Supplementary Digital Content 5

Pilot Experiment: Valuations obtained with a 50-year time frame and a test of equivalence

(This document is supplementary to the paper by Summerfield, Kitterick, and Goman entitled 'Development and critical evaluation of a condition-specific preference-based measure sensitive to binaural hearing in adults: the York Binaural Hearing-related Quality of Life System'.)

1. Introduction

1.1. Experiments 1a and 1b generated the valuation set for the YBHRQL. Valuations were elicited by an implementation of the time trade-off technique with a 10-year time frame. Participants were instructed to indicate how many years living free of the problems in the scenarios would be equivalent to living 10 years with the problems. This Supplementary Digital Content reports a pilot experiment which was undertaken before Experiments 1a and 1b. It used a different implementation of the time trade-off technique. Informants were asked to imagine that they were 30 years old with a life expectancy of 50 years. They should indicate how many of the 50 years they would give up to be free of the problems in the scenarios.[1] If they stated that they would give up *y* years, then the value assigned to the scenario was calculated as (50-*y*)/50.

1.2. The most relevant question which data from the pilot experiment address is whether the vignettes provided constraining descriptions of problems with hearing. This is an important goal because unconstraining descriptions would result in participants having an incomplete understanding of the conditions they were asked to value. As a result, their valuations might be systematically biased or unnecessarily variable. The issue was addressed by comparing the valuations of members of the general public with the valuations of clinical professionals working in cochlear-implant programmes. The rationale was that if the descriptions were unconstraining then the clinicians would exploit their knowledge of binaural hearing to fill in the gaps. As a result, they would give systematically different, or more consistent, valuations than those given by the members of the public who lacked a professional understanding of the benefits of binaural hearing. Such effects would be shown by differences in the mean or the variance of the measure of *overall utility*. Alternatively, if, as intended, the descriptions gave constraining descriptions, then the variance of valuations would not differ between the two groups and their mean values would be statistically equivalent.

1.3. The Pilot Experiment also allowed checks on whether two aspects of the results of Experiments 1a and 1b replicated despite a different implementation of the time trade-off method. First, do changes in level on the Effort & Fatigue dimension have a greater influence on binaural utility than do changes in level on the other two dimensions? Second, do students trade more years than non-students?

1.4. A further question was whether the principle of 'constant proportionality' holds; that is, whether participants traded the same proportion of the 50-year time frame in the Pilot Experiment as the 10-year time frame in Experiments 1a and 1b.

[1] This version of the time trade-off task was used by Summerfield et al. (2010) in evaluating the cost-effectiveness of unilateral and bilateral cochlear implantation for children. The rational was that, in the UK, the average age of mothers at the birth of their first child was close to 30 years and the life expectancy of 30-year-olds was approximately 50 years (Office for National Statistics, 2015). The method yielded estimates of the gain in HRQL associated with unilateral implantation that were close to estimates already in the literature and yielded plausible estimates of the gain associated with bilateral implantation.

2. Methods

2.1. Overview: With the exception of the formulation of the time trade-off task, the methods were the same as in Experiments 1a and 1b.

2.2. *Valuation:* Participants received a response booklet (Supplementary Digital Content 6) containing a consent form, a demographic questionnaire, instructions, and examples. The demographic questionnaire established the participant's age, gender, and experience of hearing loss. Thereafter, each page contained one scenario. Participants were instructed to imagine that the scenario described their own hearing. They should consider that they were 30 years old with a life expectancy of a further 50 years. They should write down how many of those 50 years that they would give up in order to be free of the hearing difficulties described in the scenario. Each participant valued all 27 scenarios which were presented in four different randomised orders counterbalanced across participants.

2.3. *Participants:* Participants were convenience samples of students from the University of York (Students), members of the public who were adult friends and family of students (Non-students [Public]), and clinicians working in cochlear-implant programmes (Non-students[Clinicians]).

2.4. *Data cleaning:* Six of 6750 valuations were missing and were imputed. There were no inconsistent traders. Zero traders were included. Table 1 lists the numbers of inconsistent traders, zero traders, and participants included in analyses.

Table 1 Numbers (N), age, and genders of participants in the pilot experiment.

Group	Participants (N)	Inconsistent traders (N)	Zero traders (N)	Included in analyses (N)	Minimum age (years)	Mean age (years)	Maximum age (years)	% female
Students	95	0	2	95	18	20.3	25	77.9
Non-students [Public]	104	0	16	104	22	47.7	79	53.8
Non-students [Clinicians]	51	0	10	51	23	45.1	62	86.3

2.5. *Derived variables:* The *binaural utility* assigned to a scenario by a participant was calculated as *(50-y)/50,* where *y* was the number of years which the participant would give up in order to be free of the problems described in the scenario. From the 27 *binaural utilities*, values of *overall utility, mean utility,* and the *influence* of each dimension were calculated for each participant using the methods described in the paper.

2.6. *Analyses:* Analyses were performed with IBM SPSS for Windows v.26.0 (2019). Effects of group, dimension, and level on binaural utility were assessed in analyses of variance (ANOVAs), as were effects of group and dimension on influence. Degrees of freedom were adjusted with Huyn-Feldt corrections if Mauchly's test demonstrated that the assumption of sphericity was violated. Levene's Test was used to determine whether there was a difference in the variance of overall utility between clinicians and members of the public. A two one-sided test (TOST) (Lakens 2017), described in Section 5 below, assessed whether mean values of overall utility were statistically equivalent between clinicians and members of the public. To do that, the test determined whether the difference between the two groups was simultaneously above a lower bound and below an upper bound. The bounds were set to ±.03 because .03 is the smallest difference in health utility that is considered to be clinically important (Horsman et al. 2003; Coretti et al. 2014).

3. Results

3.1. *Effect of Group:* Binaural utilities were analysed in an ANOVA with the between-subjects factor *Group* (Students, Non-students[Public], Non-students[Clinicians]) and a 3x3x3

arrangement of the within-subjects factors *SPiN, LOC, E&F* each with three *Levels*. There was a significant effect of *Group* ($F_{(2,247)}$=8.766, p<.001, η_p^2=.066). Students assigned lower utilities (overall utility .891, 95% confidence interval .876 to .907) than Non-students[Clinicians] (.930, .911 to .949) (p<0.05) or Non-students[Public] (.934, .919 to .949) (p<.001) whose overall utilities did not differ.

3.2. The difference in overall utility between Non-students[Clinicians] and Non-students[Public] of .004 had a standard error of .013. The variances in overall utility were .00618 (Non-students[Public]) and .00496 (Non-students[Clinicians]). Levene's Test found no difference between the variances ($F_{1,153}$=.074, p=.786). The TOST showed that the difference in the means was above the lower bound (-.03) and below the upper bound (.03) ($t_{.10,153}$=2.057, p=.0247). Thus, the valuations of the two groups were statistically equivalent. In further analyses, the two groups of non-students were combined.

3.3. *Effect of Level:* The panels in the upper row in Figure 1 show how binaural utility varied with *Group* and *Level*. For comparison, the panels in the lower row show the corresponding relationship in the data from Experiments 1a and 1b combined. There was a significant effect of *Level* on utility for each dimension: *SPiN* ($F_{(1.45,357.45)}$=136.4, p<.001, η_p^2=.356); *LOC* ($F_{(1.39,343.9)}$=157.1, p<.001, η_p^2=.389); and *E&F* ($F_{(1.29,320.9)}$=175.7, p<.001, η_p^2=.416). Utility declined as level varied from 1 to 2 and from 2 to 3 on each dimension (all p<.001). The effect of *Level* on utility interacted significantly with *Group* for each dimension (SPiN, $F_{(1.44,357.07)}$=13.764, p<.001, η_p^2=.053; LOC, $F_{(1.39,343.98)}$=22.191, p<.001, η_p^2=.082; E&F, $F_{(1.30,321.45)}$=8.435, p<.001, η_p^2=.033). The effect of level was greater for students than non-students and is shown by the divergence of open from filled data points as level changes from 1 to 3 in the upper row of panels in Figure 1.

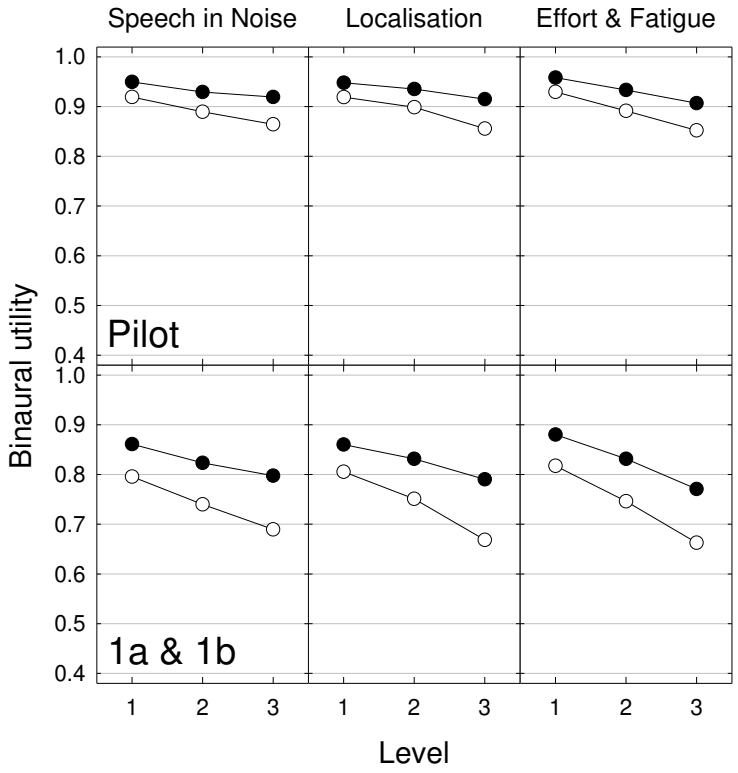

Figure 1 Relationship between binaural utility and level of dimensions (columns) for the Pilot Experiment (top row) and for Experiments 1a and 1b combined (bottom row). Each symbol plots the mean utility for one level of a dimension. Filled symbols plot results from non-students. Open symbols plot results from students. Note that the range of the abscissa has been truncated.

3.4. *Influence of Dimensions:* The heights of the bars in the upper panel of Figure 2 plot the *Influence* of each dimension for Non-students (filled bars) and Students (open bars). For comparison, the lower panel contains corresponding data from Experiments 1a and 1b combined. The measures from the pilot experiment were compared in an ANOVA with the between-subjects factor *Group* (Non-students, Students) and the within-subjects factor *Dimension* (SPiN, LOC, E&F). There was a significant effect of *Dimension* ($F_{(1.65,409.39)}$=17.59, p<.001, η_p^2 = .066). The *Influence* of E&F (.064, .056 to .073) was greater than the *Influence* of SPiN (.043, .037 to .049) (p<.001) or LOC (.048, .042 to .059) (p<.01) which did not differ. There was also a significant effect of *Group* ($F_{(1,248)}$=26.17, p<.001, η_p^2=.095). Students displayed a larger influence (averaged over the three dimensions, .065, .057 to .073) than did Non-students (.038, .032 to .045).

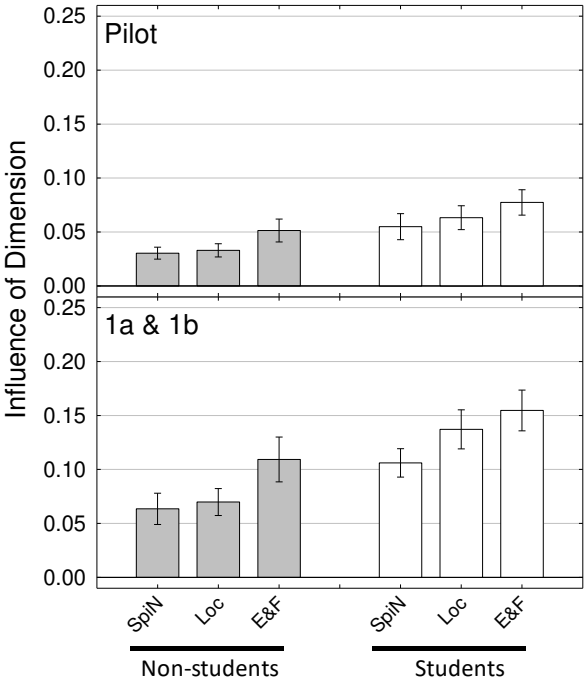

Figure 2 Influence of dimensions in the Pilot Experiment (upper panel) and in Experiments 1a and 1b combined (lower panel). Results from non-students are plotted as filled bars. Results from students are plotted as open bars. Error bars plot 95% confidence intervals.

3.5. *Differences between Experiments:* Differences in overall utility between Experiments 1a, 1b and the Pilot Experiment were compared in an ANOVA with the between-subject factors *Experiment* (1a, 1b, Pilot) and *Group* (Non-students, Students). There was a significant effect of *Experiment* ($F_{(2,446)}$=72.120, p<.001, η_p^2=.244). Overall utility was higher in the pilot experiment (.912, .899 to .925) than in Experiments 1a (.791, .772 to .809) or 1b (.776, .753 to .799) which did not differ.

4. Discussion

4.1. The statistical equivalence of overall utility between members of the public and clinicians is evidence that, as intended, the vignettes provide constraining descriptions of problems with binaural hearing.

4.2. In addition, two aspects of the results of Experiments 1a and 1b were replicated. First, participants were willing to trade more years to rectify problems with Effort & Fatigue than

problems with Localization or Speech-in-Noise. Problems with Effort & Fatigue, as described by the wording of the vignettes, are regarded as intrinsically more impactful than problems with the other two dimensions.

4.3. The second result that was replicated is that students were willing to trade more years than non-students. Further analyses (Supplementary Digital Content 7) demonstrate that, unlike Experiments 1a and 1b, the difference was not convincingly accommodated by a function relating age to average hearing utility. That is one of several reasons, discussed in Supplementary Digital Content 7, why we preferred to derive the valuation set for the YBHRQL from data gathered in Experiments 1a and 1b, despite the fact that a larger number of informants contributed valuations in the Pilot Experiment.

4.4. Finally, informants traded a larger proportion of the 10-year time frame in Experiments 1a and 1b than the 50-year time frame in the Pilot Experiment. Thus, the principle of constant proportionality did not hold. In Supplementary Digital Content 7, we speculate that the differences between the experiments in proportionality and in the relationship between the valuations of students and non-students may both relate to the different ways in which the time frame intersected with the actual life expectancy of informants.

5. Test of Equivalence

5.1. Table 2 lists the mean values of overall utility and their standard deviations for clinicians and members of the public from the Pilot Experiment. We wished to establish whether the values not only did not differ significantly but also whether they were statistically equivalent. Both criteria must be met to justify the conclusion that the vignettes were complete and constraining and thus that specialized knowledge of hearing loss is not required to produce a systematic valuation of the states of hearing defined in the YBHRQL.

Table 2 Values of overall utility and their standard deviations for two groups of participants from the pilot experiment: Non-students[Public] and Non-students[Clinicians].

Group	Mean	Standard deviation	N
Public	.934	.0786	104
Clinicians	.930	.0704	51

5.2. Levine's test showed that there was no difference between the standard deviations ($F_{1,153}$ = .074, p = .786). A conventional independent-samples 2-tailed t-test with an alpha level of .05 (5%) tested whether the difference in overall hearing utility between the groups was significant. The value of t was .744. This value is smaller than the critical value of t for an alpha level of .05 with 153 degrees of freedom which is 1.976. Thus, the hypothesis that the groups differed in overall hearing utility can be rejected.

5.3. A Two One-sided Test (TOST) (Lakens 2017) was used to determine whether the two values of overall utility were equivalent. A TOST entails two 1-tailed t-tests, each with an alpha level twice that which would be used to test for a difference between scores; so the level was set to .10 (10%). The aim is to establish whether the difference between scores is simultaneously above a lower bound and below an upper bound. The bounds should be set such that a difference falling between them would be "deemed equivalent to the absence of an effect that is worthwhile to examine" (Lakens 2017, p.356). Accordingly, we set the bounds to ±.03 because the Minimal Clinically Important Difference (MCID) (i.e. "the smallest change in a treatment outcome that an individual patient would identify as important and which would indicate a change in the patient's management", https://en.wikipedia.org/wiki/Minimal_important_difference) for values of health utility

has been estimated to be .03 for the HUI3 (Horsman et al. 2003) and is also the lowest of a range of estimates of the MCID for values of health utility obtained from the EQ5D (Coretti et al. 2014).

5.4. The mean difference in overall utility between clinicians and members of the public was .0043 with a standard error of .0130. The difference is significantly higher than the lower bound, -.03, ($t_{.10,153}$=2.635, p=.0046) and significantly lower than the upper bound, .03, ($t_{.10,153}$=2.057, p=.0247). Thus, we rejected the hypothesis that the difference between the groups is large enough to be clinically important; rather, the two measures of overall utility are statistically equivalent. Only the test yielding the larger value of p need be reported. Thus, in describing the results of the pilot experiment, above, we reported the result of the comparison with the upper bound.

5.5. The 2-tailed t-test with an alpha level of .05 is equivalent to testing whether the 95% confidence interval of the difference includes zero. The two 1-tailed t-tests each with an alpha level of .10 are equivalent to testing whether the 90% confidence interval of the difference includes neither the lower nor the upper bound. The filled square in Figure 3 plots the mean difference in average utility between clinicians and members of the public. The thick part of the horizontal line extending on either side of the filled square plots the 90% confidence interval of the mean difference. It Includes neither the lower nor the upper bound (marked by vertical lines composed of small dashes). The thin parts of the horizontal line plot the 95% confidence interval of the mean difference. It includes zero (marked by a vertical line composed of long dashes). Comparison with Figure 1 in Lakens (2017, p. 357) shows that the difference corresponds to Lakens' Case A: the two values of overall utility not only do not differ statistically but also are statistically equivalent. Table 3 lists the key values in the calculations of the confidence intervals.

5.6. In summary, this evidence justifies the conclusion that a systematic valuation of states of hearing defined in the YBHRQL does not require a specialized knowledge of hearing loss. To an adequate degree, the vignettes provide complete and constraining descriptions of problems with hearing.

Figure 3 Mean difference in overall utility between clinicians and members of the public (filled square) and its lower and upper 90% confidence intervals (heavy horizontal line) and 95% confidence interval (thin horizontal line). The vertical line composed of long dashes marks a difference of zero. The vertical lines composed of short dashes mark the lower and upper bounds of the difference in utility that is judged to be clinically important.

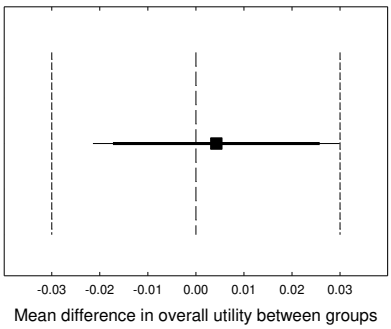

Table 3 Key values used in calculating the confidence intervals that are plotted in Figure 1.

Measure	Value
Difference between groups	.00425
Pooled variance	.00578
Standard error	.01300
Degrees of freedom	153
Value of $t_{.05,153}$	1.976
Lower 95% CL of difference	-.0257
Upper 95% CL of difference	.0299
Value of $t_{.10,153}$	1.655
Lower 90% CL of difference	-.0173
Upper 90% CL of difference	.0258

6. References

Coretti, S., Ruggeri, M., McNamee, P. (2014). The minimum clinically important difference for EQ-5D Index: a critical review. *Expert Rev Pharmacoecon Outcomes Res.* 14, 221-233.

Horsman, J., Furlong, W., Feeny, D., Torrance, G. (2003). The Health Utilities Index (HUI®): concepts, measurement properties and applications. *Health Qual Life outcomes* 1, 1-13.

IBM Corp. (2019) *IBM SPSS Statistics for Windows, Version 26.0.* Armonk, NY: IBM Corp.

Lakens, D. (2017). Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science* 8, 335-362.