Different frameworks, similar results?
A head-to-head comparison of EQ-5D-5L and CS-Base

Xin Zhang, MSc; Karin M Vermeulen, PhD; Paul FM Krabbe, PhD

Department of Epidemiology, University Medical Center Groningen, University of Groningen, Hanzeplein 1, 9713 GZ, Groningen, the Netherlands

Corresponding author: Paul Krabbe, Department of Epidemiology, University Medical Centre Groningen, Hanzeplein 1, 9713 GZ, Groningen, the Netherlands; (p.f.m.krabbe@umcg.nl).

Abstract

Objectives

We have developed a novel measurement (multi-attribute preference response, MAPR) model and a generic patient-reported outcome measure (PROM) called CS-Base based on the MAPR model. The CS-Base is preference-based and patient-centered. It comprises 12 health items (mobility, vision, hearing, cognition, mood, anxiety, pain, fatigue, social functioning, daily activities, self-esteem, independence), each consisting of 4 levels. The CS-Base is implemented in a general software application (HealthSnApp). In this study, we aimed to assess the performance of the CS-Base based on MAPR model parallel to the established EQ-5D-5L by comparing their coefficients of each level of items and the values of health states.

Methods

We conducted a cross-sectional study based on a random sample of USA patients with various kinds of diseases or complaints. The CS-Base and EQ-5D-5L (the original and an adapted 4-level version (5D)) were used to measure health outcomes. We studied the range of health items caputerd, discrimination of health states and user experiences.

Results

For both CS-Base and 5D, all the coefficients revealed a logical order and statistically significant differences. The impacts of the four items included in all three PROMs were comparable. The item 'usual activities' had the lowest impact, the other three items ('mobility', 'pain', 'anxiety') had high or moderate impacts, with slight differences between the three PROMs. High impacts were also observed for 'vision' and 'hearing', which are not included in the EQ-5D-5L and the 5D. The values of the CS-Base were more densely and evenly distributed, followed by the EQ-5D-5L. The values of the 5D were more sparsely distributed and showed some gaps. A celling effect was also observed in the 5D and EQ-5D-5L as far more respondents reported mild health states than severe health states in these two PROMs. This celling effect was minor in the CS-Base.

Conclusions

This study revealed that the CS-Base performed best in discriminating different health states, followed by the EQ-5D-5L, the 5D was the least sensitive one. In the evaluation of two MAPR instruments, the 5D was more preferred by respondents. The CS-Base, the 5D and the EQ-5D-5L each has their own advantages and biases, researchers can make a well-informed choice about which one to use according to the purpose of their study.

Keywords

Patient-reported outcome measures (PROMs), health outcome, measurement model, preference-based methods, values

Word count: 5042; Number of pages: 23

Introduction

With the advance of modern medicine, health care evolved from physician-centered to patients-centered.[1,2] Increased interest in patients' involvement in healthcare has prompted the development of patient-reported outcome measures (PROMs). A patient-reported outcome (PRO) or patient-reported outcome measure (PROM) is any assessment coming directly from patients, without interpretation by physicians or others, about how they function or feel in relation to their health condition[3]. The term PROM encompasses a broad spectrum of outcomes that include the symptoms of a disease or the side effects of a treatment (e.g., fatigue, pain, or low mood), functions (e.g., social activities, cognitive functioning, or physical abilities), and even multidimensional constructs, including health-related quality of life (HRQoL) or perceived health status.[4,5] Evidence shows that the use of information from PROMs contributes to better communication, decision making between doctors and patients and improves patient satisfaction with health outcome and care.[6,7,8,9,10]

It is crucial to include items that are relevant and important to target populations' subjective health evaluation.[11] In the development of PROMs, it's increasingly recognized that the items selection should be based on patient's input. However, many of the existing PROMs are not patient-centred in their development, but the health professionals' views are prioritized.[12,13] This could result in either omitting health items that have a high relevance to patients or accentuate irrelevant ones. Even for the widely used EQ-5D, it's content (5 items) was not selected by patients but by health researchers[14,15,16,17,18] The questions arise if its content really reflects what's important to patients and if the five items are enough to assess the overall health of patients well? Despite that the EQ-5D-5L owns the great advantage of short and simple using. A new generic health outcome measurement CS-Base has been developed[19]. The CS-Base is an electronic patient-reported outcome measure (ePROM) that runs in the mobile app HealthSnApp (www.chateau-sante.com/healthsnapp). It comprises 12 health items, each specified on four levels. All the 12 items in the CS-Base were selected by patients.

Besides health items selection, another important part of preference-based PROMs is the health valuation method. This is used to generate weights for levels of items and can further provides a quantitative measure (value) of the overall health. The value allows comparison between many different diseases groups and can be used for many areas such as calculating quality-adjusted life years, assessing cost-effectiveness of interventions, monitoring health conditions of the population, supporting clinical decision making.[5,20] Preference-based

methods are frequently used as valuation methods. Conventional preference-based methods applied in the health setting were developed by health economists and mainly based on (pairs of) hypothetic health states assessed by a sample of the general population instead of patients.[21] However, it is reasonable to assume that in many situations, a sample of unaffected respondents from the general population may be inadequately informed or lack good imagination to make an appropriate assessment about the impact of (severe) health states.[22] Besides, such conventional preference-based methods are to some extent complicate to understand by respondents, well-trained interviewers are needed to help respondents to complete the tasks. All these limitations can make the preference-based tasks complex and cognitively demanding, as a result, these tasks are likely to produce results that are less precise or that may even be biased. It is important to make the preference tasks as simple as possible.[23]

A novel preference-based measurement framework has been recently introduced. This framework is known as the multi-attribute preference response (MAPR) model.[24,25,26] In its general form, it is a probabilistic choice model that combines the Rasch model (item response theory) and the discrete choice model (i.e., discrete choice experiments). These type of choice models have a long history, commencing with Louis Thurstone's model, which was developed in 1927.[27] Other researchers have introduced extensions of the basic Thurstonian model.[28,29,30,31] There are two assessment tasks within the MAPR model, the first is a descriptive task, patients (hence, not respondents from the general population) describe the health states of themselves in this task based on a set of health items. This health items are all selected by patients themselves. The second is a preference-based task, which generates ranked preference data that is used to estimate the overall weights of the levels of the items. Currently the 'Drop-Down' (DD) method is used as the preference-based task in MAPR model. In the DD method, respondents do not need to be confronted with hypothetic health states or make trade-offs between their own health and alternative, hypothetical health states. They only focus on their own health state and select health items that hinders them most. We have used the DD method in clinical studies, and it proved to produce good results.[32] An additional benefit of the MAPR measurement framework is that the assessment tasks (Task 1: descriptive task, Task 2: preference-based task) can be performed on smartphone screens which makes the PROM user-friendly, and attractive to the users (patients) and to researchers. In addition, all responses are automatically stored and processed.

The purpose of this study is to compare the CS-Base based on the MAPR model with the EQ-5D, to see which one captures a more complete range of health items, which one gives a

more sufficient description of health conditions, and which one gives a better discrimination of different health states. Finally, we also explored the experiences of users with the different instruments.

Methods

Sample

We conducted a cross-sectional study based on a random sample of USA patients (≥ 18 years). The sample was national representative on age, gender, and education. Respondents were patients with various disease(s) or complaints (pains, mental health problems, fatigue/sleep problems, hearing or vision loss, diabetes, respiratory diseases, heart disease, eczema, gastrointestinal disease, rheumatism, cancer, stroke, epilepsy, other diseases). They were registered with Dynata, a market research company based in Rotterdam, the Netherlands. Respondents who completed the survey received a small financial compensation from Dynata. The amounts were decided based on the company's agreements with the groups of respondents. Data were collected from January to February 2022. Respondents' demographic data were provided by Dynata.

Health-outcome measures

We aimed to compare two PROMs (the CS-Base and EQ-5D-5L) in our study. For the EQ-5D-5L, two versions were used, the original EQ-5D-5L, and an adapted 4-level version. The EQ-5D-5L values are based on a hybrid model of DCE and cTTO.[14] Such methods are different from that of MAPR model, which could add to the incomparability of the two PROMs. We therefore created another experimental PROM called 5D. In the 5D, we adapted the 5-level of EQ-5D-5L to a 4-level system, to match to the MAPR measurement framework. Thus, the 5D descriptive system comprises the same 5 items as the EQ-5D, but each item consists of 4 levels. For the preference-based task of the 5D, the DD method (see below) from our MAPR model was used. In this way we were able to compare the coefficients and values of the EQ-5D-5L and the CS-Base. In addition, by adding the 5D we could also make more detailed comparisons regarding the content (items) of the EQ-5D-5L (using the 5D) and the CS-Base, as the same methodology was used.

CS-Base

CS-Base is a generic health-outcome instrument. Specifically, this instrument is an electronic patient-reported outcome measure (ePROM) that uses special software. The CS-Base was developed for measuring HRQoL and comprises 12 health items, each specified on four levels: mobility, vision, hearing, cognition, mood, anxiety, pain, fatigue, social functioning, daily activities, self-esteem, and independence.

EQ-5D-5L

The EQ-5D-5L consists of two parts, the descriptive system, and the EQ VAS (visual analogue scale).[33] The descriptive system comprises 5 items: mobility, self-care, usual activities, pain/discomfort, anxiety/depression. Each item has five response levels: no problems, slight problems, moderate problems, severe problems, unable to/extreme problems. In our study, for generating values, we did not include the VAS. The EQ-5D-5L USA value set was used to calculate values for the EQ-5D-5L in this study.[14]

5D

The 5D is an ad-hoc, experimental PROM that is fully operated in our own measurement framework. The 5D comprises the same 5 items (mobility, self-care, usual activities, pain/discomfort, anxiety/depression) as the EQ-5D-5L, but each item is reduced to four levels to make it comparable to the operation of the CS-Base. Levels are: no problems, slight problems, moderate problems, severe problems. We have dropped the 5th level of the EQ-5D-5L because this level is not that frequently selected by respondents.[34,35]

DD method

Whitin the MAPR model, respondents first performed the descriptive task (Task 1) for assessing and describing their current health status. After that they were directed to Task 2 (the DD method). In the DD method, respondents are presented with their own health state (assessed in Task 1) and asked to select the item (with a suboptimal level: 2, 3, or 4) that hindered or disturbed them the most by clicking or swiping (drop-down) this item one level lower (better) (Figure 1). We set the maximum number of selections (drop-down) to 5, i.e. respondents can make between 1 to 5 selections. Each drop-down produced a health state that could be ranked as better than the initial health state from Task 1 (there should be at least two items with levels >1, otherwise the choice was predetermined, and if an item was at level 3 or more, they could drop-down this item more than once). The initial health state (Task 1) in the DD method is ranked as the worst state. Trade-offs made in the DD method are between the levels of multiple items (i.e., is level *i* of item *x* worse than any level of another item?).

MAPR measurement model

We used the MAPR measurement model (belongs to the probabilistic choice models) for the CS-Base and the 5D. These probabilistic choice models can establish the relative merit

(value) of a subjective phenomenon. These models are indirect, producing measures using the metric scale (analogous to a yardstick). For all probabilistic choice models, respondents must perform preference-based tasks in a particular way to endorse a specific response. This then generates data for an analysis in accordance with the measurement model. The core of a preference-based task in these probabilistic measurement frameworks consists of a response task that compares at least two objects with the aim of expressing which object is most preferred (is better). From a technical perspective, these models group ordinal data obtained from respondents. The grouped data are then aggregated to infer an interval scale (metric measure: value) that is based on a mathematical (measurement) model.

The DD method works as the preference-based task (Task 2) in our MAPR model, it produces ranked health states as the ordinal data for analysis. The DD methods process preferences in this way. The value of a health state j for individual i is denoted by V_{ij}. A respondent will rank state j higher than state k if $V_{ij}>V_{ik}$. The probability that state j is chosen as the most preferred state among the entire set of J states can be written as:

$$P_{ij} = \frac{e^{V_{ij}}}{\sum_{k=1}^{J} e^{V_{ik}}} \quad (1)$$

The probability of observing a specific ranking can be written as the product of such terms, representing a sequential decision interpretation, in which the respondent first chooses the most preferred alternative, and then the most preferred alternative among the rest, etc. To process the data generated with the DD method, the rank-ordered logit model is used.[36]

Mobile app

All the three PROMs run in the mobile app HealthSnApp (www.chateau-sante.com/healthsnapp). This is a flexible tool, with interactive routines. It runs on various electronic devices (e.g., smartphone, tablet, laptop) and is highly configurable from a web-based console. The two tasks which comprise the main routines of the measurement model (see above) are performed in the mobile app. The CS-Base and 5D responses were collected by these two tasks. Following these two tasks, an available survey component on the App was used for collecting EQ-5D-5L responses (descriptive system) and posing evaluation questions.

User evaluation two MAPR instruments

After completing the second task (DD task), respondents were invited to answer five questions to enable us to assess the sufficiency of description of health and level of difficulty of the two MAPR instruments. The five questions concerned: This tool gives a good description of my health: (1) CS-Base; (2) 5D; Description of items in this tool is easy to understand: (3) CS-Base; (4) 5D; (5) Which of the two tools do you prefer? Apart from the last question, which was a binary question, all other questions were scored 0-100 (where 0 indicated *totally disagree* and 100 indicated *totally agree*).

Study design

This study comprised two arms, each of which entailed the use of all the three PROMs: CS-Base, 5D and EQ-5D-5L, the CS-Base and 5D in reverse order (Study I: CS-Base–5D-EQ-5D-5L, Study II: 5D–CS-Base-EQ-5D-5L). Respondents were directed randomly to one of the two arms by the market research company.

Analysis

Coefficients of the CS-Base and 5D were estimated using a rank-ordered logit choice model (cmrologit, Stata 17.0). The first level of each item (level 1: no problems or an optimal condition) was the reference category. Regression coefficients were estimated for the remaining three levels (2, 3, and 4) using dummy variables (12 x 3 for CS-Base, 5 x 3 for 5D). No constants were included. The derived coefficients (weightings) were used to compute the values for distinct CS-Base and 5D health states. For EQ-5D-5L, the USA value set was used to calculate the values.

Means were used to calculate the scores of the four rating evaluation questions (good description of my health: CS-Base/5D; description of items in this tool is easy to understand: CS-Base/5D). The frequency and proportion were used to describe the binary question (which of the two tools do you prefer?). For testing difference between two instruments, the t-test was used for rating questions, proportion test was used for binary question. We used theStata 17.0, and CorelDraw 22.0 software packages to compute and visualize our results.

Results

Sample and sociodemographic characteristics

The sample used for comparison of the three instruments comprised of 1,988 respondents who completed the CS-Base, 5D and EQ-5D-5L. Table 1 shows the respondents' characteristics. The mean age was 46 years (range 18 to 94). There were 1,142 female respondents (57%). The majority of the respondents (1594, 80%) are White Americans/Caucasian. Regarding the education level, more than half (1154, 58%) of the respondents were high school graduates. The most reported main complaints or diseases were pains (696, 35%), mental health problems (347, 18%), fatigue/sleep problems (313, 16%), diabetes (123, 6%), respiratory disease (121, 6%).

Coefficients

The coefficients estimation of the CS-Base was based on outcomes of 2,534 respondents who did the DD tasks (with 1,296 respondents from this study, 1,239 from a previous study[32]). The coefficients estimation of the 5D comprised outcomes of 1690 respondents. For both CS-Base and 5D, all coefficients revealed a logical order (all the coefficients are negative numbers). The more negative a coefficient is, the lower the coefficient is (indicating a higher impact) (Table 2). All coefficients showed statistically significant differences ($P<0.001$). Clear differences of coefficients were observed between levels for all items (Figure 2). All the coefficients in the CS-Base had a smaller confidence interval than those for in the 5D. Only level 4 of 'Cognition' showed a large confidence interval, which may be attributed to the small number of responses collected at that level. Four items were both included in the CS-Base and EQ-5D-5L (5D): 'usual activities' ('daily activity' in the CS-Base), 'mobility', 'pain/discomfort' ('pain' in the CS-Base), 'anxiety/depression' ('anxiety' in the CS-Base). Compared with other items, the levels of the item 'daily activity' showed the highest coefficients in all three PROMs. The item 'mobility' showed lowest coefficients in the CS-Base, while in the 5D and EQ-5D-5L, its coefficients were moderate. The levels 4 and 5 of 'pain/discomfort' had lowest coefficients in the EQ-5D-5L, while it had moderate coefficients in the CS-Base and 5D. The item 'anxiety/depression' showed moderate coefficients in the CS-Base and EQ-5D-5L, but lower coefficients in the 5D. For the item 'self-care' which only existed in the EQ-5D-5L (5D), a similar levels differentiation of the coefficients was observed in both EQ-5D-5L and 5D, it had a low coefficient for the level 2,

but a high coefficient for level 4 (and level 5 in EQ-5D-5L). Low coefficients were also observed for the two items 'vision' and 'hearing', which are not included in the EQ-5D.

Frequency of complaints

For the four items ('mobility', 'usual activities', 'pain/discomfort', 'anxiety/depression') included in both the CS-Base and EQ-5D-5L (5D), the frequencies of complaints on all of them were similar between the CS-Base and 5D, with percentages at about 30%, 40%, 65% and 60% respectively. In the EQ-5D-5L, the frequencies were a little higher than the other two instruments. Few respondents complained on level 5, with a small percentage (≤3%) for 4 items ('mobility', 'self-care', 'usual activities', 'pain/discomfort'), and a little higher percentage (8%) for 'anxiety/depression'. In all the three PROMs, 'pain' was the most frequently (≥ 65%) reported complaint. In the 5D and EQ-5D-5L, 'self-care' was the least reported complaint. In the CS-Base, 'cognition' was the least reported complaint, with a percentage of 20%. On 5 of the 8 items which are not part of the EQ-5D, a substantial number of respondents indicated that they had problems: 'self-esteem' (55%), 'social function' (43%), 'fatigue' (63%), 'mood' (47%), 'hearing' (47%). On the other 3 of the 8 items, less respondents indicated that they had problems: 'independence' (29%), 'vision' (24%), 'cognition' (20%).

Health states and values

There were 1,988 respondents that assessed their health states by all three PROMs. The number of different health states assessed in the CS-Base, 5D, EQ-5D-5L were 1,472, 329, and 483 respectively. Mean values of the health states reported in the CS-Base, 5D and EQ-5D-5L was -30.05, -13.31 and 0.67. The values for perfect health (all the items were assessed as level 1) in the CS-Base and 5D are 0.0, and 1.0 in the EQ-5D-5L. Perfect health was reported by 235, 426 and 199 respondents in the CS-Base, 5D and EQ-5D-5L respectively. The values for the worst health state (all the items were assessed as level 4) in the CS-Base, 5D and EQ-5D-5L were respectively -158.76, -61.80 and -0.57. No respondent reported the worst health state in the CS-Base, in the 5D and EQ-5D-5L, it was reported by 6 and 3 respondents respectively. The worst health state among the 1,988 respondents reported in the CS-Base is 342444443344 (value=-131.80).

The values of the CS-Base were more densely distributed (Figure 3), followed by the EQ-5D-5L. The values of the 5D were more sparsely distributed and showed some gaps (e.g.,

values at around -5, -6, -9, -10). In addition, there were more respondents reported perfect health in the 5D than in the CS-Base and EQ-5D-5L. A celling effect was also observed in the 5D and EQ-5D-5L as far more respondents reported mild health states than severe health states in these two PROMs (Figure 3), this celling effect appeared to be minor in the CS-Base.

In all the three sub-figures of figure 4, the dots were widely spread along the goodness of fit line instead of close near the line, which indicated that there are differences between these three PROMs in measuring health states. However, in figure 4A, the spread of dots is more even along the goodness of fit line than in figure 4B and 4C.

User evaluation two MAPR instruments

Next, we will report the means (SD) scores of the total sample (1,988 respondents) for the four rating questions (higher scores, better performance). For the question regarding the quality of the description of health, CS-Base scored at 63 (27), 5D scored at 62 (27). Regarding the ease of understanding, CS-Base scored 57 (31), 5D scored at 55 (31). No significant difference was found between 5D and CS-Base regarding the quality of the description of health. Regarding the question of ease of understanding, there was no significant difference between the two instruments based on the total sample, but significant differences ($P<0.001$) were observed in the separate study arms. In study I, CS-Base was scored higher than 5D (mean65 for the CS-Base, 47 for the 5D). In study II, 5D was scored higher than CS-Base (mean score was 63 for the 5D, 47 for the CS-Base). In response to the binary question regarding preference, the 5D was somewhat more preferred ($P<0.001$) based on the total sample and study II (selected by 57% of the respondents in the total sample, selected by 65% respondents in study II), while no difference was found in study I.

Discussion

A major challenge in health outcome measurement is to develop PROMs that capture the complete range of health items, and thus give a sufficient description of health conditions, at the same time these PROMs should be easy to use. Our study entailed a head-to-head comparison of the CS-Base and EQ-5D-5L (two versions: the original EQ-5D-5L and an adapted 4-level version (5D)). Results showed that the CS-Base captures a more complete range of health items and discriminates different health states better. In the evaluation of two MAPR instruments, the 5D was preferred over the CS-Base regarding the ease of use. Both instruments were regarded equally sufficient for describing health conditions.

Four items ('mobility', 'usual activities', 'pain/discomfort', 'anxiety/depression') were included both in the EQ-5D-5L/5D and CS-Base. In all the three PROMs, 'pain' was the most frequently reported complaint, this was consistent with the diseases/complaints reporting rate. The frequencies of complaints on the four items were similar between the two MAPR PROMs (5D and CS-Base), but a little higher in the EQ-5D-5L. These higher frequencies could be attributed to the extra level of the EQ-5D-5L compared to the 5D (4-level). A 5-level descriptive system can be more sensitive in identifying complaints than a 4-level system. On the other hand, a 5-level system also has its disadvantages. We observed that not many respondents complained on level 5 in the EQ-5D-5L, which is consistent with previous studies.[37,38] That's one reason we dropped the fifth level in adapting the EQ-5D-5L to the 5D. Another reason to drop the fifth level is that the phrasing of level 4 and level 5 could be perceived as problematic by some respondents, as a preference inversion was observed between level 4 and 5.[37] A preference inversion is when a respondent states a preference that contradicts the ordering of labels in an item, it means the differences between level 4 and 5 was sometimes reversed. For example, respondents could prefer being "extremely" over "severely anxious or depressed," contrary to the ordering of labels for that item. The impacts of the four items in all the three PROMs were comparable overall. The item 'usual activities' had lowest impacts in all the three PROMs. The other three items ('mobility', 'pain', 'anxiety') had moderate or high impacts, with slight differences between the three PROMs.

There are eight items only included in the CS-Base but not in the EQ-5D-5L, they are: 'vision', 'hearing', 'cognition', 'mood', 'fatigue', 'social function', 'self-esteem' and 'independence'. A substantial number of respondents (>43%) reported problems on five

items (‘hearing’, ‘mood’, ‘fatigue’, ‘social function’, ‘self-esteem’). High impacts were derived for the two items ‘vision’ and ‘hearing’, moderate impacts were derived for three items (‘mood’, ‘fatigue’, ‘social function’). In the CS-Base, the indication of the item ‘mood’ is comparable to ‘anxiety’, their impacts were also very similar. The reason can be that although “mood” includes a wider range of emotions (including anxiety, depression, happiness, sorrowness, et al), respondents could only focus on the complaints like anxiety or depression. If they feel happy or good, they could just take it as granted but won’t report such good feelings. Some of the items which are not included in the EQ-5D-5L (e.g., ‘vision’ and ‘hearing’) showed a high impact, so it’s reasonable to be kept in the PROM. For some other items (e.g., ‘self-reliance’, ‘independence’ in the CS-Base) which showed lower impact and were less reported as problematic, it can be further explored in further studies to keep them in or not. Based on the targeted population of the study, these decisions can be different.

Overall, for the CS-Base and 5D, coefficients on each of the levels of all items are comparable (e.g., level 2 was valued between minus 3 and 4 on all items, level 3 was valued between minus 7 and 8). This means that all items were assessed more or less equally important and had a similar impact in the valuation of the health condition of the group of patients in this study. For the CS-Base, this finding is expected, as in a separate study the items considered as most important were selected from a large set of candidate items.[19] In the MAPR measurement framework, standard deviations for the coefficients are different for the levels of the items. That’s because assessments and responses are based on the actual health state of individual patients, higher levels are less frequently assessed in Task 1, nor drop-down in Task 2, which resulted into larger confidence intervals for higher levels. For the EQ-5D instruments, this is different. Hypothetical health states are assessed in the EQ-5D, and these states are generated based on an experimental design. All levels for each item are presented in equal numbers, which produces equal confidence intervals for all coefficients.

According to the value distribution of the 3 PROMs, the CS-Base turned out to perform best in capturing a more complete range of health states and in discriminating between different health states. Conversely, based on the visual inspection of the value distribution, we assume the 5D to be less sensitive than the other two instruments. Celling effects were also observed in the 5D and EQ-5D-5L, but very minor in CS-Base. Celling effects were observed between more health states in the 5D than in the EQ-5D-5L. We argue that the better performance of the CS-Base can be mainly attributed to the larger number of items included in the instrument

giving a more complete description of health. Among the eight items in the CS-Base that were not included in the EQ-5D, five of them showed high or moderate impact. These items were all selected by patients, which therefore, assumably better reflects patients' perspectives on their health than instruments in which items were selected by professionals. The outperformance of EQ-5D-5L compared to the 5D can probably be explained by the larger number of levels for the descriptive system. The 5-level descriptive system of EQ-5D can give a more complete description and a better discrimination of health states than a 4-level descriptive system.

As mentioned in the instruction, many of the conventional preference-based methods have limitations including that they are mainly based on (pairs of) hypothetic health states assessed by a sample of the general population instead of patients, and these methods are to some extent complicated to perform. Compared to conventional preference-based methods, the most outstanding advantage of the DD method is that it is easy to perform. No alternative or hypothetic health states are included in this method, the patients only have to assess their own health conditions. They just need to select and swipe away the items that hinders them most. Thus, the DD method is directed more accurately at the patients' own experience and easier to perform. Meanwhile, the DD method can also be administered on smartphones or other electronic devices, which makes the tasks more convenient and attractive to users. One issue still to overcome is the fact that the conventional preference-based methods like the TTO are able to generate a utility value, while the MAPR model cannot do this directly, a further normalization step is needed to rescale the values from full health (1.0) to death (0.0).[38]

According to the binary question regarding preference, the 5D was preferred over the CS-Base, we suppose the reason may be due to its short descriptive system with only 5 items. This difference was found in the total sample and study II, but not in study I. Currently an explanation for this is lacking, the reasons hasn't been know, we are still exploring it. Regarding the question of ease of understanding, there was no difference found between the CS-Base and 5D based on the total sample. But when compared within each of two study arms, we found something remarkable. In each of the two study arms, the first presented instrument (CS-Base in study I, 5D in study II) was regarded as easier to understand. However, we cannot conclude which instrument (CS-Base or 5D) is easier based on this. We assume that such a phenomenon could be related to the anchoring bias, which means people have a tendency to rely too heavily on the very first piece of information they.[39] As Tversky

and Kahneman explained: people make estimates by starting from an initial value that is adjusted to yield the final answer.[40] One example given by Tversky and Kahneman is: participants spun a wheel to select a number between 0 and 100. The volunteers were then asked to adjust that number up or down to indicate how many African countries were in the U.N. Those who spun a high number gave higher estimates while those who spun a low number gave lower estimates. In each case, the participants were using that initial number as their anchor point to base their decision. According to the binary question, the 5D was more preferred over the CS-Base, we suppose the reason may be due to its short descriptive system with only 5 items.

Conclusion

This study revealed that the CS-Base based on MAPR model captured a more complete range of health items and giving a better discrimination of different health states as compared to EQ-5D-5L and 5D. The 5D was more preferred by respondents in the evaluation of two MAPR instruments. The adapted 4-level version appeared less sensitive than the original EQ-5D-5L in identifying patients' complaints, nor in distinguishing different health states. The CS-Base, the 5D and the EQ-5D-5L each has their own advantages and biases, researchers can make a well-informed choice about which one to use according to the purpose of their study. Good description and discrimination of health states, as well as ease of use are the main properties of a PROM. Attention should always be paid to the tradeoff between these instruments when choosing which one to include in a study.

Table 1 Characteristics of the total sample and of the separate studies (I and II)

Characteristics	Total sample (1,988)	Study I	Study II
Gender, N (%)	1,988 (100)	1,031 (100)	957 (100)
Females	1,142 (57)	601 (58)	541 (57)
Males	846 (43)	430 (42)	416 (43)
Age (yrs), Mean (SD)	46 (17)	46 (17)	46 (17)
Age (yrs), N (%)	1,988 (100)	1,031 (100)	957 (100)
18-27	293 (15)	150 (15)	143 (15)
28-37	458 (23)	254 (25)	204 (21)
38-47	369 (19)	181 (18)	188 (20)
48-57	293 (15)	147 (14)	146 (15)
58-67	294 (15)	151 (15)	143 (15)
68-77	235 (12)	126 (12)	109 (11)
≥78	46 (2)	22 (2)	24 (3)
Ethnicity, N (%)	1,983 (100)	1,030 (100)	953 (100)
Asian/Asian-American	44 (2)	19 (2)	25 (3)
Black/African-American	174 (9)	75 (7)	99 (10)
Hispanic or Latino American	112 (6)	52 (5)	60 (6)
Native American/Inuit/Alaskan	30 (2)	15 (1)	15 (2)
Native Hawaiian/Pacific Islander	17 (1)	10 (1)	7 (1)
White American/Caucasian	1594 (80)	856 (83)	738 (77)
Other	12 (1)	3 (0)	9 (1)
Education*, N (%)	1,988 (100)	1,031 (100)	957 (100)
More than high school	606 (30)	318 (31)	288 (30)
High school graduate	1,154 (58)	596 (58)	558 (58)
Less than high school	228 (12)	117 (11)	111 (12)
Main complaint/disease, N (%)	1,979 (100)	1,028 (100)	951 (99)
Pains	696 (35)	355 (35)	341 (36)
Mental health problems	347 (18)	170 (17)	177 (19)
Fatigue/sleep problems	313 (16)	161 (16)	152 (16)
Hearing or vision loss	96 (5)	53 (5)	43 (5)
Diabetes	123 (6)	66177 (617)	57163 (617)
Respiratory diseases	121 (6)	70 (7)	51 (5)
Heart disease	47 (2)	2895 (39)	1975 (28)
Eczema	46 (2)	24 (2)	22 (2)
Gastrointestinal disease	42 (2)	25 (2)	17 (2)
Rheumatism	33 (2)	17 (2)	16 (2)
Cancer	23 (1)	7 (1)	16 (2)
Stroke	18 (1)	12 (1)	6 (1)
Epilepsy	16 (1)	8 (1)	8 (1)
Other diseases	58 (3)	32 (3)	26 (3)

Table 2 Coefficients of CS-Base, 5D and EQ-5D-5L (USA)[14]

CS-Base (N=2,534)				5D (N=1,690)				EQ-5D-5L (N=1,062)		
Item levels	Coefficient	SE	Z	Item levels	Coefficie	SE	Z	Coefficient	SE	t
Mobility (2)	-3.22	0.13	-25.49	Mobility (2)	-3.92	0.20	-19.95	-0.096	0.02	-6.56
Mobility (3)	-8.95	0.19	-46.29	Mobility (3)	-8.55	0.30	-28.33	-0.122	0.02	-7.69
Mobility (4)	-15.40	0.35	-44.57	Mobility (4)	-13.68	0.53	-25.87	-0.237	0.02	-13.42
				Mobility (5)	-	-	-	-0.322	0.02	-19.85
Vision (2)	-3.25	0.12	-26.06	Self-care (2)	-4.04	0.21	-19.23	-0.089	0.01	-6.33
Vision (3)	-8.24	0.19	-43.54	Self-care (3)	-8.19	0.32	-25.72	-0.107	0.02	-6.28
Vision (4)	-14.55	0.39	-37.67	Self-care (4)	-11.39	0.60	-18.96	-0.220	0.02	-12.49
				Self-care (5)	-	-	-	-0.261	0.02	-16.30
Hearing (2)	-3.45	0.10	-35.59	Usual activities (2)	-3.41	0.16	-20.71	-0.068	0.02	-4.67
Hearing (3)	-8.66	0.16	-53.91	Usual activities (3)	-7.02	0.25	-27.99	-0.101	0.02	-6.18
Hearing (4)	-14.76	0.32	-46.31	Usual activities (4)	-9.99	0.41	-24.18	-0.255	0.01	-18.95
				Usual activities (5)	-	-	-	-0.255	0.01	-18.95
Cognition (2)	-3.28	0.14	-23.43	Pain/Discomfort (2)	-3.79	0.15	-25.21	-0.060	0.01	-4.61
Cognition (3)	-8.19	0.21	-37.77	Pain/Discomfort (3)	-8.34	0.26	-31.51	-0.098	0.02	-5.72
Cognition (4)	-12.87	0.54	-23.81	Pain/Discomfort (4)	-13.17	0.43	-30.75	-0.318	0.02	-21.05
				Pain/Discomfort (5)	-	-	-	-0.414	0.02	-24.19
Mood (2)	-3.30	0.10	-33.49	Anxiety/Depression (2)	-3.98	0.16	-24.62	--0.057	0.01	-4.03
Mood (3)	-7.89	0.16	-50.51	Anxiety/Depression (3)	-8.38	0.27	-30.81	-0.123	0.02	-6.86
Mood (4)	-13.19	0.27	-48.34	Anxiety/Depression (4)	-13.58	0.45	-30.23	-0.299	0.02	-18.48
				Anxiety/Depression (5)	-	-	-	-0.321	0.02	-21.18
Anxiety (2)	-3.13	0.09	-34.28							
Anxiety (3)	-7.44	0.14	-51.67							
Anxiety (4)	-12.94	0.22	-57.65							
Pain (2)	-3.23	0.09	-35.93							
Pain (3)	-7.54	0.14	-53.28							
Pain (4)	-13.14	0.22	-58.81							
Fatigue (2)	-3.40	0.09	-39.08							
Fatigue (3)	-7.65	0.14	-53.25							
Fatigue (4)	-12.55	0.23	-55.68							
Social function (2)	-3.44	0.10	-33.63							
Social function (3)	-7.56	0.17	-45.59							
Social function (4)	-12.71	0.29	-43.50							
Daily activity (2)	-3.46	0.10	-34.14							
Daily activity (3)	-7.65	0.17	-45.76							
Daily activity (4)	-11.72	0.35	-33.20							
Self-esteem (2)	-3.81	0.11	-35.80							
Self-esteem (3)	-7.54	0.17	-45.19							
Self-esteem (4)	-12.45	0.25	-50.58							
Independence (2)	-3.83	0.13	-28.63							
Independence (3)	-8.15	0.22	-36.41							
Independence (4)	-12.50	0.42	-29.83							

All the p values were <0.001.

Figure 1 Screen grabs in HealthSnApp for DD method operation (Task 1 to Task 2). Based on the health state assessed by patients themselves in Task 1, in Task 2, they made multiple selections (1–5 times) of items at the levels that hindered or disturbed them the most; they did this by swiping the (drop-down) level and moving the item one level lower (or better).

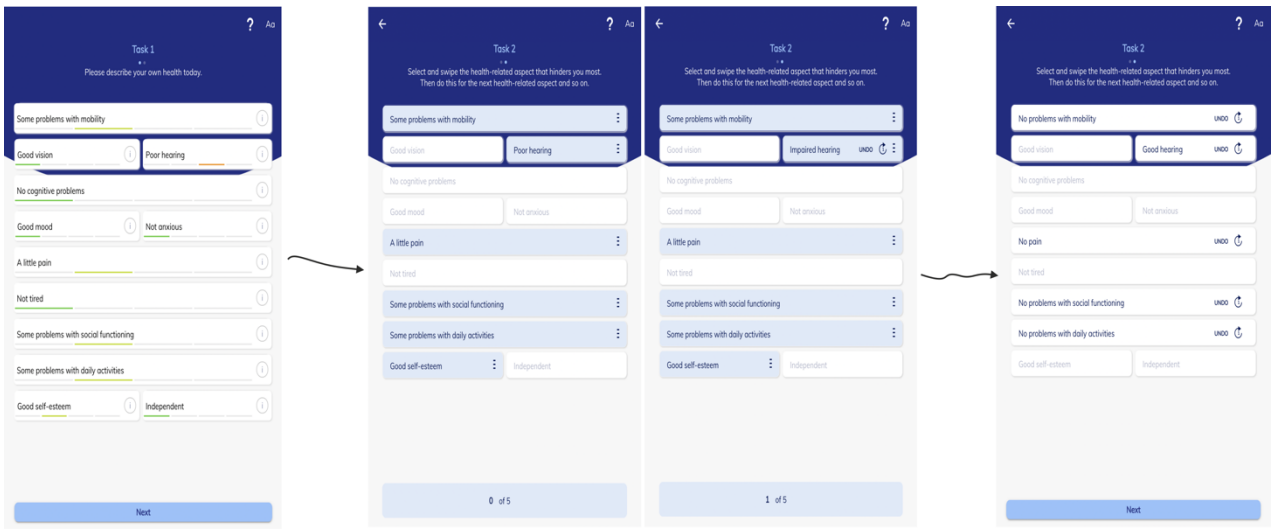


Figure 2 Distribution of coefficients derived from the CS-Base and 5D

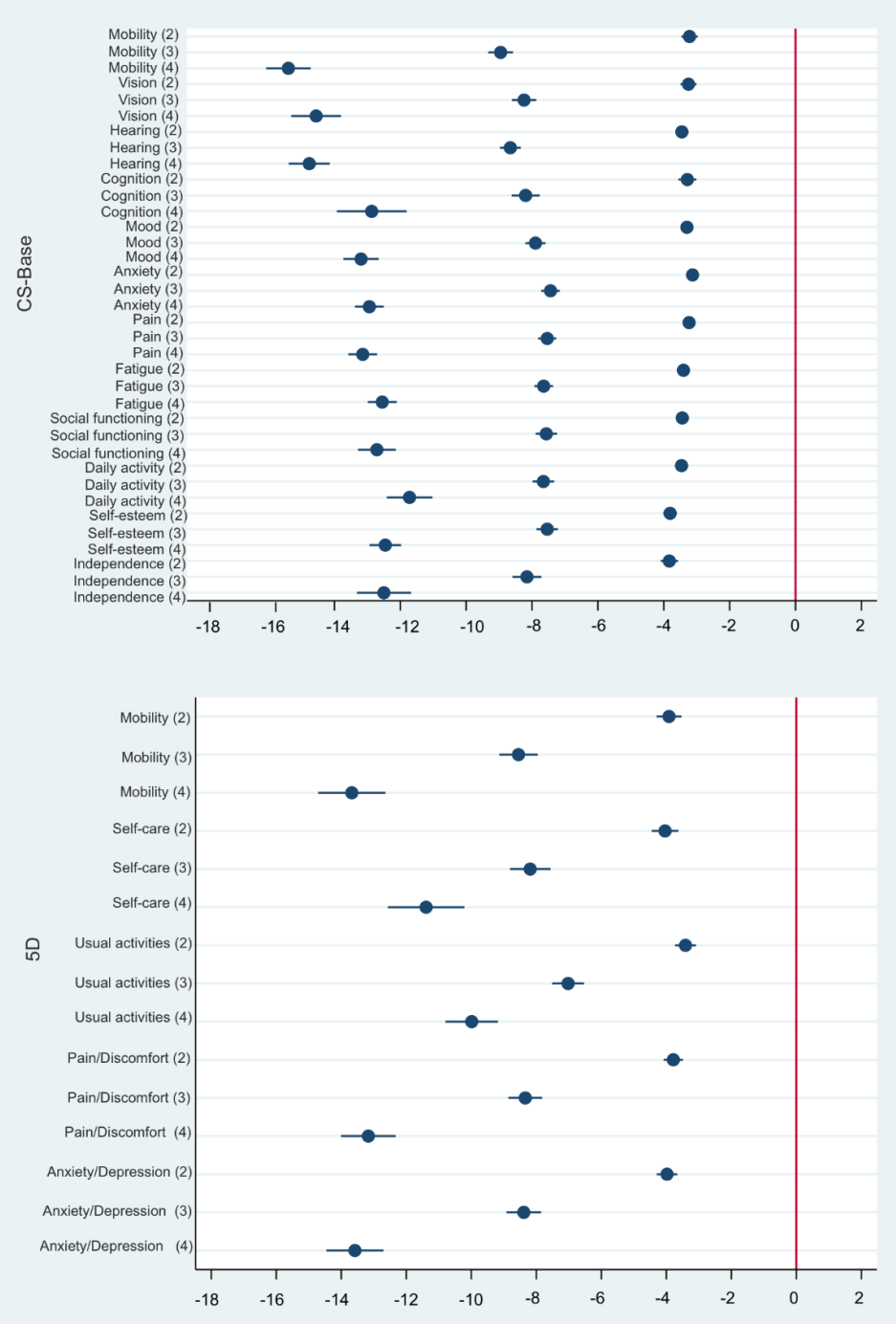


Figure 3 Distribution of values (without perfect health) for the 3 PROMs: CS-Base, 5D and EQ-5D-5L. The values for perfect health state in the CS-Base and 5D are 0, and 1 in the EQ-5D-5L. The health perfect health state was excluded from the figure presentation, so there are no bars above value "0" or "1". The number of respondents (without those reported perfect health) in each of the 3 PROMs were marked on the top-left.

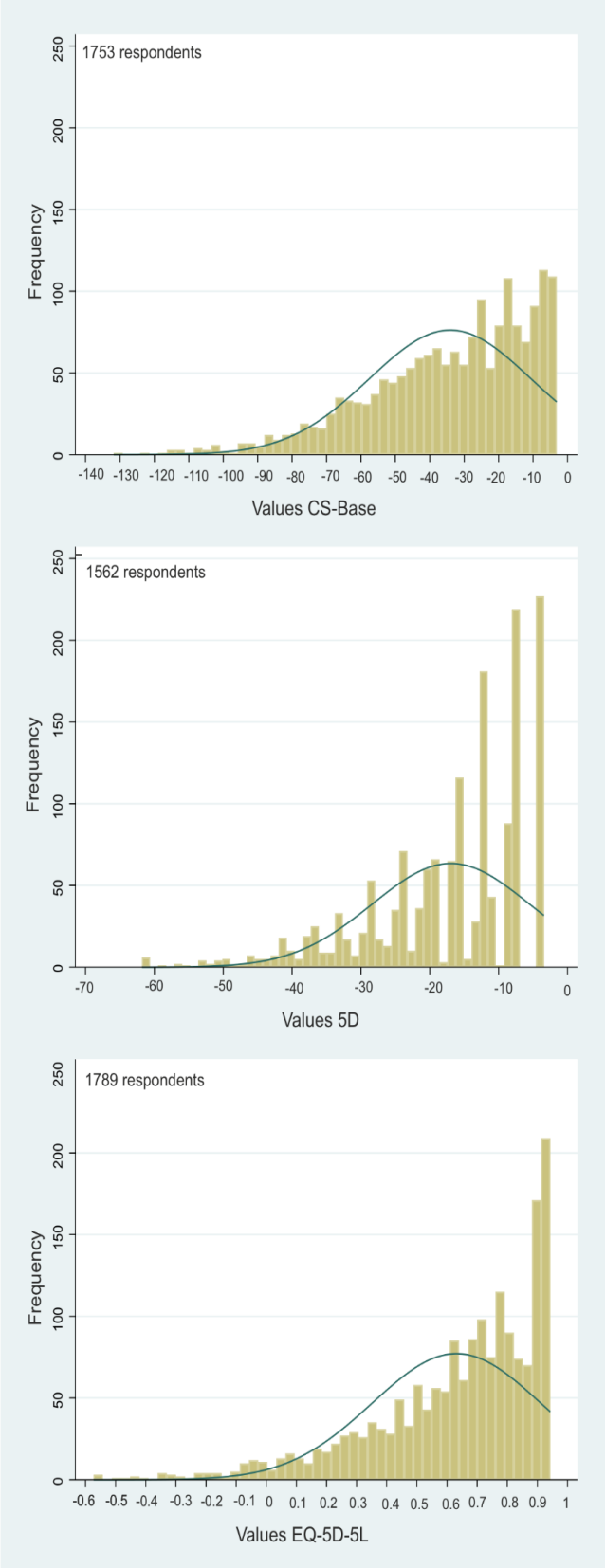
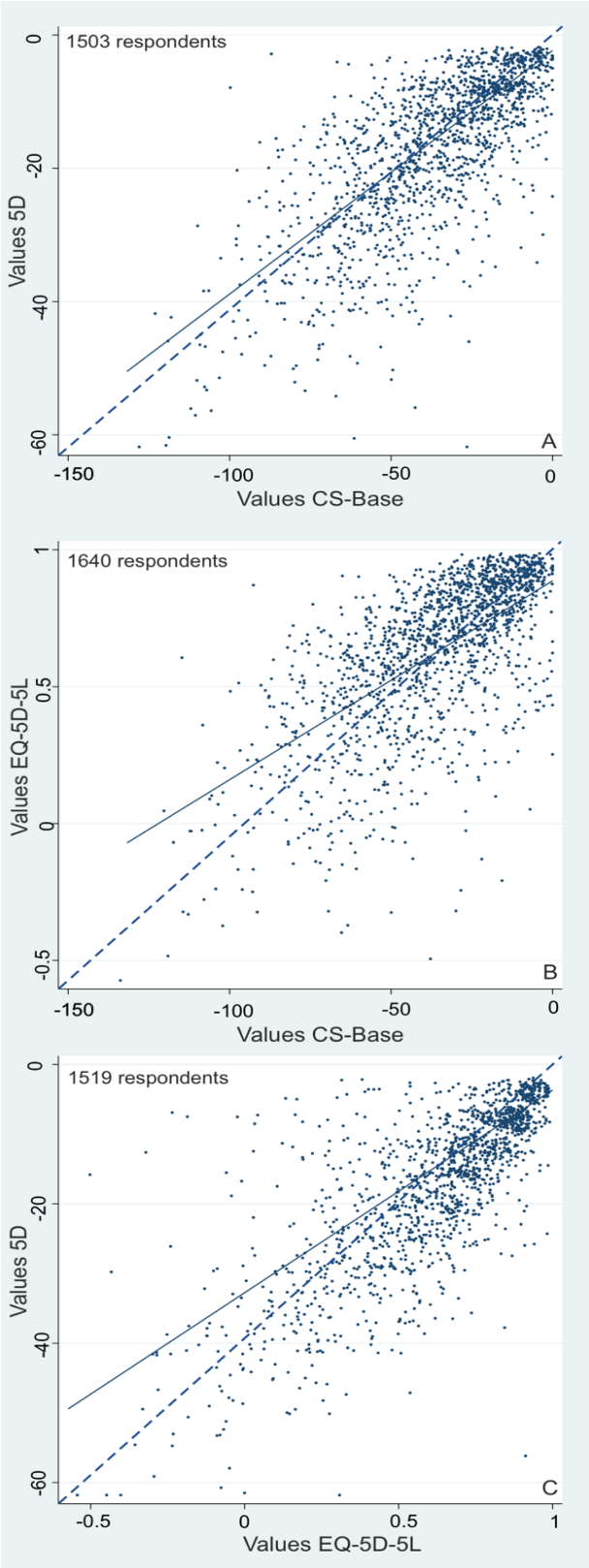


Figure 4 Scatter plots of values (without perfect health) comparing each pair of the three PROMs (dots are plotted with jitter option to reduce overplotting). Fig3A: 5D vs CS-Base; Fig3B: EQ-5D vs CS-Base, Fig3C: 5D vs EQ-5D. The perfect health state was excluded from the figure presentation. The numbers of respondents on the top-left indicate the number of respondents left in both PROMs (each pair the three PROMs) after excluding those in perfect health.



References

1 Laine C, Davidoff F. Patient-centered medicine. A professional evolution. JAMA. 1996;275(2):152-6.

2 Leyshon S, McAdam S. Scene setter: the importance of taking a systems approach to person centred care. *BMJ*. 2015;350:h985 doi:10.1136/bmj.h985

3 Krabbe PFM. *The Measurement of Health and Health Status: Concepts, Methods, and Applications from a Multidisciplinary Perspective.* London (UK): Academic Press; 2017. p7.

4 Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life: a conceptual model of patient outcomes. *JAMA.*1995;273(1):59-65.

5 Mercieca-Bebber R, King MT, Calvert MJ, Stockler MR, Friedlander M. The importance of patient-reported outcomes in clinical trials and strategies for future optimization. *Patient Relat Outcome Meas.* 2018;9:353-367

6 Nelson EC, Eftimovska E, Lind C, Hager A, Wasson J H, Lindblad S et al. Patient reported outcome measures in practice. *BMJ.* 2015;350:g7818 doi:10.1136/bmj.g7818

7 Marshall S, Haywood K, Fitzpatrick R. Impact of patient-reported outcome measures on routine practice: a structured review. *J Eval Clin Pract.* 2006;12(5):559-68. doi: 10.1111/j.1365-2753.2006.00650.x.

8 Santana MJ, Feeny D. Framework to assess the effects of using patient-reported outcome measures in chronic care management. *Qual Life Res.* 2014;23(5):1505-13. doi: 10.1007/s11136-013-0596-1.

9 Valderas JM, Kotzeva A, Espallargues M, Guyatt G, Ferrans CE, Halyard MY, et al. The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature. *Qual Life Res.* 2008;17(2):179-93. doi: 10.1007/s11136-007-9295-0.

10 Kluzek S, Dean B, Wartolowska KA. Patient-reported outcome measures (PROMs) as proof of treatment efficacy. *BMJ Evidence-Based Medicine.* Published Online First: 04 June 2021. doi: 10.1136/bmjebm-2020-111573

11 Mao Z, Ahmed S, Graham C, Kind P. The Unfolding Method to Explore Health-Related Quality of Life Constructs in a Chinese General Population. Value Health. 2021 Jun;24(6):846-854. doi: 10.1016/j.jval.2020.12.014.

12 ICHOM working groups. ICHOM website. Accessed 06 May 2022. https://www.ichom.org/standard-sets/.

13 Wiering B, de Boer D, Delnoij D. Patient involvement in the development of patient reported outcome measures: a scoping review. *Health Expect.* 2017;20(1):11-23.

14 Pickard S, Law E, Jiang R, Pullenayegum E, Shaw J, Xie F, et al. United States valuation of EQ-5D-5L health states using an international protocol. *VALUE HEALTH.* 2019; 22(8):931-941

15 Versteegh MM, Vermeulen KM, Evers SM, de Wit GA, Prenger R, Stolk EA. Dutch Tariff for the Five-Level Version of EQ-5D. Value Health. 2016;19(4):343-52.

16 Devlin N, Shah K, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of Life: An EQ-5D-5L value set for England. *Health Economics.* 2017;1-16

17 EuroQol Group. EuroQol--a new facility for the measurement of health-related quality of life. *Health Policy.* 1990;16(3):199-208. doi: 10.1016/0168-8510(90)90421-9.

18 Brooks R. EuroQol: the current state of play. *Health Policy.* 1996;37(1):53-72. doi: 10.1016/0168-8510(96)00822-6.

19 Krabbe PFM, van Asselt ADI, Selivanova A, Jabrayilov R, Vermeulen KM. Patient-centered item selection for a new preference-based generic health status instrument: CS-Base. *Value Health.* 2019;22(4):467-473.

20 Calvert M, Kyte D, Price G, Valderas JM, Hjollund NH. Maximising the impact of patient reported outcome assessment for patients and society. *BMJ.* 2019;364: k5267.

21 Gold MR, Siegel JE, Russel LB, Weinstein MC. Cost-effectiveness in Health and Medicine. New York: Oxford University Press; 1996.

22 Krabbe PFM, Tromp N, Ruers TJ, Riel PLCM. Are patients' judgments of health status really different from the general population? *Health Qual Life Outcomes.* 2011;9(31):3-9. https://doi.org/10.1186/1477-7525-9-31

23 Robinson A. Did Einstein really say that? Nature. 2018; 557: 30.

24 Krabbe PFM. A generalized measurement model to quantify health: the multi-attribute preference response model. *PLoS ONE.* 2013;8(11):1-12.

25 Krabbe PFM. A generalized measurement model to quantify health: the multi-attribute preference response model. In: Badiru AB, Racz LA (editors). *Handbook of Measurements: Benchmarks for Systems Accuracy and Precision.* Boca Raton: CRC Press, Taylor and Francis Group; 2015; p239

26 Groothuis-Oudshoorn CGM, van der Heuvel E, Krabbe PFM. A preference-based item response theory model to measure health: concept and mathematics of the multi-attribute preference response model. *BMC Med Res Methodol.* 2018;18:62.

27 Thurstone LL. A law of comparative judgment. *Psychol. Rev.* 1927;34(4):266-270.

28 Luce RD. *Individual Choice Behavior: A Theoretical Analysis.* New York (USA): Dover publication, Inc; 2005.

29 McFadden D. Economic choices. *Am Econ Rev.* 2001;91(3):351-378.

30 Marley AAJ. Some probabilistic models of simple choice and ranking. *J Math Psychol.* 1968;5(2):311-332.

31 Salomon JA. Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. *Population Health Metrics.* 2003;1-12.

32 Zhang X, Krabbe PFM. From simple to even simpler, but not too simple: a head-to-head comparison of the Better-Worse and Drop-Down methods to measure patients' health status. *Med Decis Making*. Submitted for publication.

33 EuroQol Research Foundation. EQ-5D user guides. Sep 2019. Accessed 07 May 2022. https://euroqol.org/publications/user-guides/.

34 Mulhern B, Feng Y, Shah K, Janssen MF, Herdman M, van Hout B, et al. Comparing the UK EQ-5D-3L and English EQ-5D-5L Value Sets. *Pharmacoeconomics.* 2018;36(6):699-713. doi: 10.1007/s40273-018-0628-3.

35 Janssen MF, Bonsel GJ, Luo N. Is EQ-5D-5L better than EQ-5D-3L? A head-to-head comparison of descriptive systems and value sets from seven countries. *Pharmacoeconomics.* 2018;36(6):675-697. doi: 10.1007/s40273-018-0623-8.

36 Marden JI. *Analyzing and Modeling Rank Data.* 1st ed. New York (USA): Chapman and Hall/CRC; 1995.

37 Craig BM, Pickard AS, Rand-Hendriksen K. Do health preferences contradict ordering of EQ-5D labels? *Qual Life Res.* 2015;24(7):1759-65. doi: 10.1007/s11136-014-0897-z.

38 Krabbe PFM, Jabrayilov R, Detzel P, Dainelli L, Vermeulen KM, van Asselt ADI. A two-step procedure to generate utilities for the Infant health-related Quality of life Instrument (IQI). *PLoS One.* 2020;15(4): e0230852. doi: 10.1371/journal.pone.0230852.

39 Teovanović P. Individual differences in anchoring effect: evidence for the role of insufficient adjustment. *Eur J Psychol.* 2019;15(1):8-24. doi: 10.5964/ejop.v15i1.1691.

40 Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. Science, New Series. 1974; 185 (4157): 1124-1131