Development and critical evaluation of a condition-specific preference-based measure sensitive to binaural hearing in adults: the York Binaural-related Quality of Life System

A. Q. Summerfield[1], Pádraig T. Kitterick[1,2], and Adele M. Goman[1,3]

[1]Department of Psychology, University of York, York YO10 5DD, United Kingdom.

[2]Current address: NIHR Nottingham Biomedical Research Centre, Hearing Sciences, Division of Clinical Neurosciences, School of Medicine, University of Nottingham, Nottingham NG7 2RD, United Kingdom.

[3]Current address: Cochlear Center for Hearing and Public Health, Johns Hopkins Bloomberg School of Public Health, 2024 East Monument Street, Baltimore, MD 21205, United States.

Financial disclosures/conflicts of interest: Pádraig Kitterick was supported by a grant from Advanced Bionics SARL and by infrastructure funding from the National Institute for Health Research (NIHR). Adele Goman was supported by a studentship from the Economic and Social Research Council. The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR or the Department of Health and Social Care. The authors declare no conflicts of interest.

Name and address for correspondence

Quentin Summerfield, Department of Psychology, University of York, York YO10 5DD, UK.

Tel: Intl + 44 (0)1904 738825. E-mail: quentin.summerfield@york.ac.uk

Word counts

Title page (213), Abstract (499), Main Text (10,407), References and Reference Note (1,977), Figure Legends, Table Legends, and Titles of Supplementary Digital Content Files (613).

Abstract

Objectives

The widely-used generic preference-based measures of health-related quality of life – the EuroQol Descriptive System (EQ-5D) and the Health Utilities Index (HUI3) – are limited in their response to technologies that improve hearing. The EQ-5D lacks construct validity for hearing, while the HUI3 is restricted by a ceiling effect and by using speech reception as the only evidence of the ability to hear. Consequently, neither measure consistently registers benefits from binaural hearing, such as those from bilateral versus unilateral cochlear implantation. The objectives were to test whether informants value binaural hearing, to develop a condition-specific preference-based measure sensitive to binaural hearing, to assess the psychometric properties of the new instrument, and to determine whether it meets requirements for informing judgements of cost-effectiveness: does it measure greater gains than do the generic preference-based measures, while avoiding exaggerating losses, and displaying sensitivity to side effects?

Design

Three levels of function, ranging from no difficulty to great difficulty, were defined on each of three dimensions where listening is easier or more successful when hearing is binaural rather than monaural: perception of speech in spatially-separated noise, localization of sounds, and effort and fatigue. Informants (N=203) valued the 27 combinations of levels and dimensions in a time trade-off task with a 10-year time frame to provide a value of binaural-related quality of life ('binaural utility') for each combination. A questionnaire was compiled to allow a respondent to report their level of function on each dimension so that a value of binaural utility could be assigned to them. The questionnaire and the age-standardized valuations constitute the York Binaural-related Quality-of-life System (YBRQL). Adult users of unilateral implants (N=8), bilateral implants (N=11), or bimodal aiding (N=9) undertook performance tests of spatial listening and completed the HUI3, EQ-5D, and

Speech Spatial and Qualities of Hearing (SSQ) questionnaires. They completed the YBRQL questionnaire 24 and 38 months later.

Results

Despite long intervals between measurements, the YBRQL demonstrated desirable psychometric properties: good construct validity evidenced by significant correlations with performance measures and the SSQ; a greater ability than the EQ-5D or HUI3 to distinguish unilateral, bimodal, and bilateral listening; and good reproducibility. The YBRQL did not exaggerate losses of utility but showed no sensitivity to one potential side effect (pain/discomfort). It measured a gain in utility from bilateral compared with unilateral implantation (median=.11, IQR .03 to .16) that was greater than the gain measured by the EQ-5D (.00, .00 to .00) but not the HUI3 (.00, .00 to .17).

Conclusions

The YBRQL summarizes the contribution of binaural hearing to quality of life by combining the functional status of a listener with the preferences of independent informants. It would be an efficient clinical outcome measure. In addition, if used alongside the EQ-5D or HUI3, it would provide evidence which could beneficially modulate confidence in the cost-effectiveness of interventions. Further research on its sensitivity to side effects, and on the size of the gains in utility which it measures, is needed to determine whether it could stand alone to inform resource-allocation decisions.

Introduction

Policy makers in healthcare require estimates of the cost-effectiveness of treatments to be informed by generic preference-based measures (generic PBMs) of health-related quality of life (HRQL) (e.g. Drummond et al. 2000; National Institute for Health and Care Excellence (NICE) 2013; Dubois 2016). Generic PBMs define dimensions, such as pain, mobility, and anxiety, on which better function corresponds to better health. The dimensions are intended to be universally relevant so that generic PBMs can compare the effectiveness of interventions for widely differing conditions. Respondents report their level of function on each dimension. A weighting function converts the reported levels into a value of HRQL on a scale where unity corresponds to full health and zero to being dead. On this scale, profoundly deaf adult candidates for cochlear implantation in the US have an average value of .58 (Palmer, Niparko, & Wyatt 1999), similar to values for adults who have survived heart attack (.57) or suffer chronic bronchitis (.59) (Luo et al. 2009). Following unilateral implantation, adults have an average value of .78, similar to sufferers of hypertension (.73), sinusitis (.76), and hay fever (.79).

The difference in HRQL resulting from alternative treatments can be integrated over time to estimate the number of quality-adjusted life years (QALYs) gained by one treatment compared with another. The additional QALYs can be combined with the additional costs to determine whether the cost per QALY gained (the incremental cost-effectiveness ratio) is favourable in relation to the amount that policy makers are willing for health services to pay to gain a QALY (the willingness-to-pay threshold) (e.g. NICE 2013; Neumann et al. 2014; Dubois 2016).

This article was motivated by concerns about the validity of estimates of the gain in HRQL from bilateral compared with unilateral cochlear implantation in adults. The generic PBMs which have been used to estimate the gain are the 3-level version of EuroQol Descriptive System (EQ-5D-3L; Brooks et al. 2003) and the Health Utilities Index Mark III (HUI3; Feeny et al. 2002). They, or a third method, the time trade-off (described in Methods), were used in the five studies summarised in Table 1. The estimates of the gain in HRQL vary in size between studies. More than half are small and not statistically significant, despite participants reporting improvements in their listening skills (Summerfield et al. 2006; Smulders et al. 2016).

----- Table 1 : Estimates of gain in HRQL -----

Some of the variation may stem from differences between studies in their susceptibility to biases which amplify or attenuate estimates of HRQL. Retrospective studies (Bichey & Miyamoto 2008) are prone to selection and recall biases. Scenario analyses (Summerfield et al. 2002; Kuthubutheen et al. 2015) may fail to register the impact of negative side effects. The two randomized controlled trials (Summerfield et al. 2006; Smulders et al. 2016) were under-powered to detect between-group differences in HRQL that were small in relation to the within-group variability. Supplementary Digital Content 1 expands on this critique.

The small size of many of the estimates and their lack of statistical significance also reflects limitations in the design of the generic PBMs. The five dimensions of the EQ-5D-3L – *Mobility, Self-care, Usual activities, Pain/discomfort,* and *Anxiety/depression* – are minimally sensitive to sensory disorders including hearing loss (Longworth et al. 2014) and to interventions which improve hearing such as acoustic hearing aids (Barton et al. 2004; Grutters et al. 2007) and unilateral cochlear implants (Summerfield & Barton 2019). Unsurprisingly, the EQ-5D-3L is also insensitive to the difference between one implant and two.

In comparison, the HUI3 is consistently sensitive to hearing loss (Longworth et al. 2014) and to interventions which alleviate it, including acoustic hearing aids (Barton et al. 2004; Grutters et al. 2007) and unilateral implants (e.g. Palmer et al. 1999; UK Cochlear Implant Study Group 2004; Summerfield & Barton 2019). Sensitivity occurs because the eight dimensions ('attributes') include *Hearing* (the ability to 'hear conversation') and *Speech* (the ability to be understood when speaking). However, the sensitivity of the HUI3 to a second implant is limited by a ceiling effect which arises because the two highest levels of the *Hearing* attribute (Levels 1 and 2) require the ability to hear conversation without the use of an implant/hearing aid (Horseman et al. 2003; Supplementary Digital Content 2). Thus, the highest level which a user of an implant can attain is Level 3. However, two thirds of users of unilateral implants place themselves at this level (Summerfield & Barton 2019). Therefore, only one third have sufficient headroom to report a higher level of the *Hearing* attribute when using two implants rather than one.

Faced with a range of estimates of varying quality, systematic reviews have concluded that the gain in HRQL associated with bilateral compared with unilateral implantation is sufficiently small (.03, Bond et al. 2009; .035, Health Quality Ontario 2018) that there is uncertainty about the likelihood that bilateral implantation is a cost-effective intervention (Bond et al. 2009; Lammers et al. 2011; Health Quality Ontario 2018; Theriou et al. 2019). Analysts have reached those conclusions both when they have decided on balance that bilateral implantation is a cost-effective intervention for adults (Health Quality Ontario 2018) and when they have decided that it is not (Bond et al. 2009). In turn, policy on the provision of bilateral implants varies within countries (Health Quality Ontario 2018; Boys Town National Research Hospital 2020) and between them (Vickers et al. 2016).

Such variation partly reflects differences between healthcare regimes in the costs of provision and in willingness-to-pay thresholds, but also reflects uncertainty about the size of the gain in HRQL. Herein lies a dilemma. If the true gain is of the order of .03, the likelihood is low that bilateral implantation for adults achieves acceptable value for money in many healthcare jurisdictions. If the gain is greater, but the generic PBMs are not equipped to detect it, then the decision not to support bilateral implantation for adults risks inefficiency in the allocation of health-care resources. That issue motivated the current study in which we developed and evaluated a condition-specific PBM designed to be sensitive to binaural hearing in adults.

Condition-specific PBMs

The insensitivity of generic PBMs to some conditions which self-evidently reduce quality of life and to some interventions which improve it has been noted in domains in addition to hearing

(Longworth et al. 2014). A possible solution has been the development of condition-specific PBMs (e.g. Brazier et al. 2008; Yang et al. 2011; Versteegh et al. 2012; Swinburn et al. 2013). They include a limited number of dimensions on which function is impaired by a condition and alleviated by an effective intervention. Compared with a generic PBM that is insensitive to a condition and its treatments, a condition-specific PBM would be expected to demonstrate a greater gain in HRQL associated with successful treatments for the condition.

In practice, that advantage is in tension with weaknesses which may result in condition-specific PBMs over-estimating the effectiveness of interventions (Brazier & Tsuchiya, 2010; Versteegh et al. 2012). Key among them are 'exaggeration of losses' and 'insensitivity to side effects'. Exaggeration of losses arises if, by focussing on a limited region of the health space, differences in value within the space are amplified in the minds of informants; as a result, greater differences in value are recorded between the best and worst levels of dimensions in a condition-specific PBM than would be recorded were the same dimensions valued in the context of a set of generic dimensions. The second weakness, insensitivity to side effects, arises if a condition-specific PBM includes dimensions chosen only to reflect the benefits, but not the disbenefits, of interventions for a condition. Designers of condition-specific PBMs need to demonstrate that their new instruments display the desired advantages while avoiding those weaknesses. To that end, Versteegh et al. (2012) proposed criteria which condition-specific PBMs should meet before informing resource-allocation decisions. In the General Discussion, we assess the condition-specific PBM which we developed against those and other criteria.

Phases of the study

The study involved four phases. In Phase 1, we identified dimensions on which listening is easier or more successful when hearing is binaural rather than monaural. We defined discrete levels of function on each dimension ranging from very poor to very good. Informants used the time trade-off technique (TTO; e.g. Drummond et al. 2000; York Health Economics Consortium 2016) to value states of binaural hearing defined by different combinations of the levels. The TTO is a choice-based method for eliciting values of states of health. It was used in many countries to value states in the EuroQol Descriptive System (Brooks et al. 2003). An informant imagines that a description of a less-than-perfect state applies to them and then arbitrates between two choices: either to live for a specified length of time (the time frame, *y* years) in the imperfect state, or to trade length of life for quality of life and live for a shorter time (*x* years) in full health. The informant's task is to indicate what the shorter time should be for them to be indifferent between the two choices. The value of health-related quality of life ('health utility') assigned by the informant to the state is calculated as *x/y*. In this way, values are elicited on a scale where zero corresponds to being dead and unity to full health. For example, consider that the time frame is 10 years and an informant judges that living 7.5 years in full health would be equivalent to living 10 years in a particular imperfect state of health. The utility assigned by the informant to the imperfect state is .75 (=7.5/10). In effect, the informant has equated the number of QALYs that they would live under each alternative: either 10 years at 75% of full health, yielding 7.5 QALYs, or 7.5 years at 100% of full health, also yielding 7.5 QALYs.

In our implementation of the TTO, informants imagined that descriptions of states of imperfect binaural hearing applied to them. They indicated the number of years living without the problems described in a state that would be equivalent to living for the time frame with the problems. In this way, they assigned a value of binaural-related quality of life ('binaural utility') to each state on a scale where unity corresponds to normal binaural hearing and zero to binaural hearing so bad as to be equivalent to being dead.

There is no agreed implementation of the TTO (Attema et al. 2013), although procedures used in valuing the EQ-5D-3L (Oppe et al. 2016) have been recommended (NICE 2013). Those procedures include a 10-year time frame. Accordingly, we used a 10-year time frame to obtain the primary dataset.

In Phase 2, we age-standardized the valuations. The need for age standardization arises from evidence of a systematic association between age and TTO valuations. For example, when informants valued health states in the EuroQol Descriptive System (Dolan & Roberts 2002), the number of years traded decreased as age rose to 45 years, then increased gradually as age rose to 70 years and more steeply above that age. Potentially, studies would yield different average valuations of the same health states if they recruited informants with different distributions of age. Accordingly, we first confirmed that values of binaural utility varied with age. We then weighted the data of each informant by the proportion of adults in the UK of the same age as the informant before averaging the weighted data. In that way we estimated the average values of binaural utility that would be obtained from a sample of informants who were representative of the distribution of age in the population, rather than being tied to the particular distribution of ages in our sample.

In Phase 3, we compiled a questionnaire which allowed a respondent to report the level of difficulty that they experienced on each dimension.

In Phase 4, we administered the questionnaire to adult users of cochlear implants. Participants were assigned values of binaural utility according to the combination of levels they adopted to describe their hearing. We employed these values to assess the potential of the questionnaire in two roles: as a clinical outcome measure and as an instrument for informing estimates of cost-effectiveness.

In summary, the aims of the study were, first, to confirm that people are willing, in principle, to trade length of life to improve quality of binaural hearing; second, to develop a condition-specific preference-based measure sensitive to binaural hearing in adults; third, to assess the psychometric properties of the new instrument; and fourth to determine whether it meets criteria for informing clinical and economic decisions in healthcare.

Research ethics

Ethical approval was granted by research ethics committees of the National Health Service and the Department of Psychology of the University of York.

Phase 1: Dimensions, levels, and values of binaural utility

Methods

Selection of dimensions

We sought to identify dimensions with four characteristics. They should relate to differences between monaural and binaural listening. They should describe fundamental (i.e. physiological/psychological) and universal (i.e. occurring to the great majority of people) differences. They should potentiate participation across a spectrum of activities. They should relate directly to the aims of treatments and to the benefits experienced by patients. As such, they should be supported by evidence of relevance from researchers, patients, and manufacturers of hearing technologies. A subsidiary consideration was that informants should value every combination of levels on the chosen dimensions in a single session. That constraint limited the number of dimensions and the number of levels on each dimension to three, giving 27 (=3^3) combinations of levels and dimensions.

To identify dimensions, we reviewed evidence from performance tests, self-report, and the guidance given by manufacturers of implants and hearing aids. Implanting or aiding both ears rather than one ensures that the physiologically more responsive ear is stimulated and, potentially, gives listeners access to inter-aural differences in timing and level. As a result, accuracy of localization in azimuth improves – shown for bilateral compared with unilateral implantation (Kerber & Seeber 2012), for bimodal aiding (i.e. a unilateral implant combined with a contralateral acoustic hearing aid) compared with unilateral implantation (Potts et al. 2009), and for bilateral compared with unilateral fitting of acoustic hearing aids (Byrne & Noble 1998). A second consequence is that accuracy of speech perception in noise improves, particularly when the sources of noise and speech are spatially separated – shown for bilateral compared with unilateral implantation (Litovsky et al. 2006), for bimodal aiding compared with unilateral implantation (Ching et al. 2004), and for bilateral compared with unilateral fittings of acoustic hearing aids (Dawes & Munro 2013).

Those advantages are echoed in self-reports when patients use the Speech, Spatial, and Qualities of Hearing Scale (SSQ; Gatehouse & Noble 2004) to indicate how well they perform tasks of speech hearing (in quiet and noise), spatial hearing (localisation of sound sources, distance and movement perception), and other qualities of hearing (clarity, separation, and identification of sound sources). Provision of a second implant has been associated with more positive reports in all three domains (Summerfield et al. 2006; Noble et al. 2008; Härkönen et al. 2015; Smulders 2016), while provision of a second hearing aid has been associated with more positive reports of speech hearing and spatial hearing, and of the 'listening-effort' sub-domain of qualities of hearing (Noble & Gatehouse 2006). Reductions in listening effort associated with binaural listening have also been reported by users of implants (Hughes & Galvin 2013; Noble et al. 2008; Härkönen et al. 2015) and demonstrated in performance tests by listeners with normal hearing (Rennies & Kidd 2018).

Summaries of the foregoing advantages of binaural stimulation by manufacturers of implants (e.g. Advanced Bionics 2021; Cochlear 2021; Med-El 2021) and acoustic hearing aids (e.g. Starkey 2021; Oticon Medical 2021) include some or all of improved abilities to localize sounds and to perceive speech in noisy environments, along with enhanced clarity and greater ease of listening. Those, therefore, are fundamental advantages that patients are given to expect, and often experience, from bilateral fittings. Faced with the pragmatic requirement to select three dimenions, we chose *Speech Perception in Noise* (SpiN), *Localization* (Loc),and *Effort and Fatigue* (E&F). We judged that levels of difficulty on those dimensions could be described more simply, and would be perceived to be more widely relevant, than would levels of difficulty in the fourth candidiate dimension, perception of sound quality.

Definitions of levels

We defined three levels of function on each dimension corresponding to 'No difficulty' (Level 1), 'Some difficulty' (Level 2), and 'Great difficulty' (Level 3). Each level was described by a brief vignette (Table 2). Twenty seven scenarios were formed by combining three vignettes, one for each dimension, to describe a particular combination of difficulties with binaural hearing. Considering the vignettes in the order SPiN, Loc, E&F, scenarios can be described by three numbers, referring to the difficulty on each dimension, ranging from 1:1:1 to 3:3:3.

----- Table 2 : Vignettes -----

Four considerations guided the construction of vignettes. They should involve straightforward language. The same wording should describe a state for an informant in a valuation task and should allow a patient to report their state in a questionnaire. The wording should be applicable to any condition that impairs binaural hearing and to any treatment intended to improve binaural hearing. Vignettes should provide sufficient detail to define states clearly and completely.

The fourth consideration was addressed by describing a limitation of hearing (e.g. "You have some difficulty working out where sounds are coming from."), explaining how the limitation might be manifest (e.g. "You can usually tell if a sound is coming from the right- or left-hand side, but you cannot be more accurate than that."), and setting out a consequence in the form of a restriction of everyday activities or a hurdle to be overcome (e.g. "As a result, you are not always sure who is speaking when you are in a group with several people."). In the language of the International Classification of Functioning, Disability, and Health (World Health Organisation 2001), each dimension relates to a *body function,* the consequences of whose impairment are illustrated by restrictions on *activity* and *participation.* Supplementary Digital Content 5 argues that this strategy provided complete evidence of each level of difficulty as evidenced by the statistical equivalence (Lakens 2017) of the valuations of clinicians and members of the public.

Valuations

In each experiment, participants received a booklet 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 and to value it using a version of the TTO. Each participant valued all 27 scenarios which were presented in four different randomised orders counterbalanced across participants. An experimenter was on-hand to answer questions. Examples of response booklets for Experiments 1a and 1b are included in Supplementary Digital Content 3 and 4.

Experiment 1a

Participants imagined they had 10 years to live. They decided how many years *(x)* living with no hearing problems would be equivalent to living the full 10 years with the difficulties described in the scenario. They indicated their choice by marking a visual-analogue scale which ranged from 0 to 10 with tick marks for each half year. The value of binaural utility assigned to the scenario by the participant was calculated as $x/10$. The vignettes making up each scenario were presented in a fixed order: SpiN, Loc, E&F.

Experiment 1b

In Experiment 1b, the order of presentation of the vignettes was counterbalanced among participants. In addition, half of the participants made valuations while considering that their sight was severely impaired while the other half considered that their sight was perfect. Here we report results only for participants who considered that their sight was perfect. For them, the TTO task was the same as in Experiment 1a.

Participants

Participants (Table 3) were convenience samples of students from the University of York (Students) and members of the public who were adult friends and family of students (Non-students). Students divided approximately equally between those majoring in psychology and those majoring in other disciplines.

----- Table 3 : Participants in Experiments 1a and 1b -----

Data cleaning

Two of 2997 valuations were missing in Experiment 1a and 3 of 2592 in Experiment 1b. Missing values were imputed deterministically based on the non-missing data such that *Imputed value = (S x P)/G,* where *S* was the average utility for the scenario calculated from data from other participants in the group, *P* was the average utility for the participant averaged over all scenarios, and *G* was the grand mean utility.

Some participants made inconsistent judgements insofar as they gave a lower value to the most advantageous scenario 1:1:1 than to the least advantageous 3:3:3. In line with some studies reviewed by Attema et al. (2013), inconsistent traders were excluded because they were likely to have misunderstood the instructions. Participants who declined to trade when valuing any scenario (zero traders) were not excluded, given that there is no reason to expect participants to consider that difficulties with binaural hearing are impactful enough to justify trading length of life to alleviate them. Table 3 lists the numbers of inconsistent traders, zero traders, and participants included in analyses.

Derived variables

The core data consisted of the 27 *binaural utilities* assigned by each participant. Three additional measures were calculated for each participant. The average of the 27 binaural utilities was calculated so that an *overall utility* could be associated with each participant as a measure of their willingness to trade. A *mean utility* for each level of each dimension was calculated as the average of the binaural utilities for the 9 scenarios in which the dimension was at a particular level. A measure of the *influence* of each dimension was calculated as the difference in mean utility between Level 1 and Level 3 of each dimension.

Analysis

Analyses were conducted 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.

Results

Effects of Group, Dimension, and Level

Experiment 1a

The upper row of panels in Figure 1 show how mean utility varied with *Group* and *Level* in Experiment 1a. Binaural utilities were analysed in an ANOVA with the between-subjects factor *Group* (Students, Non-students) and a 3x3x3 arrangement of the within-subjects factors *SpiN, Loc, E&F,* each with three *Levels*. There was a significant effect of *Group* ($F_{(1,108)}$=16.617, p<.001, η_p^2=.133). Students assigned lower utilities (overall utility .738, 95% confidence interval .703 to .773) than non-students (overall utility .844, .806 to .882). There was also a significant effect of *Level* on mean utility for each dimension: *SpiN* ($F_{(1.66,178.96)}$=100.7, p<.001, η_p^2=.482); *Loc* ($F_{(1.37,147.79)}$=135.4, p<.001, η_p^2=.556); and *E&F* ($F_{(1.37,148.01)}$=155.4, p<.001, η_p^2=.590). Mean utility declined as level varied from 1 to 2 and from 2 to 3 on each dimension (all p<.001). The effect of level was greater for students than non-students shown by significant interactions between *Level* and *Group: SpiN* ($F_{(1.66,178.96)}$=10.363,

$p<.001$, η_p^2=.088); *Loc* ($F_{(1.37,147.79)}$=21.013, $p<.001$, η_p^2=.163); *E&F* ($F_{(1.37,148.01)}$=4.267, $p<.001$, η_p^2=.038).

----- Figure 1 -----

The heights of the bars in the upper panel of Figure 2 plot the *Influence* of each dimension. These measures 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* ($F_{(1.73,187.16)}$=22.80, $p<.001$, η_p^2=.174). The *Influence* of E&F (.139, .119 to .160) was greater than the *Influence* of Loc (.106, .090 to .123) ($p<.05$) which was greater than the *Influence* of SpiN (.074, .062 to .087) ($p<.001$). There was also a significant effect of *Group* ($F_{(1,108)}$=21.51, $p<.001$, η_p^2=.166). Students displayed a larger influence (averaged over the three dimensions, .136, .119 to .153) than did non-students (.077, .059 to .096).

----- Figure 2 -----

Experiment 1b

The lower row of panels in Figure 1 show how mean utility varied with *Group* and *Level* in Experiment 1b. Binaural utilities were analysed in an ANOVA with the between-subject factor *Group* (Non-students, Students) and a 3x3x3 arrangement of the within-subject factors *SpiN, Loc, E&F*, each with 3 *Levels*. The effect of *Group* was close to the level of significance ($F_{(1,91)}$=3.90, p=.051, η_p^2=.041). Overall utility was .747 (.703 to .790) for students and .809 (.764 to .854) for non-students. There was a significant effect of *Level* on utility for each dimension: *SpiN* ($F_{(1.47,133.96)}$=133.62, $p<.001$, η_p^2=.595); *Loc* ($F_{(1.74,157.94)}$=121.74, $p<.001$, η_p^2=.572); and *E&F* ($F_{(1.41,128.54)}$=139.97, $p<.001$, η_p^2=.606). Mean utility declined as level varied from 1 to 2 and from 2 to 3 on each dimension (all $p<.001$). The effect of *Level* was greater for students than non-students shown by significant interactions between *Level* and *Group*: *SpiN* ($F_{(1.47,133.96}$ =8.29, p=.001, η_p^2=.083); *Loc* ($F_{(1.74,157.94)}$=7.16, p=.002, η_p^2=.073); *E&F* ($F_{(1.41,128.54)}$=4.39, p=.025, η_p^2=.046).

The heights of the bars in the lower panel of Figure 2 plot the *Influence* of each dimension. These measures were compared in an ANOVA with the between-subjects factors *Group* (Non-students, Students) and *Order* of vignettes in the scenarios (6 orders), and the within-subjects factor *Dimension* (SpiN, Loc, E&F). There was a significant effect of *Group* ($F_{(1,81)}$=11.68, p=.001, η_p^2=.126); students showed a larger influence (averaged over the three dimensions .129, .111 to .146) than non-students (.084, .066 to .103). Influence did not vary significantly with *Order* ($F_{(5,81)}$ = 1.01) nor were there any significant higher-order interactions among *Group, Sight,* and *Order.* Finally, there was a significant effect of *Dimension* ($F_{(2.00, 162.00)}$=5.02, p<.01, η_p^2=.058). The influence of *E&F* (.123, .104 to .141) was greater than the influence of *Loc* (.100, .084 to .115) (p<.01) and of *SpiN* (.097, .082 to .112) (p<.05) which did not differ significantly.

Differences between Experiments

Differences in overall utility between Experiments 1a and 1b were compared in an ANOVA with the between-subject factors *Experiment* (1a, 1b) and *Group* (Non-students, Students). There was a significant effect of *Group* ($F_{(1,199)}$=17.255, p<.001, η_p^2=.080), but not of *Experiment* ($F_{(1,199)}$=.411, p=.522, η_p^2=.002).

Discussion

Three results were common to Experiments 1a and 1b. First, the majority of participants were willing to trade years of life to improve quality of binaural hearing. Second, students traded more years than did non-students. Third, participants traded more years to rectify problems with *E&F* than problems with *SpiN* or *Loc*. In these respects, Experiments 1a and 1b replicated the results of a Pilot Experiment (Supplementary Digital Content 5) which used a different implementation of the TTO, thereby demonstrating that the three results are independent of the cognitive demands of any particular implementation of the TTO. Finally, the absence of an effect of the order in which scenarios were presented in Experiment 1b, and the absence of differences in overall utility between Experiments 1a and 1b, demonstrated that the greater influence of *E&F* than *Loc* or *SpiN* in

Experiment 1a was not a consequence of the fixed order of presentation of the vignettes. In further analyses, data from Experiments 1a and 1b were pooled.

Phase 2: Deriving an age-standardized value set

We hypothesized that a quadratic function would describe the variation in overall utility with age and would accommodate the difference between students and non-students without the need to invoke other explanatory variables. That hypothesis was tested by predicting overall utility in Experiments 1a and 1b with a linear weighted combination of the variables *Gender* (female, male), *Group* (student, non-student), and the covariates *Age* and *Age*^{2}. Significant components of the variance were explained by *Age* ($F_{(1,197)} = 12.152$, $p < .01$, $\eta_p^2 = .058$) and *Age*^{2} ($F_{(1,197)} = 10.959$, $p < .01$, $\eta_p^2 = .053$), but neither by *Gender* ($F_{(1,197)} = .456$, $p = .500$, $\eta_p^2 = .002$), nor, critically, by *Group* ($F_{(1,197)} = .677$, $p = .412$, $\eta_p^2 = .003$). This result justifies combining data from students and non-students after age-standardization. Further analyses of the relationship between age and overall utility are reported in Supplementary Digital Content 7.

Table 4 illustrates the steps entailed in deriving a value set. Participants were partitioned into age decades. Binaural utilities were averaged within each decade. Each average was weighted by the proportion of the adult population of the UK (i.e. the population from which participants had been drawn) in that decade. The weighted averages were summed. This procedure was conducted for each of the 27 binaural utilities and for the overall utility.

----- Table 4 : Calculation of age-standardized binaural utilities -----

The underlined entries in Table 5 are the 27 values that resulted, one for each scenario. Values range from 0.96 to 0.69 and are well-behaved in that they vary monotonically with level throughout the table.

----- Table 5 : Initial and expanded value set -----

Phase 3: The York Binaural-related Quality of Life System

In Phase 3 we compiled a questionnaire that could elicit a person's level of function on each dimension so that a value of binaural utility could be assigned to summarise the quality of their binaural hearing. We named the resulting combination of questionnaire and valuations the York Binaural-related Quality of Life System (YBRQL).

YBRQL Questionnaire

The YBRQL questionnaire is illustrated in Figure 3. The preamble invites respondents to indicate their level of function on the three dimensions in the listening condition which is described in bold type. Respondents choose from five levels on each dimension. Three are the levels for which values of binaural utility were obtained in Phase 1. They are described by the same wording as was used to elicit valuations. The other two levels are intermediary between 'No difficulty' and 'Some difficulty' and between 'Some difficulty' and 'Great difficulty'. The additional levels are included to allow participants to provide a granular indication of their function on each dimension.

----- Figure 3 : YBRQL Questionnaire -----

Expanded value set

To accommodate the intermediary levels of function, additional values of binaural utility were calculated by linear interpolation so as to expand the 3x3x3 matrix of 27 values (underlined entries in Table 5) into a 5x5x5 matrix of 125 values (all entries in Table 5). The in-between levels were numbered 1.5 and 2.5 in order to preserve the numbering of the original levels.

Phase 4: Evaluating the YBRQL (Experiment 2)

We assessed three psychometric properties: *Construct validity* – does the YBRQL measure difficulty with binaural listening as intended? *Discriminative ability* – is it sensitive to differences between conditions, both within and between subjects? *Reproducibility* – how similar are repeated measures *(Agreement)* and how consistently do they distinguish patients *(Reliability)* (discussed further in Supplementary Digital Content 8). We then determined whether the YBRQL *avoided exaggerating losses*, while measuring *greater gains* in utility than the HUI3 and EQ-5D-3L and displaying sensitivity to a possible *negative side-effect.*

Participants

Participants were 28 adult users of cochlear implants, domiciled in the UK, who volunteered in response to an advertisement in the newsletter of the National Cochlear Implant Users Association. They were established users of cochlear implants who were willing to travel to participate in two days of testing. Eight (Unilateral Group, mean age 70.6 years, SD 17.6 years, 1 female) used a unilateral implant only (mean duration of use of implant 10.8 years, SD 3.6 years). Nine (Bimodal Group, mean age 64.3 years, SD 9.4 years, 6 female) used a unilateral implant with an acoustic hearing aid stimulating the contralateral ear (mean duration of use of implant 7.3 years, SD 3.4 years). Eleven (Bilateral Group, mean age 64.9 years, SD 8.0 years, 5 female) used bilateral cochlear implants (mean duration of use of first implant 12.4 years, SD 5.4 years; mean duration of use of second implant 8.4 years, SD 4.4 years). The first language of all participants was English.

Stage I

Phase 4 involved three stages. In Stage I, participants completed a battery of performance tests and self-report measures (Goman 2014) over the course of two days. On one day, they used their first or only implant and completed questionnaires while thinking about themselves using only one device. On the other day, the Unilateral Group was re-tested using their only implant, and the Bimodal and Bilateral Groups were tested using both of their devices and completed questionnaires while thinking about themselves using two devices. The assignment of number of devices to days was counterbalanced as far as possible given odd numbers of participants in some groups.

Performance tests were conducted with an AB-York Crescent of Sound (Kitterick et al. 2011) consisting of nine loudspeakers positioned at a height of 1.1m in a semicircle with a radius of 149cm.

Participants sat at the point equidistant from each of the loudspeakers which were located at ±90°, ±60°, ±30°, ±15°, and 0° azimuth. Zero degrees was directly in front of the participant; negative angles corresponded to locations to the left of straight ahead. Participants made responses on a touch screen.

To measure the ability to localise, the phrase "Hello what's this?", spoken by an adult female, was presented from one of the loudspeakers positioned, in different conditions, at (1) -60°, 0°, or +60°, (2) -60°, -30°, 0°, +30°,or +60°, and (3) -30°, -15°, 0°, +15°or +30°. The location was chosen quasi-randomly on each trial such that each location was used equally often over the course of 30 trials in each condition. The intensity of the phrase was roved from trial to trial to disrupt attempts to infer the location of the source from the loudness of the stimulus at one ear. Performance was scored as the percentage of trials over the three conditions on which the source loudspeaker was located correctly.

To measure the ability to identify speech in noise, the phrase "Point to the OBJECT", spoken by an adult female, was presented from the loudspeaker at 0° azimuth. OBJECT was randomly selected from the set "cow", "house", "cup", "duck", "fork", "horse", "key", "tree", "plane", "plate", "shoe", and "spoon". The task was to report the object named in the phrase. Pink noise was presented from the loudspeaker at -90° or +90° azimuth – whichever was on the side of the participant's first or only implant. This arrangement is maximally disadvantageous if hearing is monaural and maximally advantageous if hearing is binaural. An adaptive procedure estimated a speech-reception threshold (SRT) defined as the signal to noise ratio at which the accuracy of identifying the object correctly was 70.7%. The procedure was run twice and the two estimates of the SRT were averaged.

Participants completed the EQ-5D-3L, HUI3, and SSQ questionnaires on paper. Values of health utility were obtained by using the weighting functions described by Dolan (1997) for the EQ-5D-3L and Feeny et al. (2002) for the HUI3. Participants used a 0-10 visual-analogue scale to respond to each of the 50 questions in the SSQ. An single score was calculated by averaging the 50 ratings.

There are two reasons why binaural utility from the YBRQL might differ from health utility from the HUI3 in their relationship to other measures. First, the HUI3 is restricted by the ceiling effect identified in the Introduction. Second, the HUI3 is a measure of health, not only of hearing, and reflects the combined effect of poorer function on some attributes and better function on others. Therefore, to compare like with like, we made an additional comparison between the YBRQL and the *Hearing* attribute of the HUI3 alone, using the single-attribute utility function described by Feeny et al. (2002). There is one value for each level of the *Hearing* attribute. It was obtained from informants who valued each level while all other attributes were set to their highest level. The resulting values were then scaled to range from 1.00 (Level 1) to 0.00 (Level 6). We refer to these values as measures of *hearing* utility, whereas we refer to the multi-dimensional values from the HUI3 and EQ-5D-3L as measures of *health* utility.

Stage II

Twenty four months after Stage I, the YBRQL questionnaire was mailed to participants. The Unilateral Group completed the questionnaire once. The Bimodal and Bilateral Groups completed the questionnaire twice, first considering their functional abilities when using their first or only implant (Monaural Condition), and second when using both of their devices (Binaural Condition). For the monaural condition, the preamble instructed participants to “indicate which statement best describes your own hearing when using your cochlear implant” (Unilateral Group), “when using your cochlear implant on its own” (Bimodal Group) or “when using your first cochlear implant on its own” (Bilateral Group). For the binaural condition, the preamble instructed participants to “indicate which statement best describes your own hearing when using your cochlear implant together with your hearing aid” (Bimodal Group) or “when using your two cochlear implants together” (Bilateral Group).

A value of binaural utility from Table 5 was assigned to the participant according to the levels of function reported.

Stage III

Fourteen months after Stage II, the YBRQL questionnaire was mailed to participants who completed it while considering their functional abilities with their usual configuration of devices. A value of binaural utility from Table 5 was assigned to the participant according to the levels of function reported.

Interim summary: Outcome measures

In summary, we derived values of binaural utility from the YBRQL for comparison with two sets of measures. The first set consisted of the SSQ score and the performance measures of localization and speech perception in noise; we regarded these as *gold-standard* measures of binaural hearing. The second set consisted of hearing utility from the HUI3 and health utility from the HUI3 and EQ-5D-3L; we refer to these as *reference* measures. In addition, we calculated the *binaural advantage* for each measure as the difference between values obtained with one device and two devices by the Bimodal and Bilateral Groups.

Data Analysis

Analyses were conducted with IBM SPSS for Windows v. 26.0 (2019). If all measures in a set of related analyses distributed normally, parametric tests are reported. If one or more measures departed from normality, evidenced by significant Shapiro-Wilk tests, non-parametric tests are reported. Those reports are supplemented, if informative, by results of parametric tests for measures that distributed normally. Where there was no straightforward non-parametric alternative to a parametric test, such as the analysis of covariance, the results of the parametric test are reported. However, conclusions are drawn only if significant results are corroborated by confidence intervals estimated by bootstrapping (3,000 samples per analysis; bias corrected and accelerated).

Construct validity

Construct validity was assessed by calculating correlation coefficients between binaural utility from the YBRQL and the three gold-standard measures. First, correlations were calculated with data obtained when participants thought about, and performed with, their usual configuration of devices. Second, correlations were calculated between the measures of binaural advantage. Third, for comparison with the YBRQL, corresponding correlations were calculated between the gold-standard measures and the reference measures. Correlation coefficients were classified as small (.1 to <.3), moderate (.3 to <.5), large (.5 to <.7), or very large (≥.7). Non-parametric coefficients of at least moderate size, and parametric coefficients of at least large size, were taken as evidence of construct validity.

Discriminative ability

The ability of the YBRQL to discriminate between conditions was assessed by testing two hypotheses. The first was that binaural utility from the YBRQL was higher when members of the Bimodal and Bilateral groups used two devices rather than one. Effect sizes were classified as small (≥.2 to <.5), medium (.5 to <.8), or large (≥.8) (Cohen, 1988). The second hypothesis was that the binaural advantage measured by the YBRQL was greater for the Bilateral group than the Bimodal group. This hypothesis was tested in analyses of covariance (ANCOVAs) with the score with the first or only implant as a covariate to control for differences among participants at baseline (Vickers & Altman 2001). Corresponding analyses were conducted with each of the reference measures.

Reproducibility

Reproducibility was assessed for values of binaural utility from the YBRQL. *Agreement* was quantified as the standard error of measurement (SEm), calculated as the square root of the sum of the variance due to conditions (test and retest) and the residual variance due to the interaction between conditions and participants (de Vet et al. 2006). *Reliability* was quantified as the intra-class correlation coefficient (ICC) obtained when participants considered their functional abilities with their usual configuration of devices in Stages II and III. The ICC was calculated with a 2-way mixed-effects model assessing absolute agreement for single measures (McGraw & Wong 1996). Following Koo and Li (2016), values of the ICC were classified as poor (<.5), moderate (.5 to <.75), good (.75 to <.9), or excellent (≥.9). An ICC of at least good quality was required as evidence of reliability.

Losses of utility

To test whether the YBRQL exaggerated losses of utility, we compared the YBRQL with the HUI3 in terms of the loss relative to normal hearing when participants listened with their first or only implant. For the YBRQL, the loss of utility was calculated as the difference between the maximum attainable value, .96 (Table 5), and the observed value. For the HUI3, the measure was the difference between the health utility when the hearing dimension was set to its highest level while all other dimensions were at their observed levels, and the health utility when all dimensions were at their observed levels.

Gains in utility

We tested whether the YBRQL measured larger gains than the HUI3 and EQ-5D-3L by determining whether the binaural advantages shown by the Bimodal and Bilateral Groups were significantly larger when measured as changes in binaural utility than as changes in health utility.

Sensitivity to side effects

While serious and long-lasting complications associated with cochlear implantation are rare (e.g. Figure 3 in UKCISG 2004) once clinical programs are experienced (e.g. Cohen et al. 1993), increases on the pain/discomfort dimensions of generic PBMs have been reported in the short term (Summerfield & Barton 2019), possibly arising from the surgical wound or from the receiver-stimulator irritating the scalp. Also, while tinnitus is generally reduced by implantation, it may be exacerbated in some patients and induced in others (Ramakers et al. 2015) and is associated with changes on the Anxiety/Depression dimension of the EQ-5D (Summerfield & Barton 2019). These effects, although often mild and sub-clinical, have the potential to negate gains in utility associated with improved binaural hearing. The YBRQL might be insensitive to these side-effects. Alternatively, they might be associated with increased effort and fatigue.

The data of Experiment 2 permitted only a limited assessment of sensitivity to side effects in the form of a test of the sensitivity of the YBRQL to variation among participants in pain/discomfort whether related to implantation or not. We conducted an ANCOVA with binaural utility as the dependent variable, group as a fixed factor, and the levels of the *Hearing* and *Pain* dimensions of the HUI3 as covariates. The critical test was whether both covariates were significantly related to the dependent variable. For comparison, we applied the same test to health utilities estimated with the EQ-5D-3L.

Results

Response rate

Of 31 participants who took part in Stage 1, 28 returned the YBRQL questionnaire in Stage 2, and, of them, 25 returned it in Stage 3. Non-respondents had either died or were unwell.

Construct validity

The scatter plots in Figure 4 show the relationship between the SSQ score and measures of utility. The relationship with binaural utility from the YBRQL (Figure 4A) is characterised by a moderate-to-large non-parametric correlation (Kendall's τ = .501, N=28, p<.001) and by a large parametric correlation (Pearson's r = .680, N=28, p<.001). The relationship with hearing utility from the HUI3 (Figure 4B) is characterised by a moderate non-parametric correlation (τ = .463, N=28, p=.003). The relationship with health utility from the HUI3 (Figure 4C) is also characterised by a moderate non-parametric correlation (τ = .386, N=28, p<.01). No relationship was found with health utility from the EQ-5D-3L (τ = .215, N=28, p=.146) (Figure 4D).

Panels E, F, G, and H show the corresponding relationships among the measures of binaural advantage. Binaural advantage from the YBRQL was significantly associated with binaural advantage from the SSQ (τ = .492, N=20, p=.003; r = .752, N=20, p<.001). Binaural advantage from none of the three reference measures was significantly associated with binaural advantage from the SSQ.

----- Figure 4 : Scatterplots -----

Correlations between the gold-standard performance measures and measures of utility are listed in Table 6. Higher values of binaural utility from the YBRQL were associated with better speech perception (i.e. lower SRTs) and more accurate localisation. Higher values of health utility from the HUI3 were associated with better speech perception, but not with accuracy of localisation. There was no association between health utility from the EQ-5D-3L and either localisation or speech perception.

----- Table 6 : Correlations -----

Discriminative ability

Figure 5 contains box plots of the SSQ score and of measures of utility for listening with one device (open bars) and two devices (shaded bars) by the Bilateral and Bimodal Groups. The panel beneath the plots includes the results of Wilcoxon Signed Ranks tests which compared scores with one and two devices. The SSQ score and binaural utility from the YBRQL were significantly higher with two devices than one, while there were no significant differences in hearing utility from the HUI3 or in health utility from the HUI3 and EQ-5D-3L.

Effect sizes were calculated by dividing the *z-score* by the square root of the number of observations (Field 2009) and are tabulated as values of *R* at the bottom of Figure 5. Medium effect sizes were shown by the SSQ score and the measure of binaural utility, while either small (HUI3) or negligible (EQ-5D-3L) effect sizes were shown by the other measures of utility.

----- Figure 5 : Box plots -----

ANCOVAs tested the hypothesis that binaural advantage was greater for the Bilateral Group than the Bimodal Group. *Group* (Bilateral, Bimodal) was a fixed factor. *Score with the first or only implant* was a covariate. The effect of *Group* was significant for the SSQ score ($F_{(1,17)}$=8.933, p=.008, $\eta_p{}^2$=.344), for binaural utility from the YBRQL ($F_{(1,17)}$=10.742, p=.004, $\eta_p{}^2$=.387), but not for hearing utility from the HUI3 ($F_{(1,17)}$=.224, p=.642), nor for health utility from the HUI3 ($F_{(1,17)}$=1.161, p=.296) or the EQ-5D-3L ($F_{(1,17)}$=.210, p=.653). This pattern of significance was corroborated by bootstrapped estimates of the 95% confidence intervals of the binaural advantages (taken as the estimated marginal means) and their difference (taken as the parameter for *Group* in the general linear model) (Table 7).

----- Table 7 : Discriminative ability -----

The box plots in Figure 6 show the extent to which levels of function on the individual dimensions of the YBRQL differed between monaural and binaural conditions. The Localization dimension showed a significant improvement for the Bimodal Group (Wilcoxon Signed Ranks Test, z=2.271, p<.05), while all three dimensions showed significant improvements for the Bilateral Group (SpiN, z=2.701, p<.01; Loc, z=2.579, p<.01; E&F, z=2.565, p<.05).

----- Figure 6 : Contributions of individual dimensions ----

Reproducibility

Mean values of average utility from the YBRQL differed minimally between test, .834 (SD = .068), and retest, .835 (.057). The SEm was .03. The ICC (with 95% confidence interval) was .818 (.628 to .916).

Losses of utility

The upper part of Table 8 lists median values of the losses of utility attributable to impaired hearing from the YBRQL and HUI3 when participants used their first or only implant. The losses did not differ significantly (Wilcoxon Signed Ranks Test: N=28, z=1.586, exact p=.115, 2-tailed). There is no evidence, therefore, that the YBRQL exaggerated losses in comparison with the HUI3.

----- Table 8 : Losses and gains in utility -----

Gains in utility

The lower part of Table 8 lists median values of the binaural advantages recorded by the YBRQL, HUI3, and EQ-5D-3L. For the Bimodal Group, the advantage measured with the YBRQL was not greater than the advantage measured with the EQ-5D-3L (Wilcoxon Signed Ranks Test: z=.169, N=9, exact p=.469, 1-tailed) or the HUI3 (z=.280, exact p=.422, 1-tailed). For the Bilateral Group, the advantage measured with the YBRQL was greater than the advantage measured with the EQ-5D-3L (N=11, z=2.312, exact p=.009, 1-tailed), and fell short of significance in comparison with the advantage measured with the HUI3 (z=1.689, exact p=.051, 1-tailed).

Sensitivity to side effects

Binaural utility from the YBRQL when participants used their normal configuration of devices was analysed in an ANCOVA with the fixed factor *Group* (Unilateral, Bimodal, Bilateral) and the covariates *Level of HUI3 Hearing* (Hearing) and *Level of HUI3 Pain* (Pain). There were significant effects of *Group* ($F_{2,23}$=7.830, p<.01, η_p^2=.405) and *Hearing* ($F_{1,23}$=9.111, p<.01, η_p^2=.284), but not of *Pain* ($F_{1,23}$=.154, n.s., η_p^2=.007). In comparison, an analogous analysis of health utility from the EQ-5D-3L showed the reciprocal relationship: a significant effect of *Pain* ($F_{1,23}$=13.154, p<.01, η_p^2=.364), but not of *Group* ($F_{2,23}$=1.493, n.s., η_p^2=.115) or *Hearing* ($F_{1,23}$=.923, n.s., η_p^2=.039). Thus, the measure of binaural utility from the YBRQL distinguishes patients according to differences in their ability to hear but not according to differences in the pain/discomfort which they report. The fact that the measure of health utility from the EQ-5D-3L is sensitive to differences in pain/discomfort shows that the insensitivity of the YBRQL was a true limitation rather than the result of negligible variation in pain/discomfort. Further analyses assessing the sensitivity of the YBRQL to pain/discomfort are reported in Supplementary Digital Content 2.

Discussion

Psychometric properties

Binaural utility measured with the YBRQL displayed good construct validity (Figure 4; Table 6) and discriminative ability (Figure 5; Table 7) with all three dimensions contributing (Figure 6). In these respects, it performed similarly to the SSQ, better than the HUI3, and much better than the EQ-5D-3L. It also demonstrated good reproducibility over a 14-month interval between test and retest. The value of the SEm, .03, measuring agreement, compares favourably with values for the EQ-5D and HUI3 (Palta et al. 2011) which were close to .10 (at an interval of 6 months by 250 patients following cataract surgery). The value of the ICC (.82, .63 to .92), measuring reliability, is similar to values reported for the generic PBMs (EQ-5D-3L at an interval of 3 months by 224 patients with rheumatoid arthritis: .82, .74 to .88; Macran, 2003; HUI3 at an interval of 1 month by 506 members of the general public: .77; Boyle et al. 1995; and at an interval of 3 months by 141 patients with hip fracture: .77; Jones et al. 2005).

Other statistical properties

Binaural utility avoided floor and ceiling effects and displayed acceptable effect sizes when contrasting binaural with monaural hearing (Figures 4A and 5). With a finely graded set of response options to choose from, 17 of the 20 members of the Bimodal and Bilateral Groups reported an improvement in binaural utility when using two devices rather than one. In contrast, hearing utility measured with the HUI3 showed the ceiling effect anticipated in the Introduction (Figures 4B and 5); seventeen of 28 participants reported that they functioned at ceiling (Level 3) when using their first or only implant and only 6 of the 20 members of the Bimodal and Bilateral Groups reported an improvement when using two devices rather than one. The same limitations restricted the sensitivity of health utility from the HUI3 to binaural hearing (Figures 4C and 5), while the sensitivity of health utility from the EQ-5D-3L was restricted because more than half of the participants reported levels of function corresponding to full health when using one device (Figures 4D and 5).

General Discussion

There are four respects in which it is relevant to assess the YBRQL: Its potential as a clinical outcome measure; its legitimacy for, and its practical use in, informing resource-allocation decisions; limitations that arose in its development; and desirable further developments.

Potential as a clinical outcome measure

The levels of function reported by a participant with the YBRQL questionnaire describe a profile across three fundamental dimensions of binaural hearing: speech perception in noise, localization, and the effort and fatigue entailed in listening. The value of binaural utility that can be derived from the profile provides a summary index of binaural function. The index ranges in value from .69 to .96. The lower value corresponds to great difficulties on all three dimensions; the higher value corresponds to normal performance on each dimension. The index results from the mental integration of the impact of the difficulties by independent informants. The index displays good psychometric properties of construct validity, discriminative ability, and reproducibility. Considered overall, the YBRQL provides a disciplined summary of one way of estimating the contribution of binaural hearing to quality of life. It would be an efficient and sensitive clinical outcome measure in studies that require a brief questionnaire that provides a quantitative summary of binaural function that correlates robustly with performance measures and the overall score from the longer SSQ.

Legitimacy for informing resource-allocation decisions

Legitimacy hinges on whether the YBRQL meets the criteria set out by Versteegh et al. (2012), including demonstrating larger gains in utility than are shown by generic PBMS, while avoiding exaggeration of losses and displaying sensitivity to side effects.

Assessment against criteria

The first criterion proposed by Versteegh et al. (2012) is that *empirical evidence disproves the sensitivity of existing generic instruments.* Previous studies (Table 1) and Experiment 2 (Tables 6 and 7; Figures 4 and 5) show that the EQ-5D-3L is largely insensitive to binaural hearing while the HUI3 displays restricted sensitivity. The first criterion is partially met, therefore. The second criterion is that *empirical evidence proves the superiority of the condition-specific instrument from which the condition-specific PBM was derived.* The YBRQL did not draw questions from a pre-existing instrument. However, its dimensions align approximately with the three main sections of the SSQ. The SSQ is sensitive to binaural hearing in conditions where generic PBMs are either insensitive (Summerfield et al. 2006; Smulders et al. 2016) or less sensitive (Figure 5; Table 8). To that extent, the second criterion is met.

The third criterion is that the *derived condition-specific PBM is shown to be superior to the existing generic PBM(s), not just in terms of statistical sensitivity, but also in terms of absolute differences* (Reference Note 1). The YBRQL meets the 'statistical sensitivity' part of the criterion. It displayed larger effect sizes than the HUI3 or EQ-5D-3L when comparing monaural with binaural hearing (Figure 5). It displayed a consistent pattern of significant differences when monaural and binaural hearing were contrasted both within and between groups, whereas the HUI3 displayed an inconsistent pattern and the EQ-5D-3L showed no significant differences (Table 7). However, the YBRQL failed the 'absolute differences' part of the criterion, in that the binaural advantages measured by the YBRQL for the Bimodal and Bilateral groups were not significantly larger than the advantages measured by the HUI3.

This last result is important, given that the motivation for developing the YBRQL was concern about the lack of sensitivity of the HUI3 to bilateral implantation. Note, however, that the difference between the binaural advantage for the Bilateral Group measured with the YBRQL (median=.107, IQR .035 to .162) and the HUI3 (.000, .000 to .170) fell only marginally short of significance (p=.051 when assessed with a 1-tailed Wilcoxon Signed Ranks test). Further analyses (Supplementary Digital Content 2) confirm that the significance of the difference fell in the region of uncertainty and that a more powerful study is required to resolve the issue.

Exaggeration of losses

The aim of developing the YBRQL was to avoid limitations in the sensitivity of the HUI3 to interventions which improve binaural hearing, rather than to increase sensitivity to impaired hearing. Appropriately, therefore, the YBRQL did not exaggerate the size of losses of utility due to impaired hearing when compared with the HUI3.

Sensitivity to side effects

Experiment 2 demonstrated that the YBRQL is sensitive to differences in the configuration of implants and acoustic hearing aids. Accordingly, it would be expected also to be sensitive to complications, such as the malfunctioning or sub-optimal fitting of an implant or hearing aid, that reduce binaural function. However, it is not clear whether the YBRQL is also sensitive to other side-effects, such as changes in pain/discomfort (Summerfield & Barton 2019) and tinnitus (Ramakers et al. 2015), whose impact may be orthogonal to binaural function. Experiment 2 showed that the YBRQL is insensitive to variation among participants in pain/discomfort, whether related to implantation or not. That may not matter, given that pain is reported to be an unusual long-term consequence of cochlear implantation (Celerier et al. 2017) and at least one study has reported no worsening of pain, or any other attribute of the HUI3, associated with bilateral implantation (Table 4 in Bichey & Miyamoto 2008). On the other hand, two members of the Bilateral Group in Experiment 2 reported greater pain/discomfort when using two implants compared with one (Supplementary Digital Content 2), and it would be unwise to assume that any intervention which entails surgery is immune to negative side effects. Further research is needed to establish the sensitivity, or otherwise, of the YBRQL to side effects.

Role of the YBRQL

Where do these assessments leave the YBRQL? One option would be to restrict its role to providing a measure of clinical effectiveness, while the HUI3 informs analyses of cost-effectiveness. However, that strategy is undermined by the evidence in Table 1 which motivated the development of the YBRQL, by the intrinsic limitations in the sensitivity of the HUI3 to binaural hearing identified in the Introduction and in Experiment 2, and by further limitations discussed in Supplementary Digital Content 2.

The second option would be to employ the YBRQL to supplement evidence provided by the EQ-5D or HUI3. Consider that three willingness-to-pay thresholds are relevant to resource-allocation decisions. The first is a threshold below which an intervention is unlikely to be rejected on grounds of cost-ineffectiveness. The second is a higher threshold above which special reasons are required if an intervention is to be accepted. Implicitly, there is a third, yet higher, threshold above which interventions are never accepted. The lower two thresholds may be set explicitly (e.g. Rawlins & Culyer, 2004; NICE, 2013) or inferred (Neumann et al. 2014; Dubois 2016; Cameron et al. 2018). The third threshold is necessarily inferred (e.g. Dakin et al. 2015). The range between the first and third thresholds defines a region of uncertainty. Figure 7 is a matrix of decisions that could be reached by using the YBRQL to estimate the cost-effectiveness of binaural-related gains in quality of life alongside the EQ-5D or HUI3 to estimate the cost-effectiveness of health-related gains in quality of life. The cells correspond to the nine possible combinations of two incremental cost-effectiveness ratios each of which may be more favourable than the first threshold, or in the region of uncertainty, or above the third threshold. Of particular relevance is Cell 3, where evidence from the YBRQL might prompt further investigation before accepting evidence of cost-effectiveness from the generic PBM, and Cell 4, where evidence from the YBRQL might prompt the judgement that an intervention was cost-effective despite uncertainty in the evidence from the generic PBM.

----- Figure 7 : Decision matrix -----

Limitations

Experiments 1 and 2 have limitations. First, although the choice of dimensions for the YBRQL was principled, the inclusion of additional dimensions informed by the ICF (WHO 2001) or relevant questionnaires (e.g. Hinderink et al. 2000) might result in greater sensitivity to improvements in binaural hearing. Second, the informants who contributed to the valuation of the YBRQL in Experiment 1 were a convenience sample rather than a population-representative sample and included relatively few older people. Third, the evaluation of the YBRQL in Experiment 2 was limited to interventions involving cochlear implants. Fourth, the value set for the YBRQL was obtained from informants domiciled in the UK. Valuations of health states differ between countries (e.g. Bernert et al. 2009; Gerlinger et al. 2019), at least partly because different populations place different values on the same aspects of health depending on the context in which healthcare is provided. For those reasons, policy-makers prefer decisions to be informed by preference data from their own population (e.g. NICE 2013). Researchers outside the UK considering using the YBRQL in partnership with a generic PBM to inform resource-allocation decisions might choose to generate a value set from a sample of their own population.

The fifth limitation is that, while informing analyses of the psychometric properties of the YBRQL, Experiment 2 was under-powered to detect differences between gains in utility measured with different PBMs, given the wide dispersion of utilities measured with the HUI3 and EQ-5D-3L (Figure 5), and was not designed to examine the impact of side effects. It would be desirable to recruit a larger sample of users of unilateral implants wo received a second implant, and who were tested prospectively, to address those issues. If the resolution was not satisfactory, it would be desirable to integrate the YBRQL with a generic PBM by adding dimensions to the generic PBM assessing *Speech-perception in noise* and *Localisation*. 'Bolting on' additional dimensions has been advocated where the EQ-5D is insensitive to a condition and to treatments that alleviate it (e.g. Krabbe et al. 1999; Yang et al. 2015; Finch et al. 2017). Sensitivity to a condition is enhanced, while avoiding exaggeration of losses and preserving the capacity to register side-effects. Although that approach did not yield systematic valuations when questions asking directly about the ability to 'hear' were bolted on (Yang et al. 2015), the present results demonstrate that informants are willing to trade, and produce systematic results, when questions more fully exemplify the aspects of listening which are at stake.

Conclusion

The York Binaural-related Quality of Life System provides a measure of the contribution of binaural hearing to the quality of life of adult listeners. It would be an efficient clinical outcome measure in the evaluation of interventions intended to improve binaural hearing. At its present stage of development, it should not be used on its own to inform resource-allocation decisions. However, if administered in conjunction with a generic PBM, it could beneficially modulate confidence in the cost-effectiveness of interventions. To that end, Supplementary Digital Content 9 contains the YBRQL questionnaire as an editable document which can be tailored to a particular application.

Acknowledgments

We thank Rhian Bardsley, Stephanie Bill, Sarah-Louise Buggins, Danielle Dickinson, Abigail Llewelyn, Alice Smith, Rachel Williamson, and Anna Wilson who contributed to the development of the YBRQL questionnaire and who gathered the time trade-off valuations in Experiment 1.

Author contributions

AQS, PTK, and AMG designed the experiments. AQS and AMG supervised data-gathering for Experiment 1. AMG and PTK gathered the data for Experiment 2. AQS, PTK, and AMG analysed the data. AQS and AMG wrote the paper.

References

Attema, A.E., Edelaar-Peeters, Y., Versteegh, M.M., et al. (2013). Time trade-off: one methodology, different methods. *Eur J Health Econ, 14(Suppl 1)*, S53-S64.

Barton, G.R., Bankart, J., Davis, A.C., et al. (2004). Comparing utility scores before and after hearing-aid provision. *Applied Health Economics and Health Policy, 3*, 103-105.

Bernert, S., Fernández, A., Haro, J.M., et al. (2009). Comparison of different valuation methods for population health status measured by the EQ-5D in three European countries. *Value in Health, 12,* 750-758.

Bichey, B.G., Miyamoto, R.T. (2008). Outcomes in bilateral cochlear implantation. *Otolaryngology—Head and Neck Surgery, 138,* 655-661.

Bond, M., Mealing, S., Anderson, R., et al. (2009). The effectiveness and cost-effectiveness of cochlear implants for severe to profound deafness in children and adults: a systematic review and economic model. *Health Technol Assess, 13,* 1-330.

Boyle, M. H., Furlong, W., Feeny, D., et al. (1995). Reliability of the Health Utilities Index-Mark III used in the 1991 Cycle 6 Canadian General Social Survey Health Questionnaire. *Quality of Life Research 4,* 249-257.

Boys Town National Research Hospital (2020). Statement on insurance coverage of bilateral cochlear implantation. Accessed at https://www.boystownhospital.org/knowledge-center/bilateral-cochlear-implants on 8th June 2020.

Brazier, J., Czoski-Murray, C., Roberts, J., et al. (2008). Estimation of a preference-based index from a condition-specific measure: the King's Health Questionnaire. *Med Decis Making, 28,* 113-126.

Brazier, J., Tsuchiya, A. (2010). Preference-based condition-specific measures of health: what happens to cross programme comparability? *Health Economics, 19,* 125-129.

Brooks, R., Rabin, R., de Charro, F. (2003). *The Measurement and Valuation of Health Status using EQ-5D: A European Perspective.* Dordrecht: Kluwer Academic Publishers.

Byrne, D., Noble, W. (1998). Optimizing sound localization with hearing aids. *Trends in Amplification, 3, 51-73.*

Cameron D., Ubels, J., Norstrom, F. (2018). On what basis are medical cost-effectiveness thresholds set? Clashing opinions and an absence of data: a systematic review. *Global Health Action, 11*, 1-14.

Celerier, C., Rouillon, I., Blanchard, M. et al. (2017). Pain and cochlear implantation: An unusual complication? *Otol Neurotol, 38, 956-961.*

Cohen J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Cohen, N.L., Waltzman, S.B., Fisher, S.G., et al. (1993). A prospective randomized study of cochlear implants. *New England Journal of Medicine, 328,* 233-237.

Ching, T.Y.C., Incerti, P., Hill, M. (2004). Binaural benefits for adults who use hearing aids and cochlear implants in opposite ears. *Ear Hear, 25, 9-21.*

Dakin, H., Devlin, N., Feng, Y., et al. (2015). The influence of cost-effectiveness and other factors on NICE decisions. *Health Econ, 24,* 1256-1271.

Dawes, P., Munro, K.J. (2013). Unilateral and bilateral hearing aids, spatial release from masking and auditory acclimatization. *J Acoust Soc Am., 134, 586-606.*

de Vet, H.C.W., Terwee, C.B., Knol, D.L., et al. (2006). When to use agreement versus reliability measures. *J Clin Epid., 59, 1033-1039.*

Dolan, P. (1997). Modeling valuations for EuroQol health states. *Md Care, 35,* 1095-1108.

Dolan, P., Roberts, J. (2002). To what extent can we explain time trade-off values from other information about respondents? *Social Science & Medicine, 54,* 919-929.

Drummond, M.F., O'Brien, B., Stoddart, G. L., et al. (2000). *Methods for the Economic Evaluation of Health Care Programmes*, 2nd Edition. Oxford: Oxford University Press.

Dubois, R.W. (2016). Cost-effectiveness thresholds in the USA: are they coming? Are they already here? *J Comp Eff Res, 5,* 9-11.

Feeny, D., Furlong, W., Torrance, G.W., et al. (2002). Multi-attribute and single-attribute utility functions for the Health Utilities Index Mark 3 system. *Medical Care, 40,* 113-128.

Field, A. (2009). *Discovering Statistics Using SPSS,* 3rd Edition. London: Sage.

Finch, A.P., Brazier, J.E., Makuria, C., et al. (2017). An exploratory study on using principal-component analysis and confirmatory factor analysis to identify bolt-on dimensions: the EQ-5D case study. *Value in Health, 20,* 1362-1375.

Gatehouse, S., Noble, W. (2004). The Speech, Spatial and Qualities of Hearing Scale (SSQ). *Int J Audiol, 43,* 85-99.

Gerlinger, C., Bamber, L., Leverkus, F., Schwenke, C., Haberland, C., Schmidt, G., Endikrat, J. (2019). Comparing the EQ-5D-5L utility index based on value sets of different countries: impact on the interpretation of clinical study results. *BMC Res Notes 12:18.*

Goman, A.M. (2014). *A comparison of bilateral cochlear implantation and bimodal aiding in severely-profoundly hearing-impaired adults: head movements, clinical outcomes, and cost-effectiveness.* Doctoral Thesis, University of York, UK.

Grutters, J.P.C., Joore, M.A., van der Horst, F., et al. (2007). Choosing between measures: comparison of EQ-5D, HUI2 and HUI3 in persons with hearing complaints. *Quality of Life Research, 16,* 1439-1449.

Härkönen, K., Kivekös, I., Rautiainen, M. et al. (2015). Sequential bilateral cochlear implantation improves working performance, quality of life, and quality of hearing. *Acta Oto-Laryngologica, 135, 440-446.*

Health Quality Ontario (2018). Bilateral cochlear implantation: a health technology assessment. *Ont Health Technol Assess Ser [internet], 18,* 1-139.

Hinderink, J.B., Krabbe, P.F.M., van den Broek, P. (2000). Development and application of a health-realted quality of life instrument for adults with cochlear implants: The Nijmegen Cochlear Implant Questionnaire. *Otolaryngology – Head and Neck Surgery, 123, 756-765.*

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.

Hughes, K.C., Galvin, K.L. (2013). Measuring listening effort expended by adolescents and young adults with unilateral or bilateral cochlear implants or normal hearing. *Cochlear Implants Int., 14, 121-129.*

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

Jones, C. A., Feeny, D., Eng, K. (2005). Test-retest reliability of health utilities index scores: evidence from hip fracture. *Int J Technol Assess Health Care, 21,* 393-398.

Kerber, S., Seeber, B.U. (2012). Sound localization in noise by normal-hearing listeners and cochlear implant users. *Ear Hear, 33, 445-457.*

Kitterick, P.T., Lovett, R.E.S., Goman, A.M. et al. (2011). The AB-York crescent of sound: an apparatus for assessing spatial-listening skills in children and adults. *Cochlear Implants Int., 12, 164-169.*

Koo, T.K., Li, M.Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15, 155-163.*

Krabbe, P.F.M., Stouthard, M.E.A., Essink-Bot, M-L., et al. (1999). The effect of adding a cognitive dimension to the EuroQol Multiattribute Health-status classification system. *J Clin Epidemiol, 52,* 293-301.

Kuthubutheen, J., Mittmann, N., Amoodi, H., et al. (2015). The effect of different utility measures on the cost-effectiveness of bilateral cochlear implantation. *Laryngoscope, 125,* 442-447.

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

Lammers, M.J.W., Grolman, W., Smulders, Y.E., et al. (2011). The cost-utility of bilateral cochlear implantation: a systematic review. *Laryngoscope, 121,* 2604-2609.

Litovsky, R., Parkinson, A., Arcaroli, J. et al. (2006). Simultaneous bilateral cochlear implantation in adults: a multicenter clinical study. *Ear Hear, 27, 714-731.*

Longworth, L., Yang, Y., Young, T., et al. (2014). Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: a systematic review, statistical modelling and survey. *Health Technol Assess, 18,* 1-254.

Luo, N., Johnson, J.A., Shaw, J.W., et al. (2009). Relative efficiency of the EQ-5D, HUI2, and HUI3 index scores in measuring health burden of chronic medical conditions in a population health survey in the United States. *Medial Care, 47,* 53-60.

Macran, S. (2003). Test-retest performance of EQ-5D. In R. Brooks, R. Rabin, F. de Charro (Eds.), *The Measurement and Valuation of Health Status Using EQ-5D: A European Perspective* (pp. 43-54). Dordrecht, NL: Kluwer Academic Publishers.

McGraw, K.O., Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods* 1, 30-46.

National Institute for Health and Care Excellence (2013). *Guide to the Methods of Technology Appraisal 2013. Process and Methods Guides No. 9.* London: National Institute for Health and Care Excellence.

Neumann, P.J., Cohen, J.T., Weinstein, M.C. (2014). Updating cost-effectiveness – the curious resilience of the $50,000-per-QALY threshold. *New England Journal of Medicine, 371,* 796-797.

Noble, W., Gatehouse, S. (2006). Effects of bilateral versus unilateral hearing aid fitting on abilities measured by the Speech, Spatial, and Qualities of Hearing Scale (SSQ). *Int J Audiol., 45, 172-181.*

Noble, W., Tyler, R., Dunn, C., et al. (2008). Unilateral and bilateral cochlear implants and the implant-plus-hearing-aid profile: comparing self-assessed and measured abilities. *Int J Audiol., 47, 505-14.*

Office for National Statistics (2015). Period life expectancy by sex and 5-year age group, England, Wales and regions of England 2011 to 13 and 2012 to 2014. Downloaded from https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/adhocs/005097periodlifeexpectancybysexand5yearagegroupenglandwalesandregionsofengland2011to13and2012to2014 on 10th September 2019.

Oppe, M., Rand-Hendriksen, K., Shah, K., et al. (2016). EuroQol protocols for time trade-off valuation of health outcomes. *PharmacoEconomics, 34,* 993-1004.

Palmer, C.S., Niparko, J.K., Wyatt, J.R., et al. (1999). A prospective study of the cost-utility of the multichannel cochlear implant. *Arch Otolaryngol Head Neck Surg, 125,* 1221-1229.

Palta, M., Chen, H-Y., Kaplan, R.M., et al. (2011) Standard error of measurement of 5 health utility indexes across the range of health for use in estimating reliability and responsiveness. *Med Decis Making, 31, 260-269.*

Potts, L.G., Skinner, M.W., Litovsky, R.A., et al. (2009). Recognition and localization of speech by adult cochlear implant recipients wearing a digital hearing aid in the nonimplanted ear (bimodal hearing). *J Am Acad Audiol., 20, 353-373.*

Ramakers, G.G.J., van Zon, A., Stegeman, I., et al. (2015). The effect of cochlear implantation on tinnitus in patients with bilateral hearing loss: A systematic review. *Laryngoscope, 125, 2584-2592.*

Rawlins, M.D., Culyer, A.J. (2004). National Institute for Clinical Excellence and its value judgements. *BMJ, 329,* 224-227.

Rennies, J., Kidd, G. (2018). Benefit of binaural listening as revealed by speech intelligibility and listening effort. *J Acoust Soc Am., 144, 2147-2159.*

Smulders, Y.E. (2016) *Unilateral versus simultaneous bilateral cochlear implantation in adults; a randomized controlled trial.* Doctoral Thesis, University of Utrecht, The Netherlands.

Smulders, Y.E., van Zon, A., Stegeman, I., et al. (2016). Cost-utility of bilateral versus unilateral cochlear implantation in adults: a randomized controlled trial. *Otol Neurotol, 37,* 38-45.

Summerfield, A.Q., Barton, G.R. (2019). Sensitivity of EQ-5D-3L, HUI2, HUI3, and SF-6D to changes in speech perception and tinnitus associated with cochlear implantation. *Qual Life Res, 28,* 1145-1154.

Summerfield, A.Q., Barton, G.R., Toner, J., et al. (2006). Self-reported benefits from successive bilateral cochlear implantation in post-lingually deafened adults: randomised controlled trial. *International Journal of Audiology, 45(Supplement 1),* S99-S107.

Summerfield, A.Q., Marshall, D.H., Barton, G.R., et al. (2002). A cost-utility scenario analysis of bilateral cochlear implantation. *Archives of Otolaryngology Head Neck Surgery*, *128,* 1255-1262.

Swinburn, P., Lloyd, A., Boye, K.S., et al. (2013). Development of a disease-specific version of the EQ-5D-5L for use in patients suffering from psoriasis: lessons learned from a feasibility study in the UK. *Value in Health, 16,* 1156-1162.

Theriou, C., Fielden, C.A., Kitterick, P.T. (2019). The cost-effectiveness of bimodal stimulation compared to unilateral and bilateral cochlear implant use in adults with bilateral severe to profound deafness. *Ear Hear, 40,* 1425-1436.

UK Cochlear Implant Study Group (2004). Criteria of candidacy for unilateral cochlear implantation in postlingually deafened adults II: cost-effectiveness analysis. *Ear Hear, 25,* 336-360.

Versteegh, M. M., Leunis, A., Uyl-de Groot, C. A., et al. (2012). Condition-specific preference-based measures: benefit or burden? *Value in Health, 15,* 504-513.

Vickers, A.J., Altman, D.G. (2001). Analysing controlled trials with baseline and follow up measurements. *BMJ, 323,* 1123-1124.

Vickers, D., De Raeve, L., Graham, J. (2016). International survey of cochlear implant candidacy. *Cochlear Implants International, 17:sup 1,* 36-41.

World Health Organization (2001). *International Classification of Functioning Disability and Health.* WHO: Geneva.

Yang, Y., Brazier, J.E., Tsuchiya, A., et al. (2011). Estimating a preference-based index for a 5-dimensional health state classification for asthma derived from the Asthma Quality of Life Questionnaire. *Med Decis Making, 31,* 281-291.

Yang, Y., Rowan, D., Brazier, J., et al. (2015). An exploratory study to test the impact of three “bolt-on” items to the EQ-5D. *Value in Health, 18,* 522-60.

York Health Economics Consortium (2016). Time Trade-Off. Accessed at https://yhec.co.uk/glossary/time-trade-off/ on 1st January 2021.

Reference Note

Reference Note 1. On p. 513 of Versteegh et al (2012), the wording of the third criterion should be ‘shown to be superior to the existing generic PBM’ rather than ‘shown to be superior to the existing CS-PBM’ (Matthias Versteegh, personal communication, 7th September 2018).

Figure Legends

Figure 1 Relationship between mean utility and level of dimensions in Experiment 1a (upper panels) and 1b (lower panels). Each symbol plots the mean utility for one level of a dimension (filled symbols, non-students; open symbols, students).

Figure 2 Influence of dimensions in Experiment 1a (upper panel) and 1b (lower panel). Filled bars, non-students; open bars, students. Error bars plot 95% confidence intervals of mean values.

Figure 3 YBRQL Questionnaire. The preamble in this example is worded for users of bilateral cochlear implants.

Figure 4 Relationship between SSQ score and utility when participants considered their functional abilities using their normal configuration of devices (upper row of panels). Relationship between change in SSQ score and change in utility between monaural and binaural listening for members of the bimodal and bilateral groups (lower row of panels). (Panels A and E, YBRQL; B and F, HUI3 Hearing; D and G, HUI3 Health; D and H, EQ-5D-3L). (Black circles, Unilateral Group; dark grey squares, Bimodal Group; light grey triangles, Bilateral Group).

Figure 5 Box plots of SSQ score (Panels A and C) and utility (Panels B and D) when the Bilateral Group (Panels A and B) and the Bimodal Group (Panels C and D) considered their functioning with their first or only implant (open boxes) and with two devices (filled boxes). Whiskers mark 10th and 90th percentiles; box marks 25th and 75th percentiles; heavy line within box marks 50th percentile; filled circles plot outliers beyond the 10th and 90th percentiles. (Panel E) z scores, significance levels (p), and effect sizes (R) from Wilcoxon signed ranks tests comparing scores with 1 and 2 devices.

Figure 6 Box plots of levels of function on each dimension of the YBRQL when the Bimodal and Bilateral Groups considered their function using their first or only implant (open boxes) or two devices (filled boxes). Results of Wilcoxon signed ranks tests comparing reported levels with one and two devices (*, p<.05; **, p<.01).

Figure 7 Decision matrix for interpreting binaural incremental cost-effectiveness ratios (ICERs) informed by the YBRQL together with health ICERs informed by a generic PBM.

Table Legends

Table 1 Estimates of the gain in health utility between unilateral and bilateral cochlear implantation in bilaterally-impaired post-lingually deafened adults obtained with the EuroQol Descriptive System (EQ-5D-3L), the Health Utilities Index Mark III (HUI3), and versions of the Time Trade-off technique (TTO).

Table 2 Vignettes describing levels of difficulty on three dimensions of binaural hearing.

Table 3 Numbers (N) and age of participants in Experiments 1a and 1b.

Table 4 Calculation of age-standardized binaural utilities.

Table 5 YBRQL value set. Entries are the 125 values of the expanded value set. Underlined entries are the 27 values of the initial value set.

Table 6 Kendall's coefficients of correlation among outcome measures (N=28). (**p*<.05, ***p*<.01).

Table 7 Discriminative ability: Measures of binaural advantage for the Bimodal and Bilateral Groups, and of the difference in binaural advantage (Bilateral – Bimodal), with 95% confidence intervals (95% CI) estimated by bootstrapping (3,000 samples per estimate, bias-corrected and accelerated).

Table 8 Estimates of the loss of utility due to impaired hearing and the gain in utility from using a second device estimated. (IQR: inter-quartile range.)

Supplementary Digital Content

1. SDC 1 (Gain in HRQL associated with bilateral cochlear implantation).pdf
2. SDC 2 (Levels of function reported with the HUI3 and the YBRQL).pdf
3. SDC 3 (Response booklet for Experiment 1a).pdf
4. SDC 4 (Response booklet for Experiment 1b).pdf
5. SDC 5 (Pilot Experiment and test of equivalence).pdf
6. SDC 6 (Response booklet for Pilot Experiment).pdf
7. SDC 7 (Relationship between age and binaural utility).pdf
8. SDC 8 (Tests of reproducibility).pdf
9. SDC 9 (Editable version of the YBRQL questionnaire).docx

Table 1 Estimates of the gain in health utility between unilateral and bilateral cochlear implantation in bilaterally-impaired post-lingually deafened adults obtained with the EuroQol Descriptive System (EQ-5D-3L), the Health Utilities Index Mark 3 (HUI3), and versions of the Time Trade-off technique (TTO).

Measure	Study	Design	Number of participants	Mean increment in utility (95% CI)[a]	Significance of increment
EQ-5D-3L	Summerfield et al. (2006)	Randomised controlled trial[b]	2x12	-.006 (-.091 to .078)[c]	n.s.
			24	-.063(-.120 to .005)[d]	n.s.
	Kuthubutheen et al. (2014)	Scenario analysis[e]	142	.04	p<.05
	Smulders et al. (2016)	Randomised controlled trial[f]	2x19	-.02	n.s.
HUI3	Summerfield et al. (2006)	Randomised controlled trial[b]	2x12	.105 (-.073 to .282)[c]	n.s.
			24	-.015 (-.110 to .079)[d]	n.s.
			24	.030 (-.045 to .104)[g]	n.s.
	Bichey & Miyamoto (2008)	Retrospective study[h]	23	.12 (.09 to .14)	p<.001
	Kuthubutheen et al. (2014)	Scenario analysis[e]	142	.035	p<.05
	Smulders et al. (2016)	Randomised controlled trial[f]	2 x 19	.04	n.s.
TTO	Summerfield et al. (2002)	Scenario analysis[i]	70	.03 (.02 to .04)	p<.05
	Kuthubutheen et al. (2014)	Scenario analysis[j]	142	.12	p<.05
	Smulders et al. (2016)	Randomised controlled trial[k]	2x19	.09	n.s.

[a] CI = Confidence Interval where reported or calculable from published data.
[b] 24 users of one implant were randomised to receive a second implant either immediately (treatment group, N=12) or after 12 months (control group, N=12).
[c] Comparison of groups 9 months after treatment group received second implant.
[d] Before-and-after comparison of combined group 9 months after each patient had received second implant.
[e] Scenario analysis of unilateral and bilateral implantation; informants were candidates for implantation (N=30), users of one implant (N=30), users of two implants (N=30), and clinicians (N=52).
[f] 38 candidates for implantation were randomised to receive either bilateral implantation (treatment group, N=19) or unilateral implantation (control group, N=19).
[g] Before-and-after comparison of combined group 9 months after each patient had received second implant with control for changes in annoyance due to tinnitus.
[h] Retrospective assessment of health status with one implant, and contemporary assessment of health status with two implants, by users of bilateral implants.
[i] Scenario analysis of unilateral and bilateral implantation with researchers and clinicians as informants. Time frame: current age to 75 years.
[j] Scenario analysis of unilateral and bilateral implantation; informants were candidates for implantation (N=30), users of one implant (N=30), users of two implants (N=30), and clinicians (N=52). Time frame: 30 years.
[k] 38 candidates for implantation were randomised to receive either bilateral implantation (treatment group, N=19) or unilateral implantation (control group, N=19). Time frame: life expectancy.

Table 2 Vignettes describing levels of difficulty on three dimensions of binaural hearing.

Dimension	**Level**	**Vignette**
Speech Perception in Noise	1: No difficulty	When a friend speaks to you while the TV is on or other people are chatting in the same room, you can hear your friend speaking easily, usually picking up all of the words they say.
	2: Some difficulty	When a friend speaks to you while the TV is on or other people are chatting in the same room, you can hear your friend speaking, but you can only pick out some of the words they say. This can lead to confusion if you miss an important word. Sometimes you need them to repeat themselves or to turn the volume down for you to understand them.
	3: Great difficulty	When a friend speaks to you while the TV is on or other people are chatting in the same room, you find it very difficult to hear your friend speaking. You are usually unable to pick out the words they say. This regularly leads to misunderstanding and confusion. The room needs to be completely quiet for you to understand them.
Localisation	1: No difficulty	You can work out where sounds are coming from accurately. You can point to where a sound is coming from easily.
	2: Some difficulty	You have some difficulty working out where sounds are coming from. You can usually tell if a sound is coming from the right- or left-hand side, but you cannot be more accurate than that. As a result, you are not always sure who is speaking when you are in a group with several people.
	3: Great difficulty	You have great difficulty working out where sounds are coming from. You cannot even tell if a sound is coming from the right- or left-hand side without looking around. As a result, you find it very difficult to tell who is speaking when you are in a group with several people. You are also worried about your safety outdoors because of your difficulty working out where sounds are coming from.
Effort and Fatigue	1: No difficulty	You have to concentrate a little when you are trying to hear something or someone. You can hear what people are saying with only a little effort. By the end of the day, you are not mentally or physically tired because of your hearing.
	2: Some difficulty	You have to concentrate quite hard when you are trying to hear something or someone. You have to put in some effort to hear what people are saying. By the end of the day, you are moderately mentally and physically tired because of your hearing.
	3: Great difficulty	You have to concentrate very hard when you are trying to hear something or someone. You have to put in a great deal of effort to hear what people are saying. By the end of the day, you are extremely mentally and physically tired because of your hearing.

Table 3 Numbers (N) and age of participants in Experiments 1a and 1b

Experiment	Group	Participants (N)	Inconsistent traders (N)	Zero traders (N)	Included in analyses (N)	Minimum age (years)	Mean age (years)	Maximum age (years)	% female
1a	Students	59	0	0	59	18	20.8	26	79.7
	Non-students [Public]	52	1	4	51	22	44.1	64	52.9
1b	Students	48	0	0	48	18	20.7	26	60.9
	Non-students [Public]	48	3	5	45	18	49.1	80	57.8

Table 4 Calculation of age-standardized binaural utilities.

UK population data [a]			Binaural utility data [b]						
			Students		Non-Students		Combined group		
Age Group (years)	Number in adult population	Proportion in adult population	N	Mean	N	Mean	N	Mean	Age-weighted contribution
18-20	2,390,852	0.047	52	0.748			52	0.748	0.035
21-30	8,815,943	0.172	55	0.736	15	0.730	70	0.735	0.126
31-40	8,379,311	0.163			14	0.827	14	0.827	0.135
41-50	9,070,302	0.177			27	0.847	27	0.847	0.150
51-60	8,282,160	0.161			31	0.868	31	0.868	0.140
>60	14,400,593	0.281			9	0.792	9	0.792	0.222
Total	51,339,161	1.000	107		96		203		0.808

[a] Office for National Statistics (2015); [b] From Experiments 1a and 1b.

Table 5 YBRQL value set. Entries are the 125 values of the expanded value set. Underlined entries are the 27 values of the initial value set.

Speech-in-noise	Localisation	Effort & Fatigue				
Level	Level	Level				
		1	1.5	2	2.5	3
1	1	0.9625	0.9272	0.8919	0.8556	0.8192
1	1.5	0.9323	0.8978	0.8632	0.8359	0.8087
1	2	0.9021	0.8683	0.8345	0.8163	0.7981
1	2.5	0.8758	0.8491	0.8223	0.7949	0.7674
1	3	0.8495	0.8298	0.8101	0.7734	0.7367
1.5	1	0.9321	0.9001	0.8681	0.8320	0.7958
1.5	1.5	0.9034	0.8743	0.8451	0.8147	0.7843
1.5	2	0.8746	0.8484	0.8222	0.7974	0.7727
1.5	2.5	0.8526	0.8291	0.8055	0.7768	0.7481
1.5	3	0.8305	0.8097	0.7888	0.7562	0.7235
2	1	0.9017	0.8730	0.8443	0.8084	0.7725
2	1.5	0.8744	0.8508	0.8271	0.7935	0.7599
2	2	0.8472	0.8285	0.8098	0.7786	0.7473
2	2.5	0.8294	0.8090	0.7887	0.7588	0.7288
2	3	0.8116	0.7896	0.7676	0.7389	0.7103
2.5	1	0.8889	0.8572	0.8256	0.7950	0.7644
2.5	1.5	0.8646	0.8390	0.8135	0.7803	0.7472
2.5	2	0.8403	0.8208	0.8014	0.7657	0.7301
2.5	2.5	0.8234	0.7989	0.7745	0.7442	0.7140
2.5	3	0.8065	0.7770	0.7476	0.7227	0.6979
3	1	0.8761	0.8415	0.8069	0.7816	0.7563
3	1.5	0.8548	0.8273	0.7999	0.7672	0.7346
3	2	0.8335	0.8132	0.7929	0.7529	0.7128
3	2.5	0.8174	0.7888	0.7602	0.7297	0.6991
3	3	0.8014	0.7645	0.7276	0.7065	0.6854

Table 6 Kendall's coefficients of correlation among outcome measures (N=28). (**p*<.05, ***p*<.01).

	YBRQL	HUI3	HUI3	EQ-5D-3L
	(binaural utility)	(hearing utility)	(health utility)	(health utility)
Localisation	.49**	.20	.27	.17
Speech in noise	-.30*	-.25	-.28*	-.09

Table 7 Discriminative ability: Measures of binaural advantage for the Bimodal and Bilateral Groups, and of the difference in binaural advantage (Bilateral – Bimodal), with 95% confidence intervals (95% CI) estimated by bootstrapping (3,000 samples per estimate, bias-corrected and accelerated).

	Group		
	Bimodal	Bilateral	Bilateral - Bimodal
Measure	Binaural advantage (95% CI)	Binaural advantage (95% CI)	Difference in binaural advantage (95% CI)[a]
SSQ score	.712 (.207 to 1.142)[b]	2.866 (1.679 to 4.156)[b]	2.154 (.749 to 3.842)[b]
YBRQL (Binaural utility)	.031 (.009 to .070)[b]	.098 (.067 to .127)[b]	.067 (.010 to .111)[b]
HUI3 (Hearing utility)	.102 (.018 to .183)[b]	.123 (.041 to .213)[b]	.021 (-.074 to .117)
HUI3 (Health utility)	.037 (-.017 to .087)	.065 (.022 to .122)[b]	.029 (-.050 to .112)
EQ-5D-3L (Health utility)	.024 (-.030 to .088)	.009 (-.012 to .037)	-.015 (-.100 to .056)

[a] Differences estimated by bootstrapping need not exactly equal the difference between the estimates of the Bimodal and Bilateral means.

[b] 95% confidence interval does not include zero.

Table 8 Estimates of the loss of utility due to impaired hearing and the gain in utility from using a second device. (IQR: inter-quartile range.)

		Loss of utility due to impaired hearing when using first or only implant		
Participants	N	Measure	Median	IQR
All Groups	28	YBRQL	.174	.131 to .216
		HUI3	.148	.142 to .257

		Gain in utility from using a second device		
Participants	N	Measure	Median	IQR
Bimodal Group	9	YBRQL	.020	.000 to .046
		HUI3	.000	-.046 to .191
		EQ-5D-3L	.000	.000 to .075
Bilateral Group	11	YBRQL	.107	.035 to .162
		HUI3	.000	.000 to .170
		EQ-5D-3L	.000	.000 to .000

Summerfield Kitterick Goman (Revised)

This document contains the figure legends followed by the figures of the revised paper.

Figure 1 Relationship between mean utility and level of dimensions in Experiment 1a (upper panels) and 1b (lower panels). Each symbol plots the mean utility for one level of a dimension (filled symbols, non-students; open symbols, students).

Figure 2 Influence of dimensions in Experiment 1a (upper panel) and 1b (lower panel). Filled bars, non-students; open bars, students. Error bars plot 95% confidence intervals of mean values.

Figure 3 YBRQL Questionnaire. The preamble in this example is worded for users of bilateral cochlear implants.

Figure 4 Relationship between SSQ score and utility when participants considered their functional abilities using their normal configuration of devices (upper row of panels). Relationship between change in SSQ score and change in utility between monaural and binaural listening for members of the bimodal and bilateral groups (lower row of panels). (Panels A and E, YBRQL; B and F, HUI3 Hearing*;* D and G, HUI3 Health*;* D and H, EQ-5D-3L). (Black circles, Unilateral Group; dark grey squares, Bimodal Group; light grey triangles, Bilateral Group).

Figure 5 Box plots of SSQ score (Panels A and C) and utility (Panels B and D) when the Bilateral Group (Panels A and B) and the Bimodal Group (Panels C and D) considered their functioning with their first or only implant (open boxes) and with two devices (filled boxes). Whiskers mark 10th and 90th percentiles; box marks 25th and 75th percentiles; heavy line within box marks 50th percentile; filled circles plot outliers beyond the 10th and 90th percentiles. (Panel E) z scores, significance levels (p), and effect sizes (R) from Wilcoxon signed ranks tests comparing scores with 1 and 2 devices.

Figure 6 Box plots of reported levels of function on each dimension of the YBRQL when thebilateral and bimodal groups considered their function using their first or only implant (open boxes) or two devices (filled boxes). Whiskers mark 10th and 90th percentiles; box marks 25th and 75th percentiles; heavy line within box marks 50th percentile; filled circle plots outlier beyond the 10th percentile. Results of Wilcoxon signed ranks tests comparing reported levels with one and two devices (✶, p<.05; ✶✶, p<.01).

Figure 7 Decision matrix for interpreting Binaural ICERs informed by the YBRQL together with Health ICERs informed by a generic PBM.

Figure 1

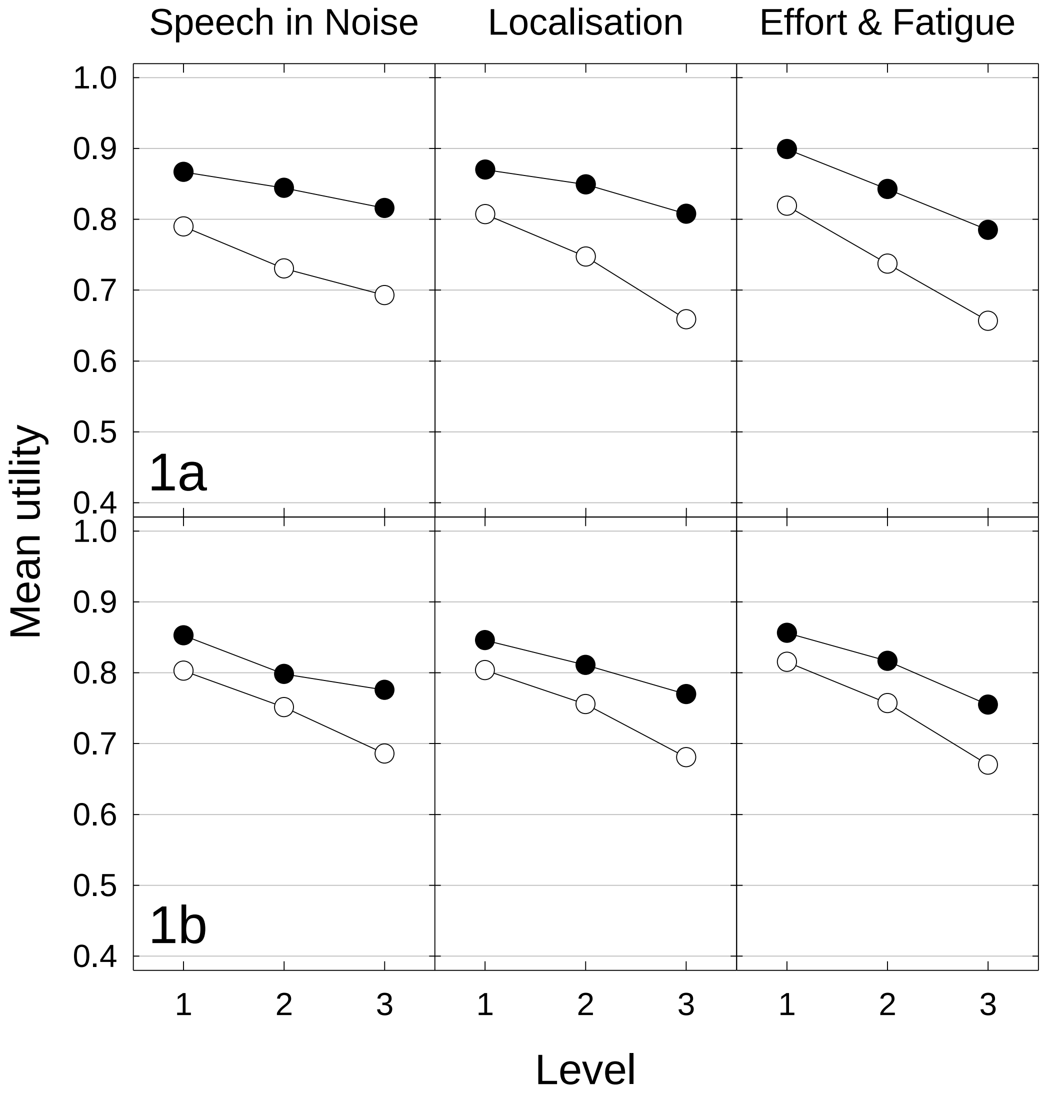

Speech in Noise
Localisation
Effort & Fatigue
1.0
0.9
0.8
0.7
0.6
0.5
0.4
1a
1b
Mean utility
1
2
3
Level

Figure 2

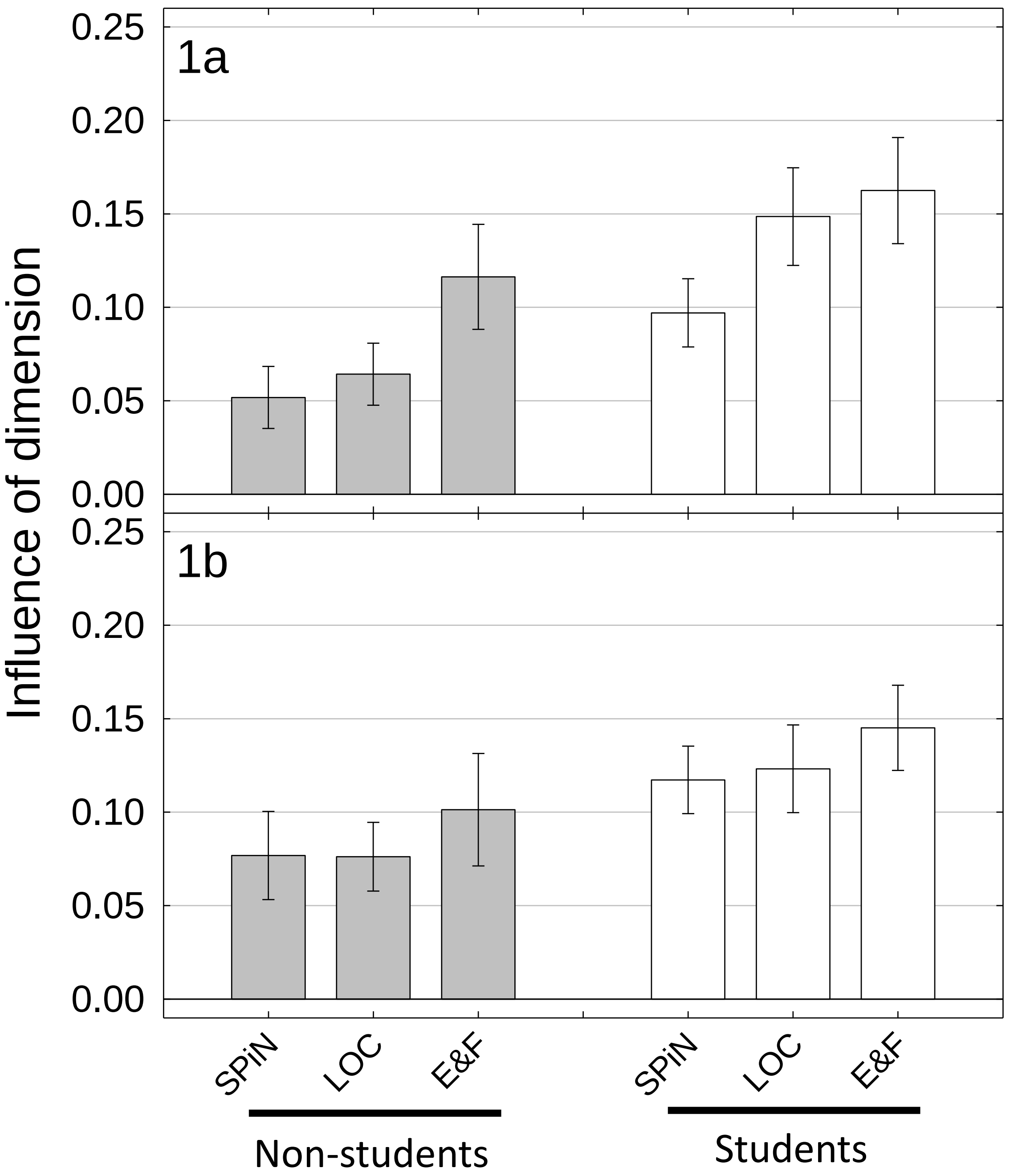
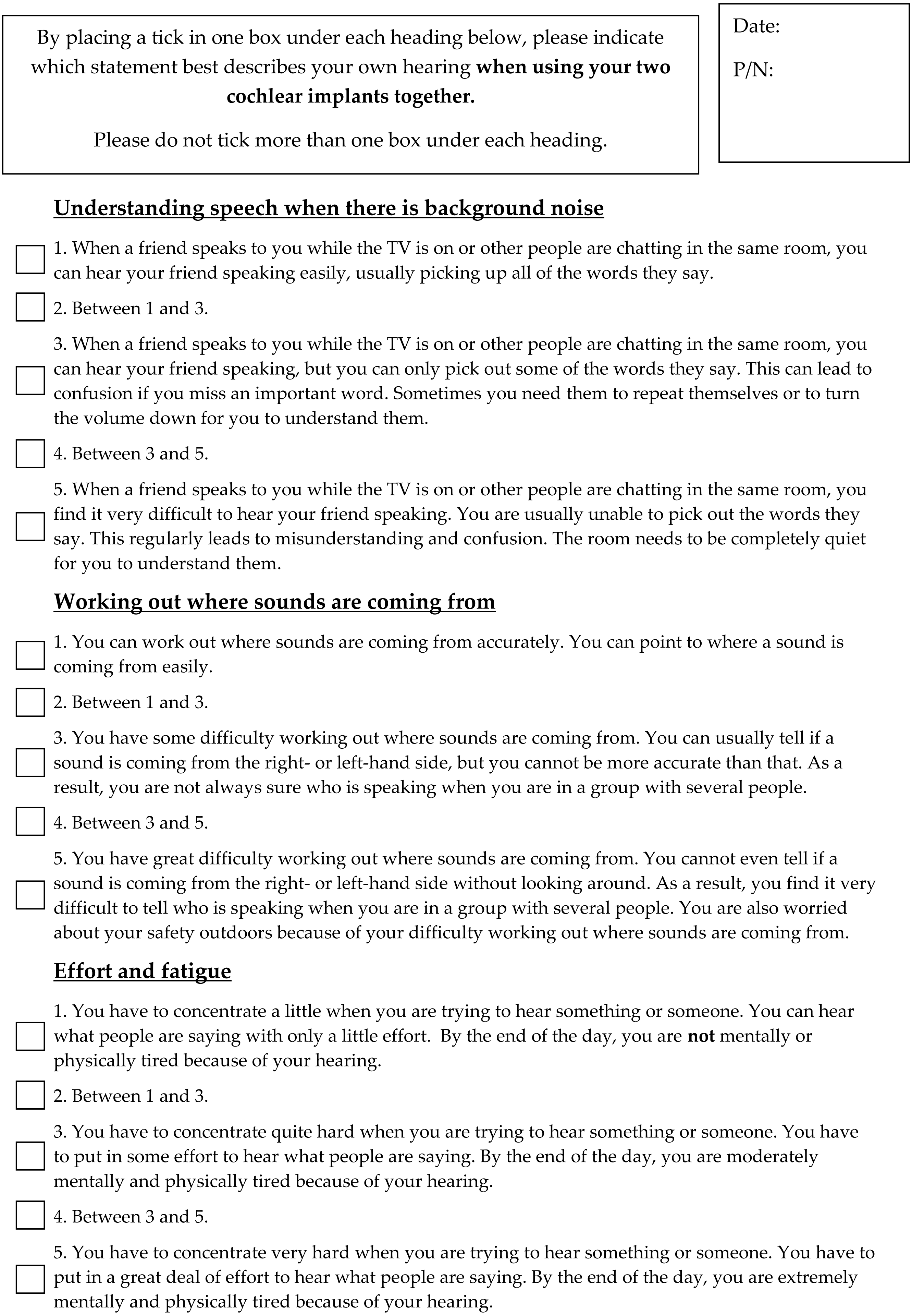

1a
1b
Influence of dimension
0.25
0.20
0.15
0.10
0.05
0.00
0.25
0.20
0.15
0.10
0.05
0.00
SPiN
LOC
E&F
SPiN
LOC
E&F
Non-students
Students

Figure 3



By placing a tick in one box under each heading below, please indicate which statement best describes your own hearing **when using your two cochlear implants together.**

Please do not tick more than one box under each heading.

Date:

P/N:

Understanding speech when there is background noise

□ 1. When a friend speaks to you while the TV is on or other people are chatting in the same room, you can hear your friend speaking easily, usually picking up all of the words they say.

□ 2. Between 1 and 3.

□ 3. When a friend speaks to you while the TV is on or other people are chatting in the same room, you can hear your friend speaking, but you can only pick out some of the words they say. This can lead to confusion if you miss an important word. Sometimes you need them to repeat themselves or to turn the volume down for you to understand them.

□ 4. Between 3 and 5.

□ 5. When a friend speaks to you while the TV is on or other people are chatting in the same room, you find it very difficult to hear your friend speaking. You are usually unable to pick out the words they say. This regularly leads to misunderstanding and confusion. The room needs to be completely quiet for you to understand them.

Working out where sounds are coming from

□ 1. You can work out where sounds are coming from accurately. You can point to where a sound is coming from easily.

□ 2. Between 1 and 3.

□ 3. You have some difficulty working out where sounds are coming from. You can usually tell if a sound is coming from the right- or left-hand side, but you cannot be more accurate than that. As a result, you are not always sure who is speaking when you are in a group with several people.

□ 4. Between 3 and 5.

□ 5. You have great difficulty working out where sounds are coming from. You cannot even tell if a sound is coming from the right- or left-hand side without looking around. As a result, you find it very difficult to tell who is speaking when you are in a group with several people. You are also worried about your safety outdoors because of your difficulty working out where sounds are coming from.

Effort and fatigue

□ 1. You have to concentrate a little when you are trying to hear something or someone. You can hear what people are saying with only a little effort. By the end of the day, you are **not** mentally or physically tired because of your hearing.

□ 2. Between 1 and 3.

□ 3. You have to concentrate quite hard when you are trying to hear something or someone. You have to put in some effort to hear what people are saying. By the end of the day, you are moderately mentally and physically tired because of your hearing.

□ 4. Between 3 and 5.

□ 5. You have to concentrate very hard when you are trying to hear something or someone. You have to put in a great deal of effort to hear what people are saying. By the end of the day, you are extremely mentally and physically tired because of your hearing.

Figure 4

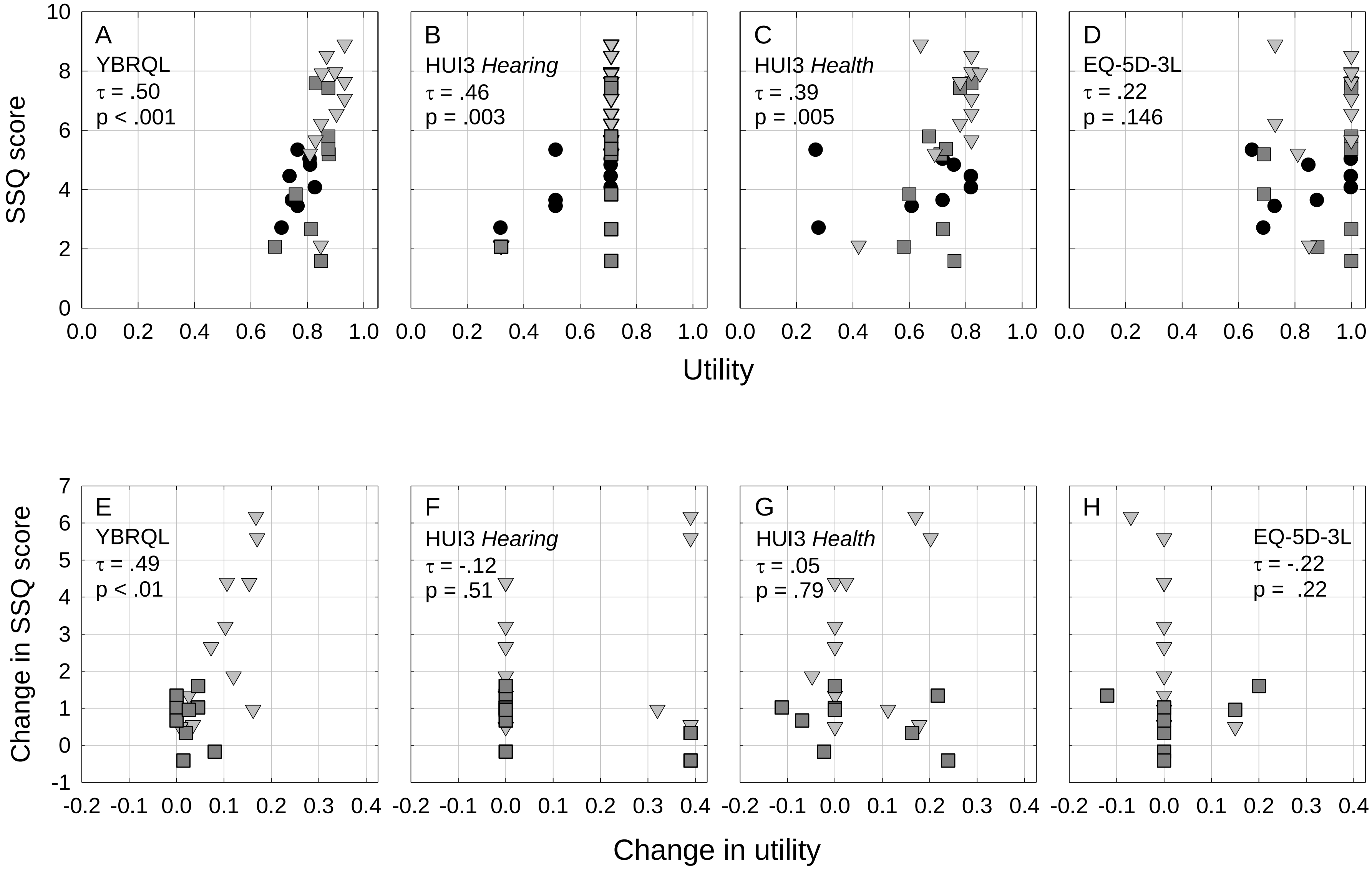
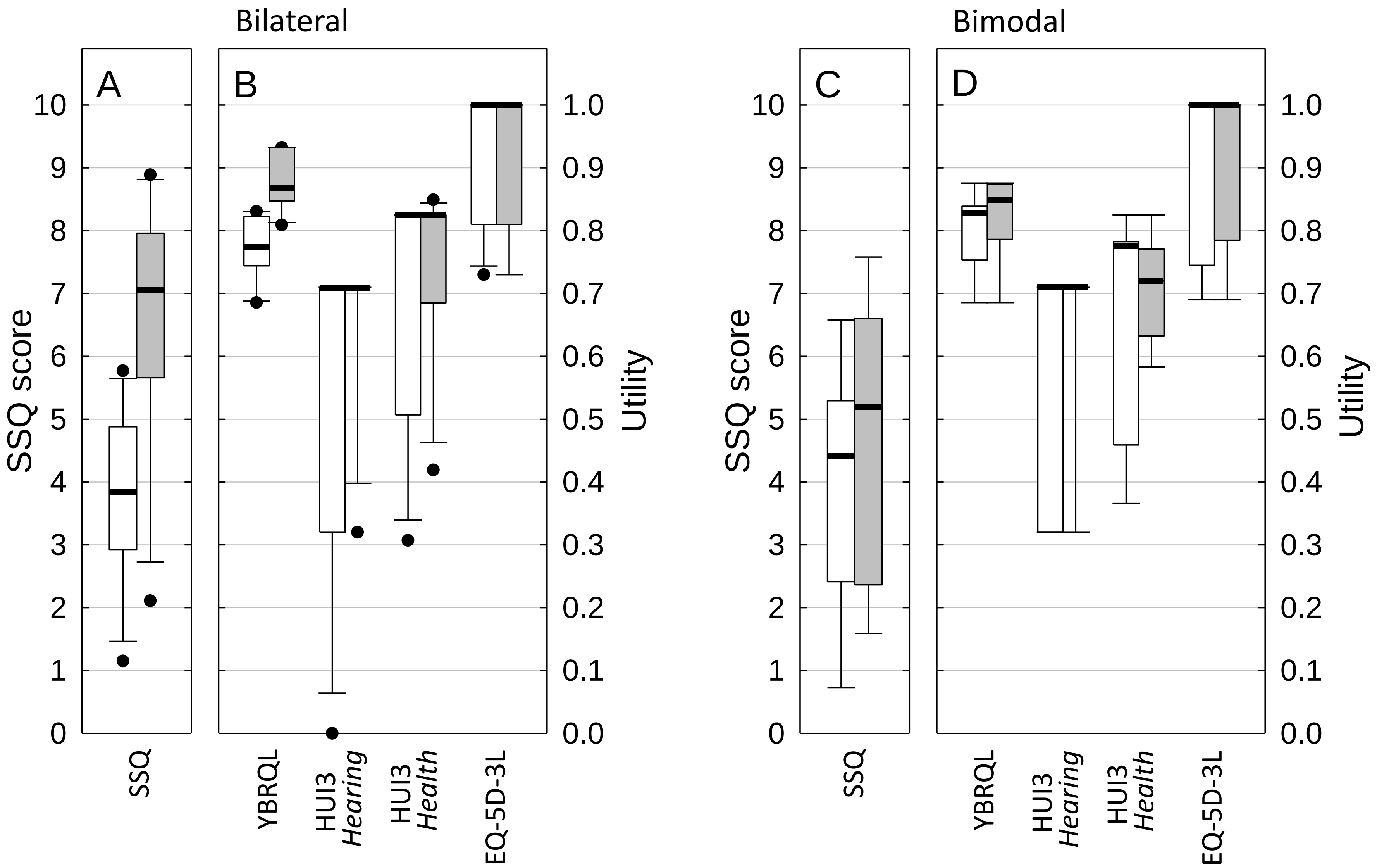

SSQ score
10
8
6
4
2
0
A
YBRQL
τ = .50
p < .001
B
HUI3 *Hearing*
τ = .46
p = .003
C
HUI3 *Health*
τ = .39
p = .005
D
EQ-5D-3L
τ = .22
p = .146
0.0 0.2 0.4 0.6 0.8 1.0
Utility
Change in SSQ score
7
6
5
4
3
2
1
0
-1
E
YBRQL
τ = .49
p < .01
F
HUI3 *Hearing*
τ = -.12
p = .51
G
HUI3 *Health*
τ = .05
p = .79
H
EQ-5D-3L
τ = -.22
p = .22
-0.2 -0.1 0.0 0.1 0.2 0.3 0.4
Change in utility

Figure 5



E	z	2.93	2.93	1.89	1.78	.447	2.19	2.21	1.41	.943	1.07
	p	.003	.003	.059	.075	.655	.028	.027	.157	.345	.285
	R	.63	.63	.40	.38	.10	.52	.52	.33	.22	.25

Figure 6

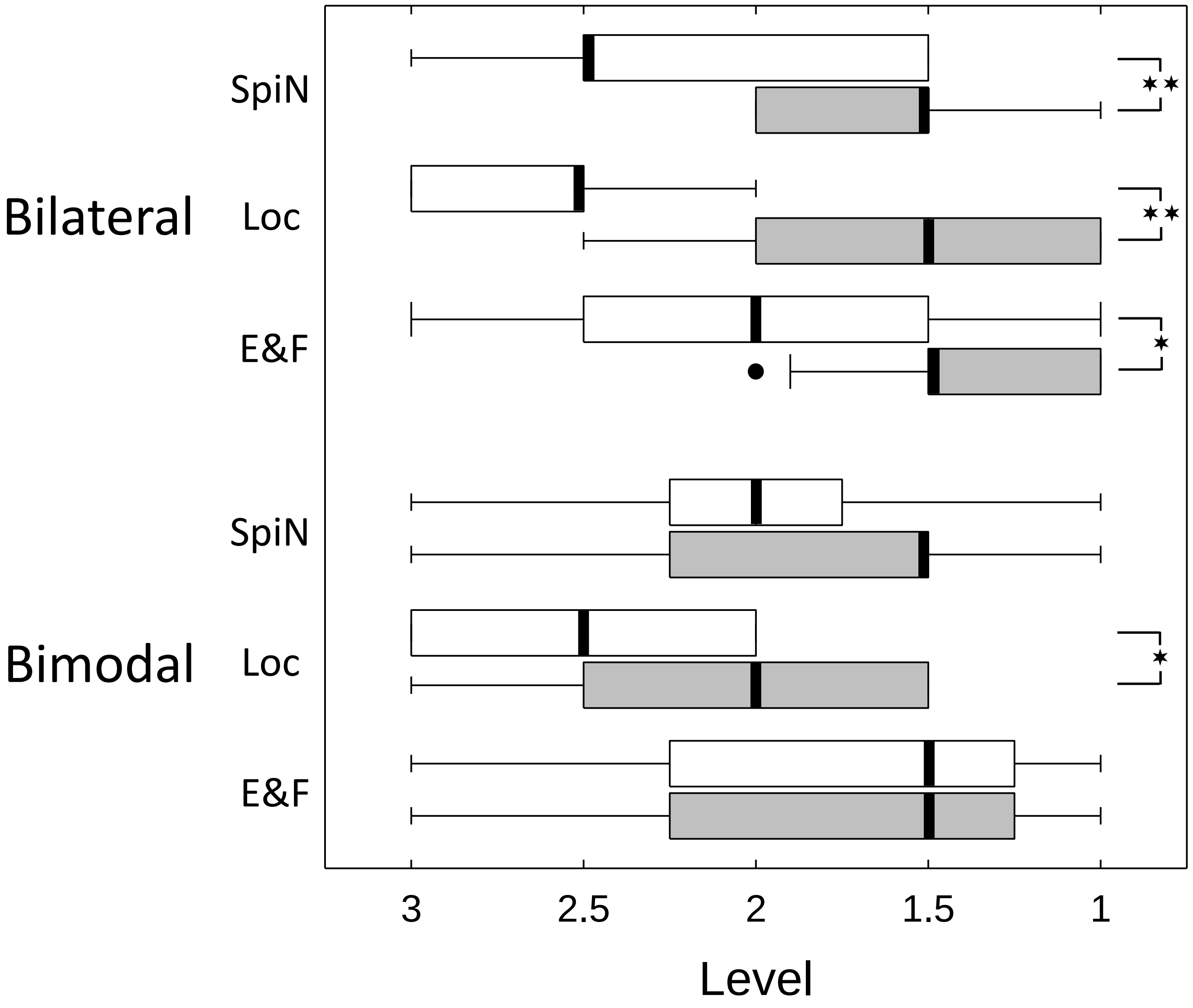


Figure 7

		Binaural		
		Binaural ICER is more favourable than WTP threshold	**Binaural ICER is in the range of uncertainty**	**Binaural ICER is less favourable than the upper boundary of uncertainty**
Health	**Health ICER is more favourable than WTP Threshold**	1. *Treatment is cost-effective.*	2. *Treatment is cost-effective.* Investigate reasons why gain in binaural-related QoL is smaller than gain in health-related QoL.	3. Investigate reasons for small or negative gain in binaural-related QoL before deciding whether treatment is cost-effective.
	Health ICER is in the range of uncertainty	4. *Treatment is probably cost-effective.*	5. *Treatment is unlikely to be cost-effective.* Investigate reasons why costs are high and/or gain in binaural-related QoL is small.	6. *Treatment is unlikely to be cost-effective.* Investigate reasons why costs are high and/or gain in binaural-related QoL is small or negative.
	Health ICER is less favourable than the upper boundary of uncertainty	7. *Treatment is not cost-effective.* Investigate reasons why costs are high and/or gain in health-related QoL is small or negative.	8. *Treatment is not cost-effective.* Investigate reasons why costs are high and/or gain in binaural-related QoL is small.	9. *Treatment is not cost-effective.* Investigate reasons why costs are high and/or gain in binaural-related QoL is small or negative.