White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/218385/

Version: Accepted Version

Article:
Rowen, D., McDool, E., Carlton, J. et al. (2025) Deriving a preference-weighted measure for people with hypoglycaemia from the Hypo-RESOLVE-QoL. Value in Health, 28 (1). pp. 125-137. ISSN: 1098-3015

https://doi.org/10.1016/j.jval.2024.10.3800

© 2024 The Authors. Except as otherwise noted, this author-accepted version of a journal article published in Value in Health is made available via the University of Sheffield Research Publications and Copyright Policy under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

Deriving a Preference-Weighted Measure for People with Hypoglycaemia from the Hypo-RESOLVE-QoL

Authors: Donna Rowen, PhD[1], Emily McDool, PhD[1], Jill Carlton, PhD[1], Philip Powell, PhD[1], Richard Norman, PhD[2]

[1] Sheffield Centre for Health and Related Research (SCHARR), University of Sheffield, Sheffield, UK
[2] School of Population Health, Curtin University, Perth, Australia

Corresponding Author Information:
Donna Rowen, PhD
Sheffield Centre for Health and Related Research (SCHARR), University of Sheffield
United Kingdom, South Yorkshire, Sheffield
Email: d.rowen@sheffield.ac.uk.
Phone: +44 114 222 0794

Precis: Generation of a novel preference-weighted measure from Hypo-RESOLVE-QoL with elicitation of UK public preferences for use in cost-effectiveness analyses for hypoglycaemia

Author Contributions:
Concept and design: Donna Rowen, Emily McDool, Jill Carlton, Philip Powell, Richard Norman.
Acquisition of data: Donna Rowen, Emily McDool, Jill Carlton.
Analysis and interpretation of data: Donna Rowen, Emily McDool, Jill Carlton, Philip Powell, Richard Norman.
Drafting of the manuscript: Donna Rowen, Emily McDool, Jill Carlton, Philip Powell, Richard Norman.
Critical revision of the paper for important intellectual content: Donna Rowen, Emily McDool, Jill Carlton, Philip Powell, Richard Norman.
Statistical analysis: Donna Rowen, Emily McDool, Philip Powell.
Provision of study materials or patients: Not applicable.
Obtaining funding: Jill Carlton, Philip Powell.
Administrative, technical, or logistic support: Not applicable.
Supervision: Not applicable.
Other (if applicable, please specify): Not applicable.

Conflict of Interest Disclosures: Rowen, McDool, Carlton, Powell reported receiving grants from the project Innovative Medicines Initiative 2 Joint Undertaking (JU) during the conduct

of the study. Rowen and Norman are members of the EuroQol Group. No other disclosures were reported.

Funding/support: This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 777460. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA and T1DExchange, JDRF, IDF, HCT. The authors are solely responsible for the content of this work, which reflects only the authors' views, and the JU is not responsible for any use that may be made of the information it contains.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Acknowledgements: We would like to thank all participants who took part in this research.

Highlights:

1. Existing preference-weighted measures do not capture all important impacts of hypoglycaemia on health-related quality of life of people living with diabetes.
2. The paper reports the development of a new preference-weighted measure capturing the health-related quality of life impact of hypoglycaemia in people living with diabetes, and its valuation using members of the UK public.
3. The new preference-weighted measure can be used to generate utilities and quality adjusted life years (QALYs) for hypoglycaemia.

Keywords: Hypoglycaemia, Hypo-RESOLVE-QoL, Diabetes, Preference-based measure, Preference-weighted measure

Abstract

Objective: Hypoglycaemia impacts the health-related quality of life (HRQoL) of people living with diabetes (PwD), and existing preference-weighted measures do not capture all important aspects. The study aimed to generate a preference-weighted measure capturing the HRQoL impact of hypoglycaemia in PwD.

Methods:

Items for the health state classification system were selected from the hypoglycaemia-specific Hypo-RESOLVE QoL measure using: relevance in cognitive interviews, translatability, suitability for valuation, endorsement by patient advisors and experts, and psychometric performance in a large survey of PwD. Second, an online valuation survey using discrete choice experiment (DCE) with survival attribute was conducted with members of the UK public. DCE data was modelled using conditional logit analysis, and results scaled to produce preference weights for the classification system on a scale where 1 is equivalent to full health, 0 is equivalent to dead, and below zero is worse than dead.

Results:

The health state classification system consists of eight items reflecting the factors of the Hypo-RESOLVE QoL (psychological, social and physical aspects). The valuation survey was completed by 1000 members of the UK public, representative for age and sex. Good understanding of DCE tasks was demonstrated. The item “do what I want to do in my life” had the largest preference weight, and “find it hard to stop thinking about my glucose levels” had the smallest.

Conclusions: This study generated Hypo-RESOLVE QoL-8D, a preference-weighted measure capturing the HRQoL impact of hypoglycaemia in PwD, with UK general public preference-weights. The measure can be generated from Hypo-RESOLVE QoL data.

1 Introduction

Assessments of patient health are increasingly undertaken using patient-reported outcome measures (PROMs) that ask patients to report their own health-related quality of life (HRQoL). The PROMs that are used can be generic, where they capture important impacts on HRQoL across conditions and hence maximise comparability, but are unlikely to include all symptoms or impacts relevant to patients with a given condition. Alternatively, PROMs can be condition-specific, where the focus is on aspects of HRQoL most likely to be relevant to the patient given their health condition.

In order to use these PROMs to understand the relative benefits from different treatments in economic evaluation, the impact of these treatments on utilities is required, where this provides an indication of the impact on HRQoL. These utilities are then used to generate quality adjusted life years (QALYs) which capture both HRQoL and length of life, and can be used to determine treatment effects and can be compared across different conditions and treatments. PROMs can be directly used to generate utilities, when scored using preference weights for different HRQoL outcomes, and these measures are referred to as preference-weighted measures (or preference-based measures, PBMs). Whilst preference-weighted measures are often used to generate QALYs and for use in economic evaluations, they can also be used to provide assessments of HRQoL changes over time and between populations.

Hypoglycaemia is a condition where glucose levels (i.e. blood sugar levels) are below a standard range, and this can be related to treatments for diabetes. Hypoglycaemia impacts on the HRQoL of people living with diabetes (PwD) [1,2]. The prevalence of hypoglycaemia amongst PwD was found to be 0.074% to 73% in a recent meta-analysis[3]. Recently, a review of existing hypoglycaemia-specific PROMs for use in PwD, designed to assess (aspects of) HRQoL, failed to determine a fully appropriate PROM [4]. Following the review, a large project was undertaken to develop both a new PROM and a new preference-weighted measure to fill this gap. The Hypo-RESOLVE QoL [5] was therefore generated, though as a PROM it does not have any associated preference weights, meaning it cannot be used to generate utilities, nor QALYs for use in economic evaluation.

The aim of this study was to generate a preference-weighted measure and scoring algorithm from the Hypo-RESOLVE QoL, capturing the HRQoL impact of hypoglycaemia in PwD, for use in economic evaluation.

2 Methods

Hypo-RESOLVE QoL

The Hypo-RESOLVE QoL (quality of life) is a new 14-item condition-specific PROM designed to capture the HRQoL impact of hypoglycaemia in PwD. The measure was developed using mixed-methods involving: semi-structured interviews to develop themes and draft items (using a semi-structured topic guide developed using a framework generated from a systematic review of condition-specific PROMs in hypoglycaemia [4,5]; cognitive debriefing interviews to refine items and assess content validity; and a large survey to assess item performance psychometrically and select the best performing items for the final measure (for full details see [5,6] and extensive supporting documents (https://doi.org/10.15131/shef.data.23295284.v2). The 14 items cover three factors: physical, social, and psychological aspects, and each item has 5 response options (none of the time, rarely, sometimes, often, most or all of the time) (see Table 1). Recent evidence shows that Hypo-RESOLVE QoL has good content validity (assessed using cognitive debriefing interviews with n=70 PwD and n=14 clinicians), and structural validity, convergent validity, internal consistency, and test-retest reliability (assessed using an online observational survey of n=1246 PwD) [6].

Derivation of health state classification system for valuation

Valuation exercises used to elicit preferences and generate preference weights, typically require consideration of all items in the classification system. However, the inclusion of all 14 items would be challenging for participants completing valuation tasks, in which they are asked to consider whole health profiles described using the instrument. A parsimonious number of items were selected from the 14-item Hypo-RESOLVE QoL to form the classification system for valuation to ensure that the two or more best performing items/ important aspects in each factor from the qualitative work used to develop the measure were retained. A range of information from the cognitive interviews (n=70 PwD and 14 healthcare

professionals, UK and Germany) and cross-sectional survey (n=1246, UK) used in the measure development was considered to inform selection [6], with preference given to items which:

- Were identified as relevant in cognitive interviews reporting on the relevance of items in UK and Germany (n=14);
- Were translatable across a range of languages (n=8);
- Met published criteria around suitability for inclusion in a preference-weighted measure [7];
- Received endorsement as a key area by a Patient Advisory Committee (PAC) comprising adults living with diabetes and representatives from International Diabetes Foundation (IDF) and Juvenile Diabetes Research Foundation (JDRF) (n=4 participants) and an advisory group comprised of experts (n=12);
- Demonstrated relatively low concentrations of ceiling or floor responses, defined as the proportion of responses at level 0 (experience problem none of the time) or at level 4 (experience problem most or all of the time) (n=1246);
- Fit within the factor (physical, social, and psychological aspects) using item response theory (IRT) (partial credit model);
- Demonstrated local independence with items in the same factor (physical, social, and psychological aspects) using item response theory (IRT) (using Mokken scale analysis based on a monotone homogeneity model);
- Had low differential item functioning (DIF) by type 1 or type 2 diabetes, or by gender using IRT;
- Demonstrated low correlations with other selected items; and
- Demonstrated a higher correlation with the number of hypoglycaemic events in the prior one week period.

Valuation of health state classification system

Valuation technique, population and recruitment

There are numerous valuation methods that can be used to generate utility values for health states, including time trade-off, standard gamble and discrete choice experiments (DCEs). Here a DCE was selected to value the classification system, with an attribute representing duration or survival in life years, which enables the modelled results to be anchored on the

scale where death equals zero. A DCE has the advantage that it can be administered quickly and cheaply in an online survey to a large sample, and its popularity is rapidly increasing for health state valuation, particularly for DCEs with choice sets consisting of two health profiles with a duration attribute [8]. Duration levels of 1, 4, 7 and 10 years were selected, as used previously in the literature [8].

Members of the UK general public were selected to value the classification system following recommendations from the National Institute of Health and Care Excellence (NICE) [9] for use in health technology assessment. Participants were recruited from existing managed online panels of people willing to answer surveys, via platform provider and market research agency SurveyEngine, and were reimbursed with points from the panel providers that are accumulated and exchanged for goods. Participants were invited from a user platform, and quotas were set to ensure sample representativeness for age and sex. A sample size of 1000 was selected to ensure each pair of profiles was valued 20 times or more [10] and at least one pair of profiles per parameter estimated was selected in the regression model.

Selecting profiles for the DCE survey

Choice sets were selected in NGene. First, a large number of potential profiles and choice sets were randomly generated (129,000). Then, a subset were selected with 3 overlapping items from the classification system, i.e. 3 items were the same in both profiles in the choice set to make the choice set an easier task (see for example [11]), and duration was allowed to overlap in addition. These potential choice sets were then optimised and selected using a design that maximises C-efficiency and takes into account the model specification required to generate preference weights. In total, 200 choice sets were selected. Across these, 41 have the same duration across both profiles. The design was tested using correlations between attributes within the pairs, number of appearances of each attribute level and regression results of models estimated on random responses.

The DCE survey

At the start of the survey, participants viewed a study information sheet and provided consent to participate. First, participants were asked questions about their sociodemographic and health characteristics, including whether they had diabetes and, if so, further questions about

treatment and experience of hypoglycaemia. Second, to provide participants with an understanding of hypoglycaemia, participants were provided with information, and a vignette describing a hypothetical example of the QoL impact of hypoglycaemia and completed the classification system imagining living in the vignette (see supplementary materials). Following a DCE exercise explanation and a practice question (with a dominant profile) with feedback explaining their response, participants completed 10 DCE tasks. Nine tasks were randomly selected (without replacement) from the 200 choice sets selected in the design, and one additional task (common to all participants) was added of a dominant choice question where one profile was better or the same in every attribute, see Figure 1 for an example choice set. Yellow highlighting was used to indicate the attributes that did not differ in the profiles in the choice set to make the task cognitively easier [11]. The 10 tasks were randomly ordered. Finally, participants completed questions about their understanding and what they thought of the survey.

The survey formatting and wording was piloted using a small convenience sample (n=9) from the host institutions volunteers list, with survey amendments made iteratively. The survey was soft launched online with 100 participants.

The research received ethical approval from the University of Sheffield Research Ethics Committee administered by the Sheffield Centre for Health and Related Research.

Analysis

Descriptive statistics of sample characteristics were used to assess the sample and sample representativeness of the UK general public. Participant understanding and engagement with the survey was assessed using self-reported understanding and whether participants correctly chose the dominant profile in the practice DCE task and the dominance task within the DCE tasks. Speedy responders (who completed the survey in less than one third of the median duration) were routinely excluded from the data by the survey host and were therefore excluded from all analyses.

Discrete choice data (excluding the practice and dominance tasks that were not part of the DCE design) was analysed using methods widely used in the literature (e.g.[11,12,13,14]),

using the conditional logit model with a model specification with dummy variables for each severity level of each item interacted with duration (life years), where level zero is the baseline, plus duration. Since the modelling approach assumes that duration is linear and continuous, this was assessed by modelling duration as a categorical variable (using dummy variables for 1,4,7,10 years) and plotting the coefficients to assess linearity [14,15]. For policy, utility weights must be logically consistent, and a fully consistent model was generated using an approach from the literature to merge logically inconsistent adjacent levels within an item (where the severity level of the item worsens but utility increases). Inconsistencies are often observed in the health state valuation literature, where this approach is commonly used (for example [13,14,16-21]).

Utility weights on the full health-dead 1-0 scale were generated using the marginal rate of substitution. The coefficient for each severity level of each item was divided by the duration coefficient, and standard errors were calculated using the Delta method. Health state utilities are generated by summing 1 and the relevant (negative) utility weights.

Model performance was assessed using log likelihood, Rho-squared, and the sign, significance and logical consistency of coefficients (where utility does not increase as QoL worsens). Robustness of the results was assessed by comparing models estimated using subsamples of the participants. Subsamples were generated by excluding participants who reported having diabetes and separately those who may not have understood/engaged (assessed separately using self-reported understanding and difficulty, responses to the dominant practice task and dominance question in main tasks and speeders/slow responders (in the top or bottom 10% of the duration distribution).

Preference heterogeneity was assessed through the estimation of main effects models including interaction effects for each main effect for sex, age, being employed/self-employed, reporting as having diabetes, having self-reported fair/poor health. The sign and significance of model coefficients was examined and compared across models. These analyses are of interest for understanding the data, but are not intended for inclusion in the value set.

A mixed logit model was also estimated (including a consistent model) to account for unobserved preference heterogeneity across respondents and to allow for all parameters within the model to be randomly distributed across respondents [22]. The coefficients are assumed to be normally distributed. Estimation was undertaken in Stata version 18. The anchored coefficients for the conditional logit and mixed logit were compared for the fully consistent models.

3 Results

Health state classification system

Table 1 presents the analyses used to select items for the classification system using traffic light analysis. For the psychological aspects factor items “anxious”, “difficult to concentrate” and “hard to stop thinking about my glucose levels” were selected. Participants responded similarly to “anxious” and “irritable”, and as irritable was less endorsed by patient advisors and experts and had disordered thresholds (not shown), “anxious” was selected. “Concern about my safety” was less relevant in cognitive interviews, and hence was not selected. Whilst “hard to stop thinking about my glucose levels” was not the best performing item in the factor across all selection criteria, it was retained due to its importance in the qualitative interviews used to generate the measure.

For the social aspects factor items “social life was negatively affected” and “inconvenience to others” were selected, as they were deemed to reflect different aspects of the factor. “Experienced interruptions during social activities” was correlated with “social life was negatively affected”, meaning only one item should be retained, and the former was not retained due to its poorer item fit and differential item functioning by diabetes type.

For the physical aspects factor items “do what I wanted to do in my life”, “felt exhausted” and “had to take time out during work activities” were selected. Item “do what I wanted to do in my life” is reverse scored (which can be confusing for participants in the valuation task when there is a mix of ordering across items) and “had to take time out during work activities” may not be relevant for retired participants, which are raised as concerns in the published criteria for classification selection [7]. These items were ultimately selected because there were no

alternative items reflecting these aspects that were important to patients as reflected in the qualitative analyses, cognitive interviews and endorsed by patient advisors. Despite some minor issues in psychometric performance, “exhausted” was selected from its importance in the qualitative development of the measure. Table 2 presents the health state classification system.

Valuation

The first soft launch (n=100) of the online survey recruited a large proportion of people with diabetes, and the information was amended in the recruitment materials to ensure the survey recruited members of the public (both with and without diabetes). These participants are not included in the final sample. A replacement soft launch (n=100) was undertaken, and since no changes were then made before recruiting the remaining 900 participants, these participants are included in the final sample.

Valuation sample: UK general public

The sample was representative of the UK population for age, sex and ethnicity, though with a larger proportion of employed participants and a lower proportion of retired participants (Table 3). The sample had a large proportion of retired participants (24%), and less than half the sample had a degree (47%). The sample had a large proportion of participants with excellent or good health (74%), with mean EQ-5D-5L of 0.77 (s.d. 0.21) (scored using[23]), though a third of participants had their daily activities limited by a health condition. The proportion of participants with diabetes (12%) is larger than found in the UK population aged over 16 (8.6%) [24] though not all people with diabetes had experienced hypoglycaemia in the prior 12 months (in total 7% of the sample had experience in prior 12 months).

Understanding and engagement

Responses to the classification system imagining the hypothetical vignette (supplementary materials, Box 1) showed no commonality in responses (supplementary materials, Table A3). The majority of participants found the survey easy to understand (75.7%) but many found it difficult to answer (48.3%) (Table 1). The majority of participants chose the dominant choice in the practice question (93.7%) and the dominance question in the main DCE tasks (87.9%).

Regression analysis

The initial model had logically inconsistent coefficients for several attributes across the lowest severity levels (see Table 4), meaning that the lower attribute levels were not perceived as being different to the reference level (level 0) for many of the attributes, and hence indicates the attribute is not valued at lower levels of impact, only at higher levels of impact (some of the time (level 3); most or all of the time (level 4)). A fully consistent model was estimated, since the preference weights must generate utilities for the health states in the classification system such that as HRQoL deteriorates the utility value does not increase. The majority of coefficients were significant in the consistent model (19 of 24). The anchored coefficients can be used to generate health state utility values, and are plotted in Figure 2 for comparison of the utility decrements by severity level and across items. For example, health state 20000300 (where this string of numeric values represents levels in each attribute) would be calculated as 1 plus -0.011 plus -0.046, which equals 0.943. The worst state has a value of 1-0.215-0.193-0.198-0.117-0.194-0.138-0.080-0.066 = -0.201. Across all possible states defined by the measure, 0.078% are below zero i.e. worse than dead.

The assumption that duration was continuous and linear was confirmed by a plot of duration coefficients from a model estimated using dummies for duration levels (supplementary materials, Figure A2). Robustness analyses, assessed by estimating the models on subsamples, do not suggest a distinguishably different pattern in responses for participants who may not have understood or engaged with the survey. The sample size of people with diabetes is small (n=121 participants), but does not indicate a clear difference in responses. Assessments of preference heterogeneity, undertaken through estimating models including interaction effects for different health and sociodemographic characteristics, identified that participants aged 18 to 24 years valued the work attribute lower than other participants, and no other significant heterogeneity was observed.

The alternative mixed logit model specification results (presented in Tables A6 and A7) also had logical inconsistencies, and the same attribute levels as the conditional logit model needed to be dropped and merged to reach a consistent model. The worst health state in this model has a value of 0.001. The majority of attribute levels had smaller but similar coefficients in the anchored consistent mixed logit model in comparison to the anchored consistent

conditional logit model. The anchored consistent conditional and mixed logit models produce similar utilities for all possible health states (Figure A3 in Supplementary Materials).

4 Discussion

This study has generated the Hypo-RESOLVE QoL-8D, a preference-weighted measure capturing the HRQoL impact of hypoglycaemia in PwD, with preference-weights from the UK general public. The development of the classification system from the Hypo-RESOLVE QoL benefitted from several sources of information including a large patient survey. The measure can be generated using Hypo-RESOLVE QoL data. The utilities that are generated can be used to generate QALYs for use in economic evaluation.

A recent review of the DCE health state literature identified that there is no single method used to anchor the values on dead, which is required for the preference-weights to be able to be used to generate QALYs [8]. This study chose to include a duration attribute for this purpose, and assumed linear time preference, a common combination in the literature, but this choice may have impacted on the findings. The study benefits from using established techniques in the literature for helping participants to manage the eight attributes, including the use of overlap in the DCE design to ensure a minimum of three items had the same level within the profiles in the choice set and using yellow highlighting to indicate this (see for example [11]).

Study limitations include that participants knew the condition that was being valued, which was necessary due to the inclusion of the glucose level attribute, but which is contrary to the valuation of generic preference-weighted measures. It has been shown that labelling the condition can impact on values in comparison to not labelling the condition [25]. Since hypoglycaemia is experienced by PwD, and with a diabetes prevalence rate of 8.6% in the UK public aged 16 years and above [24], some participants will have had prior knowledge of the condition even though they themselves do not have diabetes, and their views may be impacted by prior conceptions and misconceptions. For 7.5% of the sample they had previously experienced hypoglycaemia, and their direct knowledge and experience will have informed their values.

The sample used to elicit weights was a UK only sample recruited from an existing panel of participants willing to answer surveys. As such, participants may not be fully representative of the UK public, and since no quotas or data was collected on nation we cannot ensure that the sample is geographically representative of the four nations of the UK (England, Wales, Scotland and Northern Ireland). An online panel was used, where engagement and understanding cannot be fully determined. Measures were put in place to limit the impact of participant disengagement and potentially fraudulent activity on the quality of the data. A repeat sociodemographic question was used to screen out respondents who failed to report the same response during data collection. Additionally, duplicate IP address entries were excluded from the useable sample, as a preventative measure to limit attempts to complete the survey multiple times to access the reimbursement points. The data was checked for common patterns of responses that may indicate fraudulent responses, with no fraudulent respondents identified. Those deemed to be speeders (i.e. completed the survey in below one third of the median duration) were excluded from the sample by the survey host.

One item in the classification system "do what I wanted to do in my life" is reverse scored, which means that "none of the time" is the worst level rather than the best level contrary to the other items in the classification. This may be confusing for participants when comparing severity levels across items, but the original item wording was not amended to ensure what is valued reflects what is reported when PwD complete the Hypo-RESOLVE QoL. To aid understanding, participants saw a pop-up box in the DCE task instructions for this item stating "This health aspect is worded differently, meaning that 'none of the time' is the worst level for this aspect (it's the best level for the other aspects)". The item "had to take time out during work activities" may not be relevant for participants not currently working, and almost one quarter of the sample was retired. However, participants were asked in the DCE tasks to select the health state they prefer, and were told to imagine living in the health states, and the examples used in brackets following the item in the DCEs specified that work activities included housework, voluntary work and study. However, the item on work activities may have caused issues with engagement and caused disengagement with the survey for retired participants and other participants who were not currently working if those participants were unable to currently relate to the health states described.

The preference weights recommended for use are using the consistent conditional logit model. The conditional logit model was selected on the grounds of selecting the simpler model that is more interpretable for policy makers and users of the instrument. The added complexity of the mixed logit model had a small impact on the utility decrements and therefore the additional econometric appropriateness of accounting for preference heterogeneity was not deemed as outweighing the requirement for simplicity in this instance. The consistent model is recommended since utility must not increase as HRQoL worsens. There were a number of logical inconsistencies observed in the model including all main effects. These were mostly at levels 1 and 2, the levels with the lowest severity after the reference level (no problems), and is indicative that the attribute level is not perceived as being different to the reference level. In particular, since the DCE tasks involve consideration of trading HRQoL with length of life, this means that participants are not willing to trade life years to avoid the mildest severity levels of these items. This was observed for the items "social life was negatively affected", "feel anxious", "find it difficult to concentrate" and "find it difficult to stop thinking about your glucose levels". The latter two items in particular are items that members of the public may find it difficult to relate to. These condition-specific aspects may not have been fully understood, particularly glucose levels, and this may have impacted on the findings. Robustness analyses estimated models on separate subsamples of people with/without diabetes. Whilst no clear distinguishable pattern was observed, the sample size of participants with diabetes is too low to fully determine how lived experience impacts on the preferences of PwD.

This raises the issue of whose preferences should be used in analyses, and to inform different research questions. For example for use in economic evaluation, where comparisons of cost-effectiveness of treatments are made across conditions, general public preferences are typically recommended for many health technology agencies [26] including the National Institute of Health and Care Excellence [9]. This has the advantage of comparability and consistency of evidence and policy decisions, since public preferences are used regardless, and can be rationalised using voter and tax payer arguments and that the public have no vested interests. However, for assessments of patient health over time and across treatments for diabetes, using samples of PwD, capturing their preferences around the aspects of HRQoL

that are most relevant for them and which they fully understand from their lived experience, may be what is ultimately important. The use of patient values gives patients a voice to express what is important to them. This issue is particularly important in conditions such as this where the general public may not have sufficient understanding of the attributes, for example glucose levels and the impact on HRQoL from monitoring these at multiple timepoints during the day, every day. Further work will explore patient preferences for the Hypo-RESOLVE QoL-8D, to determine whether patient preferences differ to those of the general public. However, for use to inform health technology assessment, the public sample results reported here are recommended since this aligns with requirements by health technology agencies including NICE.

The Hypo-RESOLVE QoL-8D can be used to inform assessments of HRQoL in PwD. For use in economic evaluation, that requires comparisons across conditions and treatments, the use of a condition-specific measure such as Hypo-RESOLVE QoL-8D, can reduce comparability. However, generic preference-weighted measures do not capture all HRQoL impacts that are relevant to PwD in relation to hypoglycaemia [5,6]. Further research will compare the psychometric performance of the Hypo-RESOLVE QoL-8D in comparison to EQ-5D-5L in PwD to inform the future usage of the measure.

References

[1] Chatwin H, Broadley M, Speight J, et al. The impact of hypoglycaemia on quality of life outcomes among adults with type 1 diabetes: A systematic review. Diabetes Res Clin Pract. 2021;174:108752.

[2] Matlock KA, Broadley M, Hendrieckx C, et al. Changes in quality of life following hypoglycaemia in adults with type 2 diabetes: A systematic review of longitudinal studies. Diabet Med. 2022;39(1):e14706.

[3] Alwafi H, Alsharif AA, Wei L, et al. Incidence and prevalence of hypoglycaemia in type 1 and type 2 diabetes individuals: A systematic review and meta-analysis. Diabetes Res Clin Pract. 2020 Dec;170:108522.

[4] Carlton J, Leaviss J, Pouwer F, et al. The suitability of patient-reported outcome measures used to assess the impact of hypoglycaemia on quality of life in people with diabetes: a systematic review using COSMIN methods. Diabetologia. 2021;64(6):1213-25.

[5] Carlton J, Powell P, Rowen D, et al. Producing a preference-based quality of LIFE measure to quantify the impact of HYPOGLYCAEMIA on people living with diabetes: A mixed-methods research protocol. Diabet Med. 2023;40(3):e15007.

[6] Carlton J, Powell PA, Broadley M et al. Development of a new health-related quality of life measure for people living with diabetes who experience hypoglycemia: The Hypo-RESOLVE QoL. Diabetologia. (2024). https://doi.org/10.1007/s00125-024-06182-9.

[7] Peasgood T, Mukuria C, Carlton J, Connell J, Brazier J. (2020). Criteria for item selection for a preference-based measure for use in economic evaluation. Quality of Life Research, 30 (5):1425-1432.

[8] Wang H, Rowen D, Brazier J & Jiang L. Discrete Choice Experiments in Health State Valuation: A Systematic Review of Progress and New Trends. Applied Health Economics and Health Policy, 2023;21(2).

[9] National Institute for Health and Care Excellence. NICE, National Institute of Health and Care Excellence. NICE health technology evaluations: the manual. 2022, http://www.nice.org.uk/process/pmg36 NICE: London.

[10] Lancsar, E. & Louviere, J. (2008) Conducting Discrete Choice Experiments to Inform Healthcare Decision Making, Pharmacoeconomics, 26: 661-677.

[11] Norman R, Viney R, Aaronson NK, et al (2016) Using a discrete choice experiment to value the QLU-C10D: feasibility and sensitivity to presentation format. Quality of Life Research, 25(3), 637-649.

[12] Bansback, N., et al., Using a discrete choice experiment to estimate health state utility values. J Health Econ, 2012. 31(1): p. 306-18.

[13] Norman R, Viney R, Brazier J, et al. Valuing SF-6D Health States Using a Discrete Choice Experiment. Med Decis Making. 2014 Aug;34(6):773-86.

[14] Rowen D, Powell P, Mukuria C, Carlton J, Norman R & Brazier J (2021) Deriving a preference-based measure for people with Duchenne Muscular Dystrophy from the DMD-QoL. Value in Health, 24(10), 1499-1510.

[15] Payne K, Fargher EA, Roberts SA, et al. Valuing pharmacogenetic testing services: A comparison of patients' and health care professionals' preferences. Value Health 2011;14: 121-34.

[16] Rowen DL, Mulhern B, Stevens K & Vermaire E (2018) Estimating a Dutch value set for the paediatric preference-based CHU-9D using a discrete choice experiment with duration. Value in Health, 21(10), 1234-1242.

[17] King, M.T., Viney, R., Simon Pickard, A. et al. Australian Utility Weights for the EORTC QLU-C10D, a Multi-Attribute Utility Instrument Derived from the Cancer-Specific Quality of Life Questionnaire, EORTC QLQ-C30. PharmacoEconomics 36, 225–238 (2018).

[18] Stevens K. Valuation of the Child Health Utility 9D Index. Pharmacoeconomics. 2012 Aug 1;30(8):729-47.

[19] Mulhern, BJ; Bansback, N; Norman, R; Brazier, J. Valuing the SF-6Dv2 Classification System in the United Kingdom Using a Discrete-choice Experiment With Duration. Medical Care 58(6):p 566-573, June 2020.

[20] Pickard AS, Law EH, Jiang R, et al. United States Valuation of EQ-5D-5L Health States Using an International Protocol. Value Health. 2019;22(8): 931-941.

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

[22] Hole, A R. (2008) Modelling heterogeneity in patient's' preferences for the attributes of a general practitioner appointment. Journal of Health Economics, 27 (4): 1078-1094.

[23] Hernández Alava M, Pudney S, Wailoo A. Estimating the Relationship Between EQ-5D-5L and EQ-5D-3L: Results from a UK Population Study. Pharmacoeconomics. 2023 Feb;41(2):199-207.

[24] Public Health England. Diabetes Prevalence Model. Public Health England, 2016https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/612306/Diabetesprevalencemodelbriefing.pdf

[25] Rowen D, Brazier J, Tsuchiya A, Young T & Ibbotson R (2012) It's all in the name, or is it? The impact of labeling on health state values. Med Decis Making, 32(1), 31-40.

[26] Rowen DL, Azzabi Zouraq I, Chevrou-Severac H & van Hout B (2017) International regulations and recommendations for utility data for Health Technology Assessment. PharmacoEconomics, 35(Suppl 1), 11-19.

Table 1: Summary of psychometric and IRT data used to inform item selection for health state classification system

Factor/item	Cognitive interviews	Translatability Assessment	Peasgood et al criteria (all criteria)	Patient advisory committee endorsed item	Advisory group endorsed item	% response at ceiling (level 0)	% response at floor (level 4)	Item independence in Mokken analysis	Item fit in IRT	DIF	Correlation with other items	Correlation with number of hypos (past week)
Psychological aspects												
I felt anxious						21.7	8.3					-0.15
I felt irritable						16.8	10.2					-0.2
I felt concerned about my safety						36.8	3.7					-0.16
I found it difficult to concentrate						12.0	8.1	Violation of local independence at domain level				-0.21
I found it hard to stop thinking about my glucose levels						19.3	12.7					-0.24
Social aspects												
My social life was negatively affected						33.5	3.6				I experienced interruptions during social activities	-0.22
I experienced interruptions during social activities						26.1	2.3		Poor outift and infit at	T1 <> T2 at overall scale	My social life was negatively affected	-0.31

Factor/item	Cognitive interviews	Translatability Assessment	Peasgood et al criteria (all criteria)	Patient advisory committee endorsed item	Advisory group endorsed item	% response at ceiling (level 0)	% response at floor (level 4)	Item independence in Mokken analysis	Item fit in IRT	DIF	Correlation with other items	Correlation with number of hypos (past week)
									domain level			
I was an inconvenience to others						54.2	2.3					-0.16
Physical aspects												
I could do what I wanted to do in my life						2.0	39.7	Violation of local independence at domain level				-0.14
I had to change my plans						24.8	3.3					-0.18
I found some tasks took longer than they should						19.8	5.5					-0.18
I felt exhausted						12.0	16.0					-0.16
I found it difficult to travel to the places I needed to (e.g. driving, walking, cycling, public transport)						34.4	5.5					-0.14
I had to take time out during work						29.1	5.2	Violation of local independenc				-0.21

Factor/item	Cognitive interviews	Translatability Assessment	Peasgood et al criteria (all criteria)	Patient advisory committee endorsed item	Advisory group endorsed item	% response at ceiling (level 0)	% response at floor (level 4)	Item independence in Mokken analysis	Item fit in IRT	DIF	Correlation with other items	Correlation with number of hypos (past week)
activities (e.g. paid work, housework, voluntary work or study)*								e at domain level				
Key	% endorsed as relevant	Issues identified	Number of violations	% endorsed keeping item	% endorsed keeping item	% response	% response	Violations in factor	Problem at actor level	T1 v T2, gender	Number of item correlations ≥ 0.7	Correlation
	75-100%	no issues or minimal translation issues that can be resolved	No violations	75-100%	75-100%	> 5% < 40%	> 5% < 40%	No violations	OK	NO DIF	0	≥ 0.2
	50-74%	potentially problematic	1 violation	50-74%	50-74%	< 5%	< 5%	1 violation	Poor outfit OR infit	N/A	1	< 0.2
	0-49%	likely to be problematic	2+ violations	0-49%	0-49%	> 40%	> 40%	2+ violations	Poor outfit AND infit	Sig DIF R2 > 0.02	2+	N/A

Notes: Cognitive interviews were undertaken with 10 PwD and 14 healthcare professionals in the UK and Germany, see (Carlton et al, in submission) for further details. Psychometric analyses were undertaken using the data (n=1246) from a large UK cross-sectional survey, sample characteristics reported in supplementary materials Table A1, also see (Carlton et al, in submission) for further details. DIF = differential item functioning; IRT = Item response theory.

Table 2: Hypo-RESOLVE-QoL-8D classification system for valuation

Original item wording	Revised item stem for DCE	Level	Severity level description
I could do what I wanted to do in my life	You can do what you want to do in your life *(want)*	0	Most or all of the time
		1	Often
		2	Sometimes
		3	Rarely
		4	None of the time
I felt exhausted	You feel exhausted *(exhausted)*	0	None of the time
		1	Rarely
		2	Sometimes
		3	Often
		4	Most or all of the time
I had to take time out during work activities (e.g. paid work, housework, voluntary work or study)	You have to take time out during work activities (e.g. paid work, housework, voluntary work or study) *(work)*	0	None of the time
		1	Rarely
		2	Sometimes
		3	Often
		4	Most or all of the time
My social life was negatively affected	Your social life is negatively affected *(social)*	0	None of the time
		1	Rarely
		2	Sometimes
		3	Often
		4	Most or all of the time
I was an inconvenience to others	You are an inconvenience to others *(inconvenience)*	0	None of the time
		1	Rarely
		2	Sometimes
		3	Often
		4	Most or all of the time
I felt anxious	You feel anxious *(anxious)*	0	Most or all of the time
		1	Often
		2	Sometimes
		3	Rarely
		4	None of the time
I found it difficult to concentrate	You find it difficult to concentrate *(concentrate)*	0	None of the time
		1	Rarely
		2	Sometimes
		3	Often
		4	Most or all of the time
		0	None of the time
		1	Rarely

Original item wording	Revised item stem for DCE	Level	Severity level description
I found it hard to stop thinking about my glucose levels	You find it hard to stop thinking about your glucose levels *(glucose)*	2	Sometimes
		3	Often
		4	Most or all of the time
Duration	You live for X, then you die *(LY)*	1	1 year
		4	4 years
		7	7 years
		10	10 years

Notes: Item "I could do what I wanted to do in my life" is reverse ordered. The ordering of some of the items has been changed in comparison to the order in which the items appear in the Hypo-RESOLVE-QoL. Labels in brackets and italics show the labels used in Table 4 and Figure 2 to refer to these items.

Table 3: Sample of DCE survey respondents from the UK general public

	Characteristic	Full sample (N=1000), %	UK general population[1], %
Age	Age, mean (s.d.) in years	48.52 (17.48), range 18-89	
	Age 18-24	10.6	10.6
	Age 25-34	16.2	17.0
	Age 35-44	16.5	16.1
	Age 45-64	32.9	32.7
	Age 65+	23.8	23.6
Ethnicity	White	85.5	84.8
	Mixed ethnicity	2.3	
	Asian	6.6	
	Black	2.3	
	African	2.3	
	Other ethnicity	0.6	
	Unknown ethnicity	0.4	
Sex	Male	47.4	48.9
	Female	52.4	51.1
	Prefer not to say	0.2	
Employment status	Employed	56.0	61.7
	Retired	24.2	13.9
	Student	5.2	4.3
	Unemployed	7.2	9.3
	Long term sick	3.4	Not available
	Carer / volunteer	1.6	4.4
	Not seeking work	2.1	Not available
	Other main activity	0.3	6.5
Education	PhD or equivalent doctoral level qualification	3.4	
	Masters or equivalent higher degree level qualification	12.2	
	Bachelors or equivalent first degree level qualification	31.8	
	A-level or equivalent post-secondary level qualification	25.7	
	GCSE or equivalent secondary school qualification	23.5	
	None of the above	3.4	
Marital status	Single	36.1	
	Married/Partner	50.8	
	Separated	1.3	
	Divorced	8.2	
	Widowed	3.4	
	Prefer not to say	0.2	
House ownership	Owns house / mortgage	59.9	

	Rent from a local authority or housing association	14.4	
	Rent from the private sector	22.4	
	Other	2.8	
	Prefer not to say	0.5	
Health	Excellent	9.4	
	Very good	28.8	
	Good	35.5	
	Fair	20.6	
	Poor	5.7	
	Activities are limited by condition a little or a lot	35	
	EQ-5D-5L index scores, mean (s.d.)	0.77 (0.21), range -0.26 to 0.99)	
	EQ-5D-5L full health (state 11111)	26	
Diabetes	Has diabetes	12.3	
	Type 1 diabetes	3.8	
	Type 2 diabetes	8.0	
	Gestational / other diabetes	0.5	
	Experienced hypoglycaemia	7.5	
	Experienced hypoglycaemia in the last 12 months	6.6	
	Insulin user	6.4	
	Duration of survey, mean (s.d.) in seconds	754.44(585.0), range 176.5-5,833.0	
Difficult to answer DCE questions	Very difficult to answer	6.5	
	Quite difficult to answer	41.8	
	Neither difficult nor easy to answer	17.0	
	Fairly easy to answer	24.5	
	Very easy to answer	10.2	
Difficult to understand DCE questions	Very difficult to understand	1.4	
	Quite difficult to understand	6.9	
	Neither difficult nor easy to understand	16	
	Fairly easy to understand	41.2	
	Very easy to understand	34.5	
Practice question	Chose dominant profile outright	93.7	
Dominance question	Chose dominant profile	87.9	

Notes: [1]Statistics for UK for age and sex are from the Office for National Statistics' Mid-Year Population Estimates June 2020. The statistics on employment status are for England in the Census 2011. The census includes persons aged 16 and above whereas this study only surveys persons aged 18 and above.

Table 4: Regression analysis of DCE survey responses

VARIABLES	(1) All main effects	(2) Consistent model	Item	Level	Anchored consistent model Utility decrement and standard errors
			Want	0	0
want1_LY	-0.020*	-0.004		1	-0.011
	(0.100)	(0.661)			0.024
want2_LY	-0.003	-0.004		2	-0.011
	(0.807)	(0.661)			0.024
want3_LY	-0.059***	-0.055***		3	-0.130
	(0.000)	(0.000)			0.023
want4_LY	-0.085***	-0.091***		4	-0.215
	(0.000)	(0.000)			0.023
			Exhausted	0	0
exhausted1_LY	-0.014	-0.012		1	-0.027
	(0.208)	(0.270)			0.024
exhausted2_LY	-0.023**	-0.021**		2	-0.050
	(0.031)	(0.037)			0.023
exhausted3_LY	-0.044***	-0.042***		3	-0.098
	(0.000)	(0.000)			0.021
exhausted4_LY	-0.090***	-0.082***		4	-0.193
	(0.000)	(0.000)			0.022
			Work	0	0
work1_LY	-0.020	-0.020		1	-0.047
	(0.117)	(0.111)			0.029
work2_LY	-0.037***	-0.043***		2	-0.102
	(0.001)	(0.000)			0.023
work3_LY	-0.062***	-0.063***		3	-0.149

VARIABLES	(1) All main effects	(2) Consistent model
	(0.000)	(0.000)
work4_LY	-0.092***	-0.084***
	(0.000)	(0.000)
social1_LY	0.036***	0
	(0.003)	
social2_LY	0.027**	0
	(0.012)	
social3_LY	-0.000	-0.024***
	(0.996)	(0.004)
social4_LY	-0.032***	-0.050***
	(0.003)	(0.000)
inconvenience1_LY	0.009	0
	(0.363)	
inconvenience2_LY	-0.014	-0.020***
	(0.182)	(0.005)
inconvenience3_LY	-0.007	-0.020***
	(0.504)	(0.005)
inconvenience4_LY	-0.084***	-0.083***
	(0.000)	(0.000)
anxious1_LY	0.003	0
	(0.781)	
anxious2_LY	-0.019*	-0.022**
	(0.060)	(0.013)
anxious3_LY	-0.023**	-0.029***

Item	Level	Anchored consistent model Utility decrement and standard errors
		0.022
	4	-0.198
		0.021
Social	0	0
	1	0
	2	0
	3	-0.056
		0.019
	4	-0.117
		0.018
Inconvenience	0	0
	1	0
	2	-0.046
		0.016
	3	-0.046
		0.016
	4	-0.194
		0.018
Anxious	0	0
	1	0
	2	-0.053
		0.021
	3	-0.068

VARIABLES	(1) All main effects	(2) Consistent model
	(0.036)	(0.002)
anxious4_LY	-0.067***	-0.058***
	(0.000)	(0.000)
concentrate1_LY	-0.007	0
	(0.524)	
concentrate2_LY	0.044***	0
	(0.000)	
concentrate3_LY	-0.011	-0.029***
	(0.255)	(0.001)
concentrate4_LY	-0.023**	-0.034***
	(0.020)	(0.000)
glucose1_LY	0.051***	0
	(0.000)	
glucose2_LY	0.030***	0
	(0.005)	
glucose3_LY	0.014	-0.009
	(0.201)	(0.263)
glucose4_LY	-0.002	-0.028***
	(0.850)	(0.001)
LY	0.370***	0.425***
	(0.000)	(0.000)
Observations	18,000	18,000
Log likelihood	-4726	-4750
Rho-squared	0.242	0.239

Item	Level	Anchored consistent model Utility decrement and standard errors
		0.022
	4	-0.138
		0.019
Concentrate	0	0
	1	0
	2	0
	3	-0.068
		0.019
	4	-0.080
		0.019
Glucose	0	0
	1	0
	2	0
	3	-0.021
		0.019
	4	-0.066
		0.018

Notes: P-values are reported in parentheses. *** P-values are significant at 1% level, ** P-values are significant at 5% level, * P-values are significant at 10% level. Standard errors using the Delta method are reported for the utility decrements. Level 0 is the reference level for attributes 'want_*LY' to 'glucose_*LY' and hence

there is no utility decrement for the Hypo-RESOLVE QoL-8D items when they are at level 0. In Model 2 levels where the coefficient is zero occur because the level has been merged with the reference level due to a logical inconsistency in Model 1.

Figure 1 Example screenshot of the practice DCE task in the online survey

	Life A	Life B
You can do what you want to do in your life	Rarely	Rarely
You feel exhausted	Most or all of the time	Rarely
You have to take time out during work activities (e.g. paid work, housework, voluntary work or study)	Most or all of the time	None of the time
Your social life is negatively affected	Rarely	Most or all of the time
You are an inconvenience to others	Most or all of the time	Rarely
You feel anxious	Rarely	Most or all of the time
You find it difficult to concentrate	Often	Often
You find it hard to stop thinking about your glucose levels	Sometimes	Sometimes
How long you live with the health problems described	You live for 1 year, then you die.	You live for 4 years, then you die.
Which do you prefer, Life A or Life B?	○	○
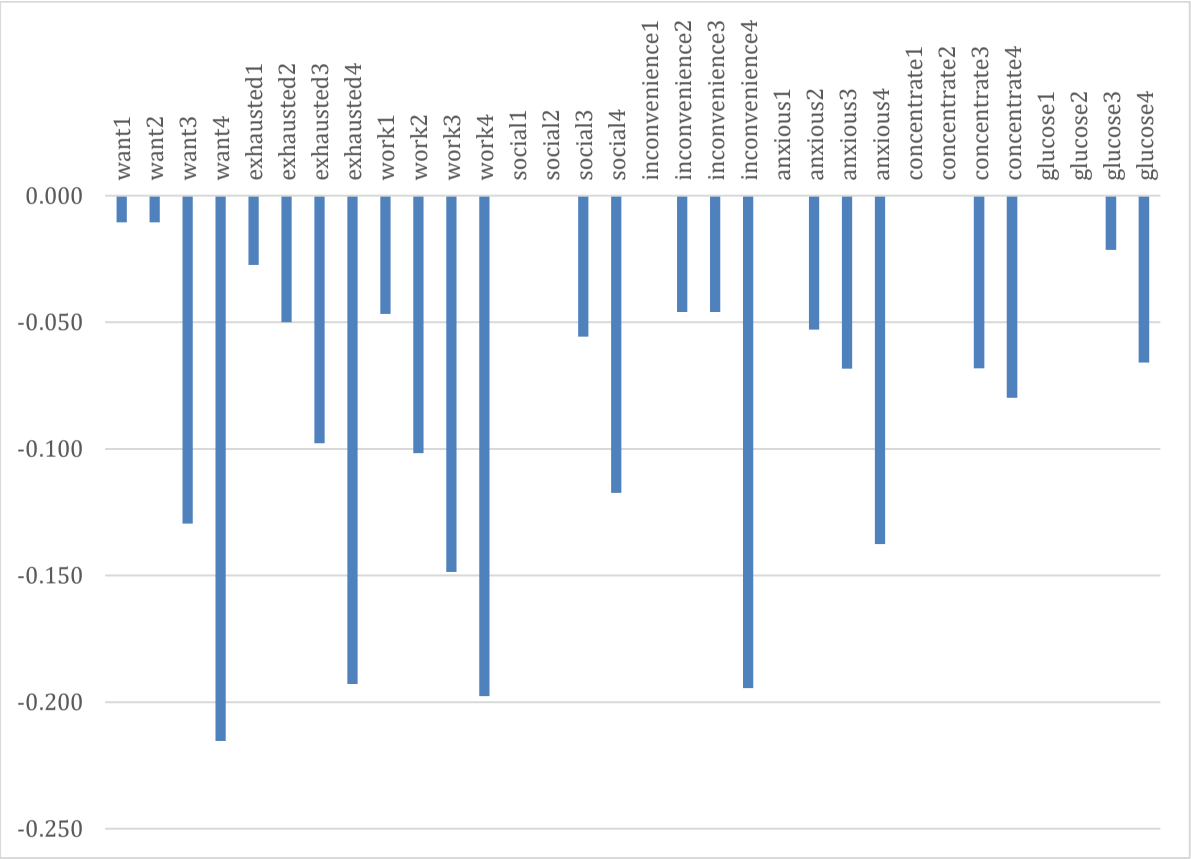
Figure 2 Plot of anchored and consistent coefficients to generate value set