Do DCE and profile-case BWS yield comparable ICECAP-A tariffs? Evidence from a French valuation study

Zexuan WANG *[1], Jonathan SICSIC[1], and Pauline CHAUVIN[2]

[1]Université Paris Cité, Liraes and Chaire AgingUP!, F-75006, Paris

[2]Université Paris Cité, Liraes, F-75006, Paris and Sorbonne University Abu Dhabi, SAFIR, Abu Dhabi, United Arab Emirates

Keywords: Best–Worst Scaling; Discrete Choice Experiments; ICECAP-A; Capability Approach; Well-Being.

Summary

Research question and motivation

An increasing body of research suggests that real well-being reflects people's capabilities, that is their real freedom to be and to achieve. Economic and social care programme evaluations usually rely on generic health-related quality-of-life instruments, which primarily focus on individuals' mental and physical health. While such measures are valuable, they offer a limited account of well-being in contexts where social participation, autonomy, dignity, and control over daily life are central. The capability approach (Sen et al., 1993) argues that an individual's true well-being depends on their real freedom to be and to achieve, and conceptualises well-being in terms of individuals' substantive freedoms, i.e. the real opportunities they have to be and do what they value.

*Corresponding author, email: zexuan.wang@u-pariscite.fr

The ICEpop Capability Measure for Adults (ICECAP-A) captures this concept by measuring five dimensions of capability well-being. Among existing capability-based instruments, ICECAP-A has been most widely adopted, with empirical evidence suggesting strong psychometric performance in capturing capability-related aspects of well-being across diverse populations. To use capability measures in economic evaluation, ICECAP-A capability states must be assigned quantitative values through population-based tariff sets. Existing studies have derived ICECAP-A tariffs using best–worst scaling case 2 (BWS-2), an approach designed to elicit preferences over capability-level combinations (Flynn et al., 2015; Farkas et al., 2021; Rohrbach et al., 2022; Wang et al., 2026). However, it remains unclear whether these tariffs adequately capture the trade-offs individuals make across different capability dimensions.

Discrete choice experiments (DCEs) are also widely used to elicit individuals' preferences in health research. A substantial body of literature has compared preferences elicited using BWS-2 and DCEs across a range of contexts — vaccination, social care programmes, and priority setting (Whitty and Oliveira Gonçalves, 2018). These comparisons highlight that methodological choices may influence both the magnitude and structure of estimated preferences, raising important questions about the robustness and interpretability of capability valuation exercises. In this context, our study provides new evidence on the robustness of ICECAP-A valuation to methodological choice, with direct implications for economic evaluation and social care policy.

Contribution to the literature

To our knowledge, this is the first study to compare DCE and BWS-2 approaches within the ICECAP-A framework. Best–Worst Scaling Case 2 (BWS-2) is a preference elicitation method in which respondents are presented with a single multi-attribute profile and asked to identify the best and worst attribute level within that profile. Unlike discrete choice experiments, no trade-off across profiles is required: the task focuses entirely on the relative importance of levels within a given state. This design naturally enforces monotonicity (i.e., a higher level is always preferred to a lower one) making BWS-2 particularly suited to normative valuation exercises. In a discrete choice experiment (DCE), respondents are asked to choose with option they would prefer (or choose) among several multi-attribute options (profiles), which allow analyzing trade-offs among attributes in a random utility maximization framework.

Such a comparison of methods is highly relevant for policymakers. Prior studies suggest that resources should be prioritised for individuals who fall

below a sufficient capability threshold, those with at least one capability dimension at level 1 or 2 (Mitchell et al., 2015; Kinghorn, 2019). Differences between tariff sets derived from alternative preference elicitation methods may lead to different assessments of whether particular capability improvements are welfare enhancing. Our paper therefore raises fundamental questions regarding public interventions aimed at improving capability well-being.

Previous studies show mixed results when comparing DCEs and BWS-2: some report broad consistency (Potoglou et al., 2011; Netten et al., 2012; Zhang et al., 2015; Coast et al., 2016), others conflicting results (Severin et al., 2013; Whitty et al., 2014; Krucien et al., 2017; Chang et al., 2025), or consistency only under extreme scenarios (van Dijk et al., 2016). Our study contributes to both the analysis of capability well-being and the growing literature comparing DCE and BWS-2 approaches, while also allowing us to compare decision time and perceived difficulty between the two task formats.

Methods

We conducted an online experiment using a representative sample of the French adult population (N = 3,546). The sample was recruited by a market research agency (Panelabs MIS Group) in July 2025 using the quota method to ensure adequate representation of specific subgroups within the French population. Quotas were defined based on age, gender, socio-professional category, and geographic region. Data were collected through an online, self-administered survey, entirely anonymously. Ethical approval was obtained from the Comité éthique de Recherche of Université Paris Cité (IRB number: IORG0010120-2025-32). The ICECAP-A instrument was translated into French based on the validated version by Clément et al. (2019).

The survey included two preference elicitation modules, the DCE and the BWS-2. We employed a within-subject design; each participant was randomly assigned to start with either the DCE or the BWS-2. The BWS-2 design was based on an orthogonal main-effects plan (OMEP) generating 32 profiles (16 original + 16 foldover), divided into four blocks of eight tasks plus one repeated task. In each task, participants selected the best and the worst level-based capability dimension within a presented capability state. The DCE module used a D-efficiency design with 32 binary choice tasks (16 + 16 foldover), divided into four blocks of eight tasks plus one repeated task. In each task, participants chose the most preferred capability state between two alternatives.

Preferences were estimated using random-coefficient models — a Mixed BWS model and a Mixed Logit model — following the approach recommended by Whitty and Oliveira Gonçalves (2018). ICECAP-A tariff sets

were derived by applying a linear transformation constrained so that the sum of Level-1 values equals 0 and the sum of Level-4 values equals 1 across all capability dimensions. Concordance between methods was assessed using Spearman rank correlation on rescaled parameters (divided by the maximum estimated parameter in each model). Non-parametric analyses compared decision times, self-reported difficulty, and response consistency across the two formats.

Results

Our first contribution concerns the properties of the resulting tariffs. Here our findings contrast with those of Krucien et al. (2017), who compared the two formats for the Glaucoma Utility Index and favoured the DCE, primarily because it produced monotonic level parameters whereas BWS did not. In our study, BWS yields monotonic parameters across all five ICECAP-A dimensions, whereas the DCE does so only for Attachment. This divergence may reflect differences in the instrument (e.g., a condition-specific health index versus a generic capability measure), in the experimental designs and in the samples. The relative performance of the two formats therefore does not appear to be a general property of the elicitation task but depends on the context of application.

Level	Stability	Attachment	Achievement	Autonomy	Enjoyment
BWS					
1	0.000	0.000	0.000	0.000	0.000
2	0.116	0.090	0.099	0.127	0.076
3	0.174	0.195	0.156	0.188	0.158
4	0.193	0.221	0.174	0.196	0.217
DCE					
1	0.000	0.000	0.000	0.000	0.000
2	0.159	0.207	0.150	0.220	0.152
3	0.130	0.275	0.178	0.196	0.295
4	0.108	0.345	0.090	0.228	0.229

Table 1: Tariffs.

Note: The sum of tariffs for the capability state "11111" equals 0, representing "no capability", while the sum of tariffs for the capability state "44444" equals 1, representing "full capability".

Values above 1 can arise in DCE-based tariffs. The normalisation of Flynn et al. (2015) fixes states 11111 and 44444 at 0 and 1, but it keeps every other state within these bounds only if level 4 receives the highest

parameter in each dimension. In our DCE, an intermediate level receives a larger parameter than level 4 in four of the five non-monotonic dimensions (Stability, Autonomy, Achievement and Enjoyment), so that some states score above full capability. This should not be read as evidence that respondents would choose a dominated state over full capability. No task in our D-efficient design pairs a state with one that dominates it, and scores above 1 are model-based predictions. They nonetheless matter in practice, because they complicate the ranking of states and the interpretation of results for policy. This underlines the importance of checking monotonicity and boundedness when deriving preference-based tariff sets. In our application, profile-case BWS is therefore better suited to deriving tariffs with these properties.

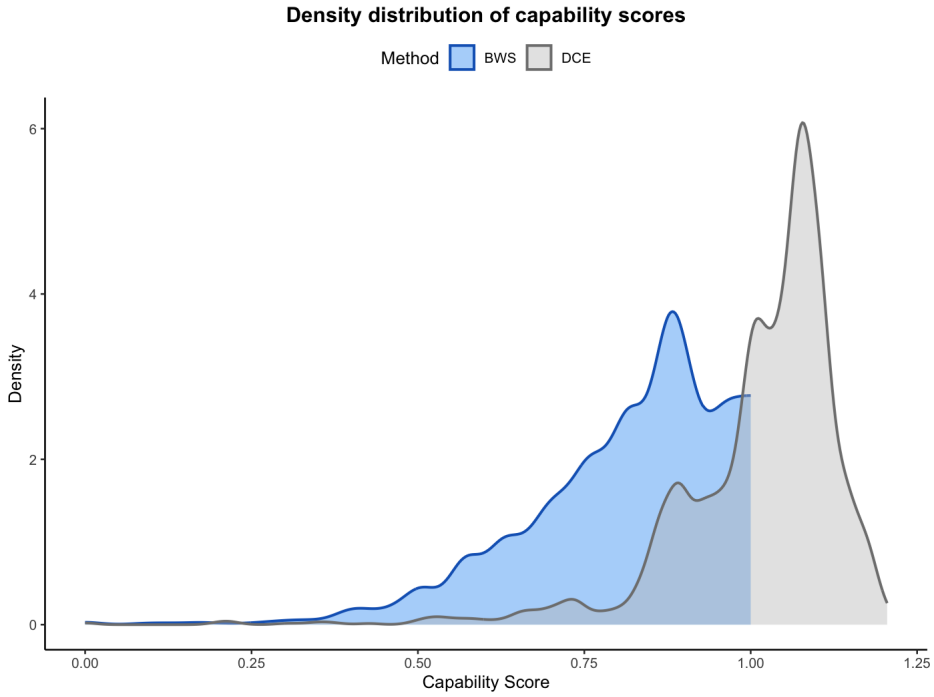


Figure 1: Capability scores.
Note: Density distribution of capability scores.

Our second contribution concerns the broader comparability of DCE and profile-case BWS. Consistent with previous studies reporting differences between the two approaches (Potoglou et al., 2011; Netten et al., 2012; Krucien et al., 2017), we find limited concordance between estimated preference structures, even after rescaling the parameters (Spearman $\rho = 0.42$). Nevertheless, both approaches identify Attachment as the most valued capability

dimension. More importantly, the mixed conditional logit and mixed BWS models achieve similar cross-format predictive performance. When predicting DCE choices, the two models obtain hit rates of 0.693 and 0.688, respectively, despite differences in their estimated parameters and predicted utility values. This apparent discrepancy reflects the fact that correctly predicting the preferred alternative does not require models to assign identical utility differences or recover equivalent preference weights. Consequently, similar predictive performance does not necessarily imply equivalent value sets. This distinction is particularly important when preference estimates are subsequently used to calculate capability scores and inform economic evaluations.

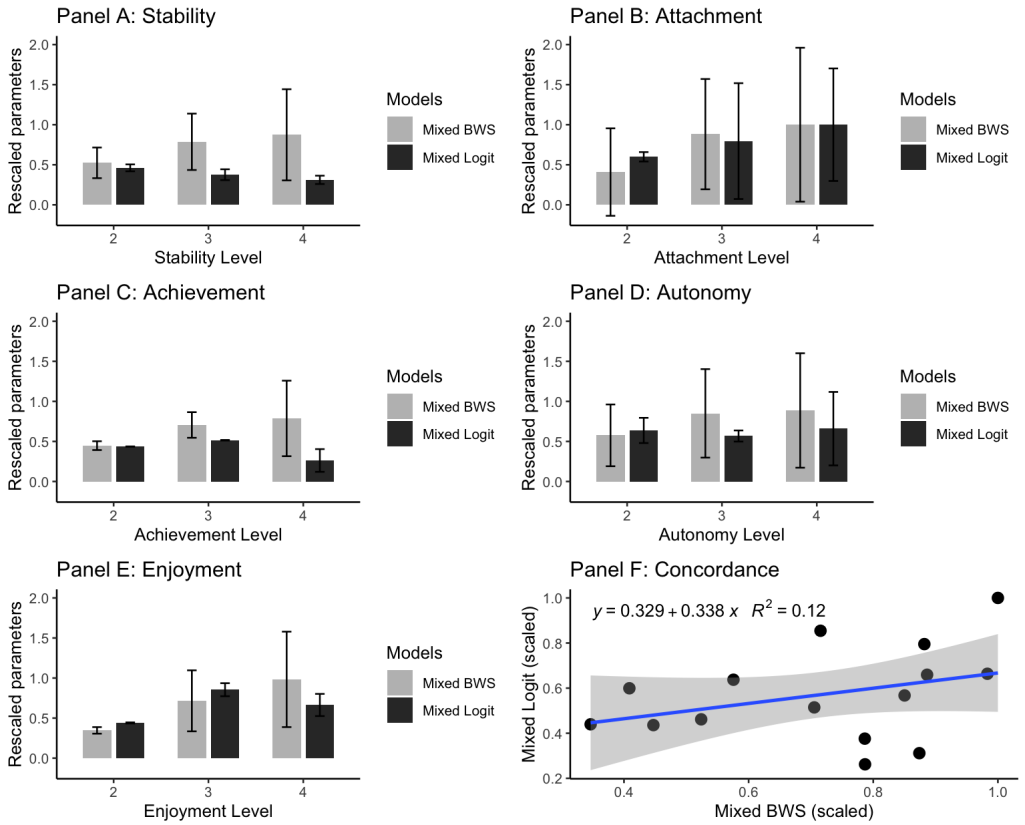


Figure 2: Concordance.

Note: The histograms represent the rescaled parameters. The whiskers indicate the upper and lower bounds of the rescaled parameters.The scatter plot shows the linear fit between the rescaled parameters derived from the mixed BWS and Logit models.

The two approaches also differ in their response burden and consistency.

DCE tasks were completed more quickly and were perceived as slightly easier, although most respondents reported no difference in difficulty between formats. Regarding response consistency, our findings are mixed: DCE shows higher raw consistency in the repeated task, whereas best choices in BWS are more consistent once chance levels are taken into account. These findings contrast with those of Krucien et al. (2017), who reported greater response consistency for DCE than for BWS, while finding similar performance in terms of completeness, defined as respondents' ability to rank all options and make a choice. Overall, our results do not establish a clear advantage for either approach in terms of response consistency.

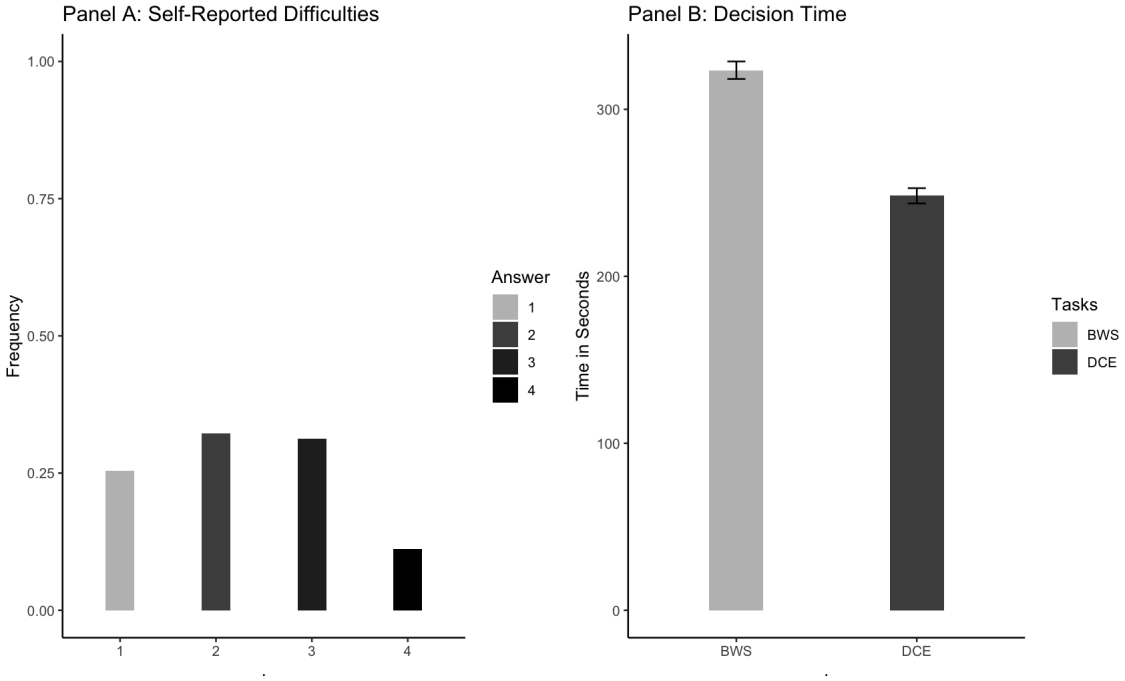


Figure 3: Profile-case BWS versus DCE.
Note: Panel A: 1 - BWS is easier; 2 - DCE is easier; 3 - Neither format is easier than the other; 4 - Both formats are equally difficult. Panel B: Decision time in seconds represents the total decision time (nine choice tasks per participant) spent on the BWS and DCE tasks. The bars indicate standard errors.

References

Chang, E., Y. Jia, X. Zhu, L. Wang, Y. Yan, K. Liu, and W. Huang (2025). Covid-19 vaccine preferences in china: A comparison of discrete choice experiment and profile case best–worst scaling. *Pharmacoeconomics-Open 9*(3), 399–413.

Clément, V., R. Trouillet, T. Blayac, and G. Ninot (2019). First results on the french version of the icecap-a questionnaire. *Journal de Gestion et d'Économie de la Santé 5*(5), 385–398.

Coast, J., E. Huynh, P. Kinghorn, and T. Flynn (2016). Complex valuation: applying ideas from the complex intervention framework to valuation of a new measure for end-of-life care. *Pharmacoeconomics 34*(5), 499–508.

Farkas, M., E. Huynh, L. Gulácsi, Z. Zrubka, Á. Dobos, L. Kovács, P. Baji, and M. Péntek (2021). Development of population tariffs for the icecap-a instrument for hungary and their comparison with the uk tariffs. *Value in Health 24*(12), 1845–1852.

Flynn, T. N., E. Huynh, T. J. Peters, H. Al-Janabi, S. Clemens, A. Moody, and J. Coast (2015). Scoring the icecap-a capability instrument. estimation of a uk general population tariff. *Health Economics 24*(3), 258–269.

Krucien, N., V. Watson, and M. Ryan (2017). Is best–worst scaling suitable for health state valuation? a comparison with discrete choice experiments. *Health Economics 26*(12), e1–e16.

Netten, A., P. Burge, J. Malley, D. Potoglou, A.-M. Towers, J. Brazier, T. Flynn, and J. Forder (2012). Outcomes of social care for adults: developing a preference-weighted measure. *Health Technology Assessment 16*(16), 1–166.

Potoglou, D., P. Burge, T. Flynn, A. Netten, J. Malley, J. Forder, and J. E. Brazier (2011). Best–worst scaling vs. discrete choice experiments: an empirical comparison using social care data. *Social Science & Medicine 72*(10), 1717–1727.

Rohrbach, P. J., A. E. Dingemans, C. G. Groothuis-Oudshoorn, J. A. Van Til, B. A. Essers, E. F. Van Furth, M. E. Van den Akker-Van, et al. (2022). The icepop capability measure for adults instrument for capabilities: development of a tariff for the dutch general population. *Value in Health 25*(1), 125–132.

Sen, A. et al. (1993). Capability and well-being. *The Quality of Life 30*(1), 270–293.

Severin, F., J. Schmidtke, A. Mühlbacher, and W. H. Rogowski (2013). Eliciting preferences for priority setting in genetic testing: a pilot study comparing best-worst scaling and discrete-choice experiments. *European Journal of Human Genetics 21*(11), 1202–1208.

van Dijk, J. D., C. G. Groothuis-Oudshoorn, D. A. Marshall, and M. J. IJzerman (2016). An empirical comparison of discrete choice experiment and best-worst scaling to estimate stakeholders' risk tolerance for hip replacement surgery. *Value in Health 19*(4), 316–322.

Wang, Z., J. Sicsic, and P. Chauvin (2026). Developing french population tariffs for the icecap-a using best–worst scaling. *Pharmacoeconomics 44*(8), 963–974.

Whitty, J. A. and A. S. Oliveira Gonçalves (2018). A systematic review comparing the acceptability, validity and concordance of discrete choice experiments and best–worst scaling for eliciting preferences in healthcare. *The Patient-Patient-Centered Outcomes Research 11*, 301–317.

Whitty, J. A., J. Ratcliffe, G. Chen, and P. A. Scuffham (2014). Australian public preferences for the funding of new health technologies: a comparison of discrete choice and profile case best-worst scaling methods. *Medical Decision Making 34*(5), 638–654.

Zhang, J., F. R. Johnson, A. F. Mohamed, and A. B. Hauber (2015). Too many attributes: A test of the validity of combining discrete-choice and best–worst scaling data. *Journal of Choice Modelling 15*, 1–13.