ScienceDirect

Contents lists available at sciencedirect.com
Journal homepage: www.elsevier.com/locate/jval

Brief Report

Test-Retest Reliability of EQ-5D-Y-3L Best-Worst Scaling Choices of Adolescents and Adults

Xiuqin Xiong, MPH, Kim Dalziel, PhD, Li Huang, PhD, Oliver Rivero-Arias, MSc, DPhil

ABSTRACT

Background: There is an increasing interest to obtain adolescents' own health state valuation preferences and to understand how these differ from adult preferences for the same health state. An important question in health state valuation is whether adolescents can report preferences reliably, yet research remains limited.

Objective: This study aims to investigate the test-retest reliability of best-worst scaling (BWS) to elicit adolescent preferences compared with adults.

Methods: Identical BWS tasks designed to value 3-level version of EQ-5D-Y health states were administered online in samples of 1000 adolescents (aged 11-17 years) and 1006 adults in Spain. The valuation survey was repeated approximately 3 days later. We calculated (1) simple percentage agreement and (2) kappa statistic as measures of test-retest reliability. We also compared BWS marginal frequencies and relative attribute importance between baseline and follow-up to explore similarities in the obtained preferences.

Results: We found that both adolescents and adults were able to report their preferences with moderate reliability (kappa: 0.46 for adolescents, 0.46 for adults) for best choices and fair to moderate reliability (kappa: 0.39 for adolescents, 0.41 for adults) for worst choices. No notable difference was observed across years of child age. Higher consistency was observed for best choices than worst in some dimensions for both populations. No significant differences were found in the relative attribute importance between baseline and follow-up in both populations.

Conclusion: Our results suggest that BWS is a reliable elicitation technique to value 3-level version of EQ-5D-Y health states in both adolescents and adults.

Keywords: adolescents, adults, best-worst scaling, EQ-5D-Y-3L, preference, test-retest reliability.

VALUE HEALTH. 2023; 26(1):50–54

Introduction

Obtaining preferences for health states is essential to generate utility values for economic evaluation and inform resource allocation decisions.[1] It is commonly accepted that preferences for adult health states should be elicited from the general adult population.[2] Nevertheless, preferences for child health states have been obtained from both adult and child samples for a variety of reasons including normative considerations and concerns about children's cognitive ability.[2] Mounting evidence suggests that child preferences differ from that of adults.[3,4] Where feasible, directly obtaining child and adolescent preferences is increasingly preferred, due to an awareness of the importance of children's own views about intervention and program outcomes.[4]

Ordinal techniques such as discrete choice experiments (DCEs) and best-worst scaling (BWS) tasks are relatively easy in terms of comprehension and administration.[5] Previous studies have demonstrated that adolescents can provide internally valid responses in DCE and BWS; for example, their responses to dominant choices are rational.[3,6] BWS tasks have been increasingly used in healthcare.[7] Profile case BWS is considered to have a lower cognitive burden than standard DCE[8] and has been used to elicit preferences from adolescents.[9]

There is a research gap related to whether children can report preferences using BWS reliably. The test-retest reliability of a valuation method, also termed repeatability, refers to its the ability to provide consistent utility elicitation over time.[10] Good test-retest reliability is important in reducing measurement error and boosting statistical power.[11] To the best of our knowledge, test-retest reliability of BWS in eliciting preferences for health states has not been examined in samples of adults or children. Beyond health state valuation and in the field of psychology, only 2 studies have explored the test-retest reliability of BWS in the measurement of facial impression and found that BWS is more reliable than Likert ratings.[11,12] This study aims to investigate the test-retest reliability of using BWS to elicit preferences for 3-level

version of EQ-5D-Y (EQ-5D-Y-3L) in adolescents compared with adults.

Methods

Two community-based samples, one of adults and the other of adolescents aged 11 to 17 years, were recruited in Spain via an online panel company in February and March 2016. Full details of the study design can be found elsewhere.[3] The 11 to 17 years age range was chosen because this is when a transitional stage of physical and mental development generally occurs for children[13] and has been used in other preferences elicitation studies.[6,14,15] Briefly, the process began with screening questions about age, sex, and region to facilitate selection of a representative general Spanish population. The first survey section included the self-completed EQ-5D-Y-3L, consisting of 5 dimensions: mobility (MO), looking after myself (SC), usual activities (UA), pain or discomfort (PD), and worried, sad, or unhappy (SW), with 3 levels in each dimension. In the second section, participants completed a profile case BWS experiment where participants were presented with single profiles EQ-5D-Y-3L health states and were asked to indicate the dimension level they considered best and worst (see Appendix 1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.07.007 for example BWS task). Adolescents and adults completed the survey from their own perspective. A full factorial design was adopted dividing the 243 health states into 20 blocks to generate the profiles for the BWS experiment. Each block included 13 BWS tasks except for one block that included 14. The complete experimental design has been published elsewhere.[3] Each participant was randomly allocated to complete one block. The final section of the survey asked participants about their sociodemographic characteristics.

All participants were invited to repeat the survey within a week and were allocated to their original block of the BWS tasks. To analyze the test-retest reliability, we used the sample completing both the baseline and the follow-up surveys. This study received ethics approval from the Ethics Committee at the University Hospital Nuestra Señora de Candelaria in Santa Cruz de Tenerife, Spain (#PI-26/14).

We calculated (1) simple percentage agreement and (2) kappa coefficient[16] as measures of test-retest reliability at the input data level. Each participant was given 13 or 14 BWS choice tasks represented by a single EQ-5D-Y-3L profile.[3] Participants were asked to choose a level in one of the dimensions (MO, SC, UA, PD, and SW) that they considered best and a level of a different dimension that they considered worst. For example, for best choices, each participant had 13 of the 14 observations at each survey, with 1 observation for each choice task (1 health state). The simple percentage agreement between the 2 repeated surveys was calculated dividing the number of matched answers by the number of all answers available. For kappa estimation, we compared the dimension chosen at baseline and follow-up given that the level for each dimension was the same between the 2 surveys. Unweighted kappa was calculated as the choices were categorical variables with 5 categories (MO, SC, UA, PD, and SW).

We estimated simple agreement and unweighted kappa separately for best choice and worst choice given that there is literature indicating that worst choices may be less reliable than best choices.[4,9,17] Strength of agreement for kappa statistic was judged according to the recommended classification from Landis and Koch[16]: $<$ 0.00 indicates poor agreement, 0.00 to 0.20 indicates slight agreement, 0.21 to 0.40 indicates fair agreement, 0.41 to 0.60 indicates moderate agreement, 0.61 to 0.80 indicates substantial agreement, and 0.81 to 1.00 indicates almost perfect agreement.

We also compared BWS marginal frequencies and relative attribute importance (RAI) between baseline and follow-up to explore similarities in the obtained preferences as supplementary measures of test-retest reliability.[18,19] These measures investigated test-retest reliability evaluating the stability of choices and model coefficients over time.[18,19]

The marginal frequency was computed by dividing the number of times a dimension level was chosen as best (or worst) by the number of times that dimension level was available for selection. The Pearson correlation coefficients of the marginal frequencies at the 2 time points were calculated.

The RAI was calculated based on recommended methods.[20] First, we used a conditional logit model and the pooled best-worst data to estimate latent scale values associated with each dimension level, where the choice responses were treated as a binary dependent variable (1 and 0 for being chosen or not respectively).[21] We used a linear additive utility function (see Equation 1) and assumed that the value of the worst choices was the negative of the value for a best choice. Therefore, we used variables dummy coded for each dimension assigning 1 to best and −1 to worst. Level 1 for each EQ-5D-Y-3L dimension was used as reference level. All standard errors are cluster-robust, which allows for arbitrary correlation between the error terms at the individual level.

$$V = \beta_1 MO2 + \beta_2 MO3 + \beta_3 SC2 + \beta_4 SC3 + \beta_5 UA2 + \beta_6 UA3 + \beta_7 PD2 + \beta_8 PD3 + \beta_9 SW2 + \beta_{10} SW3$$

The beta coefficients in Eq. (1) are not directly interpretable and comparable because they represent within-attribute importance referring to the reference levels and must be interpreted in the context of all the other attributes presented to respondents.[20] To aid in their interpretation and comparison across different groups, we used attribute-based normalization to obtain the RAIs, with attribute importance calculated as a proportion of the reference attribute importance.

$$RAI_y = \frac{\beta y}{\beta x}$$

RAI_y is the RAI score for attribute Y. Attribute X is the reference attribute, and in this study, this is "worried, sad, or unhappy" given that it was the least important dimension from the pooled best-worst model in both samples. β_y and β_x are the coefficients for the level 3 of attribute Y and attribute X, respectively. For example, RAI_MO = β_MO3/β_SW3, RAI_SC = β_SC3/β_SW3, with other attributes following the same process.

All analyses were conducted in Stata SE 16.[22]

Results

The baseline survey included 1006 adults and 1000 adolescents, with 470 adults and 323 adolescents completing the repeated survey (average 3.36 days for adults and 2.93 days later for adolescents, detailed frequency distribution in Appendix Table 2.1 in Supplemental Materials found at https://dx.doi.org/10.1016/j.jval.2022.07.007). The sample completing both baseline and follow-up were broadly representative of the general Spanish adult population in terms of gender and age, with slightly higher male and older population (Appendix Table 2.2 in Supplemental Materials found at https://dx.doi.org/10.1016/j.jval.2022.07.007).

The simple percentage agreements were similar between adolescents and adults and were slightly higher for best choices than worst choices (adolescent best, 0.571, worst, 0.513; adult best, 0.570, worst, 0.531). There were no notable differences in

Table 1. Kappa for best choice and worst choice between baseline and follow-up.

Sample	Kappa for best choice (95% CI)	Kappa for worst choice (95% CI)
Adults	0.46 (0.45-0.47)	0.41 (0.40-0.42)
Adolescents	0.46 (0.44-0.47)	0.39 (0.37-0.40)
11-12 years	0.44 (0.41-0.47)	0.39 (0.36-0.42)
13-14 years	0.49 (0.46-0.52)	0.41 (0.38-0.44)
15-17 years	0.45 (0.43-0.47)	0.37 (0.35-0.39)

Note. Landis and Koch[16] proposed the following standards for strength of agreement for the kappa coefficient: ≤ 0 = poor, 0.01-0.2 = slight, 0.21-0.40 = fair, 0.41-0.60 = moderate, 0.61-0.80 = substantial, and 0.80-1 = almost perfect.
CI indicates confidence interval.

agreement among different age groups of adolescents (Appendix Table 3.1 in Supplemental Materials found at https://dx.doi.org/10.1016/j.jval.2022.07.007).

Table 1 presents the estimated kappa for adults and adolescents. For best choice, the kappa was 0.46 for adults and adolescents indicating moderate test-retest reliability. For worst choice, the kappa was 0.41 for adults indicating moderate reliability and 0.39 for adolescents indicating fair reliability. Adolescents had almost the same test-retest reliability in best choice and slightly worse reliability in worst choice than adults. The test-retest reliability of worst choices was worse than best choices in both adults and adolescents. Adolescents as young as 11 to 12 years old had moderate test-retest reliability in best choices (kappa = 0.44) and fair reliability in worst choices (kappa = 0.39). The kappa estimates were generally similar in different age groups. Similarly, the test-retest reliability of worst choices was worse than best choices in all age subgroups.

The sample with longer baseline completion time (minimum total completion time, adult, 2.24 minutes, adolescents, 1.66 minutes; median BWS tasks completion time, adults, 4.7 minutes, adolescents, 4 minutes) had higher absolute agreement and kappa estimates, which indicates better test-retest reliability (details in Appendix Table 3.2 and 3.3 in Supplemental Materials found at https://dx.doi.org/10.1016/j.jval.2022.07.007).

The marginal frequencies between baseline and follow-up were similar for both adolescents and adults, with PD being the most frequently chosen dimension as both best and worst. Baseline and follow-up marginal frequencies were highly correlated (correlation coefficients > 0.9). Correlation coefficients were slightly higher for best choices than worst choices, for both adolescents (correlation for best, 0.996, worst, 0.993) and adults (correlation for best, 0.999, worst, 0.986). Please see Appendix Table 3.4 in Supplemental Materials found at https://dx.doi.org/10.1016/j.jval.2022.07.007 for the detailed marginal frequency results.

The RAI score results of baseline and follow-up were presented in Table 2. Take the RAI score of 1.72 for dimension PD from adolescents at baseline for example; it can be interpreted as respondents consider PD to be 1.72 times more important than SW on an average. There were no significant differences in RAIs between baseline and follow-up for both adolescents and adults.

Discussion

To the best of our knowledge, this is the first study reporting test-retest reliability of BWS for health state preference elicitation by both adolescents and adults. We found that adolescents aged 11 to 17 years were able to self-report preferences for EQ-5D-Y-3L health states with a level of reliability similar to adults. The results suggest that it is reliable to directly elicit preference from adolescents as young as 11 to 12 years old using profile case BWS valuation tasks.

Previous studies have explored test-retest stability of BWS for adult responses.[23,24] Nevertheless, test-retest stability only measures consistency for a few tasks within a survey in comparison

Table 2. RAI scores and differences between baseline and follow-up.

Sample and EQ-5D-Y-3L dimension	Baseline		Follow-up		RAI difference (95% CI)	*P* value
	RAI	SE	RAI	SE		
Adolescent (n = 323)						
Mobility	1.49	0.09	1.34	0.07	0.14 (−0.09 to 0.37)	.229
Looking after myself	1.26	0.08	1.18	0.06	0.08 (−0.12 to 0.28)	.448
Usual activities	1.49	0.09	1.42	0.07	0.06 (−0.17 to 0.29)	.606
Pain or discomfort	1.72	0.10	1.56	0.07	0.17 (−0.07 to 0.41)	.171
Worried, sad, or unhappy*	1.00		1.00			
Adult (n = 470)						
Mobility	1.17	0.05	1.17	0.05	0.00 (−0.14 to 0.13)	.958
Looking after myself	1.09	0.05	1.12	0.04	−0.03 (−0.16 to 0.09)	.613
Usual activities	1.29	0.06	1.31	0.05	−0.02 (−0.16 to 0.13)	.825
Pain or discomfort	1.42	0.06	1.40	0.05	0.02 (−0.13 to 0.17)	.801
Worried, sad, or unhappy*	1.00		1.00			

Note. Coefficients obtained from conditional logistic regression model; SE calculated using the delta method.
CI indicates confidence interval; RAI, relative attribute importance; SE, standard error.
*Worried, sad, or unhappy was the reference attribute.

with test-retest reliability, which provides a complete capture of preference reliability measured through a follow-up survey.[25]

Moderate test-retest reliability of BWS was found for adults for both best and worst choices, with kappa ranging from 0.41 to 0.46. Moderate test-retest reliability was found for adolescents in best choices, and fair reliability was found for adolescents in worst choices, with kappa ranging from 0.39 to 0.46. Compared with evidence reported previously by DCE (see Appendix Table 4.1 in Supplemental Materials found at https://dx.doi.org/10.1016/j.jval.2022.07.007 for detailed kappa results for DCE reported in previous studies in healthcare area), the kappa we reported suggests that BWS has comparable or slightly less reliability in adults.[19,26-32] For example, Xie et al[26] reported kappa of 0.528 for DCE with duration in valuing SF-6Dv2 health states. Gamper et al[27] reported kappa of 0.411 (in France) and 0.605 (in Germany) for valuing Quality of Life Utility-Core 10 Dimensions (QLU-C10D) health states. Bryan et al[19] reported kappa of 0.65 for preference measurement in treatment of knee injuries. To the best of our knowledge, no previous study reported kappa for DCE for adolescents. Nevertheless, caution is required when comparing kappa across different valuation techniques and studies. BWS tasks focus on choices among dimension levels whereas DCE focuses on choices among health states (combined dimension levels). This may lead to BWS being less cognitive demanding for certain population such as adolescents. In addition, the interpretation and comparison of kappa should be exercised with caution because there are other factors that can influence kappa coefficients including prevalence, bias, and nonindependence of ratings.[33]

Besides kappa, the high correlation between baseline and follow-up marginal frequencies of BWS choices and the nonsignificant differences between baseline and follow-up RAI added to the evidence of stability of preferences obtained by BWS in our current study. This is similar with previous DCE studies too.[19,31] For example, Bryan et al[19] reported that the coefficients from models between test and retest were similar and had overlapping 95% confidence intervals.

The interval between the initial survey and follow-up in our study is approximately 3 days, which is short enough to avoid any significant changes in preferences or health status and long enough to minimize memory effects. Previous studies investigating test-retest reliability adopted intervals ranging from several days to several months.[12,19,29] One study compared the test-retest reliability at 2 days and 2 weeks to inform interval selection and found no statistically significant differences in the test-retest reliability for the 2 time intervals, although the study was with adults.[34] The memory effect can be partially tested by comparing the time taken to complete each experiment. If memory effects exist, the time taken for the follow-up experiment is hypothesized to be shorter. Unfortunately, we only collected time to complete the survey at baseline that precluded this analysis, which will be a valuable consideration for future experiments. Nevertheless, we found that people with longer baseline completion time had better reliability. The reason may be that longer completion time signifies careful thinking and thus increased reliability. Another interesting finding is that adolescents took shorter time to complete the BWS tasks than adults. Similar results were seen in previous studies[6,35]; nevertheless, further investigation may be needed as to why this is the case. Given that in our study longer completion time appears to be associated with a higher kappa, we speculate that adults may tend to think more carefully about their choices.

Another factor that may affect the reliability of preferences is preference construction that occurs during an elicitation task.[36] In the retest survey, individual's preferences may be affected by the thoughts provoked by the initial evaluation tasks. This may partly explain why the responses between the initial and follow-up surveys are never 100% the same. Considering this issue, the true reliability may be higher than our estimates.

We found that best choices were slightly more reliable than worst choices. This echoes with previous research findings that worst choices tend to be less consistent.[4,9,17] Therefore, caution should be taken when combining best and worst choice responses. Further research is warranted to investigate the implications and options for managing differences in best and worst values when eliciting health state preferences.

Our study has several strengths. The sample size of our study is relatively large among similar studies evaluating test-retest reliabilities. Second, we included both adolescents and adults, enabling the comparison between them. In addition, the test-retest reliability was assessed at different levels, including choice-set level (eg, simple agreement and kappa) and level of parametric models (eg, RAI estimates), making the conclusion more robust. Nevertheless, our study is not without limitations. A higher percentage of the adult participants than adolescent participants completed the follow-up survey. This may be related to internet accessibility on a day-to-day basis, which would be unlikely to correlate with reliability and preferences. Although the unequal samples may imply more precise estimates for adults, variability around estimates in terms of 95% confidence interval of the kappa suggests that the impact was minimal and unlikely to affect our conclusion. Additionally, the test-retest reliability of BWS may be different in valuing health states of other multiple-attribute utility instruments, and future similar studies using other multiple-attribute utility instruments would be valuable.

Conclusion

Our study adds to the evidence that adolescents as young as 11 to 12 years old can complete BWS tasks reliably.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2022.07.007.

Article and Author Information

Accepted for Publication: July 7, 2022

Published Online: August 13, 2022

doi: https://doi.org/10.1016/j.jval.2022.07.007

Author Affiliations: Health Economics Unit, School of Population and Global Health, The University of Melbourne, Melbourne, Australia (Xiong, Dalziel, Huang); National Perinatal Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, England, UK (Rivero-Arias).

Correspondence: Oliver Rivero-Arias, MSc, DPhil, National Perinatal Epidemiology Unit, Nuffield Department of Population Health, Old Road Campus, University of Oxford, Oxford, England OX3 7LF, United Kingdom. Email: oliver.rivero@npeu.ox.ac.uk

Author Contributions: *Concept and design*: Dalziel, Rivero-Arias
Acquisition of data: Dalziel, Rivero-Arias
Analysis and interpretation of data: Xiong, Dalziel, Huang, Rivero-Arias
Drafting of the manuscript: Xiong, Huang
Critical revision of the paper for important intellectual content: Xiong, Dalziel, Huang, Rivero-Arias
Statistical analysis: Xiong, Rivero-Arias
Provision of study materials or patients: Rivero-Arias
Obtaining funding: Rivero-Arias
Administrative, technical, or logistic support: Dalziel, Rivero-Arias
Supervision: Dalziel, Huang, Rivero-Arias

Conflict of Interest Disclosures: Drs Dalziel and Rivero-Arias reported receiving grants from the EuroQol Research Foundation (developers of the EQ-5D-Y instrument used in this study) outside the submitted work. Dr Rivero-Arias is a member of the EuroQoL Group and a shareholder and director of Maths in Health (MiH), a consultancy company providing expertise on health economics and outcomes research. He also reported receiving grants from Instituto de Salud Carlos III and the European Regional Development Fund during the conduct of the study. No other disclosures were reported.

Funding/Support: This study was supported by a grant from Instituto de Salud Carlos III and the European Regional Development Fund (PI14/00619). Xiong is supported by China Scholarship Council (201906010310).

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

Acknowledgment: The authors thank all adolescent and adult participants in Spain who completed the questionnaire.

REFERENCES

1. Brazier J, Rowen D, Karimi M, Peasgood T, Tsuchiya A, Ratcliffe J. Experience-based utility and own health state valuation for a health state classification system: why and how to do it. *Eur J Health Econ*. 2018;19(6):881–891.
2. Rowen D, Rivero-Arias O, Devlin N, Ratcliffe J. Review of valuation methods of preference-based measures of health for economic evaluation in child and adolescent populations: where are we now and where are we going? *Pharmacoeconomics*. 2020;38(4):325–340.
3. Dalziel K, Catchpool M, Garcia-Lorenzo B, Gorostiza I, Norman R, Rivero-Arias O. Feasibility, validity and differences in adolescent and adult EQ-5D-Y health state valuation in Australia and Spain: an application of best-worst scaling. *Pharmacoeconomics*. 2020;38(5):499–513.
4. Ratcliffe J, Huynh E, Stevens K, Brazier J, Sawyer M, Flynn T. Nothing about us without us? A comparison of adolescent and adult health-state values for the Child Health Utility-9D using profile case best-worst scaling. *Health Econ*. 2016;25(4):486–496.
5. Ali S, Ronaldson S. Ordinal preference elicitation methods in health economics and health services research: using discrete choice experiments and ranking methods. *Br Med Bull*. 2012;103(1):21–44.
6. Mott DJ, Shah KK, Ramos-Goñi JM, Devlin NJ, Rivero-Arias O. Valuing EQ-5D-Y-3L health states using a discrete choice experiment: do adult and adolescent preferences differ? *Med Decis Making*. 2021;41(5):584–596.
7. Cheung KL, Wijnen BFM, Hollin IL, et al. Using best–worst scaling to investigate preferences in health care. *Pharmacoeconomics*. 2016;34(12):1195–1209.
8. Rogers HJ, Marshman Z, Rodd H, Rowen D. Discrete choice experiments or best-worst scaling? A qualitative study to determine the suitability of preference elicitation tasks in research with children and young people. *J Patient Rep Outcomes*. 2021;5(1):26.
9. Ratcliffe J, Huynh E, Chen G, et al. Valuing the Child Health Utility 9D: using profile case best worst scaling methods to develop a new adolescent specific scoring algorithm. *Soc Sci Med*. 2016;157:48–59.
10. van Agt HM, Essink-Bot ML, Krabbe PF, Bonsel GJ. Test-retest reliability of health state valuations collected with the EuroQol questionnaire. *Soc Sci Med*. 1994;39(11):1537–1544.
11. Burton N, Burton M, Fisher C, Peña PG, Rhodes G, Ewing L. Beyond Likert ratings: improving the robustness of developmental research measurement using best–worst scaling. *Behav Res Methods*. 2021;53(5):2273–2279.
12. Burton N, Burton M, Rigby D, Sutherland CAM, Rhodes G. Best-worst scaling improves measurement of first impressions. *Cogn Res Princ Implic*. 2019;4(1):36.
13. Sawyer SM, Azzopardi PS, Wickremarathne D, Patton GC. The age of adolescence. *Lancet Child Adolesc Health*. 2018;2(3):223–228.
14. Ratcliffe J, Flynn T, Terlich F, Stevens K, Brazier J, Sawyer M. Developing adolescent-specific health state values for economic evaluation. *Pharmacoeconomics*. 2012;30(8):713–727.
15. Ratcliffe J, Couzner L, Flynn T, et al. Valuing Child Health Utility 9-D health states with a young adolescent sample: a feasibility study to compare best-worst scaling discrete-choice experiment, standard gamble and time trade-off methods. *Appl Health Econ Health Policy*. 2011;9(1):15–27.
16. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174.
17. Chen G, Xu F, Huynh E, Wang Z, Stevens K, Ratcliffe J. Scoring the Child Health Utility 9-D instrument: estimation of a Chinese child and adolescent-specific tariff. *Qual Life Res*. 2019;28(1):163–176.
18. Liebe U, Meyerhoff J, Hartje V. Test–retest reliability of choice experiments in environmental valuation. *Environ Resour Econ*. 2012;53(3):389–407.
19. Bryan S, Gold L, Sheldon R, Buxton M. Preference measurement using conjoint methods: an empirical investigation of reliability. *Health Econ*. 2000;9(5):385–395.
20. Gonzalez JM. A guide to measuring and interpreting attribute importance. *Patient*. 2019;12(3):287–295.
21. Mühlbacher AC, Kaczynski A, Zweifel P, Johnson FR. Experimental measurement of preferences in health and healthcare using best-worst scaling: an overview. *Health Econ Rev*. 2016;6(1):2.
22. StataCorp. *Stata Statistical Software: Release 16*. College Station, TX: StataCorp LLC; 2019.
23. Krucien N, Watson V, Ryan M. Is best–worst scaling suitable for health state valuation? A comparison with discrete choice experiments. *Health Econ*. 2017;26(12):e1–e16.
24. Xie F, Pullenayegum E, Gaebel K, Oppe M, Krabbe PFM. Eliciting preferences to the EQ-5D-5L health states: discrete choice experiment or multiprofile case of best–worst scaling? *Eur J Health Econ*. 2014;15(3):281–288.
25. Janssen EM, Marshall DA, Hauber AB, Bridges JFP. Improving the quality of discrete-choice experiments in health: how can we assess validity and reliability? *Expert Rev Pharmacoecon Outcomes Res*. 2017;17(6):531–542.
26. Xie S, Wu J, Chen G. Discrete choice experiment with duration versus time trade-off: a comparison of test–retest reliability of health utility elicitation approaches in SF-6Dv2 valuation. *Qual Life Res*. 2022. https://doi.org/10.1007/s11136-022-03159-2.
27. Gamper E-M, Holzner B, King MT, et al. Test-retest reliability of discrete choice experiment for valuations of QLU-C10D health states. *Value Health*. 2018;21(8):958–966.
28. Bijlenga D, Bonsel GJ, Birnie E. Eliciting willingness to pay in obstetrics: comparing a direct and an indirect valuation method for complex health outcomes. *Health Econ*. 2011;20(11):1392–1406.
29. Skjoldborg US, Lauridsen J, Junker P. Reliability of the discrete choice experiment at the input and output level in patients with rheumatoid arthritis. *Value Health*. 2009;12(1):153–158.
30. Bijlenga D, Birnie E, Bonsel GJ. Feasibility, reliability, and validity of three health-state valuation methods using multiple-outcome vignettes on moderate-risk pregnancy at term. *Value Health*. 2009;12(5):821–827.
31. Ryan M, Netten A, Skåtun D, Smith P. Using discrete choice experiments to estimate a preference-based measure of outcome—an application to social care for older people. *J Health Econ*. 2006;25(5):927–944.
32. San Miguel F, Ryan M, Scott A. Are preferences stable? The case of health care. *J Econ Behav Organ*. 2002;48(1):1–14.
33. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther*. 2005;85(3):257–268.
34. Marx RG, Menezes A, Horovitz L, Jones EC, Warren RF. A comparison of two time intervals for test-retest reliability of health status instruments. *J Clin Epidemiol*. 2003;56(8):730–735.
35. Prevolnik Rupel V, Ramos-Goñi JM, Ogorevc M, Kreimeier S, Ludwig K, Greiner W. Comparison of adult and adolescent preferences toward EQ-5D-Y-3L health states. *Value Health*. 2021;24(9):1350–1359.
36. Lloyd AJ. Threats to the estimation of benefit: are preference elicitation methods accurate? *Health Econ*. 2003;12(5):393–402.