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.

Introduction

Obtaining preferences for health states is essential to generate utility values for economic evaluation and inform resource allocation decisions. It is commonly accepted that preferences for adult health states should be elicited from the general adult population. 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. Mounting evidence suggests that child preferences differ from that of adults. 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.

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

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. Good test-retest reliability is important in reducing measurement error and boosting statistical power. 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. This study aims to investigate the test-retest reliability of using BWS to elicite 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. The 11 to 17 years age range was chosen because this is a transitional stage of physical and mental development generally occurs for children and has been used in other preferences elicitation studies. 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](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 that included 13 BWS tasks except for one block that included 14. The marginal frequency was computedit 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. First, we used a conditional logit model and the pooled best-worst data to estimate latent scale values associated with each dimension, where the choice responses were treated as a binary dependent variable (1 and 0 for being chosen or not respectively). 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.

\[
V = \beta_1 MO + \beta_2 MO3 + \beta_3 SC + \beta_4 UA2 + \beta_5 UA3 + \beta_6 PD2 + \beta_7 PD3 + \beta_8 SW2 + \beta_9 SW3
\]

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. To aid in their interpretation and comparison across different groups, we used attribute-based normalization to obtain the RAI, with attribute importance calculated as a proportion of the reference attribute importance.

\[
RAI_Y = \frac{\hat{\beta}_Y}{\hat{\beta}_X}
\]

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. \(\hat{\beta}_Y\) and \(\hat{\beta}_X\) are the coefficients for the level 3 of attribute Y and attribute X, respectively. For example, RAI_MO = \(\hat{\beta}_MO3/\hat{\beta}_SW\), RAI_SC = \(\hat{\beta}_SC3/\hat{\beta}_SW\), with other attributes following the same process.

All analyses were conducted in Stata SE 16.

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](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](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...
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.

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

Note. Landis and Koch16 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.

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.

<table>
<thead>
<tr>
<th>Sample and EQ-SD-Y-3L dimension</th>
<th>Baseline</th>
<th>Follow-up</th>
<th>RAI difference (95% CI)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RAI</td>
<td>SE</td>
<td>RAI</td>
<td>SE</td>
</tr>
<tr>
<td>Adolescent (n = 323)</td>
<td>1.49</td>
<td>0.09</td>
<td>1.34</td>
<td>0.07</td>
</tr>
<tr>
<td>Mobility</td>
<td></td>
<td></td>
<td>1.18</td>
<td>0.06</td>
</tr>
<tr>
<td>Looking after myself</td>
<td>1.26</td>
<td>0.08</td>
<td>1.42</td>
<td>0.07</td>
</tr>
<tr>
<td>Usual activities</td>
<td>1.49</td>
<td>0.09</td>
<td>1.56</td>
<td>0.07</td>
</tr>
<tr>
<td>Pain or discomfort</td>
<td>1.72</td>
<td>0.10</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Worried, sad, or unhappy*</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adult (n = 470)</td>
<td>1.17</td>
<td>0.05</td>
<td>1.17</td>
<td>0.05</td>
</tr>
<tr>
<td>Mobility</td>
<td>1.09</td>
<td>0.05</td>
<td>1.12</td>
<td>0.04</td>
</tr>
<tr>
<td>Looking after myself</td>
<td>1.29</td>
<td>0.06</td>
<td>1.31</td>
<td>0.05</td>
</tr>
<tr>
<td>Usual activities</td>
<td>1.42</td>
<td>0.06</td>
<td>1.40</td>
<td>0.05</td>
</tr>
<tr>
<td>Pain or discomfort</td>
<td></td>
<td></td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Worried, sad, or unhappy*</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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. 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 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, For example, Xie et al26 reported kappa of 0.528 for DCE with duration in valuing SF-6Dv2 health states. Camper et al27 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 al19 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.

Besides kappa, the high correlation between baseline and follow-up marginal frequencies of BWS choices and the non-significant 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. For example, Bryan et al19 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. 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. 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; 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. 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. 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: xxxx
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

Supplemental Material
Conflict of Interest Disclosures: Drs Dalziel and Rivera-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 Rivera-Arias is a member of the EuroQol Group and a shareholder and director of Mathis 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