0.31)</td>
</tr>
<tr>
<td>ICECAP-O (0-1). Mean (SD)</td>
<td>0.78 (0.16)</td>
</tr>
<tr>
<td>ASCOT (-0.171-1). Mean (SD)</td>
<td>0.80 (0.16)</td>
</tr>
<tr>
<td>Health GRS</td>
<td>190</td>
</tr>
<tr>
<td>- Excellent. %</td>
<td>1.6%</td>
</tr>
<tr>
<td>- Very good. %</td>
<td>3.2%</td>
</tr>
<tr>
<td>- Good. %</td>
<td>45.3%</td>
</tr>
<tr>
<td>- Fair. %</td>
<td>39.5%</td>
</tr>
<tr>
<td>- Poor. %</td>
<td>10.5%</td>
</tr>
<tr>
<td>ADL limitations (0–6). mean (SD)</td>
<td>0.89 (1.15)</td>
</tr>
<tr>
<td>SF-12 PCS (0-100). mean (SD)</td>
<td>35.0 (9.8)</td>
</tr>
<tr>
<td>SF-12 MCS (0-100). mean (SD)</td>
<td>53.3 (10.2)</td>
</tr>
<tr>
<td>QoL GRS</td>
<td>188</td>
</tr>
<tr>
<td>- Excellent. %</td>
<td>3.2%</td>
</tr>
<tr>
<td>- Very good. %</td>
<td>6.3%</td>
</tr>
<tr>
<td>- Good. %</td>
<td>44.7%</td>
</tr>
<tr>
<td>- Fair. %</td>
<td>39.5%</td>
</tr>
<tr>
<td>- Poor. %</td>
<td>5.3%</td>
</tr>
<tr>
<td>Mastery (5-25). Mean (SD)</td>
<td>15.9 (3.8)</td>
</tr>
<tr>
<td>CCCQ. (15-75). Mean (SD)</td>
<td>52.5 (6.9)</td>
</tr>
</tbody>
</table>

Table 2. The variance components and indices

<table>
<thead>
<tr>
<th>Questionnaire</th>
<th>Between subject variance</th>
<th>Within subject variance</th>
<th>Residual</th>
<th>ICC (95% CI)a</th>
<th>SEM b</th>
<th>SEM (%) c</th>
<th>SDC d</th>
<th>SDC (%) e</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQ-5D-3L</td>
<td>0.064</td>
<td>0.000</td>
<td>0.017</td>
<td>0.79 (0.72-0.85)</td>
<td>0.130</td>
<td>8.2%</td>
<td>0.361</td>
<td>22.6%</td>
</tr>
<tr>
<td>ICECAP-O</td>
<td>0.016</td>
<td>0.000</td>
<td>0.004</td>
<td>0.79 (0.71-0.84)</td>
<td>0.066</td>
<td>6.6%</td>
<td>0.184</td>
<td>18.4%</td>
</tr>
<tr>
<td>ASCOT</td>
<td>0.014</td>
<td>0.000</td>
<td>0.006</td>
<td>0.71 (0.60-0.78)</td>
<td>0.075</td>
<td>6.4%</td>
<td>0.209</td>
<td>17.9%</td>
</tr>
</tbody>
</table>

aICC, intraclass correlation coefficient for agreement (between subject variance / total variance)
bSEM, standard error of measurement (V within subject variance)
cSEM expressed in percentages of corresponding scale
dSDC, smallest detectable change in scale points (1.96 * v2 * SEM)
eSDC expressed in percentages of corresponding scale
rejected: different from what we expected, the ICECAP-O was as strongly correlated to the health GRS as the EQ-5D-3L and only 0.08 points stronger to the QoL GRS.

The same findings concerning our hypotheses hold for the ASCOT compared with the EQ-5D-3L, although the strength of the correlations was somewhat different. Furthermore, our hypothesis concerning ADL limitations was rejected: the ASCOT was as strongly correlated as the EQ-5D-3L to ADL limitations.

**Table 3. Correlation coefficients**

<table>
<thead>
<tr>
<th></th>
<th>Correlation coefficients (T12 scores)</th>
<th>Change scores T18-T12</th>
<th>EQ-5D-3L</th>
<th>ICECAP-O</th>
<th>ASCOT</th>
<th>EQ-5D-3L</th>
<th>ICECAP-O</th>
<th>ASCOT</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQ-5D-3L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICECAP-O</td>
<td>0.41</td>
<td>-</td>
<td>-0.01</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASCOT</td>
<td>0.41</td>
<td>0.63</td>
<td>0.09</td>
<td>0.31</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health GRS</td>
<td>0.47</td>
<td>0.48</td>
<td>0.14</td>
<td>0.06</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADL limitations</td>
<td>-0.43</td>
<td>-0.29</td>
<td>-0.04</td>
<td>-0.26</td>
<td>-0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>impact of physical limitations (SF-12 PCS)</td>
<td>0.60</td>
<td>0.30</td>
<td>0.26</td>
<td>0.23</td>
<td>0.04</td>
<td>0.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>impact of emotional influences (SF-12 MCS)</td>
<td>0.36</td>
<td>0.53</td>
<td>0.50</td>
<td>0.02</td>
<td>0.22</td>
<td>0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QoL GRS</td>
<td>0.39</td>
<td>0.47</td>
<td>0.12</td>
<td>0.15</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mastery</td>
<td>0.33</td>
<td>0.50</td>
<td>0.48</td>
<td>-0.02</td>
<td>0.17</td>
<td>0.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Client-centeredness (CCCQ)</td>
<td>0.02</td>
<td>0.17</td>
<td>0.22</td>
<td>0.09</td>
<td>0.17</td>
<td>0.34</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Responsiveness**

At T18, on average 175 days after T12, 41 of the 190 participants had dropped out (Figure 1). Changes scores for 149 participants were calculated if complete data were available. The strength of the correlations between change scores of the EQ-5D-3L, ICECAP-O and ASCOT and the other measures was weak (Table 3); the strongest correlation was found between change in client-centeredness and change in ASCOT index scores (0.34). The evaluation of hypotheses concerning responsiveness of the PBMs is presented in Table 4.

As in the evaluation of the construct validity, hypotheses concerning the correlation of changes in ICECAP-O with changes in global rating scales were rejected. Also hypotheses about changes in ADL limitations and client-centeredness were rejected.

In contrast, hypotheses concerning the correlation of changes in ASCOT with changes in global rating scales were not rejected. Only the hypothesis about the correlation of changes in ASCOT with changes in ADL limitations was rejected.
Table 4. Hypotheses for the construct validity and responsiveness of the EQ-5D-3L, ICECAP-O and ASCOT

<table>
<thead>
<tr>
<th>Hypotheses:</th>
<th>ICECAP-O scores are 0.1 ...... than EQ-5D-3L scores</th>
<th>ASCOT scores are 0.1 ...... than EQ-5D-3L scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hypotheses accepted:</td>
<td>Construct validity</td>
</tr>
<tr>
<td></td>
<td>(T12 scores)</td>
<td></td>
</tr>
<tr>
<td>1a</td>
<td>more strongly correlated to ASCOT</td>
<td>Yes</td>
</tr>
<tr>
<td>2a</td>
<td>less strongly correlated to health GRS</td>
<td>No</td>
</tr>
<tr>
<td>3a</td>
<td>less strongly correlated to ADL limitations</td>
<td>Yes</td>
</tr>
<tr>
<td>4a</td>
<td>less strongly correlated to impact of physical limitations (SF-12 PCS)</td>
<td>Yes</td>
</tr>
<tr>
<td>5a</td>
<td>more strongly correlated to impact of emotional influences (SF-12 MCS)</td>
<td>Yes</td>
</tr>
<tr>
<td>6a</td>
<td>more strongly correlated to QoL GRS</td>
<td>No</td>
</tr>
<tr>
<td>7a</td>
<td>more strongly correlated to mastery</td>
<td>Yes</td>
</tr>
<tr>
<td>8a</td>
<td>more strongly correlated to client-centeredness (CCCQ)</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Discussion

In this paper, we compared the test-retest reliability, construct validity and responsiveness of the EQ-5D-3L, ICECAP-O and ASCOT in a sample of frail older adults. In summary, differences in reliability parameters were small and the assessment of construct validity and responsiveness indicates that the EQ-5D-3L in general focuses more on health while the ICECAP-O and ASCOT are associated with broader aspects of quality of life.

We consider the test-retest reliability parameters for all three scales as good, with ICC values above 0.70 and SEM values of less than 10% of the scale. These values mean that the discrimination between frail older adults is not much affected by measurement error and that relatively small changes can be considered true changes rather than changes due to measurement error.

Because the SEM of the EQ-5D-3L is slightly larger than those of the ICECAP-O and ASCOT, a somewhat larger sample size is required to detect changes at group level beyond measurement error with the EQ-5D-3L. We reported the SDC in individuals in this article. For economic evalua-
tions, however, changes at a group level are relevant. The SDC at group level should be reduced by a factor of \( \sqrt{n} \). For example, SDCs beyond measurement error in a group of 100 frail older adults would be 0.036, 0.018 and 0.021 on the EQ-5D-3L, ICECAP-O and ASCOT, respectively.

That these small changes are detectable beyond measurement error does not automatically imply that these changes are considered important; important changes could be either smaller or larger than the SDC. A review of studies in several patients groups found a mean estimate of 0.074 of the minimal important change for the EQ-5D-3L, although values below 0.036 were reported as well. Minimal important change estimates for the ICECAP-O and ASCOT have not been reported yet.

The construct validity and responsiveness of PBMs can be considered good as well because most of our hypotheses were accepted. Although there is some overlap between the three measures, the strength of associations with other constructs differs. Compared with the EQ-5D-3L, the ICECAP-O and ASCOT were more strongly correlated to each other and to instruments that measure broader aspects than health. The EQ-5D-3L was more strongly correlated to the impact of physical limitations (SF-12 PCS) than did the other two PBMs. Longitudinally, change in ASCOT was of the three measures most strongly correlated to change scores of instruments that measure broader aspects than health.

The differences in correlations of the self-perceived quality of life and self-perceived health scales with the EQ-5D-3L on the one hand and the ICECAP-O and ASCOT on the other hand were not as pronounced as we expected. An explanation could be that older adults take their health into account when giving a global rating of their quality of life. The similar correlation size of the ICECAP-O, ASCOT and EQ-5D-3L with the Health GRS and with (changes in) ADL limitations, suggest that the domains included in the ICECAP-O and ASCOT do not directly measure health but capture some of the effects of health that are important to older adults.

We found that change in self-perceived quality of life, change in mastery and change in client-centeredness were most strongly correlated to the ASCOT. Therefore, the ASCOT may be most responsive to changes in quality of life from a broad perspective. The ICECAP-O may be most responsive to changes in the impact of mental health and in ADL limitations, and the EQ-5D-3L to changes in the impact of physical limitations. The strength of correlations between change scores, however, was weak. This may be due to the limited amount of change that our sample experienced within 6 months. Further research is necessary before strong conclusions can be drawn about the responsiveness of the measures.

This was the first study with a head-to-head comparison of the measurement properties of the ICECAP-O and ASCOT to the measurement properties of the EQ-5D-3L in frail older adults living at home. Other studies also found differences between the EQ-5D-3L and ICECAP-O.¹²–¹⁴
or between the EQ-5D-3L and ASCOT\(^6\) in associations with other measures. With this study the properties of the ICECAP-O and the ASCOT can also be compared. Although reliability parameters and cross-sectional correlation coefficients were similar for the ICECAP-O and ASCOT, the correlation coefficients between change scores suggest differences in responsiveness to change.

While interpreting the results of the current study some limitations have to be considered. First, for the validity assessment we were restricted to measurement instruments that were available in the ACT study. We chose the measures that were theoretically most closely linked to the PBMs. These measures, however, have a more specific scope than the generic PBMs and are not preference-weighted. These characteristics of the comparator instruments may explain the moderate absolute size of correlations with the PBMs. Weighted scores based on preferences of the generation population do not necessarily correlate with unweighted scale scores.\(^5\)\(^6\) For example, the maximal contribution of the pain/discomfort domain to the index score of the EQ-5D-3L is much higher than that of the usual activities domain (see Fig. 2), which implies that the EQ-5D is probably more strongly correlated to quality of life measures that focus on pain than those that focus on ability to perform usual activities. The index score of PBMs reflects both level of functioning/capability and the importance of domains for quality of life according to the general population. Brazier\(^5\)\(^6\) argues that such index scores should preferably be validated using revealed or hypothetical preferences as ‘criterion standard’. This is, however, challenging due to the specific features of health care markets. We adopted a pragmatic approach using non-preference-based comparator instruments to assess the validity of the ICECAP-O and ASCOT relative to the EQ-5D-3L and confirmed expected differences between the PBMs.

Second, the order in which the instruments were administered may have influenced the results. Because the ASCOT was administered last, respondents may have been more tired responding to this measure. We do not think that this order effect has influenced results of the test-retest reliability much because the instruments were administered in the same order at both moments and the retest was a short administration. Construct validity might have been influenced by this order effect, as the EQ-5D-3L was preceded by questions about self-perceived health and the ICECAP-O and ASCOT by questions about self-perceived quality of life.

Third, although we compared the measurement properties of three PBMs, the index score scales itself are not directly comparable. The measures differ in intended scope and in valuation methods used to obtain preference weights. As a result, the scales are anchored to different points. For example, a score of 1 on the EQ-5D-3L (full health) is not the same as a score of 1 on the ICECAP-O (full capability). Therefore, it is not recommended to use the PBMs interchangeably in economic evaluations; the index scores represent different concepts.

We recommend that researchers choose one of the measures based on the specific purposes
of a study and the targeted population because our findings show that the test-retest reliability does not differ profoundly between the PBMs and that the three measures are correlated to different constructs. Choosing a measurement instrument that is most relevant to the target population and the objectives of an intervention reduces the probability of making type II errors. When the main objective of an intervention or care service is to improve or maintain health, the EQ-5D remains the most appropriate choice. When the aims are broader, the ICECAP-O and the ASCOT seem to be more appropriate. The results of cost-effectiveness analysis are indeed influenced by the reliability and content of the outcome measures, as was shown in a recent economic evaluation in which personal health budgets were considered cost-effective using the ASCOT and not cost-effective using the EQ-5D.57 Two other studies showed that conclusions of the economic evaluation of integrated care in frail elderly58 and telehealth for patients with long term conditions59 did not differ using the ICECAP-O or EQ-5D, although at the same level of willingness to pay (WTP), interventions had a higher probability of cost-effectiveness using the ICECAP-O.

More research is necessary to investigate whether the ICECAP-O and ASCOT are responsive to the impact of various care services, and to important changes in the quality of life of older adults. Studies that estimate the minimal important change of the ICECAP-O and ASCOT would be helpful. Qualitative research may be helpful to include the perspective of older adults in assessing the content validity of the measures. A few studies have already used a qualitative or mixed approach for the ASCOT and the adult population version of the ICECAP.9,13,60,61 Furthermore, in order to help researchers select the most appropriate instrument, we would suggest further work that concentrates on differences between the ICECAP-O and ASCOT. Finally, although health is not the only relevant quality of life domain for older adults, it is an important one.9,62 More research is recommended to gain a better understanding of the extent to which the ICECAP-O and ASCOT are sensitive to changes in (self-perceived) health. Are the ICECAP-O and ASCOT capable of indirectly picking up effects of health on quality of life? Or is it necessary to use these PBMs complementary to PBMs based on health-related quality of life?

Taken together, the findings of this study support the adoption of the ICECAP-O and ASCOT as outcome measures in economic evaluations of care interventions for older adults that have a broader aim than health-related quality of life because they are at least as reliable as the EQ-5D-3L and are associated with aspects of quality of life broader than health.
REFERENCES


29. Flynn TN, Joanna Coast BA. Change in capability-related quality of life resulting from hip or knee replacement: results from a cohort study using the ICECAP-O instrument. Unpubl Manuscr; : 1–16.


