Supplemental Materials

A. Sample size calculations

The sample size was based on the power calculations with the following assumptions. In the Polish EQ-5D-5L valuation, for worse than dead (WTD) observations only, the observed standard deviations (SDs) of the level sum score (LSS) and utility amounted to 4.2 and 0.27, respectively, while the Pearson linear correlation coefficient amounted to -0.06 [2]. Detecting an increase of (the absolute value of) correlation by 0.2, i.e., the change of the correlation to -0.26, with a power 75% using a significance level of 0.1 (increased to equate more the probabilities of type I and type II errors) requires 168 observations (i.e., severity-negative utility pairs) per arm. Assuming the proportion of WTD observations about 15% (based on Polish data) and 10 states per person, this number of observations requires approx. 110 respondents per arm.

Eventually, in the project we interviewed 120 respondents per arm, there were 12 tasks per respondent, no observations were lost, and the proportion of WTD observations was larger ($> 20\%$ in all the arms, 46.5% in arm B), which increased the power. In consequence, the actually observed difference of 0.2 in Pearson correlation coefficient between arms A and B was significant with p-value equal to 0.001.

B. Health state selection

The health states were selected as follows. We started with 10 blocks of 10 health states each, exactly as in EQ-VT. Each such block contains a mix of mild, moderate, and severe states; 55555 is in each block. From each block, 4 states with the largest LSS were picked (except for 55555; there was just one tie, which was randomly broken) and a pool of 40 states was created. Each of the 10 blocks was duplicated, and in each copy two empty slots for health states were created. The slots were filled in in the following way. From the pool of states, health states were one by one manually assigned to the blocks while trying to keep the blocks as heterogeneous as possible (technically speaking, by assigning a state to a block for which the minimal Manhattan distance to states already in this block was as large as possible).

The resulting 20 blocks of states were put in a queue. Each block was assigned to one interviewer, and it was used for four consecutive interviews for each of the arms A–D in random order. Such a randomization procedure was used, to make sure that each block was equally used for all arms.

C. Numeracy testing questions

We used two sets of questions to measure the numeracy skills. The first three questions were based on [3]:

1. 'What is the most likely number of heads to be obtained in 1000 coin flips using a fair coin?'
2. 'Convert 1% to a proportion, i.e. type how many people out of 1000 it is.'
3. 'Convert the proportion "1 out of 1000" to a percentage.'

with the correct answers being, respectively: 500, 10, 0.1%. We decided to slightly modify the first question. In [3], the authors simply asked for the number of heads in 1000 coin flips. We deemed that because the outcome is a random variable, and effectively any answer between 0 and 1000 is possible with non-zero probability, the questions needs to be asked in a more precise way. In [3] the authors accepted answers from the symmetrical range encompassing 95% of the probability mass, which seems arbitrary.

The subsequent questions were based on the Berlin Numeracy test with the specific questions being selected in an adaptive way based on previous answers [1]. The details can be found here: `http://www.riskliteracy.org/files/BNT%20Versions.pdf` (last access 15$^{\text{th}}$ Nov 2023).

D. The distribution of values per arm and per interviewer

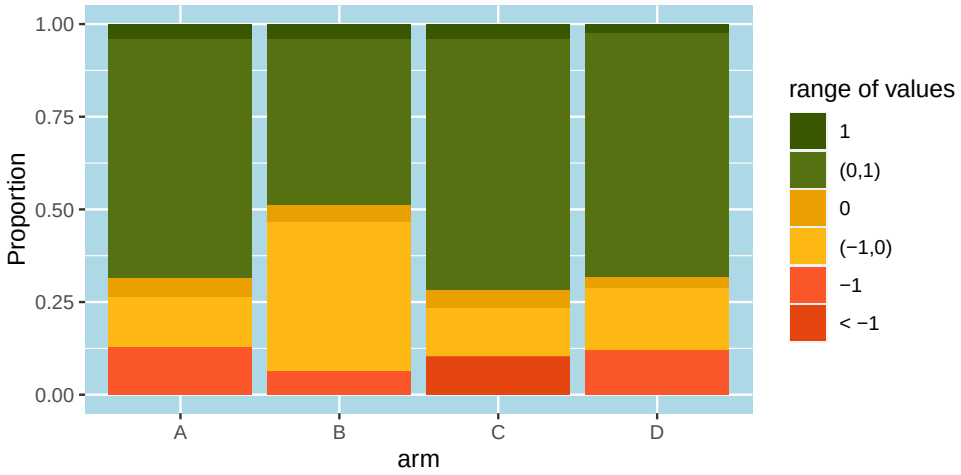

Figure S1: The utility ranges obtained for study arms.

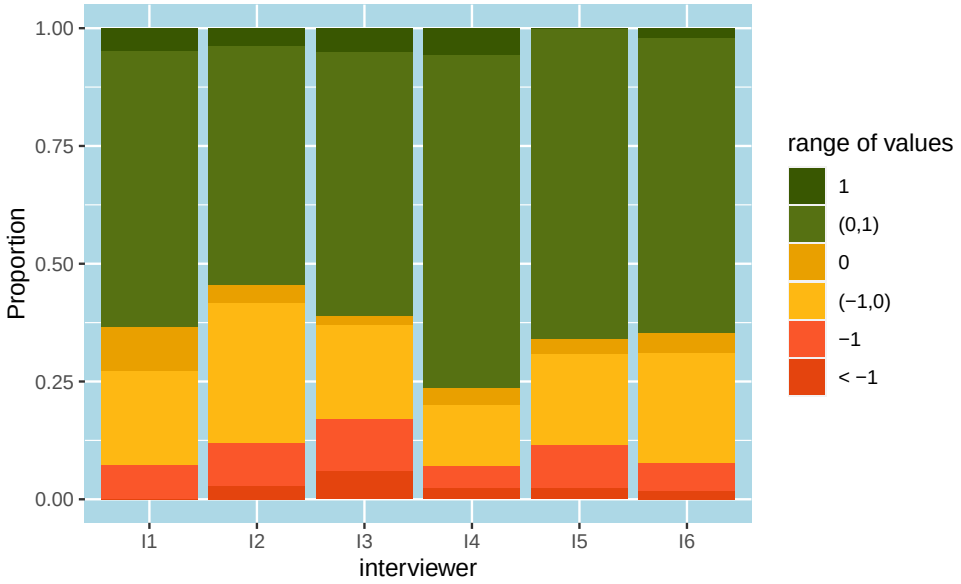

Figure S2: The utility ranges obtained for individual interviewers (all arms pooled).

E. The comparison of value sets produced using data from study arms

We built econometric models, as typically done in valuation studies, to extrapolate the results obtained in the sample to all 3125 EQ-5D-5L health states, split by arm. We used the tobit regression with censoring (at -1 for arms A, B, and D and at -10 for arm C, weighted by observed SD to account for heteroscedasticity). Incremental dummies were used for individual levels, and the dummies were dropped in case of non-intuitive sign of estimated coefficient. We deemed the sample size to be insufficient to interpret individual coefficients per dimensions/level between the arms. Hence, in Table S1, we report the estimated utilities of six states, to allow for comparisons of dimension importance, overall range of utilities, relative importance of levels 2–5, and the proportion of WTD states. The illustration of how the estimated value sets compare between arms is presented using scatter plots in Figs. S3, S4, and S5. Arms B and C lower the index value for the 55555 health state (the pits state) and they increase the proportion of EQ-5D-5L states that have negative index values. From the individual dimensions perspective, the impact is largest for PD in arm C: worsening this single dimension to level 5 reduces the state value by > 0.8. Looking at the relative importance of levels, arm B makes levels 2–4 relative to level 5 more important, while arm C impacts the results in the opposite way.

Characteristic	Arm A	Arm B	Arm C	Arm D
MO level 5 disutility	0.253	0.315	0.292	0.304
SC level 5 disutility	0.306	0.258	0.327	0.276
UA level 5 disutility	0.233	0.287	0.378	0.300
PD level 5 disutility	0.408	0.456	0.837	0.420
AD level 5 disutility	0.330	0.462	0.319	0.349
$u(55555)$	-0.479	-0.588	-0.931	-0.520
% of states with $u < 0$	18.2%	38.6%	31.3%	21.6%
22222 to 55555 relative disutility	23.7%	29.6%	11.1%	28.8%
33333 to 55555 relative disutility	42.2%	59.6%	22.1%	38.0%
44444 to 55555 relative disutility	86.8%	95.5%	76.4%	88.4%

Table S1: Characteristics of per-arm value sets.

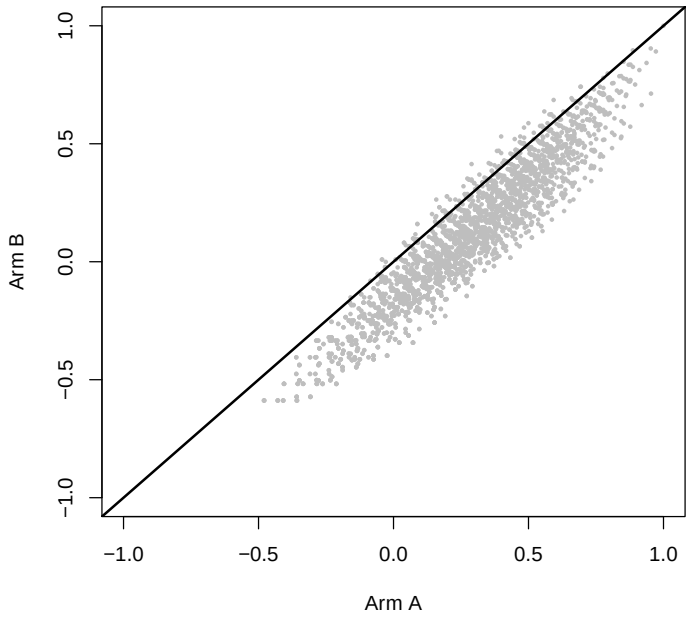

Figure S3: Scatter plot comparing the value sets obtained based on data from arm A vs. those from arm B.

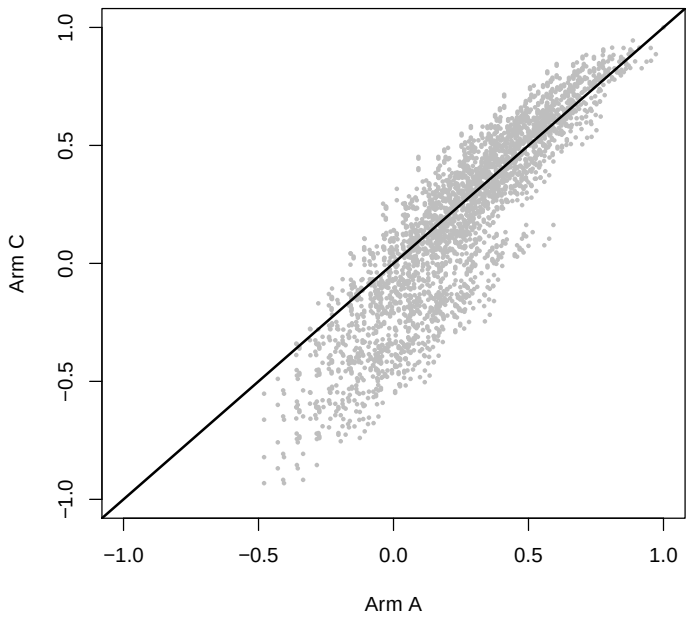

Figure S4: Scatter plot comparing the value sets obtained based on data from arm A vs. those from arm C.

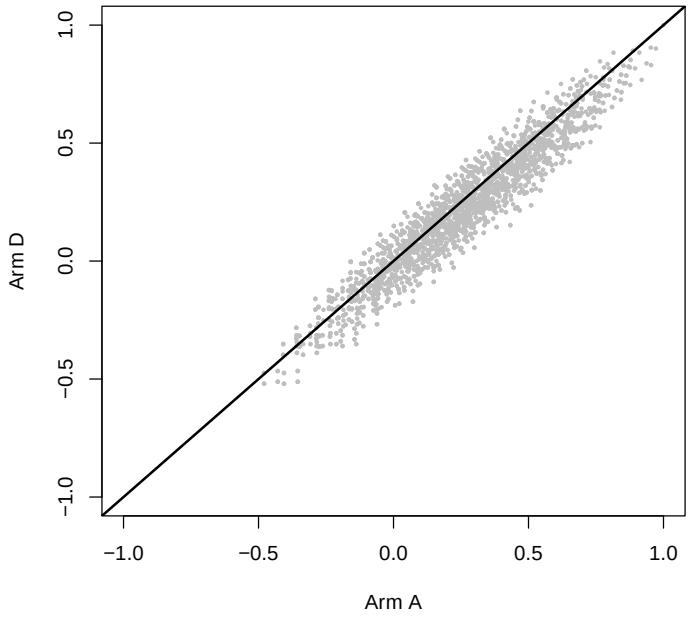

Figure S5: Scatter plot comparing the value sets obtained based on data from arm A vs. those from arm D.

F. Individual-level regression of utility by level sum score for worse-than-dead states

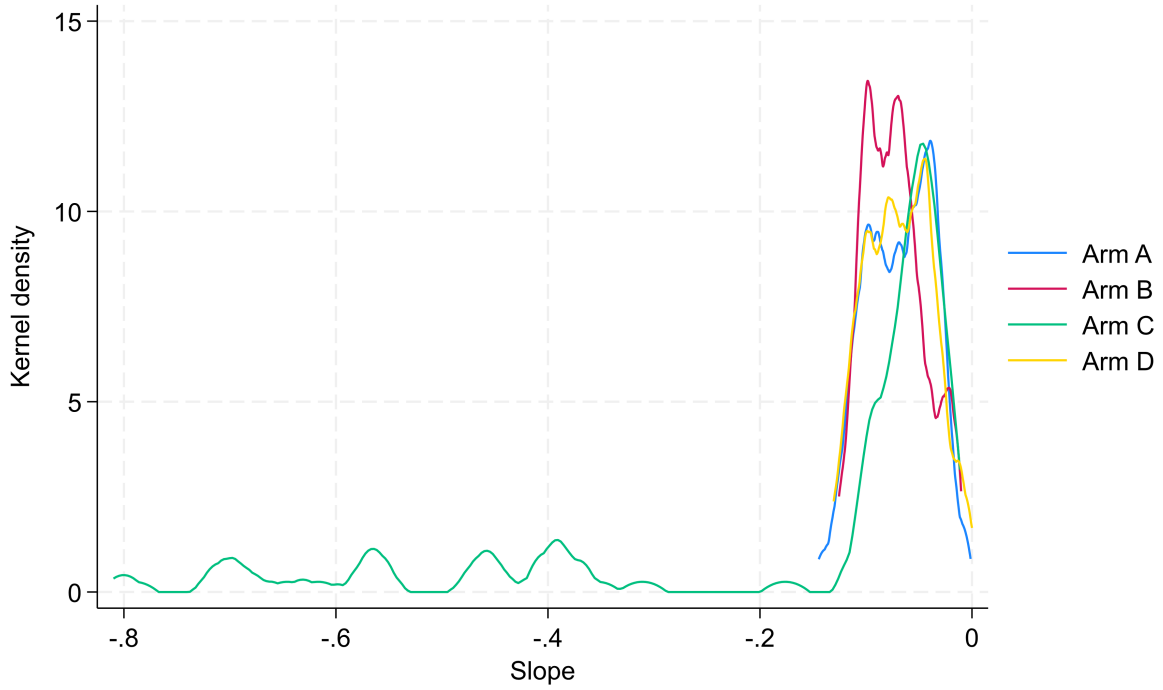

Figure S6: The individual slopes when regressing utility by the level sum score. Visualised per arm using kernel density plots.

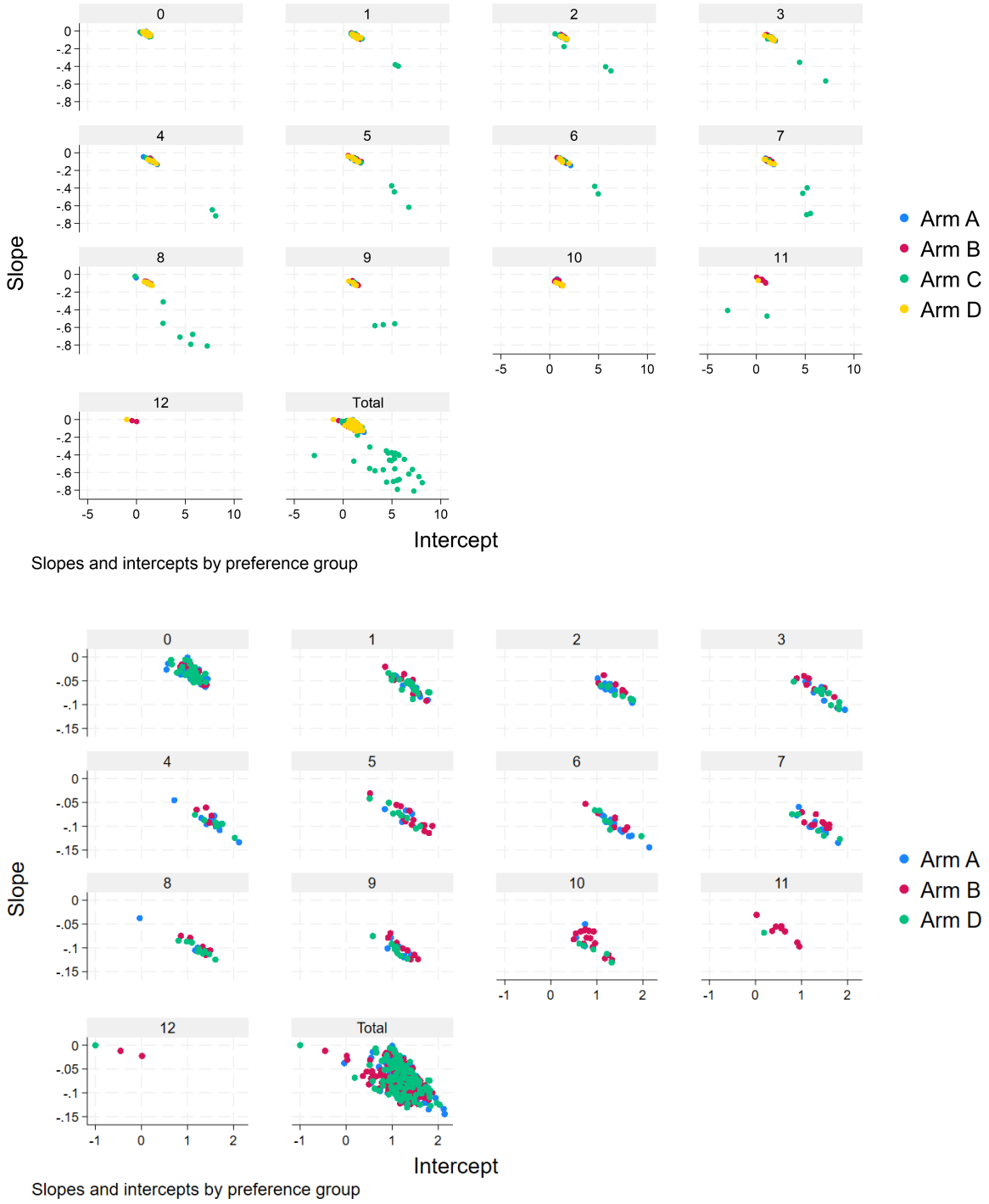

Figure S7: Slopes and intercepts for regressions of utility by the level sum score. Presented per arm, separately for subgroups of respondents depending on the number of states with strictly negative utility. The upper plot for all the arms, the bottom plot zoomed in with arm C removed.

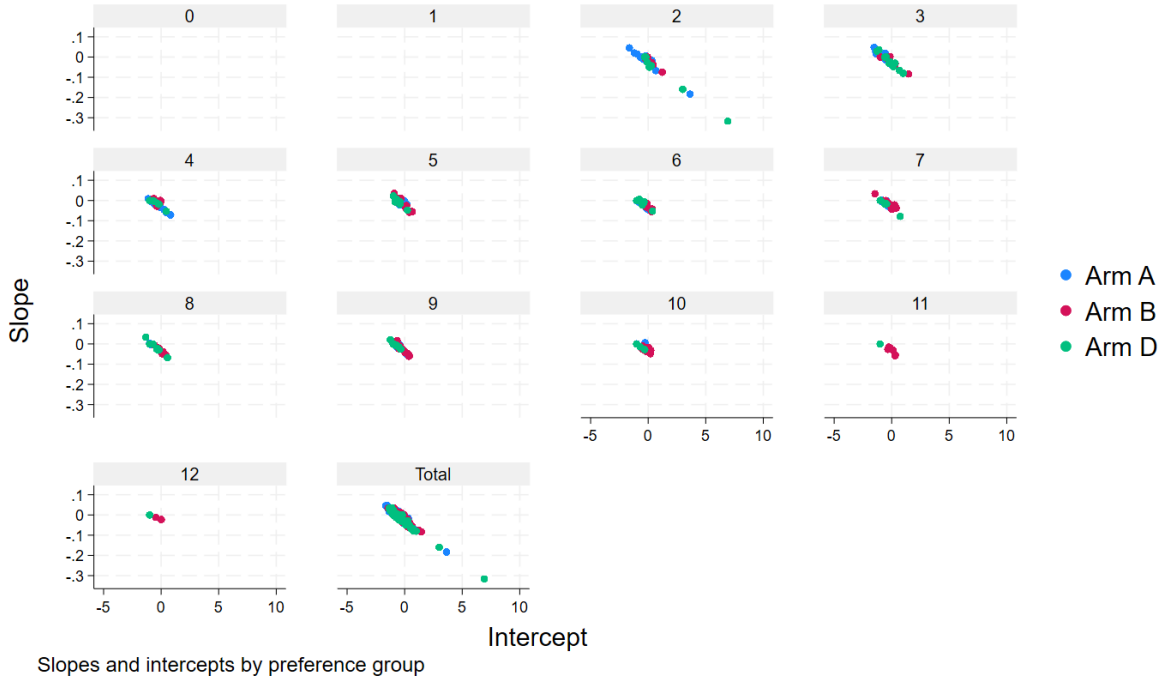

Figure S8: Slopes and intercepts for regressions of utility by the level sum score for WTD states only. Presented per arm, separately for subgroups of respondents depending on the number of states with strictly negative utility.

G. Arm perceived difficulty

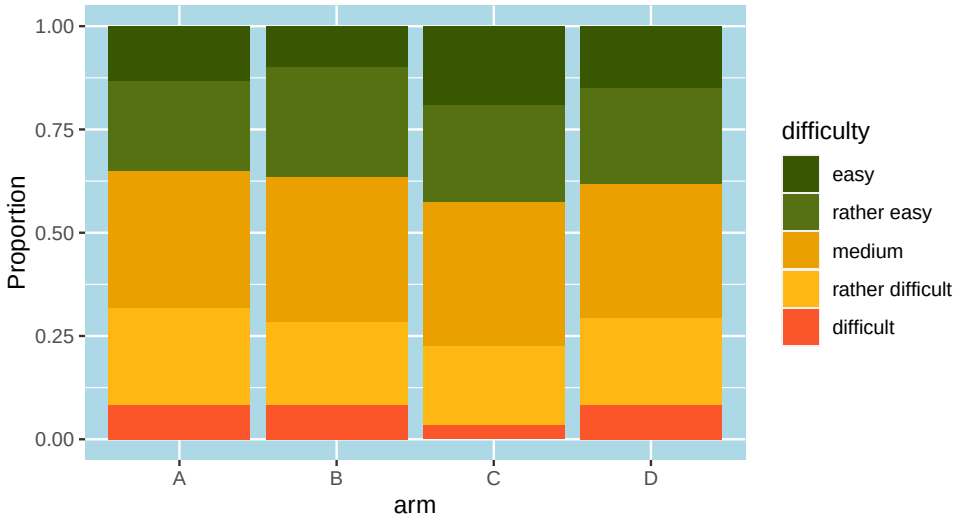

Figure S9: The difficulty of each arm as perceived by the respondents.

References

[1] Cokely E, Galesic M, Schulz E, Ghazal S, Garcia-Retamero R (2012) Measuring Risk Literacy: The Berlin Numeracy Test. Judgment and Decision Making 7:25–47

[2] Golicki D, Jakubczyk M, Graczyk K, Niewada M (2019) Valuation of EQ-5D-5L Health States in Poland: the First EQ-VT-Based Study in Central and Eastern Europe. Pharmacoeconomics 37:1165–1176

[3] Woloshin S, Schwartz L, Moncur M, Gabriel S, Tosteson A (2001) Assessing Values for Health: Numeracy Matters. Medical Decision Making 21:382–390