Appendix A: Methodological study

Note: The research detailed in this Section has been published in: **Stahl-Timmins, W.**; Pitt, M. & Peters, J. 2010. *Graphical presentation of data for health policy decisions: An exploratory online decision task experiment to measure effectiveness.* Information Design Journal 18:3.

A – 1 Introduction

The evaluation of information graphics is key to understanding how data (or information) is communicated by them. As informing policy is a fundamental part of HTA, it is essential to be able to perform this evaluation appropriately. As the research methods used in information design and HTA are quite different, a methodological study is used to show the value and limitations of quantitative measurements in assessing the communicative benefits of information graphics. It uses a simplified decision problem suitable for a general public audience, to provide as large a sample size as possible.

As mentioned in Chapter 1.2, The foundations of HTA lie in evidence-based medicine, using scientific tests to inform policy decisions (Banta, 2003). There are few people, if any, with specialist training in visual communication in the field. For any new graphical technique to be accepted in HTA, empirical testing will be necessary to show the advantages and disadvantaves of visual presentation, in the language of empirical research familiar to the field. However, what testing there is in the field of information design tends to be largely small sample, and qualitative in nature. This kind of research is less familiar in HTA, in which the standard evidence base is large, quantitative trials.

The experiment described here investigates a method of providing quantitative feedback on alternative information presentations. In contrast to most studies, which investigate particular graphical representations of information, it focusses on the process of testing an information graphic.

A – 1.1 Relevant published research

Empirical, controlled studies of the effectiveness of information graphics are widely distributed through the research literature, and spread across many

fields of application (such as health, management, business, finance, design, and engineering, amongst others.) A systematic literature review in this area is therefore challenging. Table A – 1 shows details of some empirical studies of information graphics. These have been sourced from bibliographic searches as well as a request for information made through the PHD-DESIGN email list on JISCmail. (PHD-DESIGN on JISCmail, 2009). They address visual information presentations in management (4 papers), health (2 papers) and finance (1 paper). These papers cover quite a wide period of time, which potentially causes difficulties with outdated technologies being used to create and distribute information graphics. However, since many static, black and white, printed graphics are currently used in printed health technology assessment reports, the research methods used to test them are still likely to be relevant.

Author(s)	Year	N	Measurements used
Benbasat & Dexter	1985	35	accuracy time preference
Elting et al.	1999	34	time accuracy preference
Feldman-Stewart, Brundage & Zotov	2007	216	accuracy time
Frownfelter-Lohrke	1998	290	accuracy time
Lim & Benbasat	2000	79	'perceived equivocality' preference
Remus	1984	53	accuracy
Remus	1987	54	accuracy

Table A – 1
Empirical studies of information graphics

A – 1.2 Measurements used

Three commonly used measurements in empirical studies of information graphics' effectiveness are:

Response time: how long it takes for a participant to make a decision, or take some other action, based on information presented in different formats.

Accuracy: how close a participant's response is to an already established "best answer" (or correct answer).

Preference: each participant chooses between multiple information presentations (such as numerical vs. graphical format, or between different graphical presentations).

These variables are then analysed with a range of different statistical techniques, including ANOVA (analysis of variance), t-tests, and regression analysis, among others.

One study (Benbasat & Dexter, 1985) also recorded qualitative data, in the form of participants' comments, about the different information presentations

given. Although these comments were referred to in the discussion section, no formal qualitative analysis was included in the paper.

The two companion papers by Remus (Remus 1984; Remus 1987) provide evidence in support of information graphics being more appropriate for displaying complex data. They note that:

> In low complexity environments decision makers using tabular displays are in the aggregate better able to weigh the appropriate factors in the decision making process… However, in intermediate levels of environmental complexity, the composite rules show graphical displays to be significantly better than tabular displays.

This is disputed by Bertin (Bertin 1981). The study detailed in this chapter asks whether this finding is applicable to HTA.

In general, the area of application for the papers vary. Remus addresses a production scheduling management decision problem, and Frownfelter-Lohrke uses information presentations that are designed to support predictions on the financial condition of companies in future. Of these, Benbasat & Dexter use different presentation methods to support a marketing budget optimisation decision across different territories. Elting et al. situate their work in the context of deciding when to stop a clinical trial based on results presented in different formats.

The study detailed in this section is intended to provide a method for evaluating the relative performance of different information presentations, which might be applied within specific contexts and to particular audiences.

A – 1.3 Research questions

The aim of the study is to see how effectively a quantitative study, analysed using a statistical model, can explain the actions of people using graphical information presentations. This is defined in reference to the following four research questions:

1) What effect does increasing complexity have on the relative difference in participants' performance between the two presentation methods?

If this application situation is similar to Remus' (Remus 1984; Remus

1987), the difference in performance (time or accuracy) between the two presentation methods may be larger in the more complex tasks.

2) Can time and accuracy be shown to be interdependent?

In some experimental tasks, decision time (i.e. the length of time someone takes to make a decision) may affect their accuracy. The experiment detailed here recorded both time and accuracy, with people allowed to decide for themselves how long they spent on the tasks. A statistically significant negative correlation between these two variables would suggest that, for these tasks, time can be traded for accuracy. If such correlations can be demonstrated using qualitative data like this, studies of this kind may be able to suggest which of a set of alternative presentations would be most effective in time-pressured situations.

3) Can it be shown that people perform better, in terms of time and/or accuracy, depending on their preference?

Experimental studies of graphical displays have recorded participants' preferences. (Benbasat & Dexter, 1985; Elting et al., 1999; Lim & Benbasat 2000). It might seem logical to assume that the preferences of the people using a presentation method might affect their performance, but that can not be assumed. The Elting et al study in fact found that: "Despite the superior accuracy of the icon display, none of the participants preferred that method, and eight voiced considerable contempt for the display." (Elting et al. 1999). The experiment detailed here was used to test whether a participant's preference has a demonstrable effect on their performance in this instance. In situations where there is a positive preference effect, it may be worth the increased production time to give an audience some level of choice in how information is displayed to them.

4) Which characteristics can be shown to affect performance or preference with one or other presentation method? (age, gender, familiarity with health technology assessment, socio-economic group, continent of residence.)

There may be characteristics that could be used to predict a person's accuracy, preference, or the time taken to complete the tasks. Any of these could be used to assess a person's likelihood to respond to different presentation methods, perhaps tailoring the presentation to some degree, according to the most likely match for their specific characteristics.

A – 2 Methods

The study was performed as a randomised, web-based experiment, with a general public audience. The decision problem that the participants addressed was a treatment optimisation task, which was appropriately simplified for a general public audience. This non-specific audience is used for ease of recruitment, given the study's focus on testing the method of research rather than an actual information presentation.

Participants were shown data on three different (fictitious) drugs, with different costs and effectiveness, and given a limited budget to spend on these. Their task was to choose which drugs to give to an imaginary cohort of people, to minimise the number of deaths caused by the condition from which they suffered. The task gradually increased in complexity, as the participants were asked to assign treatments to increasing numbers of patient subgroups, working within the constraints of their budget.

The study employed a randomised, crossover design, in which each person was randomised to be shown either a 'numerical' or 'graphical' presentation first. They completed three tasks, with increasing complexity, with the first presentation method. Then, they were given the alternative presentation, with a new set of data, for three more tasks. Irrespective of which presentation method was assigned to a person, the data for the six tasks was given in the same order.

The two different presentation methods were visually similar, with the same colours and typeface used in each (see Figure A – 1). Colour can be regarded as a graphical element, but the decision was taken to use it in both numerical and graphical presentations, to avoid people choosing their preference based on which was more colourful, rather than on the effectiveness of the presentation method.

The participants' time to decision was recorded for each of the six tasks, as well as the number of patients that would die with that treatment option (the effectiveness measure for this experiment). Participants were asked for personal details after the third task – their age, gender, the continent on which they normally live, their familiarity with HTA on a four-point scale, and four questions for assessment of socio-economic group, from the UK's national office for statistics self-coding questionnaire. (Office for National Statistics, 2009) (see Figure A – 2)

Figure A – 1
The 'numerical' display (above) and the 'graphical' display (below).
Task 3 of 6 is shown in each case

Participants were given feedback on their performance after completing each task (see Figure A – 3). They were shown the time that they took to reach a decision, as well as the mean time to decision of the other participants that had taken part in the study to that point in time. Also, the number of 'lives saved' was shown, calculated as the difference between their choice and the worst treatment choice. The number of lives that could have been saved, with the optimal treatment choice, was also shown.

It could be argued that showing participants their performance in this way

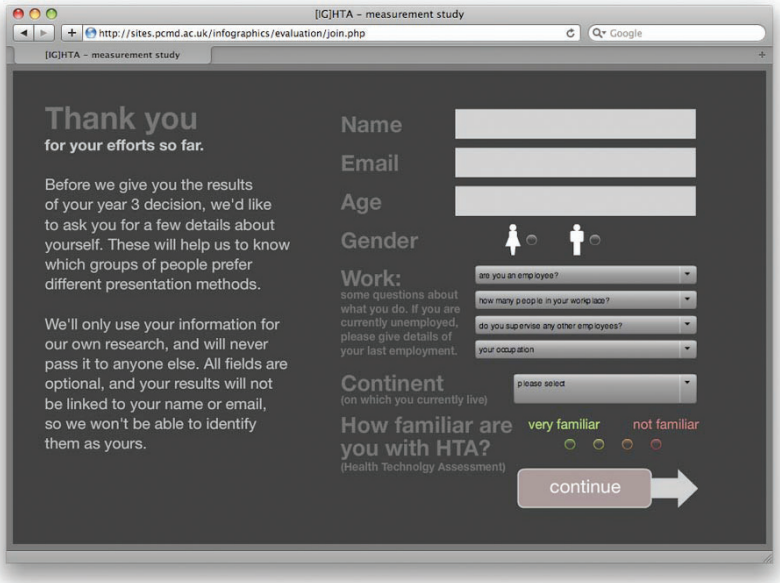


Figure A – 2
The details collection layout (shown after task 3)

might affect a person's preference for one or the other presentation methods. However, it gave the advantage that people were invited to choose whether to focus their efforts on saving more lives, or taking less time to make their choices (desirable in this case as we would like to test the interaction between time and accuracy). Also, the instant feedback was designed to engage people's attention, encourage them to continue with the study, and possibly repeat it, so that the effect of multiple attempts could be analysed.

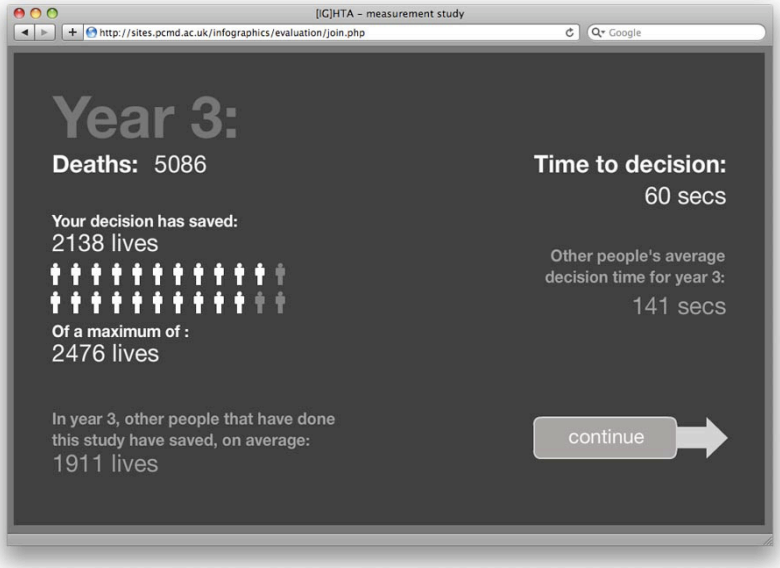


Figure A – 3
The feedback layout (shown after every task)

In the graphical presentation application situation evaluated by Remus, the graphical presentation method was found to be more favourable when the data was more complex (Remus 1984; Remus 1987). The decision task produced for our experiment was designed to be increasingly complex from task to task. The first and fourth tasks (the initial tasks with each presentation method) had only two choices of treatment that were possible, within the constraints of the budget given. The second and fifth had five and six possibilities respectively, and the third and sixth tasks (the last task with each presentation method), had over 600 different treatment combinations.

It could be argued that, in this experiment, a definite "best answer" could easily be calculated, and so the human decision-making process is unnecessary. However, the completely artificial situation allows the experiment to show objectively how close to a perfect decision the participants can come with the different presentation methods. In a real HTA decision-making situation, it is much more difficult to assess the possible ramifications on the entire health service and the people within it. The artificially-created decision problem used in this experiment serves to allow comparison with a precisely calculated best decision, based on very limited information and few criteria.

A – 2.1 Sampling

The effect size that would be measured by the study was largely unknown, but expected to be quite small. To provide as large a number of participants as possible, a web-based application was designed for participants to use, and a sample of the internet-using general public was employed.* It was acknowledged that the use of this very large population would limit how much the results could be generalised. However, as the main objective was to determine how accurate a quantitative experiment and statistical model for the effectiveness of a graphical presentation would be, this was an acceptable sacrifice.

* The test is still available for viewing online at http://sites.pcmd.ac.uk/infographics/evaluation/index.php at the time of writing.

Online, web-based research studies have a few important differences to ones using more traditional, face-to-face or paper-based research methods. Birnbaum notes that: "Internet research has two potential problems, sampling and control." (Birnbaum 1999).

In terms of sampling, it is always difficult to know, in open web-based studies like this, whether each entry is from a unique participant. This study starts with a simple question on whether this is a participant's first time attempting it, so that a person does have the opportunity to try again without affecting the results. However, there is no way of knowing if this question has been answered correctly or truly, so each participant's computer's IP address and the web browser used were recorded, so that possible duplicate entries could be removed in this way also. Excluding entries that had both the same IP address and browser software could have removed some genuine results (for example, multiple people on an organisational network). However, it is much more unlikely that duplicate entries would have been included in the analysis using this method.

In terms of control, participants were asked at the end of the study if they had

been distracted in any of the tasks, and whether they had written anything down or made any calculations to help them answer correctly. Not all participants would have reached the end of the study, and responded to this question. However, its inclusion does mean that participants who have not been shown these questions can be excluded from the time-based analysis, as they may have been distracted and not given the chance to report on it.

The sampling method used was similar to "respondent-driven sampling" (Salganik 2004; Wejnert & Heckathorn 2008) sometimes referred to as "snowball sampling". Each participant was asked to pass the website link on to other people (see Figure A – 4). However, unlike Salganik's method, no attempt was made to record links and relationships between participants. The primary purpose of this experiment is not to sample a particular population, but to assess the feasibility of quantitative analysis in such studies. The chain-referral sampling method used here simply serves to increase the number of participants.

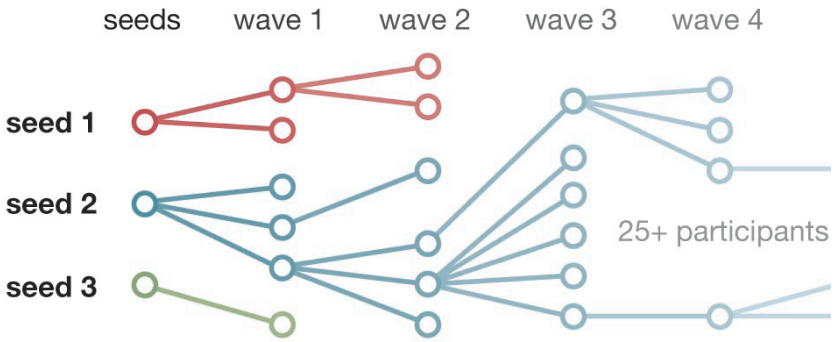


Figure A – 4
"respondent-driven sampling"

The initially sampled "seeds" chosen were members of the the PHD-DESIGN mailing list (PHD-DESIGN on JISCmail 2009), as well as the PENTAG technology assessment team (personal acquaintances). Each seed was sent an email with a web link to the online application. The link was also given in a presentation at the DD4D conference (an interdisciplinary design/statistics conference held at the OECD headquarters in Paris, June 2009.)

Those that completed the first three tasks were asked for their name and email address. An automated email was then sent to them, asking them to give the link to anyone that they thought would be interested in taking part. Email addresses and names were not recorded, due to ethical approval regulations of Exeter university, so no follow-up of participants was possible.

A – 2.2 Analysis methods

Analysis was largely performed using statistical tests on the quantitative data collected by the web server during the experiment. Qualitative results were analysed using a framework analysis (Ritchie & Spencer, 1994).

A – 3 Results

A – 3.1 Descriptive statistics

Study duration

36 days (15th June - 21st July 2009)
244 submissions were recorded by the web server during this time.

Exclusion

Of these submissions, **23** were excluded from the analysis, as the participant had not clicked "yes" when asked whether it was the first time they had tried the study.
24 were excluded as possible duplicates, as they were submitted from a computer which had both the same IP address and the same browser software as an earlier submission.
1 entry was excluded, as the system did not properly record the participant's randomisation, leaving a total of **196** unique participants for the analysis.

Randomisation

The participants were randomised to receive either a graphical or numerical presentation of information first, using the PHP rand() function.
99 participants received the graphical presentation first.
97 received the numerical one first.

Completion

See Table A – 2 for the number of people that completed each 'year' of the study (each task).

	N-G group	G-N group	Total
Task 1	78	77	155
Task 2	71	70	141
Task 3	68	68	136
Task 4	55	64	119
Task 5	55	62	117
Task 6	53	60	113

Table A – 2
Completion for tasks

106 people completed the entire study, giving a preference at the end (see Figure A – 5). In this figure, graphical tasks are shown in orange, and numerical tasks in blue. The width of the arrows is proportional to the number of people remaining. Note that after their personal details are asked for after the third task (y3), the participants in each group swap to the other presentation. The crossing arrows at the bottom show the numbers of people that completed all 6 tasks, and gave a preference for one or the other presentation method.

After the third task (y3), people were asked for some personal details. The frequencies of the results of these questions are presented in Table A – 3.

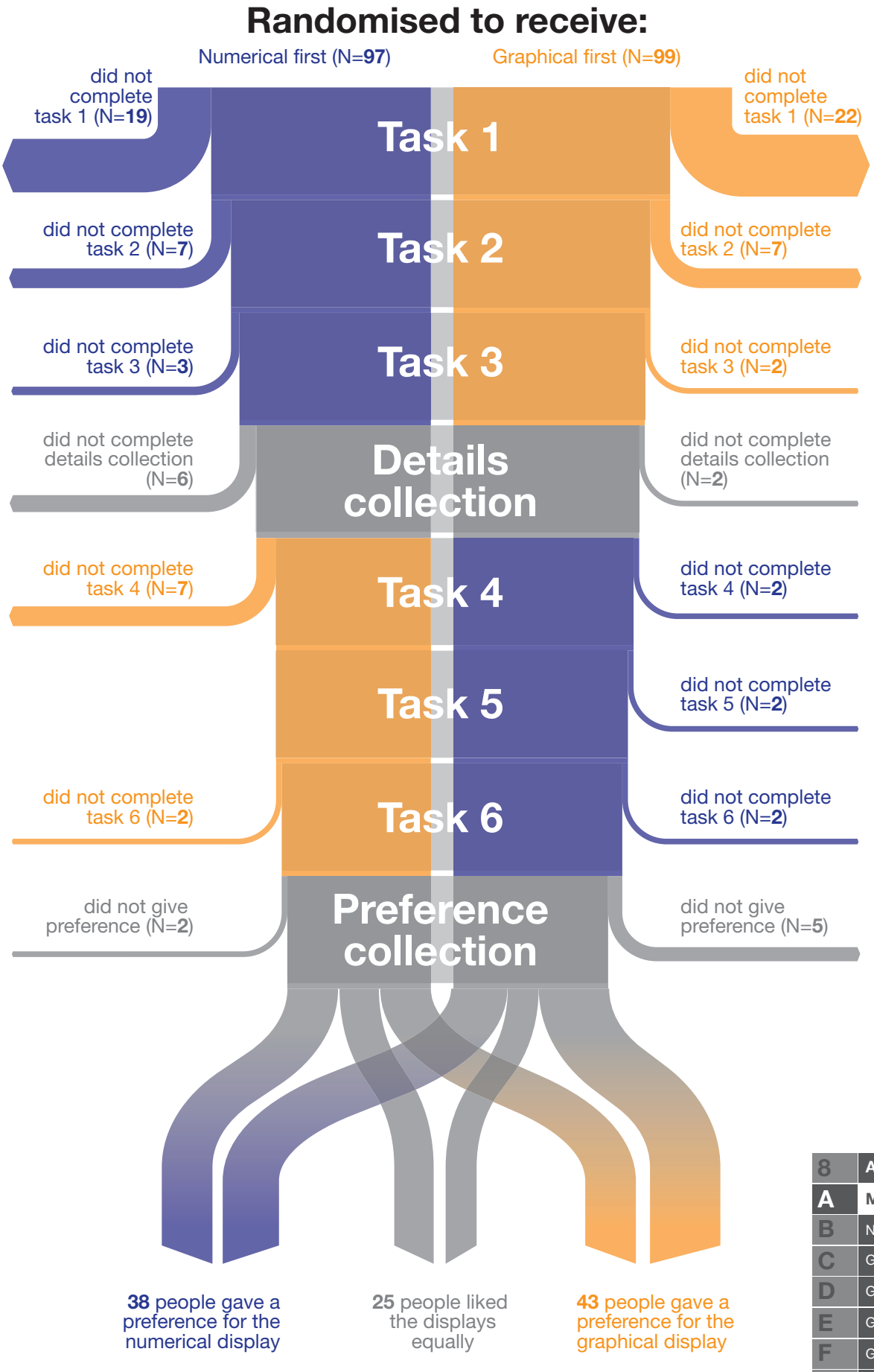


Figure A – 5
Dropout and preferences of participants

	N-G group	G-N group	Total
Gender			
Men	**22** (36.7%)	**30** (46.2%)	**52** (41.6%)
Women	**32** (53.3%)	**26** (40.0%)	**58** (46.4%)
Not reported	**6** (10%)	**9** (13.8%)	**15** (12.0%)
Age group			
20-29	**15** (25.0%)	**16** (24.5%)	**31** (24.8%)
30-39	**16** (26.7%)	**14** (21.5%)	**30** (24.0%)
40-49	**7** (11.7%)	**9** (13.8%)	**16** (12.8%)
50-59	**14** (23.3%)	**5** (7.7%)	**19** (15.2%)
60-69	**3** (5.0%)	**5** (7.7%)	**8** (6.4%)
70-79	**0** (0.0%)	**1** (1.5%)	**1** (0.8%)
Not reported	**5** (8.3%)	**15** (23.1%)	**20** (16%)
Familiarity with HTA:			
Green (very familiar)	**5** (8.3%)	**5** (7.7%)	**10** (8%)
Yellow	**3** (5.0%)	**4** (6.2%)	**7** (5.6%)
Orange	**11** (18.3%)	**5** (7.7%)	**16** (12.8%)
Red (not familiar)	**41** (68.3%)	**45** (69.2%)	**86** (68.8%)
Not reported	**0** (0.0%)	**6** (9.2%)	**6** (4.8%)
Socio-economic group*:			
1	**32** (33.3%)	**30** (46.2%)	**62** (49.6%)
2	**1** (1.0%)	**1** (1.5%)	**2** (1.6%)
3	**0** (0.0%)	**0** (0.0%)	**0** (0.0%)
4	**1** (1.0%)	**1** (1.5%)	**2** (1.6%)
5	**0** (0.0%)	**1** (1.5%)	**1** (0.8%)
Not reported	**26** (43.3%)	**32** (49.2%)	**58** (46.4%)
Location:			
Africa	**1** (1.7%)	**0** (0.0%)	**1** (0.8%)
Asia	**1** (1.7%)	**2** (3.1%)	**3** (2.4%)
Australia	**4** (6.7%)	**6** (9.2%)	**10** (8.0%)
Europe	**37** (61.7%)	**40** (61.5%)	**77** (61.6%)
North America	**12** (20.0%)	**9** (13.8%)	**21** (16.8%)
South America	**1** (1.7%)	**1** (1.5%)	**2** (1.6%)
Not reported	**4** (6.7%)	**7** (10.8%)	**11** (8.8%)
Task understanding:			
Green (understood tasks well)	25 (50.0%)	32 (56.1%)	57 (53.3%)
Yellow	20 (40.0%)	14 (24.6%)	34 (31.8%)
Orange	3 (6.0%)	10 (17.5%)	13 (12.1%)
Red (did not understand tasks)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Not reported	2 (4.0%)	1 (1.8%)	3 (2.8%)

* Of the 67 people that fully answered this question, 58 (86.6%) gave their occupation as "modern professional", automatically putting them in group 1 (of 5).

Table A – 3
Participants' characteristics

Mean accuracy results

(See Figure A – 6)

Average deaths (accuracy measure) were quite similar between the two different groups across all six tasks.

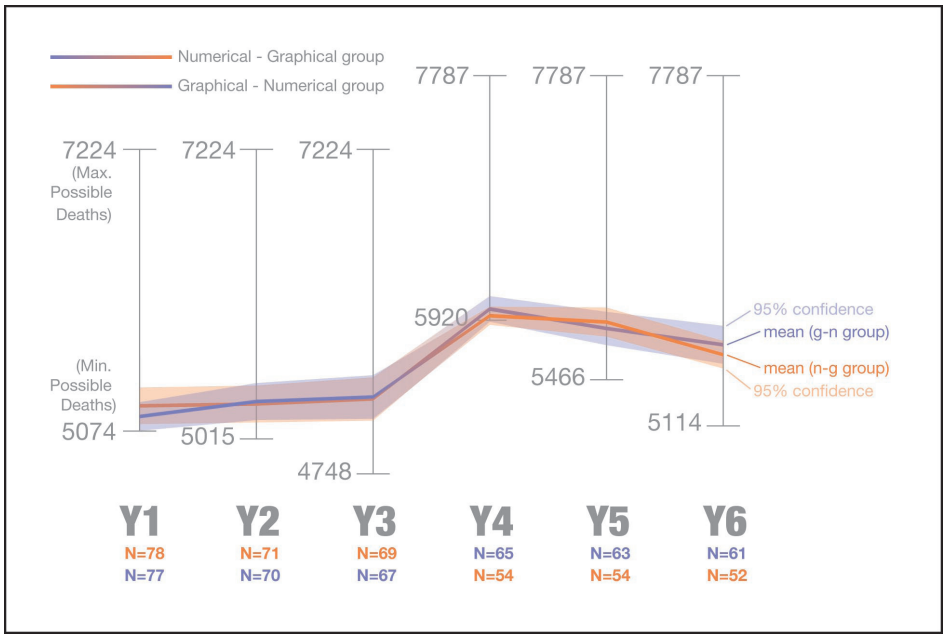


Figure A – 6
Mean deaths, by randomised group

Mean time results

(See Figure A – 7)

Average times to complete the six tasks were similar, with the exception of the third task ("year 3"), for which participants given the graphical presentation tended to take longer, although the 95% confidence intervals still overlap.

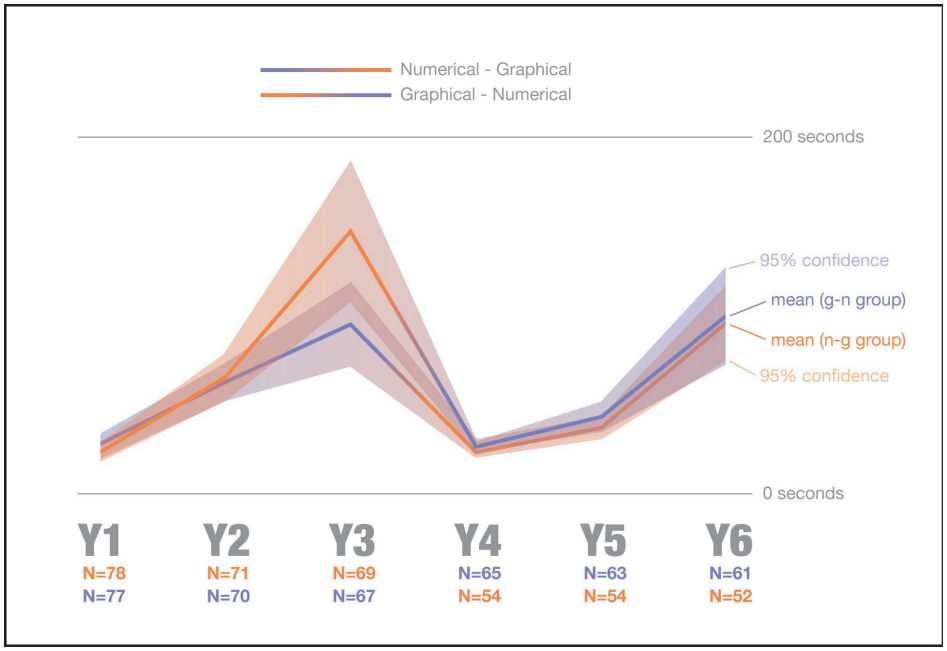


Figure A – 7
Mean times, by randomised group

A – 3.2 Effects of increasing complexity

1) What effect does increasing complexity have on the relative difference in results between the two presentation methods?

The increasing complexity of the decision tasks seemed to have very little difference on the relative performance of the participants in the two groups (see figures A – 6 and A – 7). The only result that was substantially different between the presentation methods overall was the mean time taken for the third task, which was noticeably longer for the graphical presentation than for the numerical presentation. However, even this difference is not observed in the other complex task (task 6). Neither presentation seems to be strongly advantageous in complex situations than more simple ones in this context.

A – 3.3 Time and accuracy dependancy

2) Can time and accuracy be shown to be dependant on each other?

Focussing on the third task, for all participants (the first complex task), a scatter plot of time against accuracy suggests a possible correlation between the two (see Figure A – 8).

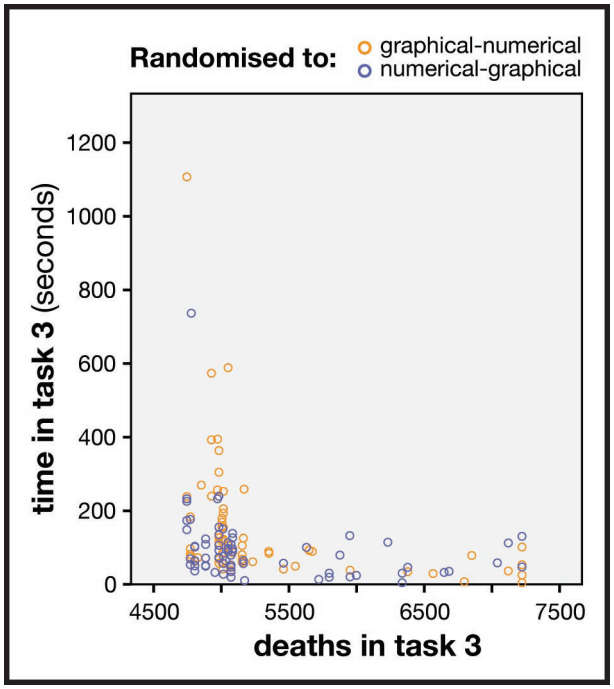


Figure A – 8
Scatter plot of times against deaths in task 3

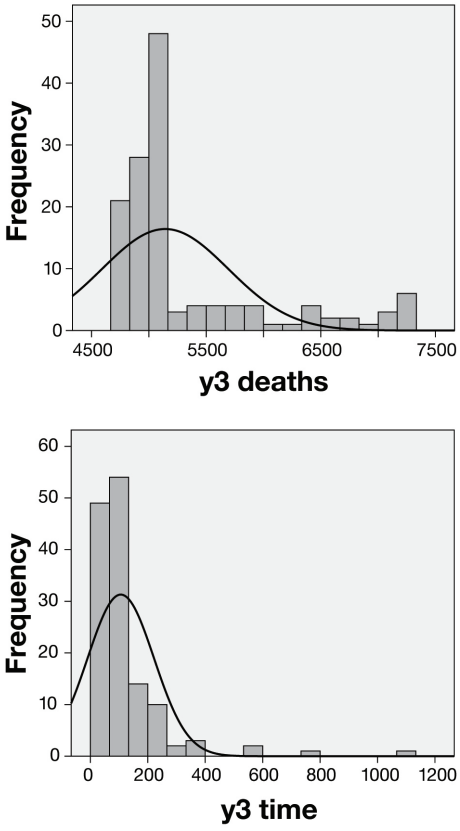


Figure A – 9
Frequencies and normal curve, deaths and time in 3rd task (y3)

Neither of the two variables have a normal distribution (Figure A – 9). For the y3 deaths variable (accuracy measure for the third task), tests of normality performed with PASW statistics 17 show that the distribution of results was

significantly not normal (Kolmogorov-Smirnov < 0.05, Shapiro-Wilk < 0.05). The same was true for the y3 time variable (time taken in the third task).

A non-parametric test should therefore be used to determine whether there is a significant correlation between the two variables. A Spearman correlation test reveals that there is a significant positive relationship between the time spent on the third task and how many lives were saved ($r_s = .484$, $p < 0.05$).

To determine which of the methods might be more suitable in more time-pressured situations, the participants' results can be split into three equal-sized groups, by how quickly they came to their decision in the third task. Shown as a box plot, this gives the results shown in Figure A – 10. The box plot indicates that when the time taken to decide was short, the answers chosen based on the information provided by the numerical display were more accurate, on average. However, as the decision time became longer, the two display methods seemed to produce similar results.

In the case of those that took the longest time to make a decision, the graphical presentation produced a much smaller variability than the numerical presentation, although the median result was similar.

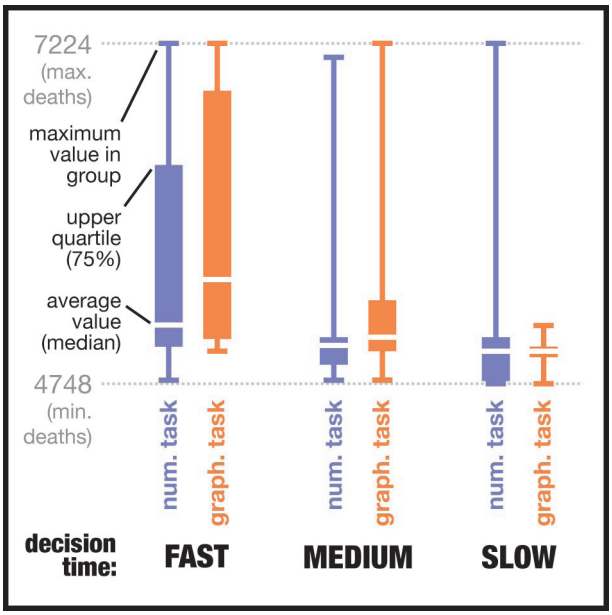


Figure A – 10
Deaths in task 3, by presentation and speed

A – 3.4 Preference effects

3) Can it be shown that people perform better, in terms of time and/or accuracy, depending on their preference?

Box-whisker plots can again be used to show the differences between those people that preferred the graphical presentation and those that preferred the numerical one. Figure A – 11 shows the number of deaths (accuracy) in the third task ("year 3"), divided into which presentation method was given, and which presentation method was preferred at the end of the experiment.

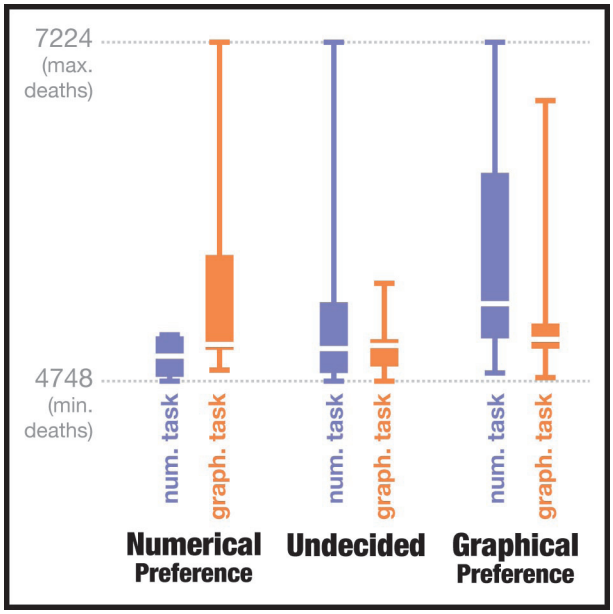


Figure A – 11
Deaths in third task, by presentation and preference

This seems to show that people were able to achieve fewer deaths with their preferred presentation method.

This can also be shown with a statistical test. As the data is not normally distributed, another non-parametric technique is used, this time a Mann-Whitney test. The data is split into two groups for the test: Those that preferred the graphical presentation and those that preferred the numerical presentation. As two tests are conducted, a Bonferroni correction is applied to avoid inflating the type 1 error rate, and significance is therefore reported at $p < 0.025$.

For those using the graphical presentation, although those that preferred the graphical presentation (Mdn = 2171 lives saved) did, on average, save more lives using that presentation method than those that preferred the numerical presentation (Mdn = 2207.5 lives saved), the difference was not significant ($U = 230$, $p > 0.025$).

For those using the numerical presentation, however, those that preferred the numerical presentation (Mdn = 2293 lives saved) did save significantly more lives using that presentation method than those that preferred the graphical

presentation (Mdn = 1910 lives saved). $U = 42.5$, $p < 0.025$. This represented a large effect ($r = -0.63$)

The results for each person in the third task can also be compared to their results in the sixth task (year 6) with a diagram that combines a box-whisker summary with a link diagram showing the actual results for each person in the third and sixth tasks (the two complex tasks with each presentation method). (See Figure A – 12a & 12b).

As many of the link lines cross each other, the diagram shows no clear direction of change, between tasks 3 and 6. For example, of the people that preferred the graphical presentation and were given it first, there do not appear to be noticeably more that perform worse (or better) with the numerical presentation in task 6 than they did in task 3. The only exception is those that preferred the numerical presentation, and received it first, who seemed to generally perform worse in task 6 (i.e. when they were given the graphical presentation).

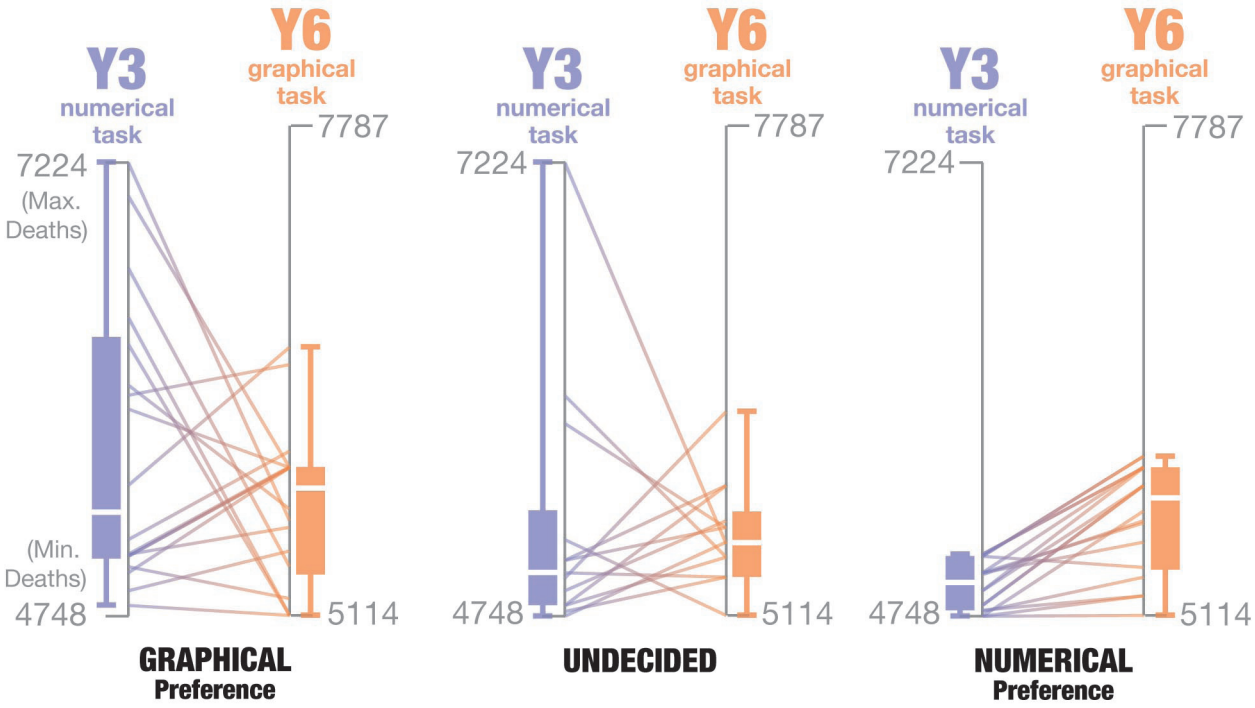


Figure A – 12a
Deaths in most complex tasks, by preference (numerical-graphical group)

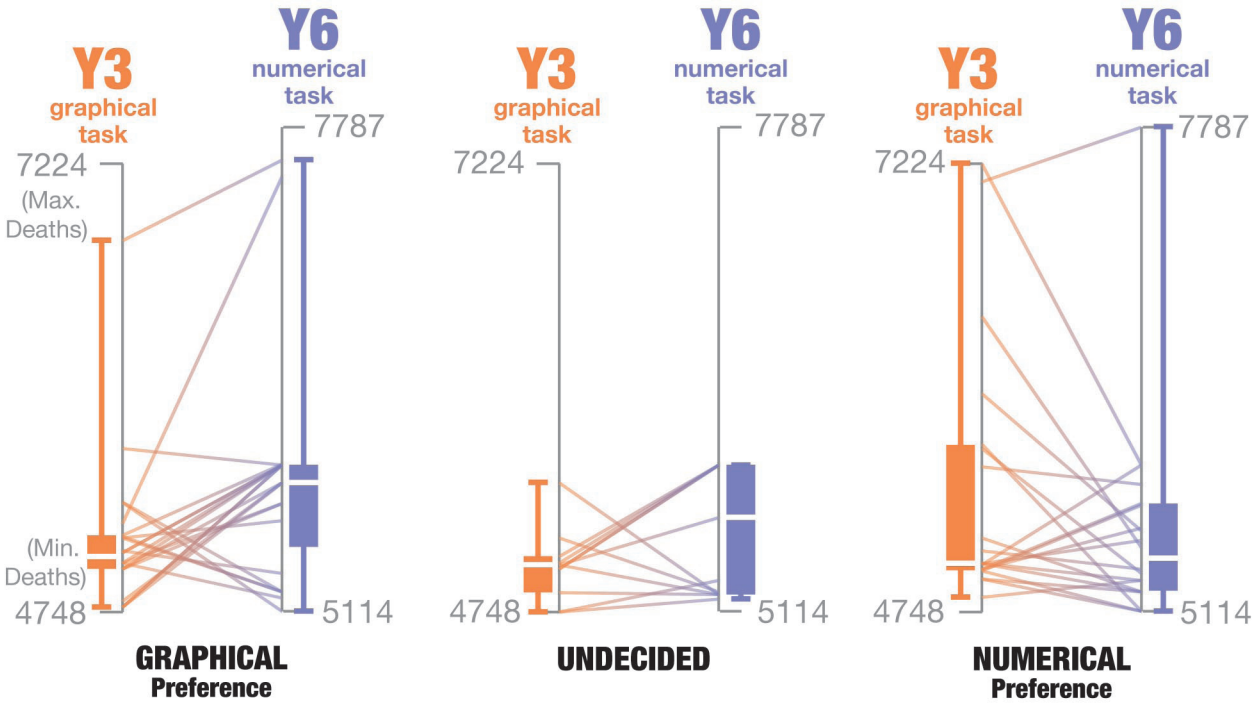


Figure A – 12b
Deaths in most complex tasks, by preference (graphical-numerical group)

Box-whisker diagrams can also be used to show the effects of preference on participants' time in this third task (see Figure A – 13). This chart suggests that participants tended to spend longer with their favoured presentation.

Participants that preferred the numerical presentation tended to spend longer when performing the third numerical task (Mdn = 112 seconds) than with the third graphical task (Mdn = 86.5 seconds) This difference is not statistically significant, however (U = 161.5, p > 0.025).

Those that preferred the graphical presentation also tended to spend longer if they were given the graphical task with the third set of data (Mdn = 95 seconds) than if they had the numerical task (Mdn = 72 seconds. This difference is statistically significant (U = 129, p < 0.025).

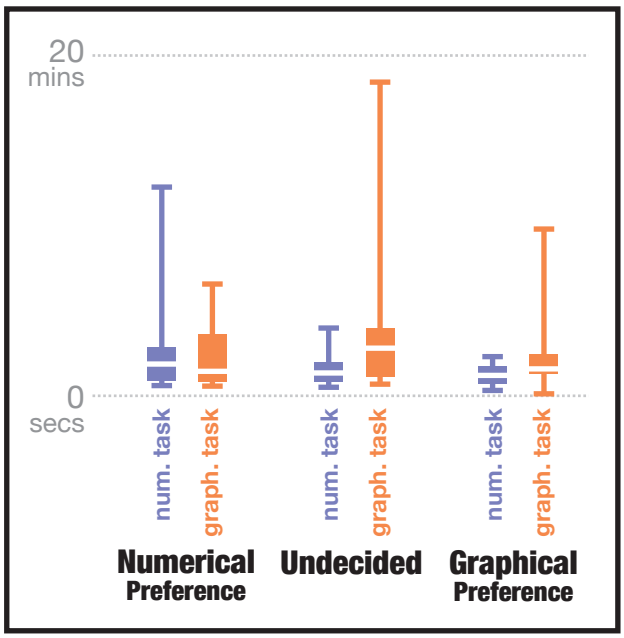


Figure A – 13
Time to decision in third task,
by presentation and preference

A – 3.5 Participant characteristics

4) Which characteristics can be shown to affect performance or preference with one or other presentation method? (age, gender, familiarity with health technology assessment, socio-economic group, continent of residence).

Two of the characteristics for which data was collected, socio-economic group and continent of residence, could not be used for the analysis. In the socio-economic group self-coding questionnaire, 85% of the people that responded put their occupation as "modern professional", automatically putting them

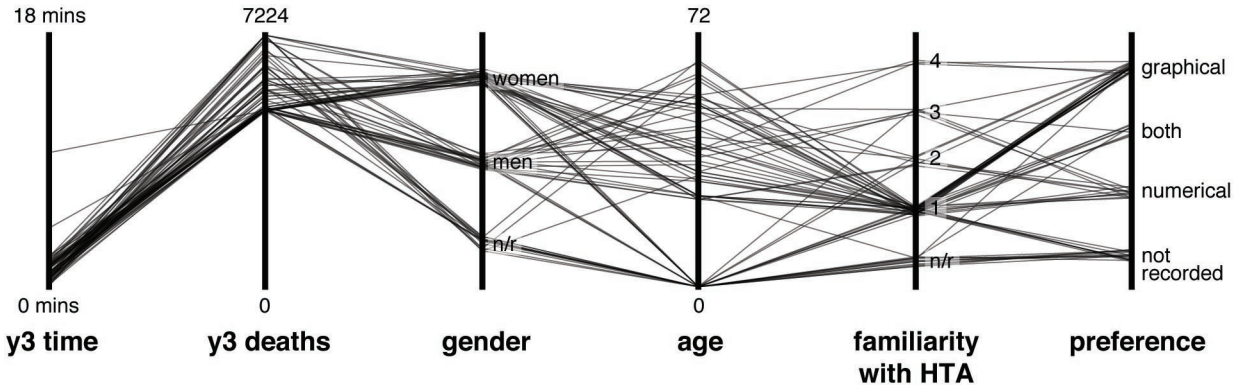


Figure A – 14a
Characteristics of those with lower decision accuracy (deaths above median in task 3)

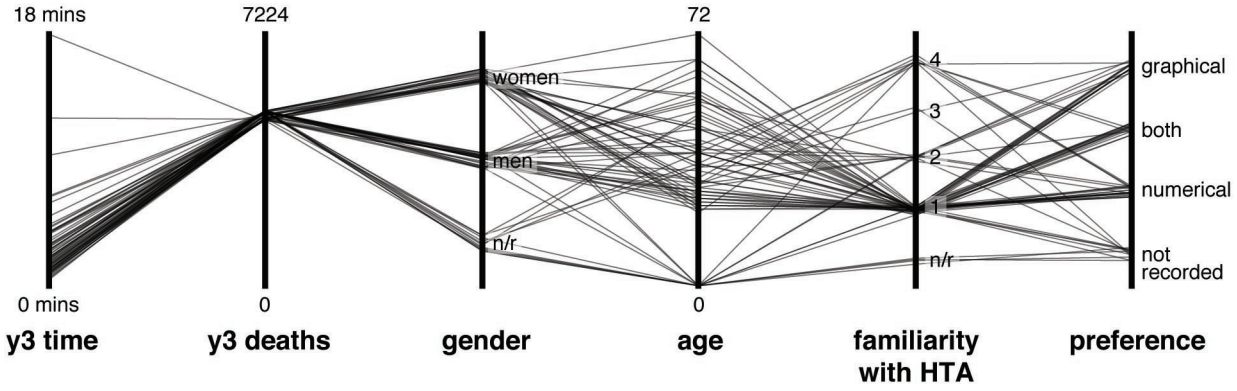


Figure A – 14b
Characteristics of those with higher decision accuracy (deaths below median in task 3)

in socio-economic group 1. In the end, over 90% were placed in this group, making any comparative analysis impossible. Similarly, the vast majority of participants reported that they lived in Europe, leaving other groups too small for meaningful analysis.

A set of parallel co-ordinates does not clearly show that any characteristic is likely to be common among those that were in the highest or lowest 50% of deaths in task 3 (see Figure A – 14a & 14b). In this figure, each person

is represented by a line that touches each of the vertical bars at a point representing that characteristic. Any very distinct difference between the two groups (such as those with more deaths being often younger) should be discernable to the eye. However, this is not the case.

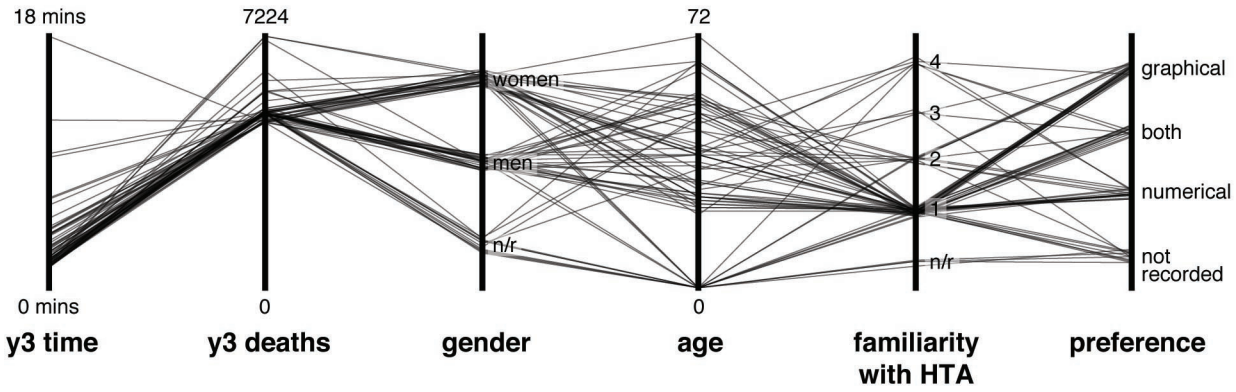


Figure A – 15a
Characteristics of those with lower decision speed (time above median in task 3)

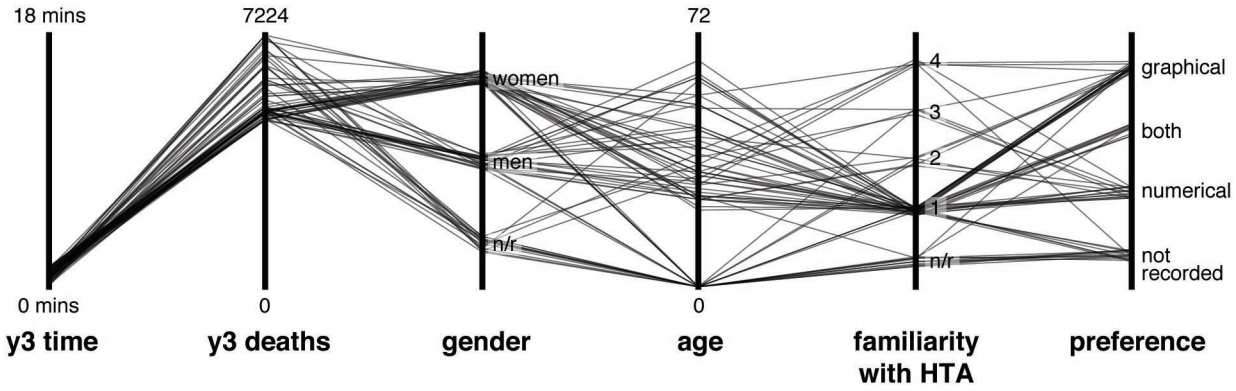


Figure A – 15b
Characteristics of those with higher decision speed (time below median in task 3)

Similarly, no clear pattern can be discerned for people that took a long or short time to complete task 3 (see Figure A – 15a & 15b).

A more detailed view can be shown using other techniques. Looking first at the effects of participants' gender, a set of box-whisker diagrams indicate that the number of deaths in task 3 are similar between the men and women that took part in the study, no matter which presentation method they used (see Figure A – 16).

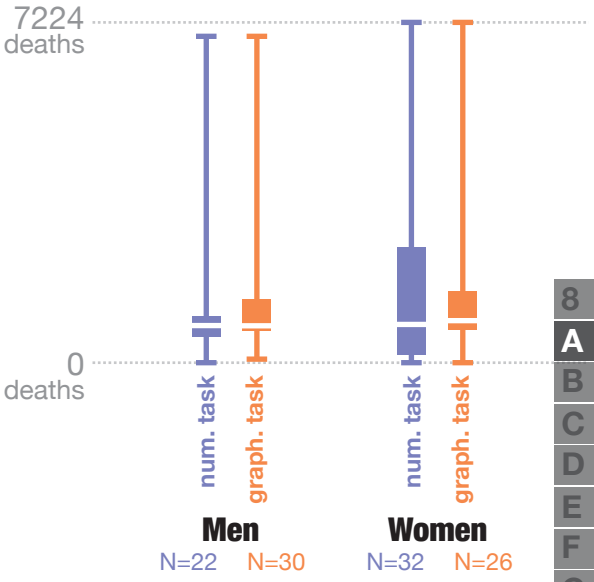


Figure A – 16
Deaths (accuracy measure) by gender

For the gender variable, a Mann-Whitney test indicates that men (Mdn = 2206 lives saved) did not seem to differ significantly in terms of decision accuracy from women (Mdn = 2171 lives saved). U = 1451, p > 0.05. The difference would have represented a very small effect (r = -0.03).

However, this test not giving statistically significant evidence of a difference is not the same as proving equivalence. Statistical equivalence testing is, however, challenging in this case. It would require definition of a 'zone of indifference' within which effects could be thought of as negligible, from a scientific or clinical point of view (Wellek, 2003). As it is impossible to have a clear idea of how accurate a participant should be within a particular time in this simplified decision context, this judgement can not be made in this case. Also, equivalence tests are based on calculating sample mean and confidence intervals, which may be less reliable in non-normally distributed data like that from this experiment. For these reasons, box-whisker diagrams showing range, median and quartiles are used in place of equivalence tests from here on.

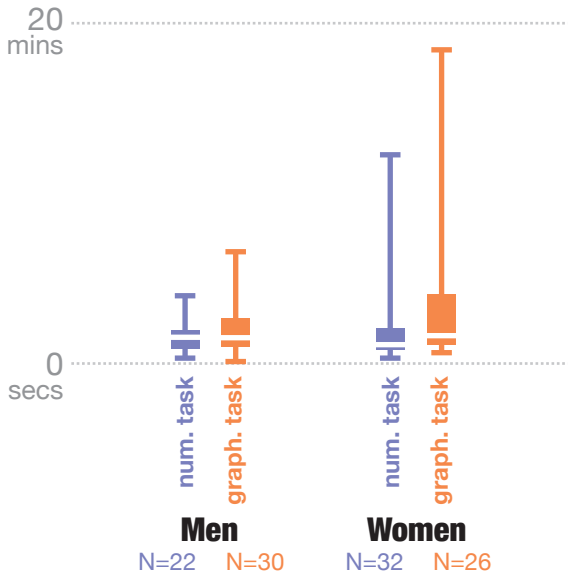


Figure A – 16
Time to answer by gender

Men and women also took generally similar times to answer in task 3, as shown in Figure A – 16. Women may have taken very slightly longer to answer with the graphical presentation, but the difference is unlikely to be significant.

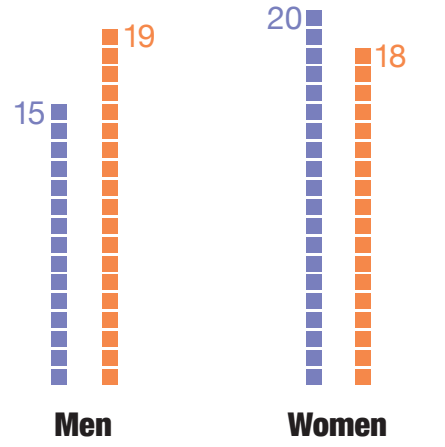


Figure A – 17
Preference by gender

Gender also had no significant impact on preference for one or other of the displays, as might be expected from Figure A – 17 (Chi-square (1, N = 72) = 0.5, p = 0.471).

Familiarity with HTA was self-reported by participants on a 4-point scale. Most participants (72%) reported the lowest degree of familiarity with HTA. As suggested in Figure A – 18, those that were more familiar with HTA were not significantly more accurate. A Kruskal-Wallis test suggests that deaths in task 3 were not significantly affected by familiarity with HTA ($H(3) = 3.721$, $p > 0.05$).

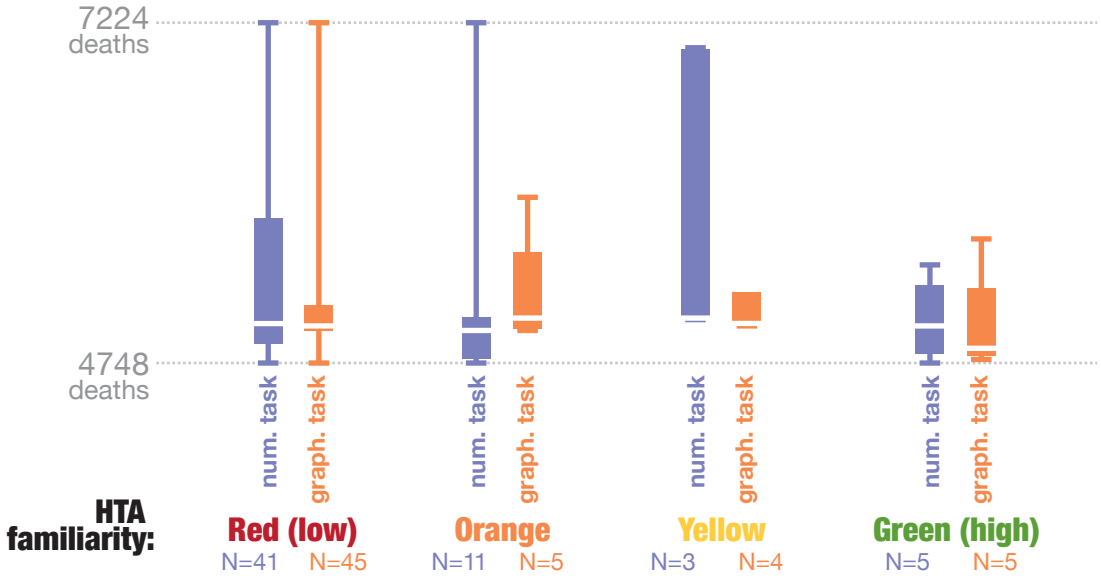


Figure A – 18
Deaths (accuracy measure) by HTA familiarity

Similarly, in the case of time to answer in task 3, no clear trend is observable for those with more familiarity with HTA (see Figure A –19).

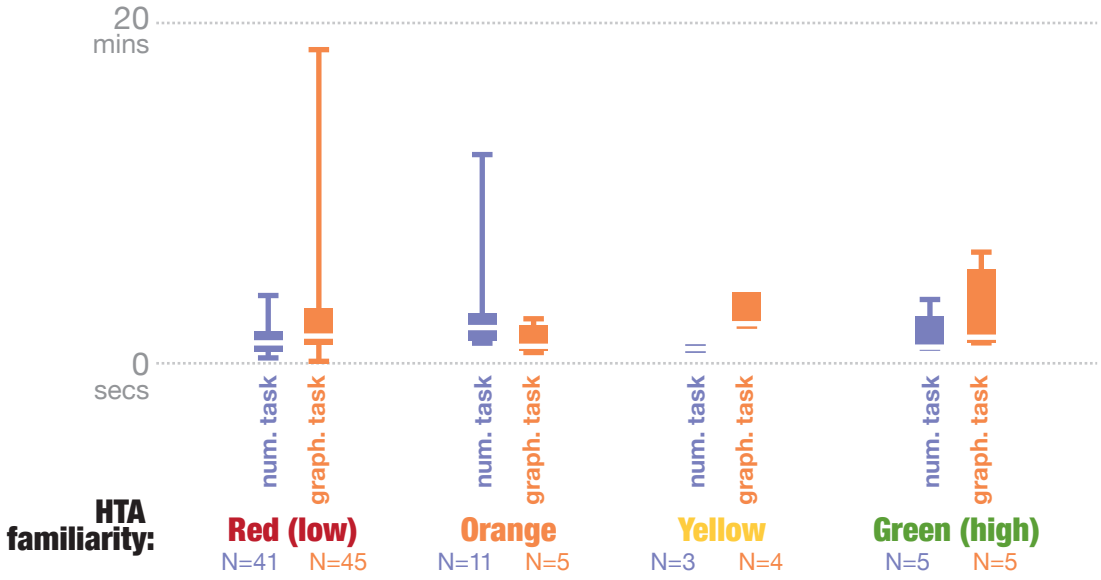


Figure A – 19
Time to answer by HTA familiarity

Preference for one or other of the presentations does not seem to be affected by HTA familiarity, as shown in Figure A – 20. A chi-squared test is not appropriate for this data, as four of the eight groups have expected counts of less than 5.

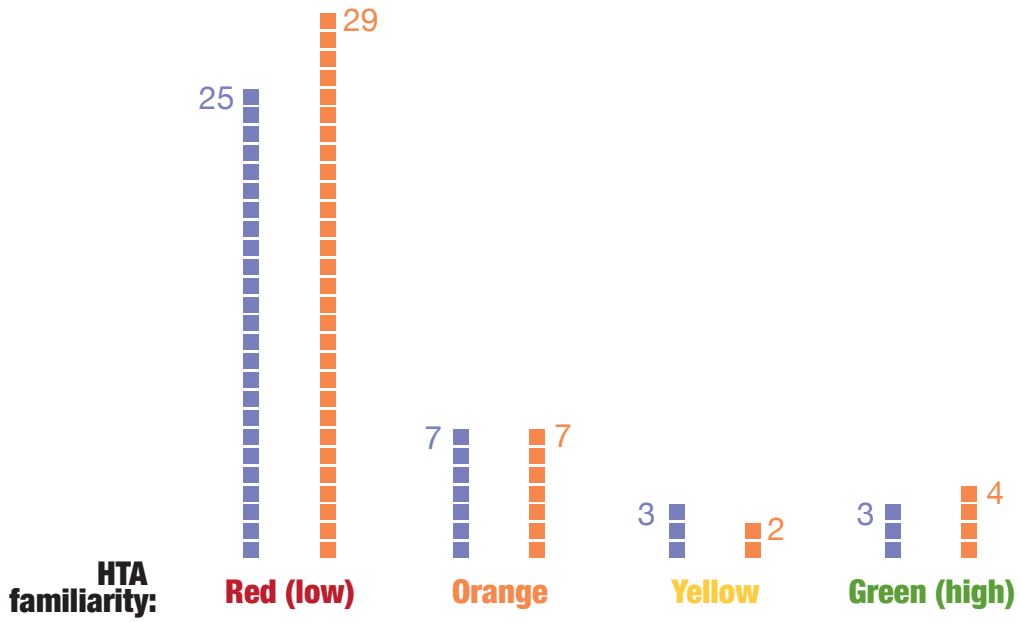


Figure A – 20
preference by HTA familiarity

While participants of a different ages seemed to have largely similar decision accuracy, only those below about 45 were able to achieve the very highest decision accuracy in task 3 (see Figure A – 21). This difference is enough that Spearman's rho suggests age significantly affects decision accuracy (rho = 0.021, p = 0.021). However, this scatter plot does not suggest that different aged people had higher decision accuracy with one or other presentation method.

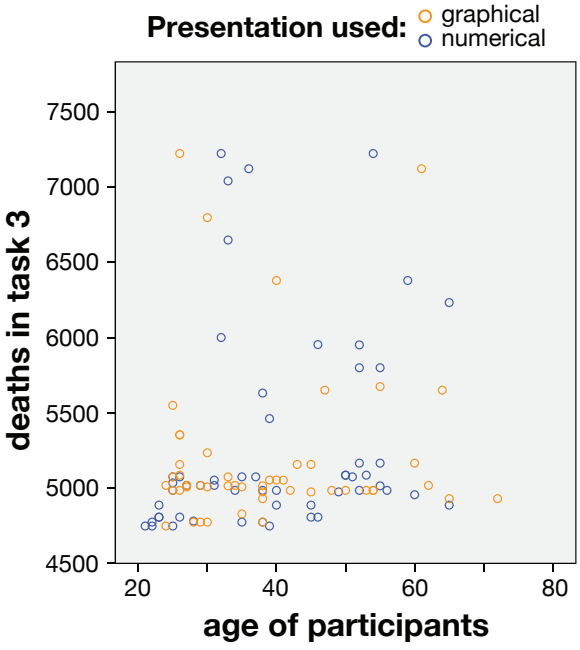


Figure A – 21
Deaths (accuracy measure) by age

However, the age of participants did not seem to affect time to answer in task 3, as shown in Figure A – 22. Given the relationship between time and decision accuracy, this suggests that the effect of age on accuracy can not be explained by those younger people taking longer to answer.

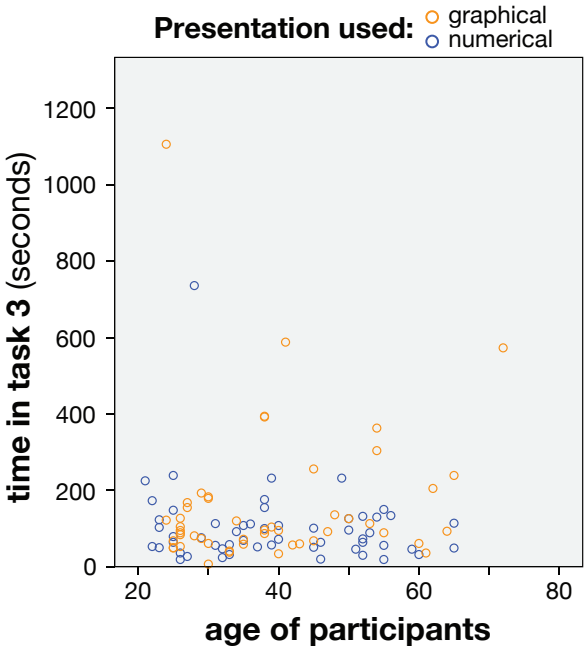


Figure A – 22
Time to answer by age

Age did not seem to affect overall preference, as shown in Figure A – 23.

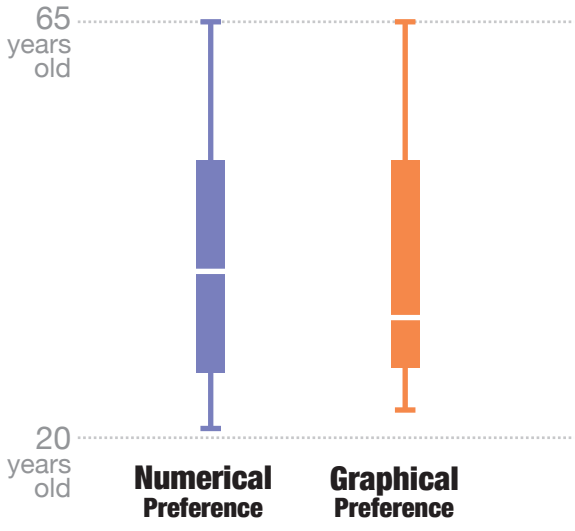


Figure A – 23
age by preference

A – 3.6 Qualitative feedback

The one way in which participants were encouraged to give qualitative feedback was through "comments" boxes, which they were invited to fill after performing all of their tasks. For the analysis, these responses were categorised into a thematic framework*, as follows:.

* Framework analysis is a method for quickly analysing qualitative data, developed for practical use in health resaerch. See Ritchie and Spencer (Ritchie & Spencer, 1994) for full details of the analysis technique.

- speed / time
- accuracy
- comparison
- confusion / noise
- ignoring graphical elements
- ignoring numbers
- layout/design comments
- budget
- decision strategies
- subgroups
- familiar presentations
- colours
- real situation
- mental computation
- rationality
- instructions / task understanding
- learning effects

Quite a few people mentioned that either the graphical or the numerical presentation would enable them to make a quicker decision, but there was little consensus as to which. Very few commented on the accuracy of their decisions.

In general, the graphical presentation was mentioned as helpful for making comparisons, and the "budget bar" was generally liked. However, a number of people found the graphical presentation more confusing. The duplication caused by presenting the data in both numerical and graphical form in one presentation seemed overwhelming for some participants.

Some people mentioned that they ignored the graphical elements, and focussed solely on the numbers, even when given the graphical presentation. Another person said that they "did not pay much attention to the numbers", relying on the visual elements, even when the numerical presentation was

shown (paying close attention to when the budget box went red).

There were a number of comments that suggested the graphical presentation was too complex, and that the type size was too small. Several people mentioned that a "holding tank" for best guesses so far would be useful.

Decision strategies were largely trial and error. However, some tried to be as equal as possible, not disadvantaging one subgroup over another. Some preferred to give more expensive treatment to higher risk patients, on ethical grounds.

Several participants noted that, in reality, more information would be needed for this kind of decision.

Different people mentioned that either the graphical or numerical presentation methods required more cognitive strain than others, depending on their preference.

Instructions and interface were described as many things from "clear and helpful" and "very good" to "irritating" and "poor".

Several people mentioned that, by the time they had seen the second information presentation, they felt more comfortable with the decision problem, and expected this to affect their performance.

A – 4 Conclusion

It seems that online, task based experiments may be able to provide quantitative evidence of the suitability of different information presentations for decision tasks. They could certainly be used in time-critical situations to determine which of two or more possible ways of displaying numerical data would allow people to make good decisions more quickly. However, quantitative evidence alone gives a limited view of the reasons for differences in performance, and can probably be strengthened by qualitative research.

This experiment was conducted on a simplified decision task, with a general public audience. For testing health policy information graphics, a more focussed sample would be needed, using experts in the field. However, the method of using a randomised quantitative experimental study, delivered through the web (probably with the addition of a log-in system so that the

researcher can be sure who is participating in the study) could provide a cost-effective way of carrying out future research.

1) What effect does increasing complexity have on the relative difference in results between the two presentation methods?

It seems that the increasing complexity of the decision task did not lead to any major differences in relative performance between the two presentations. The first complex task (year 3) did seem to take longer with the graphical presentation than with the numerical presentation. However, this difference was not evident in the second complex task (year 6), so may be an indication of the learning required for an unfamiliar information presentation, as well as the added complexity induced by duplicating the information in numerical form on the graphical presentation.

This finding seems different to those presented by Remus (Remus 1987) which suggested that graphical presentation methods are superior when complex information is presented. This suggests that generalisations about information graphics as a whole are not useful. Individual presentations of information are perhaps likely to be more or less suitable, depending on the qualities of the individual presentation design, the context of use, or the intended audience. This supports the claim made by Sless: that assessing the understanding of a document's users is more important than specifying exact typefaces, font sizes etc. (Sless, 2008,).

2) Can time and accuracy be shown to be dependant on each other?

In the specific, and rather contrived example tested here, the longer a participant spent on the tasks, the better their decision, on average. If we imagine that this is a real situation, we might conclude, in reference to Figure A – 10, that the numerical presentation was superior if the time for people to make decisions like this was very limited (especially if it was less than about 61 seconds, the cut-off for the fastest group). However, if the time available was longer (above about 3 minutes), the graphical presentation might be recommended, as it produced much more reliable results (participants scores did not vary as much).

3) Can it be shown that people perform better, in terms of time and/or accuracy, depending on their preference?

In this experiment, the two different information presentations (called

numerical and graphical) were broadly similar in terms of average performance (measured in terms of decision accuracy and time taken to decision). There were also similar numbers of people that preferred each presentation. The interaction between decision accuracy and preference was notable, however (see Figure A – 11), and statistically significant in the case of the accuracy of those that preferred the numerical presentation.

Participants also tended to spend longer with their favoured presentation method. Given the correlation between length of time spent on the tasks and the participants' accuracy, this may explain why the increased accuracy can be observed when participants use their preferred presentation.

The effect of preference on accuracy may also relate to the fact that participants were shown their scores against the average to that point in the study after each task. Remembering which presentation led to their most positive scores may have affected participants' preferences, which were collected at the end of the experiment. However, they were not shown an aggregate score with the numerical and graphical presentation methods, and would have had to remember how well they had done for quite some time for this to be the case.

4) Which characteristics can be shown to affect performance or preference with one or other presentation method? (age, gender, familiarity with health technology assessment, socio-economic group, continent of residence.)

The participants sampled were not sufficiently varied to give findings on the effects of socio-economic group or continent of residence. However, The groups for three of the predictors (gender, age, and familiarity with HTA) were large enough that even small to medium differences would have been picked up. In this case, gender and familiarity with HTA did not have any significant effect on their performance (time or accuracy).

Age seemed to have an effect on accuracy, in that only those under 45 were able to achieve the lowest possible number of deaths in task 3. This was not explained by decision time. One possible explanation would be that the younger participants were more familiar with web-based tools, or computer games with similar optimisation premises.

If our method of testing information presentations was applied in another situation, a particular participant characteristic might have a significant effect on the participants' performance with one presentation over the other. In such situations, it might be possible to use this information to provide appropriate

presentations to appropriate people, or at least to use an information presentation that would suit the majority of the people that were expected to use it.

A – 4.1 Recommendations for future research

We were able to draw some statistically significant conclusions from this experiment, and the method can in general be recommended for comparative testing of alternative information presentations. However, some subgroups were only just large enough for statistical tests, even with nearly 250 submissions to the web application - probably at least 220 unique people starting the test initially. It may well be that the target audience for an information presentation is a smaller group than this, in which case qualitative methods would probably enable a more comprehensive evaluation. It should also be noted that the differences, while statistically significant, may not be strong enough to make a practical difference in a real-world context.

If a quantitative, online test was to be used to assess information presentations for health policy decisions, a more focussed sample of expert users would be needed. A log-in system would be needed, and recruited participants could be individually sent a password to allow them to take part in the study. This would also allow for more sophisticated characteristic-based randomisation, but probably result in a reasonably large non-response rate.

The relationship between the time taken by each participant and their accuracy in this experiment caused difficulties for the analysis. It could not be determined how much importance each individual person had placed on time or accuracy, and therefore each outcome measure had to be treated separately. In future, it would make sense to treat one of these two variables as an absolute (i.e. how many times can a participant do something in a set amount of time, or how long does it take for a participant to reach a certain result). This might depend on which of these two variables is the primary concern, given the context of the research.

Even though the characteristics of the participants collected in this experiment were not associated with their performance, it should not be assumed that they should not be collected in other studies. Given the possible correlation between a person's preference and their performance with the different presentations, it might be productive to assess the learning styles of people using graphical and numerical information presentations. (see http://www.vark-learn.com/ and Gardiner, 1983).

Appendix B: NICE interview data

B – 1 Script for telephone interviews

NICE technical advisors
May 2009

Introduction

Thank you for agreeing to talk to us about the presentation of information for decision makers at NICE technology appraisals. As you may know, my PhD focusses on methods for displaying information in health technology assessment, and I'm looking particularly at the numerical data presented in TAR reports for NICE appraisal committees.

I understand that, as someone with experience of being a NICE technical lead for appraisal committees, you are in the ideal position to help us. We've identified several different situations in which new methods might provide benefits. Your answers to our questions will help us to focus on parts of the reports that could be better presented with new techniques.

Questions

- Interview should be about 15 mins long.
- Check it's okay to record (will only be used for our own research purposes)
- How long have you been a NICE technical lead?
- What do you see as your role in terms providing / summarising information for decision makers at appraisal committees?
- In your experience, which information in TARs is the most complex? It might be difficult to summarise, or it might prove confusing to decision makers.
- Can you think of a specific report in which this has been the case? (name of

intervention, approximate date)

- Have you ever had difficulty trying to fit HTA information onto a single page for presentation?

- Some of our techniques can be used to compare different information together. Can you think of any time that decision makers have asked for two or more separate pieces of information to be presented together? Or have you ever felt the need to provide this?

- Approximately how often does this occur?

- Can you think of any examples of time being limited for decision makers to absorb a large amount of information?

- Is there any information in TAR reports that is presented in an amalgamated form, which some decision makers will only need parts of?

- Can you think of any information in HTA which is difficult to understand for decision-makers for any other reasons?

- Any questions?

Outro

Thanks again for your time. We'll be developing several graphical tools in response to these suggestions, and testing them over the following months. Your responses will help us to ensure that we design with the end-user in mind.

We hope to disseminate the results of our research at HTAi 2010 and in the IJTAHC journal, as well as writing them into my PhD thesis. We're also hoping to collaborate with Leicester to provide some graphics in a live HTA report.

B – 2 Sample NICE interview transcript

Information needs interview 3

07 July 2009

Interviewer: Will Stahl-Timmins

Interviewee: ----------

0:00 interview begins

I'll start a recorder. I'll run the transcript past you as well if that's okay, just to make sure I've got it right. So, without further delaying I shall begin. I wanted to ask, first of all, how long you've been a NICE technical lead.

I have been at NICE for four years. I was a technical lead for two and a half of those years, and I'm now a technical advisor.

Ah, you're a technical advisor as well, just like ---------- I think is, as well.

1:00

Yes.

Okay, so you're one of the senior technical leads, really. What do you see as your role in terms of providing or summarising information for decision-makers at appraisal- well, what do you see as the role of a technical lead?

Really it's about taking the information that we get, a range of summaries- a range of submissions, and summarising them, so that they're sufficiently short and digestible and easy to understand, so that the committee members then feel sufficiently comfortable to make a decision.

I see. And that's information provided from the TAR teams and from the manufacturers.

Yeah, and for an MTA it might be from other professional groups and stuff like that as well.

2:00

Great. Okay, thank you! In that case, I'll ask you: In your experience, which (particularly numerical) information in TARs, particularly, is the most complex. It might be difficult to summarise, or it might be confusing to decision-makers.

I guess in terms of quantity of information, then the problem is usually with the clinical effectiveness, and summarising that, because that's something that you really do have to get across in a single slide, or a single page. Where it's like an STA or close to licensing, and we only have one or two trials, that's obviously very easy. But when you're looking at an MTA of devices, or older

drugs, then the sheer quantity of trials and studies that you get, and it can be across different subgroups or indications, makes that a much harder task to do. That's usually the biggest job for me, finding a way summarising that kind of information.

3:00

So particularly MTAs or where there's a lot of subgroups, or there's some other compounding factor, it starts to get really difficult to put that information across in the short space of time available at the committee.

Yeah. Really, it's about getting that information written down. So in the overview, or in the slide presentation, because that's not the information that the committee wants to linger on. But you have to get the ultimate message across. But really, you're always aware that you're meant to be getting onto the cost-effectiveness as quickly as possible. But you can't not give them all of the clinical effectiveness data.

That's really interesting and useful information. I was wondering if you might have anything in your mind that's a specific example of a report which it was difficult to do that for?

4:00

I think there's the PenTAG assessment report they did for Cochlear implants, which was obviously just very large, looking at both adults and children, severe to profound deaf[ness]. That had a lot of studies in it with very wide- a lot of outcomes. That took a lot of summarising and there is an overview which I did for that which shows, in the end, how I did it. Kind of taking the PenTAG stuff and making it a bit smaller. There are also other MTA assessment reports. There's the corticosteroids for asthma, that was done by SHTAC. And then there's another one. A spinal cord stimulation one, which we did last year, which was a ScHARR one. It should have four different pain conditions, and looking at spinal cord stimulation for that.

5:00

Oh, I see. So, different severities, or different degrees of pain.

Different- I can't remember the word now.

I'll have a look at that in more detail. It's great to have some examples so that I can see what you mean by the difficulty in presenting clinical effectiveness information.

What you will be able to see is, for all of them, you'll be able to see what we had in the assessment report, and then there'll also be an overview, so you'll also be able to see how the technical team handled it- modified it. We also have lead team presentations. Which, if you like, take it to the next level. How it then gets summarised into PowerPoint slides.

It would be really interesting to see some of those, actually.

6:00

...by the lead team. The lead team presentations aren't in the

public domain, but we still have them all, so they could be dug out for you if you wanted.

Oh, wow. That would be really amazing. If I think of some specific ones where I'd be really interested to see how it's been done.

Yeah, if you find a couple of good examples that you want, yeah.

That would be brilliant. Okay then. Well in that case I will move on to my next questions. I don't want to take up too much of your time. That was a really interesting response to that one. So, I'll move onto question five, which is: Have you ever had difficulty fitting HTA information onto a single page for presentation. I guess that's pretty much what you were explaining there- about summarising information onto a singe page. But was there any other time that you've had difficulty fitting information onto a single page?

7:00

No, that's really my main difficulty. With the models you find that it seems to just chunk up more intuitively, so you can divide it by cost- you can have a slide of cost, and you can have a slide of health-related quality of life, and then you can have a slide of incremental costs and QALYs, and if you really want to go full way you can have a slide of model structure. There's a much more intuitive way of actually breaking it up, and it's kind of what the committee want to focus on. So I don't really have the same problem summarising the economic stuff.

Well, that's a good answer in itself, because it means that we can focus a bit on that, perhaps. I'll move onto question six. Some of our techniques can be used to compare different information together. So I'm wondering if you could think of any time when decision-makers have asked for two or more separate pieces of information presented together, or you've ever felt the need to take different bits of a report which are quite widely spaced, and present them together?

8:00

There's something very useful. Sorry, it's absolutely chucking down with rain here... PenTAG did something really good with the economic model- they did it for cochlear, and I've actually tried to encourage other groups to do it as well. Whereby where they're summarising the model results they will, in a single table include the key costs, and the incremental QALYs from the different models that you get. And I find that very useful way of doing it, and have encouraged other people to do it, and have actually copied it into my overviews and stuff.

Ah, that's exactly the kind of thing I'm looking for, so that's really useful to know. Thank you. I'll have definitely another look at the cochlear implants one. That was just happening as I started here, so I noticed it happening around me, but I'll look in more detail at those particular bits of it. Does that happen often, that you feel this would be useful?

9:00

Now that the MTA is dying out, I don't see as many of them.

The endangered species of the MTA?

Exactly. [laughs] It's the kind of thing that much more relevant to MTAs where you get multiple models.

Okay, well that's fine. I see what you mean. Okay, I'll go on and ask if there's any examples you can think of time being very limited for decision-makers to absorb a lot of information.

I think my default position is that they always have limited time. And you're always trying to summarise it in a way that minimises the amount of time that people have to spend on it. I think committee meetings are the ultimate example for that, but with overviews, they would only get the weekend to look at it before the committee meeting, so you're always mindful of that.

10:00

Okay, are there any points in the committee meeting where you know it's going to slow down because there's a lot of information to be absorbed?

The committee meeting will always start with the lead presentation from one of the committee members, which is meant to summarise it, and that will take, what, 20 minutes? And that's meant to give a good overview of the evidence submitted, and then identify the key issues for consideration, which then get discussed and considered by the committee. So, it's not much time. I do have to- because it's now all in public, you do have to make sure everything's in there.

11:00

Yes, of course, it's sort of showing everything, but having such a short space of time to do it in, so you can't linger. I guess it does help to have the reports in advance though, so maybe a lot of that absorbing information might happen before that anyway.

Oh, yeah. Well, you hope. Yeah. You know, they should.

Okay. Right, question number eight is a bit of the reverse of one of my earlier ones, actually. I was wondering if there was any information in TAR reports, that's presented in an amalgamated form? Which some decision-makers might only need parts of? So is there anything that you feel needs to be brought out just for particular people? Or do you need to present absolutely everything to everybody?

Usually, no. It would usually all get presented to everyone. There isn't... no.

12:00

That's fine... That's an answer that I'm having from everybody, actually, so far, so that's good. That means that that particular set of things is something that we don't need to look at. But perhaps some other things are. That's one of the things that I really wanted to get out of the telephone interviews was a bit of focussing, so that's good. Okay, we're coming to the last few questions, but I just wanted to ask if there's any information in HTA which is difficult to understand for decision-makers for any

other reasons than I've just said?

13:00

I mean, obviously for us, they all come from different backgrounds, so not all of them have the same set of skills that they're using to interpret the information and that brings with it it's own challenges, because you're not just- you can't necessarily summarise the cost-effectiveness information in terms which a health economist would understand, and are sometimes the most succinct. You kind of have to do it in the old plain English.

Any kind of- we're looking mainly at numerical data, that might be overwhelming, or not easy to understand or anything like that.

Numerical data-

Don't worry, there may not be... It's not a particular issue. It's just, I always like to finish with an open question in case I've forgotten something really obvious.

14:00

Yeah, no, I mean there are things that are probably difficult for people to understand, like the probabilistic sensitivity analysis and all of that, but that's not really because of how it's presented, that's because none of them are health economists. And even I, not being a health economist, look at all the little CEAC clouds and think: "Well, that's very pretty, but what is it really?"

Everybody up to this point has mentioned the probabilistic sensitivity analyses.

I mean, it's very nice, but I'm never quite sure, and I have heard committee members say: "well, so it's 47% likely to be cost-effective. 47%, so what's the threshold? You know what I mean? Who knows?

So it's not necessarily a presentation issue, but an issue with the complexity of the methodology? I suppose?

Yeah. I think so.

15:00

It's also interesting to understand that, as far as I'm aware, different ways of doing a sensitivity analysis can end up with slightly different results, so, which one do you use? I guess it comes down to doing absolutely everything, and then seeing-

Yeah, that has happened.

Anyhow. Do you have any other questions? I've been asking you several, but you're more than welcome to ask any to me, if you like.

No, I don't have. No, I don't think so. I think I've told you all my good examples.

Okay, that's brilliant. I'll have a closer look at the one's you've

mentioned. I'm getting quite a few out of doing the interviews so far.

16:00

I think sometimes the assessment groups present cumulative analyses without always taking them apart, where they incorporate multiple things and present you with single things. But, that's not really about presentation. Well, it's kind of about presentation because you end up with a single value which include three different things, and then the committee only want to think about one of those things, and you can't disentangle it, the one thing that the committee are interested in, from all the other things in the analysis.

I see, it's almost like you need to separate things out, in a clear way, so that each individual bit can be brought out separately, before it's all brought together.

Yeah.

17:00

There's an amalgamation. Okay, well, that's really interesting as well. You may have gathered before now that what we're hoping to do is find graphical tools for presenting information. I'm sorry about all the cloak and dagger stuff and not mentioning that, but as soon as you start talking about graphics, people start thinking about line graphs, but we're really interested in where the graphics aren't, if you see what I mean? But that's hopefully what we will be presenting, some new ways of presenting things graphically, which might (or might not) help. And I'm designing a few over the next few months, and this is part of how we like to design with the end-user in mind. We like to have a really close relationship with them, and be part of that process of (presenting?) evidence, that's why we decided to talk with the technical leads. So, if it's okay, I'll type up the transcripts, which does take me a little while, but I will get back to you with those, and hopefully you see some of the results at some point soon. We're going to disseminate them at HTAi 2010 in Dublin, and probably in the IJTAHC journal, and obviously it'll be part of my PhD thesis as well. And we're collaborating with a project with Leicester to produce some graphics in a live MTA, so you might see our work in- we'll definitely stay in touch. If you think of anything else that you've missed out, feel free to get in touch.

18:00

I'll let you know.

You've got my contact details. Thank you very much for your time, ----------

interview ends

B – 3 NICE interview terms

Technical terms, data types and information presentation issues identified during analysis of NICE telephone interviews.

3.01 background of condition

Not mentioned in any interviews.

3.02 clinical effectiveness (general)

This section of the reports pools together the available evidence, to show how effective the technology or technologies being assessed are. Problems with presentation arise where there is an unusually large number of trials relevant to an appraisal, or where there is no evidence that relates (directly) to the condition.

3.03 survival

There are various ways of showing survival, which may affect the eventual results "Exponential" and "Weibull" were both mentioned. It's probably quite a common outcome of interest, so it's perhaps not surprising that it was mentioned several times. May be problematic where survival of different subgroups are shown.

3.04 QoL/utilities

Quality of life is one of the most important considerations in the appraisal process. The measurements used for quality of life can affect the outcome of a trial, and different utility measurements are often included in sensitivity analysis, sometimes having a large impact on the outcome.

3.05 multiple outcomes

"Outcomes" is often used as a shorthand for "outcome measures". Also the terms "indication" and "licensed indication" are used to describe a measure of the effectiveness of an intervention, although these are subtly different in that it is these that are specifically what the intervention in question is allowed to treat, and outcome measures may be broader than this.

If many different outcomes are important to consider for an intervention, it can make certain parts of the analysis difficult to present, particularly where there are many subgroups. It may be difficult to show a combined analysis of

several technologies together if there are multiple outcomes.

3.06 subgroup analysis

Large numbers of subgroups can lead to very complex data to report, particularly when compounded by multiple outcomes, multiple sensitivity analysis criteria or multiple technologies. It is possible (in cases like the Hepatitis B MTA) that different treatment sequences can be appropriate for different subgroups.

3.07 costs

The cost of the treatment can be a key driver of the results of an analysis. They can also be uncertain, as prices for treatment and administration change over time. It is not unknown for decision-makers to disagree with costs at the appraisal stage. It may be important for decision-makers to see the magnitude of costs.

3.08 economic analysis / cost-effectiveness (general)

The term "economic analysis" is used quite interchangeably with "economic model" so it's sometimes quite hard to tell whether the word "model" is being used to refer to the entire economic analysis, or the structure of a model.

It's important to consider the economic analysis in light of where the inputs have come from. They can be very complex, and have lots of numerical data outputs that are difficult to summarise.

3.09 model structure

Decision makers like to see how a model is working, and know where its inputs come from. Knowing the model structure is important, but it needs to go deeper than that.

In MTAs, there are sometimes multiple different models from manufacturers to consider.

3.10 one-way sensitivity analysis

Sensitivity analysis can lead to great complexity, especially where multiple sensitivity criteria are repeated for different subgroups or technologies. It is possible to reduce this by just presenting the results of a few "scenarios". The results taken from the effectiveness reviews are often presented as the "base

case scenario". Other scenarios may be taken from other trials or groups of trials.

3.11 probabilistic sensitivity analysis (PSA)

The idea of taking not a single value from the evidence but using a probabilistic distribution, and sampling from it many times for different variables to create a picture of how uncertain the output of a model is. This can be conceptually difficult for decision-makers - not only in understanding what is being done, but in how to incorporate this into their decision. How certain do the results have to be for them to be able to decide?

3.12 cost-effectiveness acceptability curves (CEACs)

Decision-makers have difficult interpreting these, especially for multiple comparisons. Again, the problem of incorporating uncertainty into a decision is a difficulty.

3.13 Incremental cost-effectiveness ratio (ICER)

The ICER is a number, nominally representing the cost of buying a year of perfect health for one person. It is affected by many other variables all through the analysis, and the committee will probably spend longest discussing those things that affect it most. It can be tempting to focus on the ICER almost exclusively, but it's important to see where it comes from, what affects it, and have an idea of how uncertain it is.

3.14 multiple/mixed treatment comparisons

These can be complex, particularly if there are also multiple patient subgroups / large sensitivity analyses / multiple outcome measures. Presenting them seems to be a challenge - CEACs are difficult to interpret with multiple technologies being assessed.

3.15 sequential treatments

Treatment sequences are difficult when the clinical evidence doesn't include sequencial treatment. Different subgroups may require different sequences. The terms "dominated" and "extendedly dominated", which are used to describe treatment sequences that are not worth investigating further are not well understood.

3.16 absent information

Missing information is always a challenge in visual presentation. It was only mentioned twice in the interviews. Once in relation to the Hepatitis B MTA not having clinical evidence on sequencial use, and once on treating anaemia in chemotherapy patients, where it was impossible to include information on the fact that the different chemotherapy drugs are more or less likely to cause a kind of anaemia that is responsive to erythropoietin.

3.17 changing assumptions at committee

Examples were mentioned where the committee disagreed with the type of utilities used, the costs for administration that were used, and the assumptions about long term disease progression. It is possible that a committee member will pick up something that no-one else has noticed, that changes the result presented to the committee. Care has to be taken if an assumption has been changed, as going back and discussing a subgroup issue, for example, might be based on the base case scenario, not the one that they've decided on.

3.18 how informed are decision-makers?

two interviewees mentioned that there was an assumption that the committee members have read the reports before the appraisal, although there is no way of telling whether this is the case.

3.19 methodology

Methodological issues, such as how the meta-analysis or cost-effectiveness modelling was done, can slow down committee meetings. It was hinted that different methodologies might have different results, but this may not be an issue that graphical presentation can help with.

3.20 base case assumptions

These often come from the published evidence, and act as a starting point for the cost-effectiveness modelling. Manufacturer's base case scenarios may be different from the technology assessment group's (and each other's).

3.21 uncertainty

Often presented with CEACs or scatter plots. The uncertainty in a review can come from the effectiveness data, around subgroups, costs, and probably other things. It can be problematic to take account of uncertainty in a decision,

however.

3.22 belief / trust

A lot of trust is placed in the technology assessment group. If they say they've done something, the committee believes that it has been done to a high standard. However, they don't like to see things as a "black box", and like to know where the results have come from.

3.23 current use of graphics

probabilistic sensitivity analyses from reports often include graphics, but they are not included in the presentations given by the technical team at NICE. Although they are useful for the technical team's understanding, they don't feel they are useful for presenting to decision makers.

CEAC graphs for multiple comparisons are described as "rather peculiar" and difficult to understand. One person even wondered if committee members understood CEACs and scatter plots at all.

3.24 MTAs vsSTAs

Summarising the evidence is more challenging for the technical teams in MTAs, where the technology assessment report is not the only source. MTAs can have larger volumes of complex data, especially those with many trials, subgroups, or multiple indications. Sequencing is difficult for STAs - it's hard to answer questions about where it should go in a sequence. Manufacturers are increasingly often asked to resubmit their analysis in STAs.

3.25 different backgrounds

Appraisal leads that are used to the HTA process might do the lead presentation themselves, or technical advisors might help them to various degrees. Not all the people involved with the appraisals are health economists, so much of the information presented is difficult to interpret. Different people may find different pieces of information more or less complex.

3.26 face validity

Decision-makers like to be able to check the "face validity" of results, rather than just being presented with the outputs of the analysis. Face validity can be dangerous, however, if a reasonable-seeming ICER is presented that is based on very poor or no evidence.

3.27 scenario analyses

some scenarios are often picked out in an analysis, such as the different manufacturers' base cases. However, too many scenarios presented can be confusing, particularly in terms of which one was being looked at.

3.28 quality of evidence

even very bad quality evidence can lead to plausible-looking outputs, such as ICERs.

3.29 Interim analyses

mifamurtide for osteosarcoma was mentioned as a time when the analysis had to be repeated several times.

3.30 discrete event simulation

A lenalidomide appraisal was mentioned, in which the manufacturers had used discrete event simulation, along with other unusual methodological choices. This was reportedly quite confusing for the committee, and took some time to go through.

3.31 pair-wise comparisons

a table showing costs, QALYs, incremental costs, incremental QALYs is useful for pair-wise comparisons, but becomes exponentially more data heavy with more than pair-wise comparisons.

3.32 large volume of trials

the PenTAG cochlear implants assessment was mentioned as having a large volume of trials.

3.33 quality-adjusted life years (QALYs)

QALYs are one half of the ICER calculation, representing the quality of life attributed to an intervention. An incremental QALY represents the amount of quality of life gained by adopting a given intervention. Several of the interviewers thought that it was important to see how many QALYs were gained, rather than just relying on the ICER, which incorporates them.

3.34 mixed treatment comparisons

where the intervention contains more than one treatment, often in different sequences, it can be difficult for decision-makers to come to an understanding. It can take some time to explain.

3.35 probability

what happens if a treatment is 47% likely to be cost-effective? Even if a probabilistic sensitivity analysis or uncertain result is understood, it may be difficult to translate this into a definite decision.

3.36 licensing

a manufacturer's decision problem may be different to what NICE is interested in (which is generally an intervention's licensed indication).

If the appraisal is an STA which is close to licensing (ie, it is a very new intervention), and there are very few trials, it is much easier to present the data.

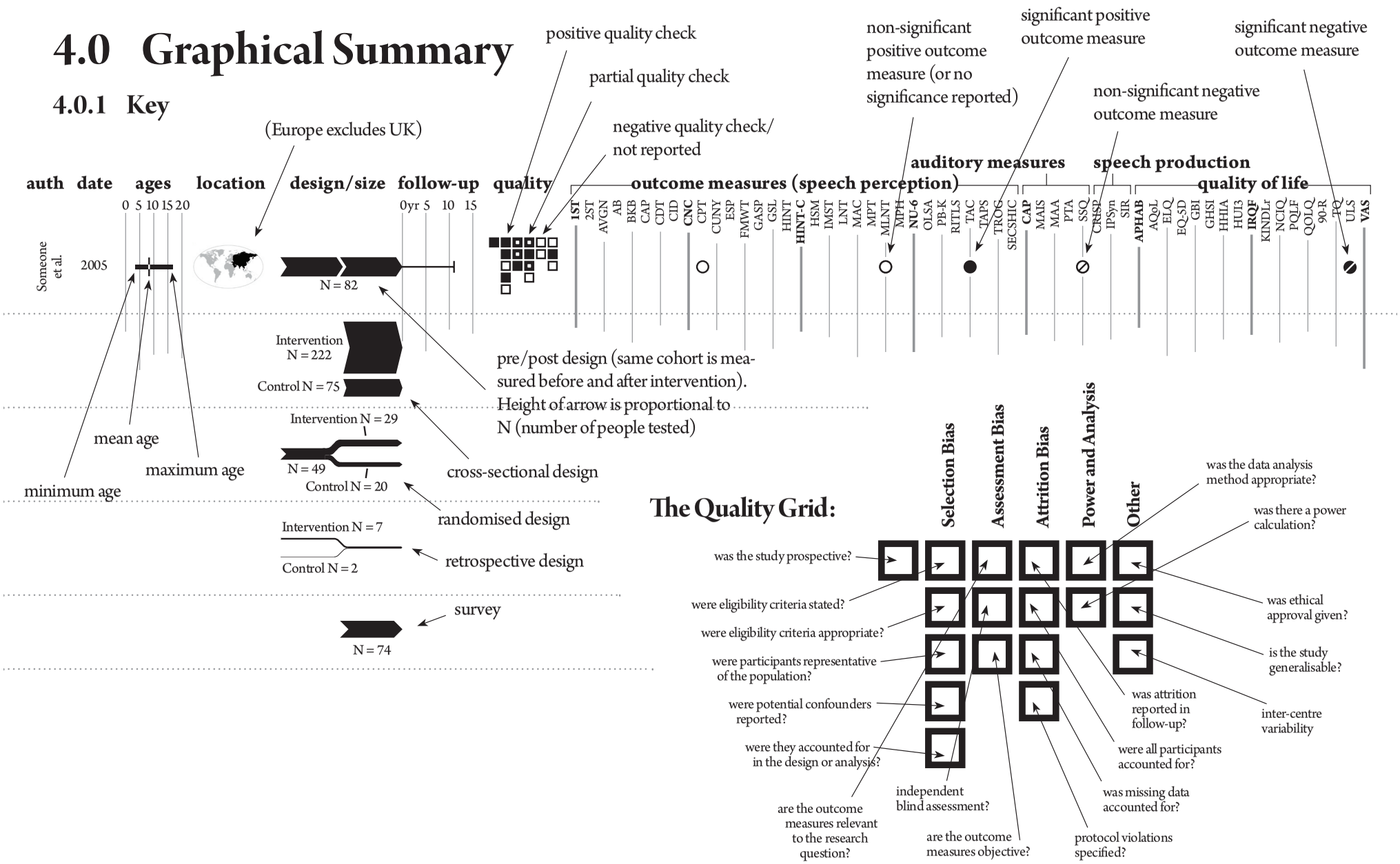
Appendix C

GOfER graphic

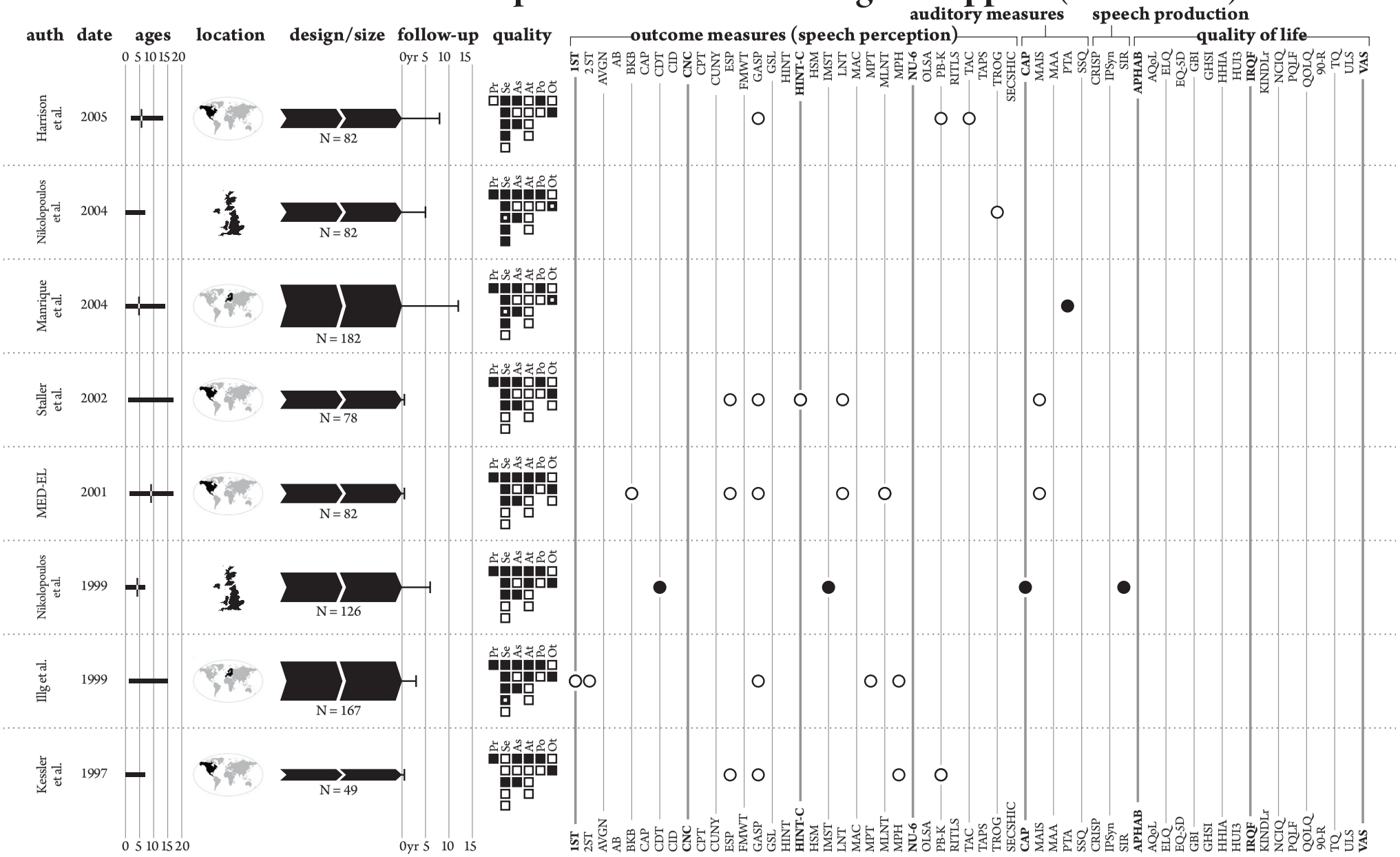
The version of the GOfER graphic that was tested in Chapter 5 is presented over the next five pages. The graphic has been scaled to 90% of the tested size, to fit within the required document margins for this thesis.

4.0 Graphical Summary

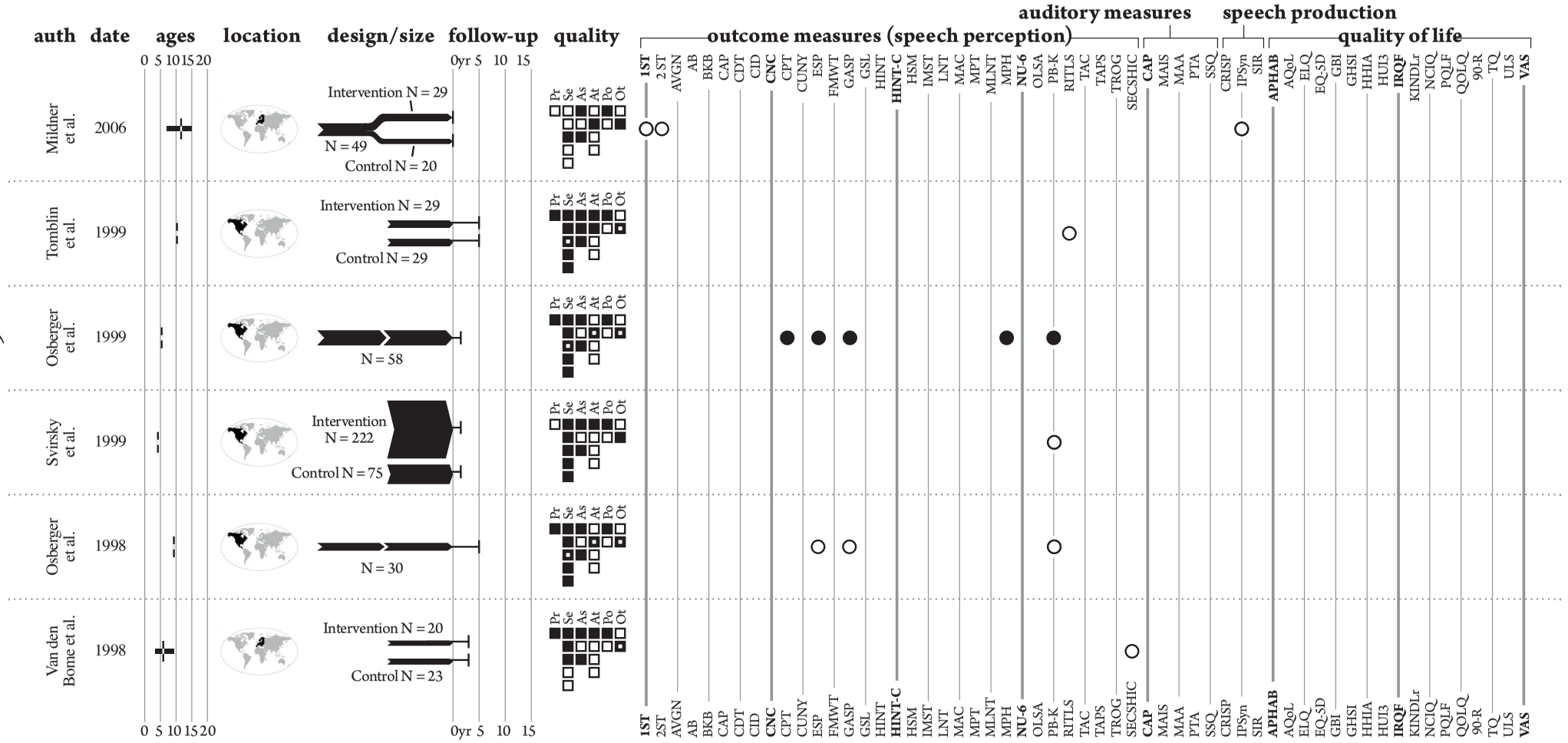
4.0.1 Key



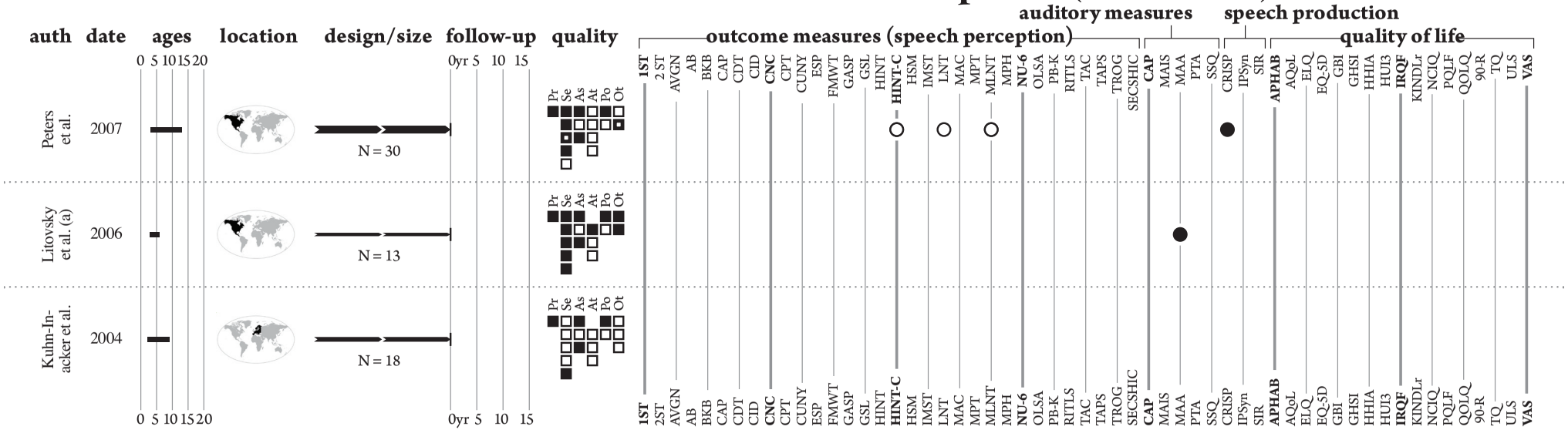
4.0.2 Unilateral cochlear implants vs. non-technological support (in children)



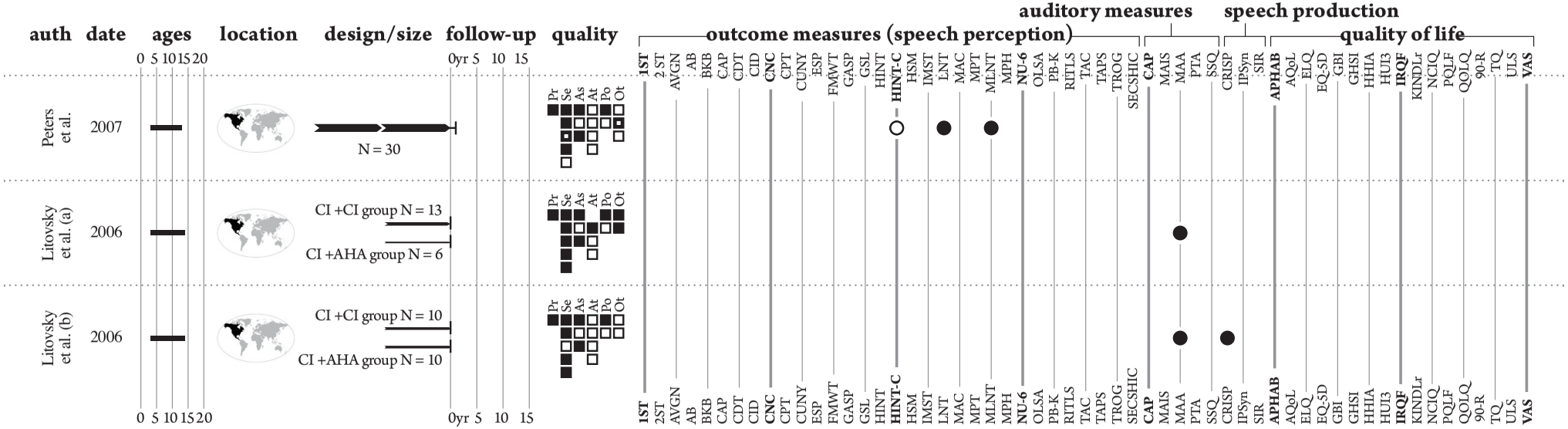
4.0.3 Unilateral cochlear implants vs. acoustic hearing aids (in children)



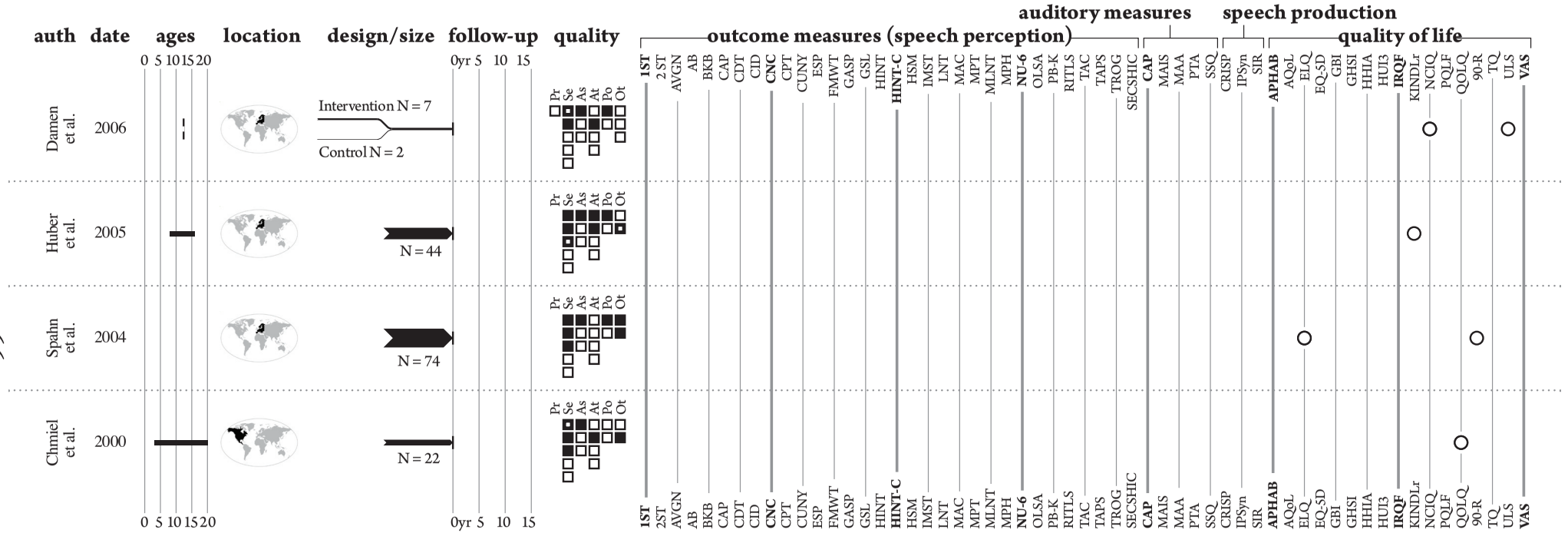
4.0.4 Unilateral vs. bilateral cochlear implants (in children)



4.0.5 Bilateral cochlear implants vs. unilateral + acoustic hearing aid (in children)



4.0.6 Quality of life with cochlear implants (in children)



Appendix D

Script for GOfER test interviews

24th Feb 2010

Intro

Thank you for agreeing to help us to test the new information presentation format that we have developed. The answers that you give us today will help us to refine our techniques, and inform how we use graphical techniques in presenting HTA research in the future.

I should tell you before I begin what we plan to do with the research. The analysed results of all our interviews will be presented in my PhD thesis, along with the transcripts in full, but anonymised. They will also be presented in a poster at HTAi in Dublin, and possibly a follow-up paper. I will not give them to any third party, or use them for any other purpose.

I'd like to check that you are happy to be video recorded? It will enable me to concentrate on the tests, and allow us to proceed more quickly, as I will not have to make extensive notes.

Context

I've designed, in consultation with the original authors, a graphical summary of the systematic review of clinical effectiveness for cochlear implants, produced by PenTAG in 2007. This was a somewhat problematic review to interpret, as a very large number of outcome measures were used by the included studies.

I'll be giving you, in turn, both the graphical summary and the original report, and asking you to pull certain information out of them for me. The idea of the test is a "speak aloud protocol", where you tell me about your thought process as you look for the information I ask for. Have you done one of these tests before?

It can be harder than it sounds. As a practice exercise, could I ask you to tell me how many windows are at the front of the place where you live? Please tell me

about how you are thinking about your answer.

When you're answering my questions about this review, please don't worry if you have difficulty finding any of the information. The presentation method is what is being tested here, not you. You are also welcome to ask questions as we go through. I've provided some basic stationery that might be useful. (paper, pen, pencil, rubber, coloured sticky tabs, post-it notes)

Personal information collected

Before we begin, may I collect some personal information about you? You are welcome to answer only those questions that you feel comfortable with.

- What is your past experience with systematic reviewing?
- What is your past experience with HTA?
- How familiar are you with the cochlear implants report?
- How long did you spend on reading the executive summary, roughly? (if they didn't, give them 5 mins with it)
- Did you complete the learning style preference questionnaire for me? (if not, ask them to complete it now)

--**First presentation given**-- (graphic or report, randomly selected)

Okay, we'll begin the test. Here is the first information presentation. [This is just the section of the report that details the] / [this graphic presents only the data on] clinical effectiveness of cochlear implants in children. Please take a few minutes to familiarise yourself with the information that it contains.

--5 mins to look through--

Do you have any initial impressions of the data here?

As I mentioned, I'll be asking you to pull certain information out of the [report] / [summary]. Remember that I'm testing the information presentation, not you. It would be very helpful if you could speak your thought processes aloud while you are performing the tasks as much as possible.

Task 1

Which trial has the largest N?

Task 2

How many of the trials were conducted in the UK?

Task 3

Which trials used the Lexical Neighbourhood Test?

Task 4

Can you tell me about selection bias in the Peters et al. (2007) trial please?

General questions

How do you feel now about the quality of the evidence overall?

--Other presentation given-- (graphic or report, whichever not given before)

Again, please take a few minutes to familiarise yourself with this presentation method. It contains mostly the same data, but with a few [omissions] / [additions].

--5 mins to look through--

Okay, I'll ask you to perform some more tasks. As before, if you could speak your thought processes aloud, I'd be very grateful.

Task 5

Which trial had the longest follow-up, and how long was this?

Task 6

How many of the trial reports were published in 2005 or later?

Task 7

In which trials were all participants accounted for?

Task 8

Of the unilateral cochlear implants vs non-technological support trials, which reported at least one significant outcome measure, and which measures were these?

General questions

Has your opinion of the quality of the evidence changed at all?

--Combination presentation given--

This is a combination of the two presentations, which is how the summary might be used in a TAR report. I'm going to ask you to do a few more tasks. Please use whichever method you find easiest to find information I ask you for.

Task 9

How many trials used a cross-sectional study design?
--probe if needed: why did you choose the [graphical summary] / [report]--

Task 10

Which outcome measures were used by Nikolopoulos et al. in their 1999 trial?
--probe if needed: why did you choose the [graphical summary] / [report]--

Task 11

Which trial (or trials) have the lowest mean age? (of those that report this). How old is this?
--probe if needed: why did you choose the [graphical summary] / [report]--

Task 12

Which trial seems to have the highest quality, according to the checklist used in the report?
--probe if needed: why did you choose the [graphical summary] / [report]--

General comparative questions

- How do you feel about the use of a graphical summary for this systematic review of clinical effectiveness?

- If you think it is useful in this case, do you think that graphical summaries like this would be useful in other clinical effectiveness reviews?
- If it was used for a different reviews, do you think it would need different information presented?
- How much do you think this would vary from review to review?
- Here is a list of the 12 tasks you've performed. Do you think they are representative of things you would do to understand a systematic review of clinical effectiveness?
- Is any information missing from the graphical summary, that would be useful to you?

  -- suggestions if needed:

  --would knowing about which brands of implants and hearing aids had been tested be useful?

  --would knowing the degree of deafness of the participants have been useful?

- Is there any information in the graphical summary that you don't feel is needed there?
- Would you find it useful to have an interactive version of the graphical summary, which could be sorted by study size, design, outcome measure, etc.?
- Would you find this more or less useful than a spreadsheet containing the same information?

Outro

It only remains for me to thank you for your help. Tests like this are invaluable for showing us how experts in the field like you use reports, and how they can be improved.

Would you like me to contact you when these results are published?

Do you have any questions at all that I can answer?

Thanks again.

Appendix E

Sample interview transcript from GOfER tests

interview transcript 7

Interview date: 26th July 2010

Interviewer

Participant

Timings

Questions/tasks

Participant's actions

<6s> - shows where participant is quietly reading or gathering information. In this case, for 6 seconds.

p32 / p vi is used as shorthand for the participant turning to a particular page.

question is used for shorthand to indicate that the participant is starting to look at the currently visible question card.

three dots (...) indicate an unintelligible syllable.

a hyphen at the end of a word (word-) indicates an unfinished sentence.

3:00

Okay, could I start by collecting a bit of personal information?

Yep

Just answer those questions that you're comfortable with. First of all, what's your past experience with systematic reviewing?

I've been doing systematic reviews here at ScHARR for probably

about 10 years, actually. Something like that.

So, exetensive.

well, I suppose so, yes. I always feel as though I'm just beginning, but yes.

Perhaps we all are, in one way or another.

Yes.

And that's all been work for HTA?

It's mostly been. There's one that we were doing last year that was. Oh, no, that was HTA, non-NICE. There was one we're working on at the moment that's MRC I think, but yes

But pretty much NICE HTA?

Yes.

And did you have any experience of the HTA process before starting here?

No.

And you didn't do any other systematic reviewing work?

No.

That's great. And you've already said that you're not familiar with the Cochlear implants report, which is great.

4:00

So, what I'll do is I'll give you the first information presentation. And I'll give you about 5 minutes to look through it and familiarise yourself with where the information is. I know it's not going to be very long, to familiarise yourself with a whole systematic review section of a TAR.

Right. Okay. That probably replicates the committee members' experience.

That's very true

They read it on the train

This is what I'm hoping. This is the-

takes report section from interviewer

The review dealt with both children and adults. This is just the section with cochlear implants in children.

Yep, okay.

Feel free to open that and have a look through.

Okay.

Opens report, p vi

So I'm supposed to be looking at the

p vi

systematic review, and not the modelling bit, presumably.

p37

Yes. That's only the systematic review bit of the report

5:00

Yes

p38

Okay

p39 (summary of implant brands in included reports and on contract to NHS)

And, as I say, I'll be asking you 4 questions. So I'll ask you to pull information out of the report.

Okay

Bits of numerical data, basically, I'll be asking for.

Okay, fine. So you're looking at quite a lot

p40

of different

p39

models of

p40

implant

yes, there were different ones

p41 <6s> p42 <6s> p41 <4s>

okay

p42 <5s> p43 <4s> p44 <2s>, goes to turn page and changes mind <2s> p45

6:00

<28s>

Right. A lot of it's-

We're looking at page 45

45, yes. The way of limiting the number of studies is novel to me.

<5s>

Ah, yes, they took an arbitrary number out. Because there were so many, they took a random sample

Yes.

<3s> p46, p45

7:00

p46 <6s> p47

right

<8s> p48

yes, I'm rather glad I didn't have to do this one.

Yes, I think-

It probably wasn't their happiest hour, was it?

No.

Interesting, but messy. Mmm.

p49 (speech perception measures table) <7s>

Oh, gosh. Oh my goodness, yes, right, okay.

p50

Are you looking at the-

The number of outcome measures, yes

p49, points to table

And just for one aspect

p50

of it, yes. Yes, that's fairly-

p51 (quality of life measures table)

oh, god.

<4s> p52

yes. Tricky one, isn't it?

p53, turns head, then report, to view landscape

8:00

<2s> p54, p55 <4s>

okay. They reported a sort of- I suppose, with this kind of thing,

p56

the study designs are so various, it's perhaps not so obvious,

p57, turns report to view portrait

but with straight forward study designs, you do wonder why people haven't cottoned on to the benefits-

p58

the merits of reporting things in the way that systematic reviewers will score them highly for. It seems a bit dim, really.

<4s> p59 <10s> p60

right

<5s> p61, turns report to view landscape <2s> p62

9:00

<3s> p63 <3s> p64 (first "visual summary of outcomes") <6s> p65

ah, yes, right.

<3s>

yes

<4s> p64 <3s> p65

so- okay, so positive significant outcomes

p66

are a bit thin on the ground

yes

p67, turns to look portrait.

right. okay.

okay?

right, yes

just coming up to about 5 minutes now,

oh, okay.

so, if you've got a general view of how it's layed out

p68

I think I do, yes

p69

yes

which bits are where

okay.

can we move onto doing some tests?

yes.

we'll have a chance for some more general discussion at the end.

yep, that's fine

closes report

10:00

that's question number 1

interviewer gives question 1 [10:02]

question

Which trial- No, I didn't

looks at interviewer

register that at all.

Oh, you can use the report to-

Oh, sorry.

opens report, flicks through pages near start.

It's not a memory test.

No, no, no, no

Oh, okay. So-

finds landscape page (p53?) <2s> keeps hand in page, while flicking back through earlier pages. Opens at landscape page again. <4s> Turns page <2s> half-turns 6 pages, then lets them fall back.

Okay, so it's the Manrique et al, 182.

Excellent.

Or is there another table?

Have a look. There may be another- because it was split into 4-

I may have come past them

p52, tuns report portrait, p52, p?? (a few later on) turns head to view landscape,

There's 4 different questions really.

Yes

turns one page back,

being asked

so that's the-

p52

4 different comparisons

11:00

p??, back one page, on a few pages to p63?, p65, p64, p52

that's still non-technological support. okay, so.

p68? portrait

Acoustic hearing aids

p69?, views landscape,

okay

p70?

they're smaller

p72? views portrait

please do feel free to fold over the corners or something.

p74? views landscape

I should have provided sticky tabs.

It's okay.

p75?

... ...

p76 p78 <2s> p77 falls down while participant turns to p79, views portrait, p80, p81

I'm not sure this is ... ... smaller again.

p82

okay

p85, p??

and-

<2s> turns back 1 page <2s> turns page <6s> turns page

so It's

turns back to p53?

still

still the same one?

still the same one, yes.

participant completes response [11:59]

12:00

I wanted you to check the others as well. Great, thank you very much. And your process for finding that was, you noticed the information was in the data tables

Yes. Yes, I looked down them, and found the one that seemed to jump out, and checked that there was nothing further down that one

so you had that number in your mind, while you were-

opens report, turns to landscape page (p53?)

well, I looked down this one, hit on that

points to just left of centre of page

and then thought: is there anything bigger than that, further

down-

So you were holding that number, 182, in your mind, and comparing it as you went through.

Yep.

Excellent, that's great. We'll move onto another one, if that's okay?

lets p53 fall. Now on p52?

Interviewer gives question 2 [12:36]

question <1s>

oh, gosh.

p53

Again, it's something I didn't notice first off.

<2s>

1

p54 <2s>

2

p60?

I should have done the pages after all.

p61?, p63?, p64, p65, flicks through some more pages, stops on a landscape page <3s>

13:00

flicks forward through more pages, going backward at one point. Stops on another landscape page <2s> flicks forward again, stops on a portrait page.

So only two, in only one of the four interventions, then.

Completes response [13:22]

[transcriber's note: after finishing this question, participant appears to have fingers of left hand marking 4 places in the report, perhaps ready for the next question?]

excellent. Very swift, actually, as well.

Oh, right.

It's almost as if you've dealt with these reports before.

I suppose I've written them, yes.

And that was a similar process, to the last one, just looking through the data.

Yes.

Okay, we'll go on to number 3.

Interviewer gives question 3 [13:37]

question

right. ooh.

transfers held pages to right hand, so that p53 is visible <5s> question, p53

one

flicks the edges of the pages between p53 and the first finger held inside the report, as if to judge their thickness, or perhaps planning to turn page and changing their mind. <2s> question <2s> p54

two

14:00

<4s> p55

Sorry, it was which trials

p54

Which trials, which trials,

p53

all right. Not which number. So, Staller

p54

and MED-EL

p55

and-

<4s> turns to next place held by a finger, probably p69 <4s> p70, next finger, probably p81 <2s> last finger, probably p88

I think just those two

<4s> p89

just those two then, yes.

completes response [14:38]

and, again, it's a process of looking at those first data tables

yes, running down the list, yes.

great. Okay, we'll move onto the last one with this report section, I'm sure you'll be pleased to hear.

Yep, fine.

Interviewer gives question 4 [14:56]

question

selection bias in Peters

p53

right

15:00

p54, turns to place marked by finger (p69?) <2s> p70, next finger (p81) lifts report with both hands

okay.

<8s> looks puzzled, question, p81

selection bias-

<13s> draws in breath, hesitates.

I don't know. The fact that they're not telling you how many are pre-ligually deaf, is an issue. Or the-

<10s>

The information may be somewhere else in the report.

Oh, right. Okay, yes, fine.

16:00

So in the text somewhere.

p82

Discussion of-

p81, p80 <2s>

okay

<18s>

You're looking at page 80 there, in the text

I am, yes.

Just doing that for the tape, don't worry.

Sorry, no, yes, I realise that. I was just thinking, I haven't a clue what the implications of this are for selection bias. ... ... yes, no. Pass.

looks at interviewer.

participant gives up [16:50]

Okay, that's fine. No problem at all. But that's the last one with that, so you're finished with that. What do you think of the quality of the evidence overall

17:00

in the report, from what you've seen there? You said it was quite- there were lots of outcome measures.

Yes. It's confusing. There are a lot of outcome measures, there are a lot of small studies, that don't reach statistical significance, so it doesn't seem like great evidence. I don't know how easy it is to research this, in that I don't know how unusual this would be to how difficult recruiting would be.

Of course, yes.

But nonetheless, something where the number is 30, and you think.

I guess, as a surgical intervention, you'd have to- people might be worried about going in for it, or something like that.

Well I think there's also issues in the deaf community, aren't there, about cochlear implants? And whether

18:00

it's appropriate to say deafness is a problem, we've got a physical fix.

That's very true. They're part of the deaf community, and have been taught that deafness is not a disability, but a difference. That's a very good point. Excellent. Well, thank you for going through that. Could I take that back from you, and I'll give you the other presentation that I'm comparing it against, which is this one here

gives report back, takes graphic from interviewer.

So, again, I'll give you about 5 minutes to have a look through that,

p vi

and find out where the information is.

p32 <13s>

I always have to work really hard with things like this, because I think in words, not diagrams.

Ah, that's interesting. The learning styles will tell me a bit more about that.

19:00

But it's interesting that you know that.

Well. I don't know what that

points off screen

learning styles thing will tell you, but about 15, 16 years ago, the only other time I've done a learning syles thing, it came out very very skewed. It was ever so, ever so over towards learning by reading.

Probably an excellent strength for a systematic reviewer.

Both laugh

Possibly, yes. It certainly wasn't just slightly off-centre. It was way over. So, whenever it comes to putting diagrams into reports, I always have to make myself. Or put some stuff in tables. I only do it when I realise I have far too much clogging up the text.

p33

So-

p32 <7s>

This, I guess, once I'm used to it

p33, p32, p33 <2s>

you know, is fine. But it will take me a bit of time to get used to it. It's certainly good and clear on the size,

20:00

and location, that's neat. And very easy to- Yes, and the age. It's the quality bit that takes a bit of thinking about.

Yes, it's difficult to fit everything on the-

Yes, and it's something I might be distinctly slower about

p33

than some of the others. I bet the modellers would find something like that a lot easier to (go?) with.

p32 <3s> p33 <3s>

This is nice

places hand over outcome measures grid

This is easier to see.

moves head back and forth between the outcomes and the rest of the graphic

Oh, that's interesting as well, because you've got two of the bigger studies coming up with significant results, but not the other one.

But not the last one-

Not that one. Which is between those two in size.

p34

21:00

Oh, that's a small one ... ...

Yes. I guess that just reflects- they might have had a larger weight of- a very obvious difference.

Yes, absolutely, yes. Well, it could be

p33

that it's something that the tool they were using.

p34

They might lend themselves more to- yes.

<3s>

Yes, that's an interesting way of presenting it, actually

p35

Is it similar to anything else you've seen before?

No, not at all, completely different.

Okay.

Which presumably was the idea, was it?

Looks at interviewer

Yes, it works well for my PhD in that way.

No, I've never seen anything like this.

p36

But I shouldn't bias you. Do be as candid as you possibly can. We've looking for critical feedback. There are certain flaws in it, which I'm sure you'll find as you go through.

22:00

Well, essentially,

p32

I think it's pretty neat, as I say.

p33

The-

<2s> p32

And, although it takes a little while to get used to that

points to quality grid

in terms of the detail, the general

p33

picture of the blacker it is, the better it is

p34

is straightforward, isn't it?

but I guess the disadvantage of that is, if there's one that's particularly important, that might not stand out more than the others.

That's true, yes, you can't weight it towards-

p32

the fatal flaw, or whatever

yes, exactly.

p34, p35

I need to find a way of doing that.

Yes. Yes. Oh, that's good.

Okay, so you've taken about

p32

4 minutes to have a look through that. Are you happy to start some more tasks?

Yes, fine.

closes document

I'll give you 4 more tasks with this.

Interviewer gives question 5 [22:56]

question <2s> p32

23:00

oh, right, we're here. Yes, I hadn't really focussed in on

p33

the length of follow-up. Yes, that makes it much easier

p34

to spot

p35

doesn't it? So it's

p36, p33

very obviously Manrique

Completes response [23:14]

Excellent, thank you. And the next one. You got that from looking at the length of line in the follow-up section. Yes, I see. And you scanned down all of the pages. That one obviously stood out fairly clearly again.

Yes, definitely, yes.

Interviewer gives question 6 [23:35]

question <2s> p33 <2s>

Okay, so we're on 1

p34

2

p35

3, 4, 5, 6, 7

p36

8, 9.

question

so 9.

Completes response [23:52]

And yes, that's easy to pick out.

Okay, but you had to pick out the numbers,

p vi

there is no

p32

visualising.

p33

No, but they're very clearly there.

24:00

Okay.

So that's fine.

Okay. Number 7.

Interviewer gives question 7 [24:06]

question <2s> p33 <3s>

This one involves the quality bit, doesn't it?

p32 <16s>

Attrition bias, okay, so it's that one

points to quality grid key.

You're looking at the second one down in the fourth column

Yes.

p33

And-

holds finger on quality grid for top trial

Okay

question

so

p33, runs left index finger quickly down the quality grids, stops near bottom, moves fingers out of the way to see author name

that's MED-EL, Nikolopolous, Illg

p34

Mildner, Tomblin,

25:00

p35

Litovsky. Litovsky twice, (I've already done that?).

p36

Damen, Huber, and Chmiel.

Completes response [25:20]

Excellent.

p33

Right, okay. That's a bit more fiddly to pick out.

It is, yeah. That's definitely one of the weaknesses of this presentation. But you were able to go through and see which (ones were there?)

Yes. Oh, yes, definitely. But I mean, I'm able to do it, but it just takes a bit more time and thought.

Yes.

Interviewer gives question 8 [25:40]

question <2s> p33, question <2s> p33 question <2s> p33

okay

places RH index finger on black dot in Manrique trial, question,

non-technological support

p34, p33, replaces finger on Manriqe's dot

okay, so look, we've got two trials with

question

significant

p33

measures. So-

traces to left with eyes

The abbreviation is fine

26:00

Manrique on PTA,

moves RH index finger to LH index finger, already on left-most black dot for Nikolopolous

and Nikolopolous on CDT,

moves RH index finger to next dot

IMST,

moves RH index finger to next dot

CAP and

moves RH index finger to next dot, pauses, turns page to view portrait, with both index fingers still in place on first and last black dots

SIR.

completes response [26:14]

And that was, again, a process of looking for the black dots-

and then looking back up to see which column.

You were okay with the fact that there was a difference between those in the gaps and those on the lines?

Yes, that was okay, yes. What are the significance of the darker lines? Or is that just to break it up.

Just to break it up, yes. Nothing particularly different, it's just to make it easier for the eye to follow them. Great, well that's the end of the questions in that part. So, has your opinion of the quality of the evidence changed at all since you've been doing that?

<2s> p34

27:00

<2s> p35

I don't think so, but I do think it's easier to see with this, actually. It's a good way of presenting it.

Good. Okay then, well, I'll just move onto the last bit of the test, then, which is to give you access to both of those in one document, so you can now choose which one to use, so-

takes combination document from interviewer.

There's that one.

opens report near end, flicks backwards through

It's just got the graphical summary at the front

flicks to the front

and the rest of the report behind

oh, I see, yes. Okay.

So it's before that.

fine

Okay. Okay to give you another 4 tasks with that?

Mhm, fine

Interviewer gives question 9 [27:43]

question, p33, p32 <8s> p33

28:00

<3s> p32 <5s> p33=

Right, I see, got it. Okay.

p34

so that's 1

p35 <2s> p34, p35 <3s>

... ...

p32 <2s> p34

okay, so they're still cross-sectional, they're just very weenie are they?

Yes.

That's 1, 2

p35

3, 4

p36

okay, 4.

completes response [28:36]

p35

What were you looking for then, as a cross sectional one?

Well, I was looking at

p32

this

points to cross-sectional design on key

and looking at this and looking for something that looked

p34

like that. I was slightly fazed by the much thinner ones, but then I realised that it would be, because of this weight of size.

Yes, that's quite a common thing. People seem to be-

29:00

everyone seems to get that one, because it's exactly the same as the key, but because it's such an unusual skewing and shape, a lot of the other ones are more difficult to see.

But it's easier to pick them out from here than flick from table to table in the-

I guess in a sense, it is a slightly unfair test in a way, because in the report they were broken down into sections, and it might be that the graphic would get broken into sections as well, if it were presented in the report. So it might have been more fair to ask questions about one single comparison. But I had to do the same for everyone, unfortunately.

Yes.

Which is the way it goes. Okay I'll move onto another task.

interviewer gives question 10 [29:49]

Question 10.

<2s> p32 <2s> question, p33 <4s>

okay, Nikolopolous. So they used

30:00

TROG.

question

Oh, no, sorry,

p33

wrong trial.

<3s>

Right, so CDT,

turns report to view portrait, points to dots with LH index finger, and traces line up to titles with RH index finger

IMST, CAP, and SIR

completes response [30:16]

Yes, great. And, again, that was a process of looking at dots, and following lines.

Yes.

Interviewer gives question 11 [30:24]

question <2s> p33 <2s> question <2s> p32 <3s> p33 <5s> p34 <3s> p35 <3s> p36, question, p33 <3s>

okay, so. Lowest mean age. Nikolopolous seems slightly below the 5 year

p34

31:00

band, and so does Svirsky.

completes response [31:05]

p35, p36

And I guess if I wanted to know for definite

p33

which one of those two, then I'd have to look at the other tables to get the precise decimal point=

yes, this is one thing about graphical presentation methods.

Sometimes, if it is a very close value, it does get difficult to see which one's which. It's not too bad here, because they're quite close to the line, but there are about 3 of them. And if they were in the middle of nowhere, it would be quite hard to see which ones they were.

Yes.

I suppose you could argue that that actually is quite interesting. Because it might look like a very large difference in the number, or you could quite clearly see, whereas in fact not knowing is quite right, because it is very close.

Yes. It's interesting. It brings home how many don't give you the mean.

Oh, that's a good point, yes. No-one else has mentioned that.

A fair few of them give you the range, but no-

but no mean.

no mean, yes.

And of course the mean could be anywhere within that pretty much.

32:00

Yes, oh yes, there's no natural-

moves hand quickly from left to right, left, right

So it might indicate something different to what is actually the case. Very good point.

Yes.

Excellent. Well, I shall move onto the very last question.

Okay.

Interviewer gives question 12 [32:13]

question <2s> p33 <6s> p34 <8s> p35, p34 <3s> p35 <10s> p34 <4s>

... ...

p35 <2s> p34

where there's a sort of

p35, points to page

blank there, as opposed to a- is that a non-applicable-

absolutely right, yes.

Right.

33:00

Okay, so

p34

... ...

p35

I suppose the- that Litovsky

points to quality grid near bottom of page

is presumably the-

p36

best quality, is it?

Excellent, is that the Litovsky A or B? A?

Yes.

Completes response [33:21]

Excellent. Well, that's the end of the tasks, I'm sure you'll be very pleased to hear.

Okay, no problem.

I noticed that you entirely chose to use the graphical summary for those tasks.

Yes.

Why was that?

I guess it seemed easier than-

opens report near the middle.

Yes, it's quicker to get the information out of than that, actually, isn't it?

Mmm. Do you have any idea why that might be?

p35 <2s>

Just because it's all in one place?

It's more summarised. It's all in one place, and it's more condensed, isn't it?

Okay, I guess you can see more on a single page, perhaps.

34:00

I think so, yes. Which is interesting, because, I wouldn't have predicted I'd have thought that, not thinking I was very visual about how I handle data. But yes. It works, doesn't it?

Good. Sorry, how are we doing for time?

I'm fine for time. Have you got a train to catch?

No, no, I'm staying over. I'm going to do some more tests tomorrow. Could I ask you a couple of more general questions now?

Mmm.

About how you see these possibly being used?

p33?

You seem to think it was useful for this report here. How applicable do you think it would be for other systematic reviews, in other reports?

Yes, I don't see why it shouldn't be.

35:00

The one thing you're not getting is the effect size.

Yes, of course.

Other than that, it pretty much tells you everything you need to know, doesn't it?

There are a few bits of information missing, actually. Such as the brand of the implant used

Oh, of course. Yes, sorry. Because I'm not totally into this area

I understand. And also degree of deafness of the participants isn't reported there. So there's a few things that it doesn't show you.

Yes, okay

Which it might. And I suppose a different intervention might have different information that was needed to display. But you wouldn't have that many outcome measures, I would imagine.

No, no, quite. No, which would give you a bit more space to play with for something else, wouldn't it? But it might be that on another- or with other things there might not be so much variety around these

points to middle of page. Possibly indicating the left edge of the outcome measures.

Yes. mhm.

But you think it might be useful for another review.

Yes.

36:00

I think it would be well worth trying on another review, yes. Definitely.

Okay, but it might have some different information presented. And how much do you think it would vary from review to review? I guess there are some things that you'd want in all reviews, like the design and size, follow-up, and things like that would be.

Yes, you'd want the authors, the date, the location is always useful to know.

Particularly because, in NICE we're looking at the UK situation more than other places.

Yes, quite. So, depending on what the intervention was, exactly, you might feel that it was more or less relevant. I've done quite a bit with osteoporosis. Quite a lot of the work is done in Japan, but they use smaller doses, because they're smaller people, and things like that.

Different populations-

Different populations.

37:00

They're genetically very different. So you're not sure if things will work in a different kind of way.

Interesting.

So you'd definitely. Author, date, study design and size. The ages I suppose isn't always relevant.

Yes. This is particularly because it's in children.

I think the follow-up is always worth knowing about, and quality. And then the summary of outcomes.

But yes, some way of displaying the weighting would be really really good. I have to think about that in future versions.

Yes. It might be quite difficult getting it all in one-

You might be able to have a sort of a colour thing. So it would

be darker if it had a higher weighting, and lighter ones have less weighting, or something like that, which might mean that you can see- why haven't I thought of that before? That could possibly work. But then you have to work out how certain you are about it, which is another dimension-

It's tricky, isn't it? yes.

38:00

Well, I'll continue working on it, hopefully, as long as they'll let me. Oh, and the 12 tasks that you performed. Were those representative of things that you'd need to do, to understand a systematic review of clinical effectiveness?

Yes, yes, I think so, yes.

The sort of thing that you'd approach, you'd need to know that kind of information. Okay. Oh, yes, and the last ones. Would you find it useful to have an interactive version of this graphical summary, presented on screen, which could be sorted by, for instance, you could click on the top and it would sort it by the size of the trials, or which outcome measures were used, or the ages, or something like that. So you could have a list from lowest age to highest age, or a list from smallest trial to largest trial, or-

Well, it would be fun, I suppose. Yes. It would be useful

39:00

if it wasn't too hard a thing to do I guess. I'm not sure quite how much effort it would merit putting into it.

That's interesting

But it has potential.

Do you think that would be more or less useful than having just a spreadsheet with the numbers in it? Which you would sort.

I think it would be more useful. I'm not very into spreadsheets with numbers. I would find it more intuitive to do this

points at graphic

You'd rather get the information out of it like that. okay. Interesting. Well, that's great. That's the end of the test. It only remains for me to thank you very much for your help.

Okay, fine. A pleasure, no problem.

Interview ends [39:48]

Appendix F

GOfER test data

This section presents detailed results of each task in the GOfER test. Times are given as a proportion of overall task time, to account for more or less talkative participants (see Chapter 5.2.3). Figure F – 1 shows a key, to be used for interpreting these performance summary figures. The pink tab to the right of the page can be used to quickly find this key if necessary.

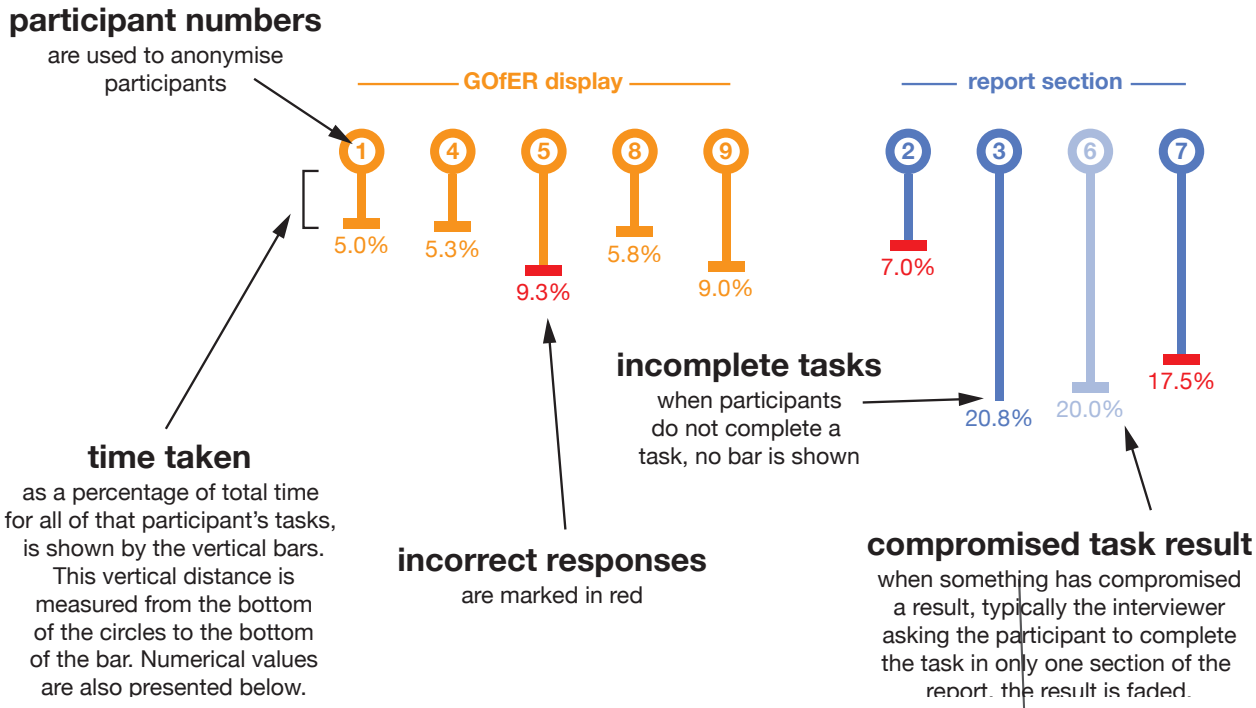


Figure F – 1
Key for interpreting performance summary figures

F – 2 Presentation 1

After initial questions were complete, the participants were given either the GOfER display or the report section.

F – 2.1 Presentation 1 familiarisation

Each person was initially given up to five minutes to look through their assigned first presentation.

The five people given the GOfER display first tended to spend some time looking at the key, then move on to looking over the data pages, often returning to the key to check details. Four of them used the full five minutes to look through, but participant 4 only spent about a minute and a half, looking at the key and each data page once only.

The four people given the report section first had more varied familiarisation strategies. They mostly started with the contents page, and then either read the early pages, or flicked through other parts of the report looking at data tables. Participant 3 used their five minutes to find and mark each comparison section with the coloured sticky tabs that were provided. Three of the participants used the full five minutes, with participant 2 spending two minutes reading the contents and flicking quickly through the report.

There was generally very little discussion during this time, as the participants were not specifically asked to 'think aloud'.

F – 2.2 Task 1

Q: Which trial has the largest N?
A: Svirsky, 1999 (N=297)

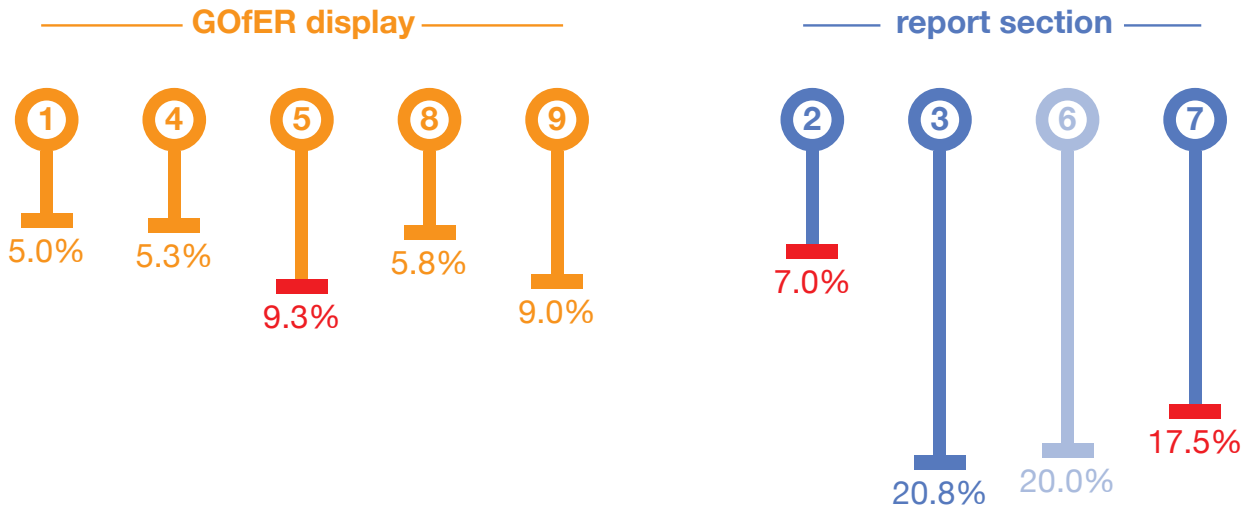


Figure F – 4
Task 1 results

Figure F – 4 shows a summary of the time to decision and accuracy of each of the participants for this task.

Considering all nine participants' decision times, the GOfER display does seem to reduce the time taken to answer for this task (report mean = 16.3% of total

task time, GOfER mean = 6.9% of total task time, two-sample t(7) = 3.1, p = 0.016) However, there were several incorrect responses.

Participant 5 (using GOfER) and Participant 7 (using the report) seem to have simply missed the largest trial. Participants 2 and 6 (both using the report) used a single comparison section rather than all five. Participant 2 didn't realise that there were multiple sections, so this is marked as an error. Participant 6 claimed that they couldn't answer, as they couldn't find an overall summary of all comparisons, not wanting to look through each in turn. After 4 minutes, the interviewer asked them to use a single comparison section, hence the result is marked as compromised in the diagram.

Removing the two people from the analysis that used the report to look at only one comparison section, the test is still significant (report mean = 19.1% of total task time, GOfER mean time 6.9% of total task time, two-sample t(5) = 6.8, p = 0.001).

The number of errors is then similar in both the GOfER and report arms, with one participant in each missing the largest trial, possibly due to it being reported in both displays in two separate arms.

F – 2.3 Task 2

Q: How many of the trials were conducted in the UK?
A: Two

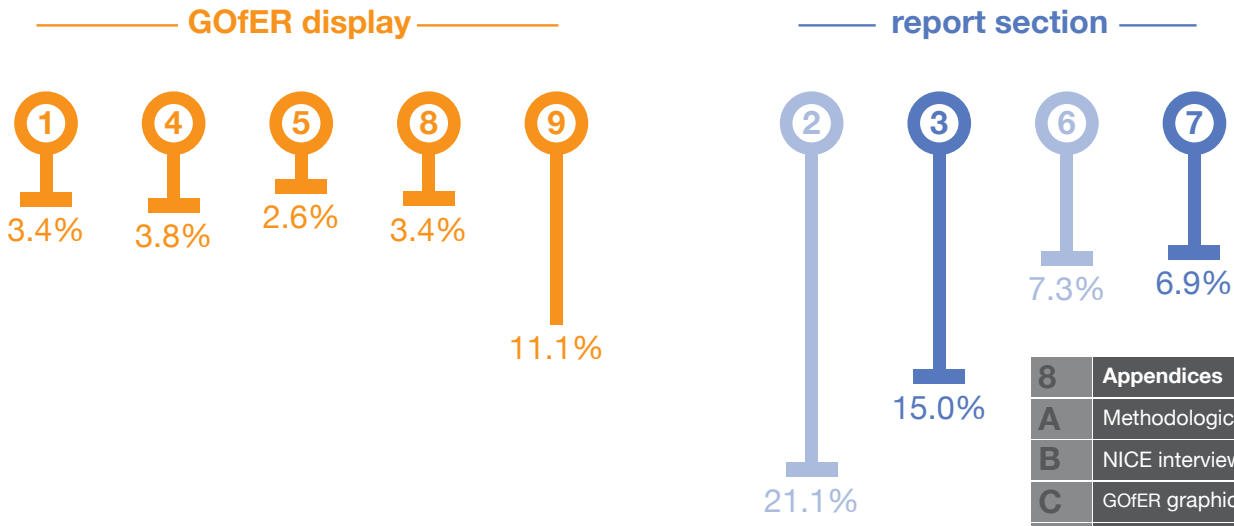


Figure F – 5
Task 2 results

Figure F – 5 shows a summary of the time to answer and accuracy of each of the participants for this task.

Considering all participants' task completion time, the GOfER display seems to reduce task time, but does not reach statistical significance (report mean task time = 12.6% of total task time, GOfER mean task time = 4.8% of total task time, two-sample $t(7) = 2.2$, $p = 0.062$)

All participants answered accurately, with the exception of participant 9 (using GOfER), who was not familiar with the shape of the UK. They mentioned that if the unfamiliar outline did represent the UK, then the answer would be two. Two people looked only at the first section in the report (participants 2 & 6), which happened to contain both of the UK trials. If there had been any in other sections, their responses would have been incorrect.

Participant 3 (using the report) may have taken a particularly long time because they initially looked for an overall summary of all the trials. After abandoning this search, they were able to complete the task correctly. Participants 2 and 6 (both using the report) also looked for an overall summary, expecting that one would be present.

F – 2.4 Task 3

Q: Which trials used the Lexical Neighbourhood Test (LNT)?
A: Staller 2002, MED-EL 2001, Peters 2007.

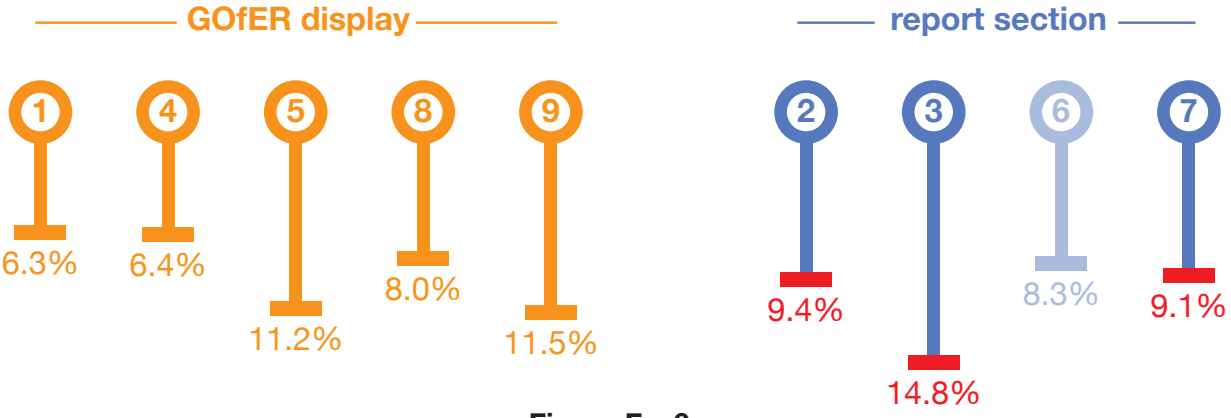


Figure F – 6
Task 3 results

Figure F – 6 shows a summary of the time to decision and accuracy of each of the participants for this task. While the timing for this task with both displays is fairly similar, the difference in accuracy is striking.

Participants 2 and 3 (using the report) were confused by the summary of

outcome measures table on p49. This table gives the name of the developer of the LNT test, and states that only two of the trials in the review used it. One of the three trials required by the answer to this question (Peters 2007) used LNT for two of three age groups in the trial, using a test called MLNT for the youngest age group. The report authors may have categorised this trial as using MLNT rather than LNT, resulting in LNT being marked in this table as used in only two trials, or an incorrect value might have been given in the table.

Participant 6 asked whether they could again just use one of the comparison sections of the report, answering correctly within that.

Participant 7 looked at both p69 and p81, where the information on Peters 2007 is given, but did not report that this trial used the measure.

All five participants that used the GOfER display were able to identify the three trials that used the LNT outcome. They all noticed that the two Peters trials presented in different comparison sections were likely to be the same study. Two of them mentioned that the alphabetical arrangement of outcome measures was helpful. Four of them used fingers, pencils or other aids to help them trace down the lines on the page to the correct outcome measures.

F – 2.5 Task 4

Q: Can you tell me about selection bias in the Peters et al. (2007) trial please?

A: Were eligibility criteria stated? – Yes

Were eligibility criteria appropriate? – Yes

Were the participants representative of the population? – Partially

Were potential confounders reported? – Yes

Were they accounted for in the design and analysis? - No

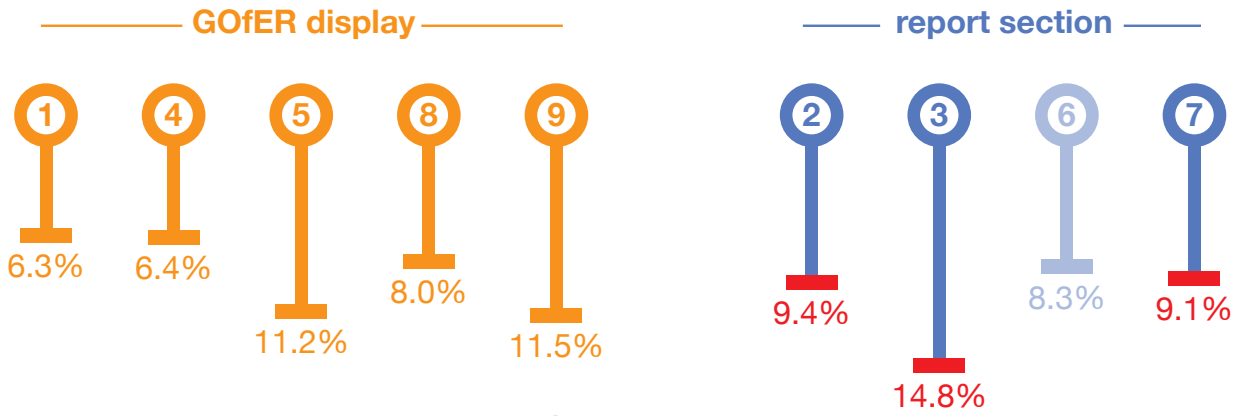


Figure F – 7
Task 4 results

The quality grid in GOfER was designed to give an overview of the quality of the trials, rather than for people to be able to draw specific information out of the grids in this way. To complete this task with GOfER, participants had to flick between the key and the data.

It is perhaps surprising, then, that there was no significant difference between the decision times of the two groups, with the mean task times being virtually identical, at 14.8% of total time for the report, and 14% for GOfER. There was more variability in the GOfER group, however, with a standard deviation of 6.9%, compared to the report group's SD of 2.2%. This may have been caused by participants 1 and 8 (both using GOfER) spending a long time considering the relative importance of the different quality checks in relation to the study design.

People using the report generally spent a long time looking for the Peters trial. Two looked for a list of references, which was not provided. As it happens, this reference list in the published version of the report was not presented in alphabetical order in the version printed in the HTA journal, and had over 250 references, so it is unlikely that this would have helped them to find the trial more quickly.

F – 2.6 General reliability of evidence in presentation 1

At the end of the first section of the interviews, the participants were all asked for their opinion of the overall reliability of the evidence in the review that they'd been working with. There was a marked difference between the responses given by the two groups.

Those that had been using the GOfER display tended to look at a much broader range of information when considering this question. They often paid close attention to the quality grids, but also mentioned the size and design of studies, the outcome measures, and the location of the studies.

Two of those that had been using the report (participants 2 and 6) said that they had no idea at all, and found it hard to get an overall sense of the evidence. Participant 3 turned to each quality table in turn, looking for text relating to each one. By the end of this lengthy process, they knew that the evidence was "moderate or poor" in each section, but few specific reasons for this.

The last person (participant 7) mainly focussed on trial size and the lack of studies reporting statistical significance as their marker of reliability. They had looked at the "visual summary of outcomes" for the first comparison while familiarising themselves with the report section, which showed this information. This may be why they focussed on these two criteria.

F – 3 Presentation 2

The first presentation was then taken away from participants, who were then given whichever display they had not previously received.

F – 3.1 Presentation 2 familiarisation

Again, most people used the full five minutes allowed to look through the second presentation. Two of the participants (4 and 9) that were given the report, having been using the GOfER display for the first four tasks, seemed dismayed at its size and complexity. Three of the participants (2, 6 and 7) that had just been given the GOfER display said how much more interesting they found it. Participant 7 (now using GOfER) immediately saw a pattern that they said they had not noticed with the report, with two of the larger trials on the first data page (marked p33) producing significant outcomes, but not the third.

F – 3.2 Task 5

Q: Which trial had the longest follow-up, and how long was this?

A: Manrique 2004 – 12 years.

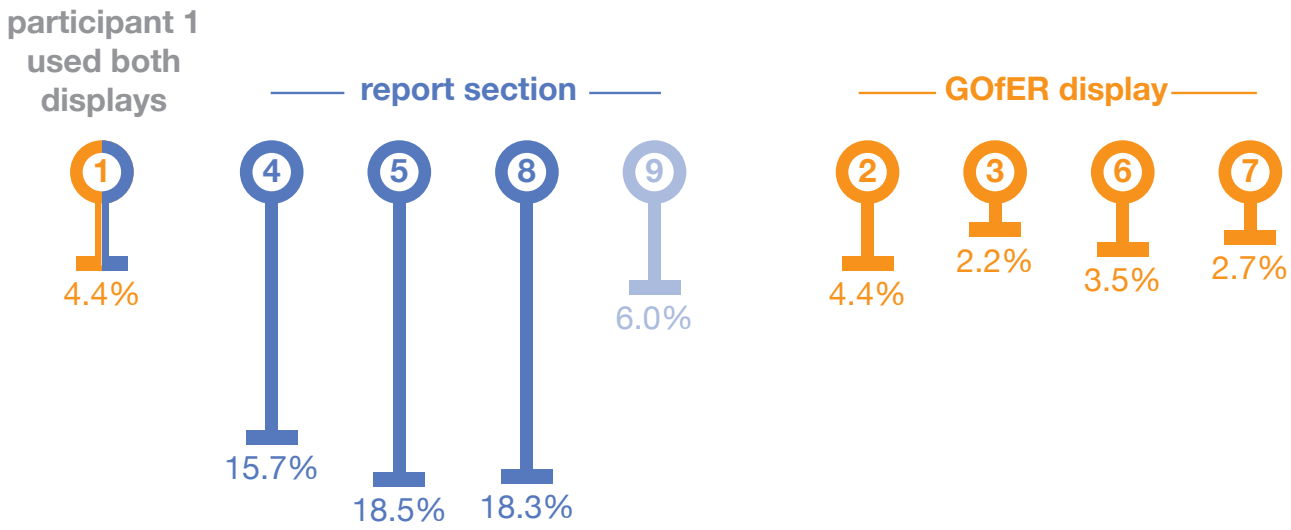


Figure F – 8
Task 5 results

While all participants were able to provide the correct answer, participants 4 and 5 (both using the report) only looked at the first comparison (the section comparing single cochlear implants with non-technological support). As it happened, this section did contain the correct trial. Participant 9 (report) was asked to complete the task just in the second of the five comparison sections. They answered correctly within this section. The people using GOfER had to guess the value from the visual display, which fell between two scale points of 10 and 15 years. All gave the correct figure of 12 years.

The time difference between those with GOfER and those with the report is striking, and statistically significant (report mean task time = 14% of total task time, GOfER mean task time = 3% of total task time, two-sample $t(6) = 3.8$, $p = 0.009$). Participant 1 was excluded from this test, as they used both presentation methods (this task was given at a different point in that interview).

Participant 7 (GOfER) mentioned that they hadn't really noticed length of follow-up in the report at all, and that it was much easier to see this in the GOfER display. Participant 1 (using both presentations) mentioned that, while it was easy to see which was largest in this case in the graphical display, it might be more difficult if there were a few close values.

F – 3.3 Task 6

Q: How many of the trial reports were published in 2005 or later?

A: Seven (without double counting the two studies that appear twice, in different comparisons).

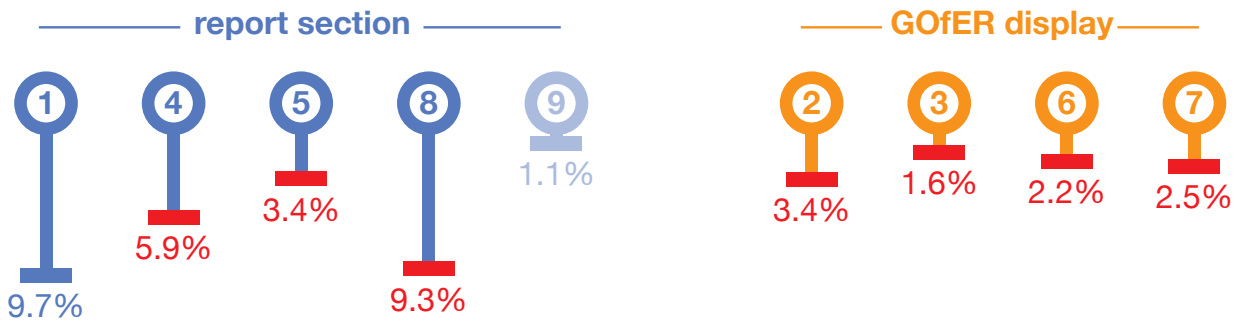


Figure F – 9
Task 6 results

Task accuracy was very low here for both presentations. The most common mistake, made with both GOfER and the report, was to count two of the trials twice each (Peters 2007 and Litovsky (a) 2006). It seems that neither GOfER nor the report very clearly showed where studies were presented multiple times, in different comparison sections.

F – 3.4 Task 7

Q: In which trials were all participants accounted for?

A: MED-EL 2001, Nikolopolous 1999, Illg 1999, Mildner 2006, Tomblin 1999 (partial/unclear), Osberger 1999 (partial/unclear), Osberger 1998, Litovsky A 2006, Damen 2006, Huber 2005, Chmeil 2000.

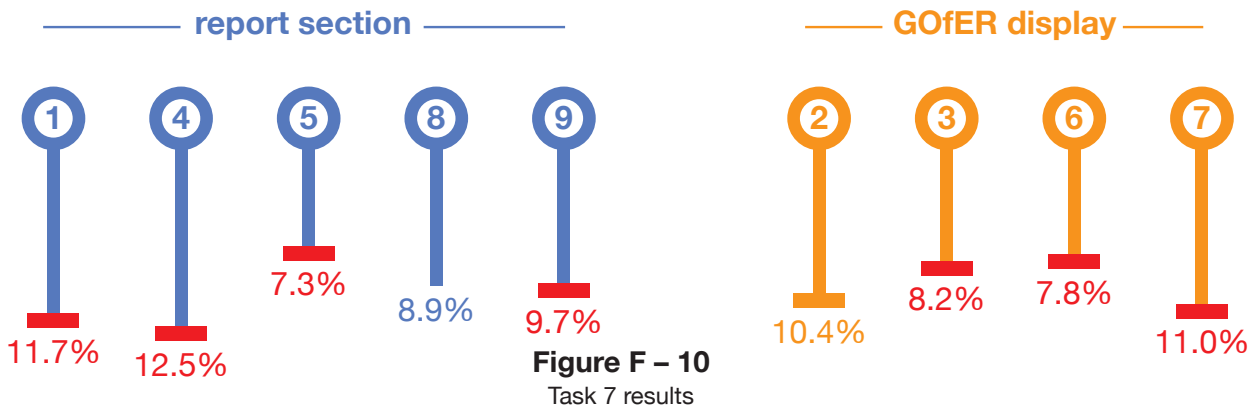


Figure F – 10
Task 7 results

Task accuracy was again very low for this task. A common error, made by two of those using the report and one using GOfER, was to miss the two trials with 'partial' success in accounting for participants. However, this may have been due to their interpretation of what the task was asking for, rather than an inability to see this information from the displays. A range of other errors were made, including missing the fifth comparison section, looking at only one comparison section, and mistaking the labelling of the quality checks.

F – 3.5 Task 8

Q: Of the unilateral cochlear implants vs non-technological support trials, which reported at least one significant outcome measure, and which measures were these?

A: Manrique 2004 – PTA; Nikolopoulos 1999 – CDT, IMST/IOWA, CAP, SIR.

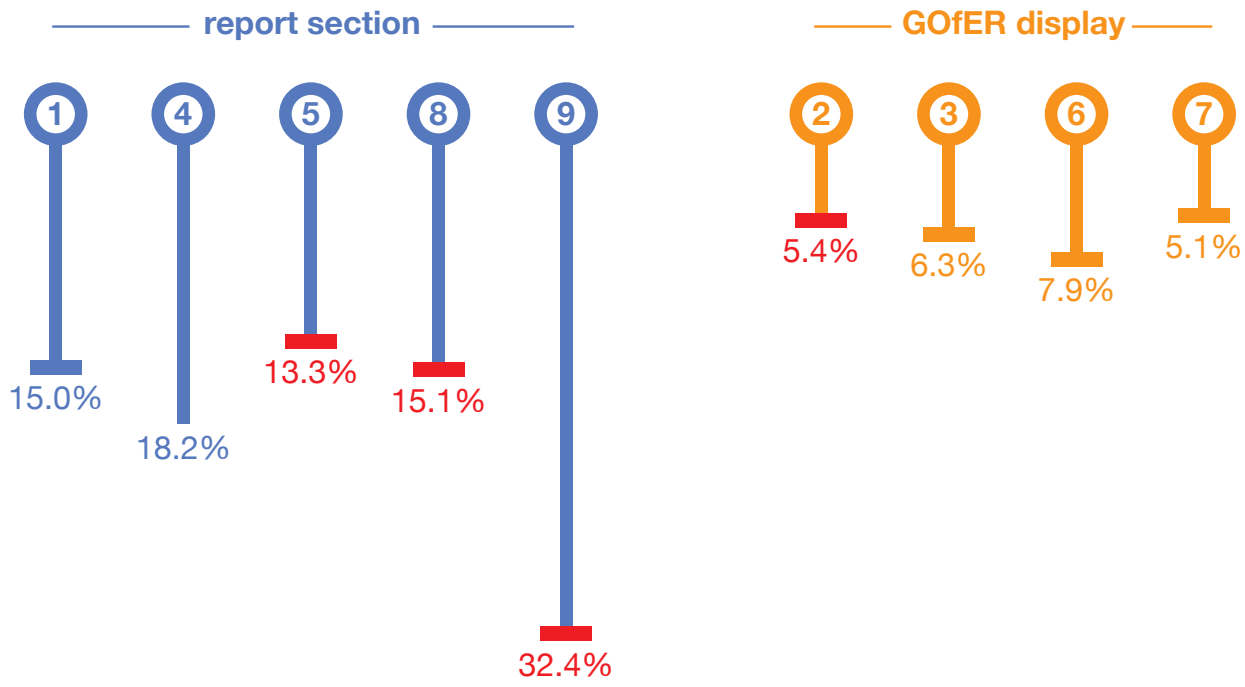


Figure F – 11
Task 8 results

This is a particularly important task for the analysis, as it is the only one that specified a particular treatment comparison. The report was split into sections for each treatment comparison, surrounded by textual content and other information, and the GOfER display was not. In hindsight, basing tasks on single treatment comparison areas, like this one, would have been a fairer comparison of the presentation methods.

This question seems to show an advantage for GOfER in terms of both decision accuracy and task time. While the sample of nine participants does not provide enough power to perform a categorical analysis (ie item response) statistical test on the decision accuracy, one of five participants with the report does not compare favourably with the three of four correct responses using GOfER. The time to complete data shows a statistically significant difference between the two displays (report mean task time = 18.8% of total task time, GOfER mean task

time = 6% of total task time, two-sample $t(7) = 3.2$, $p = 0.016$).

Those using the report almost all tried to use the "summary of results" section on page 61 – 63. They were frequently frustrated by the lack of repeating headers, and changing column widths between the pages, making it hard to see what any of the figures presented past the first page represented.* The only participant that successfully completed this task with the report used the 'visual summary' tables at the end of the section instead of the 'summary of results' table.

F – 3.6 General reliability of evidence in presentation 2

The participants were asked whether the second presentation had affected how reliable they thought the trials in the review were. None of the five people that used the report second felt that their understanding of the overall reliability of the trials had been enhanced by their seeing it. Four of the five said it was very difficult to get any kind of overview using it.

The four that had just used the GOfER display generally thought that they had a better understanding after seeing it. Two people mentioned that they were better able to perform within-study comparisons, but also between-study comparisons with the GOfER display. The one person (participant 7) that didn't think their understanding had changed greatly did say that they thought it was easier to see the overall reliability with the GOfER display.

decision accuracy, one of five participants with the report does not compare favourably with the three of four correct responses using GOfER. The time to complete data shows a statistically significant difference between the two displays (report mean task time = 18.8% of total task time, GOfER mean task time = 6% of total task time, two-sample $t(7) = 3.2$, $p = 0.016$).

F – 5 Combination format

The participants were then given a document which combined both the report section and the GOfER display. GOfER was presented at the front, envisaged as a kind of "visual executive summary".

F – 5.1 Combination format familiarisation

Most participants spent a few seconds flicking through this document, sometimes asking if everything was the same as that which they had already seen.

F – 5.2 Task 9

Q: How many trials used a cross-sectional study design?

A: Five from GOfER / two from report (the graphic represented cross-sectional and non-randomised designs similarly, but with no follow-up for cross-sectional designs).

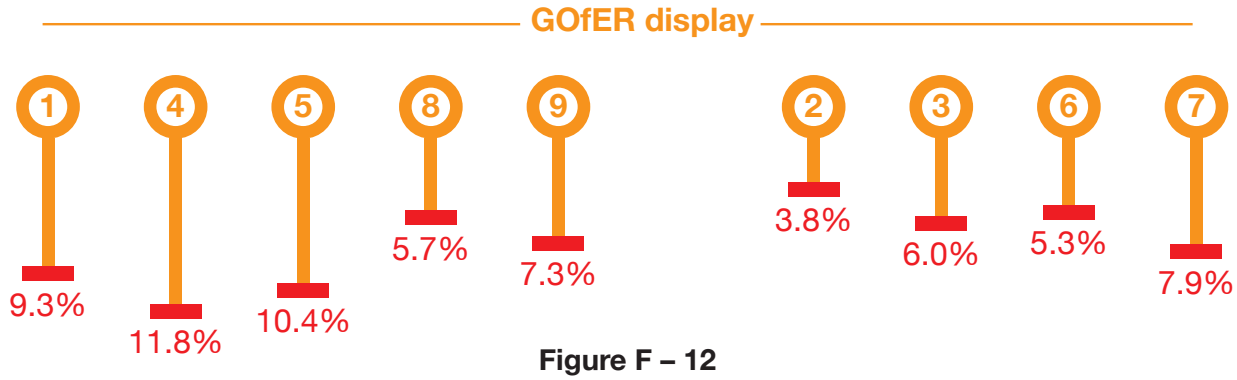


Figure F – 12
Task 9 results

All of the participants chose to use the GOfER display for this task, but none of them gave the correct answer. All of the participants recognised the Svirsky trial (which was the example used on the key), but seven of the participants didn't realise that the four smaller trials were also cross-sectional, perhaps because they looked so different to the larger Svirsky trial. Three people (participants 3, 4 and 6) were not sure how the 'survey' design (single arrow) was different to the 'cross-sectional' design (two separate arrows). Two people (participants 1 and 7) did correctly identify three of the four smaller cross-sectional trials, but both missed Tomblin 1999, the only one that comes before the larger Svirsky trial used in the key.

The accuracy issues with this task can largely be attributed to poor explanation in the key, and possibly insufficient definitions of study designs.

F – 5.3 Task 10

Q: Which outcome measures were used by Nikolopoulos et al. in their 1999 trial?

A: CDT, IMST/IOWA, CAP, SIR.

Figure F – 13
Task 10 results

Again, every person chose to use the GOfER display for this task. They generally answered quickly and accurately. Participant 2 was confused by the labelling of the outcome measure categories, and only reported the speech perception measures used.

Four of the six participants given this task used fingers or an object to point to dots on the grid, and used them to trace to the titles or author names.

F – 5.4 Task 11

Q: Which trial (or trials) have the lowest mean age? (of those that report this). How old is this?

A: Nikolopoulos 1999 & Svirsky 1999 – 4.2 yrs (4-4.5 acceptable).

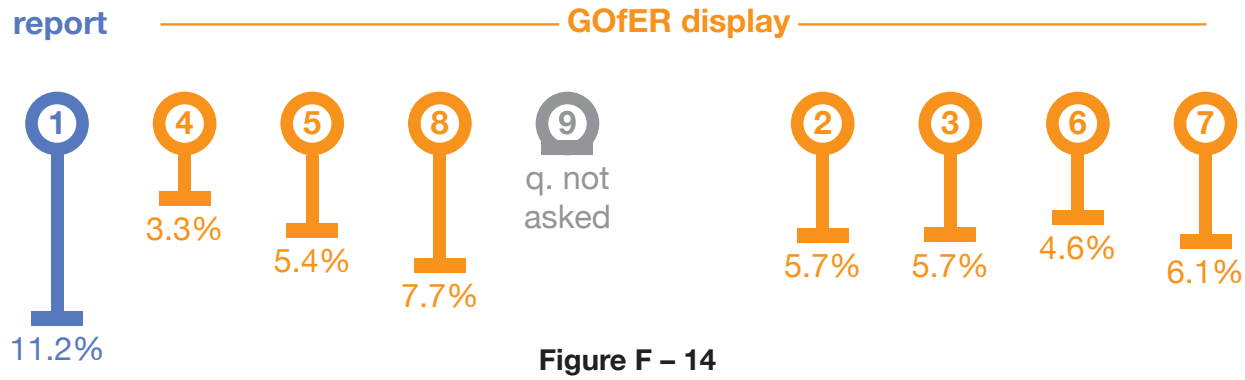


Figure F – 14
Task 11 results

All eight of the people that had a choice selected the GOfER display again for this task. Participant 1, who was given this task in a different part of the interview, only had access to the report. All participants gave responses that lay within the acceptable range shown above.

Five of the participants were also asked how they felt about not knowing the exact numerical value of the lowest age, looking at the GOfER display. Participant 2 thought that it was actually useful to have it emphasised how close some of the values were. Participants 3, 5 and 7 thought that it was probably acceptable not to know, but would be happy to look the value up in the report if they did need to know for any reason. Participant 8 thought that decision makers would probably want to know the exact value, and suggested showing it on the display in brackets.

F – 5.5 Task 12

Q: Which trial seems to have the highest quality, according to the checklist used in the report?

A: Must mention Litovsky (A) 2006, and optionally Illg 1999 (less filled quality grid, much bigger trial) or Tomblin 1999.

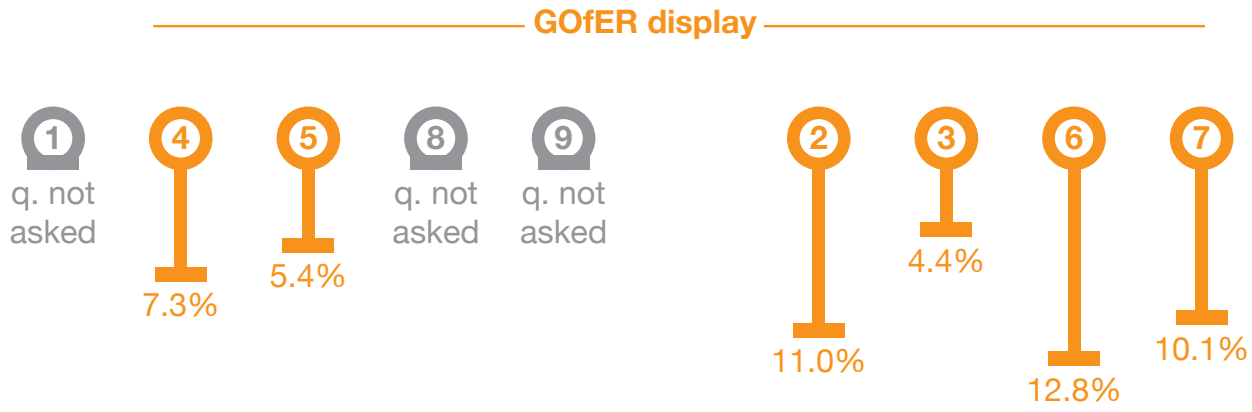


Figure F – 15
Task 12 results

Each person given this task again chose to use the GOfER display rather than the report, and all managed to complete the task successfully.

The strategy used by every person was to count the number of squares that were black, using a 'first pass' to decide on a few 'dark-looking' quality grids, and then counting squares to decide on a final answer.

Three of the six people given this task (participants 2, 3 and 4) mentioned that some questions might be more important than others in considering the quality of the studies, but did not take this into account while giving their responses.

F – 5.6 Reason for selecting chosen presentation in combination format document

At some point after one of the four combination format tasks in the interviews, the participants were asked why they had chosen to use a particular presentation method for their responses.

If anyone had chosen to use the report for one of the tasks during this section of the interviews, the responses could have been compared. As it happens, every person given the opportunity to use the GOfER display did so for every task, so this comparison can not be made from these interviews.

Six of the nine people that were asked this question mentioned that it was easier to find the necessary information to answer overall questions that applied to all studies, irrespective of which treatment comparison the study gave information on, in the GOfER display. It might be seen as an unfair advantage in this experiment that the GOfER display was not separated into sections and surrounded by text, as it could be if used in a report like this one. However, two of these six people mentioned that the GOfER display also had advantages in terms of being able to quickly get a sense of the data visually, such as by looking at how dark a quality grid seemed, or scanning down the length of follow-up bars to quickly identify the largest. Two other people mentioned that the GOfER display gave more information in a more condensed space, allowing them to see more about each individual trial in one place.

The other three people had slightly different reasons. Participant 4 found the report tables inconsistent, both between different pages of the same table and between sections of the report. Participant 8 thought it was easier to retain the information in GOfER, and found that they didn't have to write on a separate sheet of paper to count studies meeting a certain criteria. Participant 9 just mentioned that the GOfER display "was faster", and that they "knew that it [the study design] was a picture".

Appendix G

SOC graphic

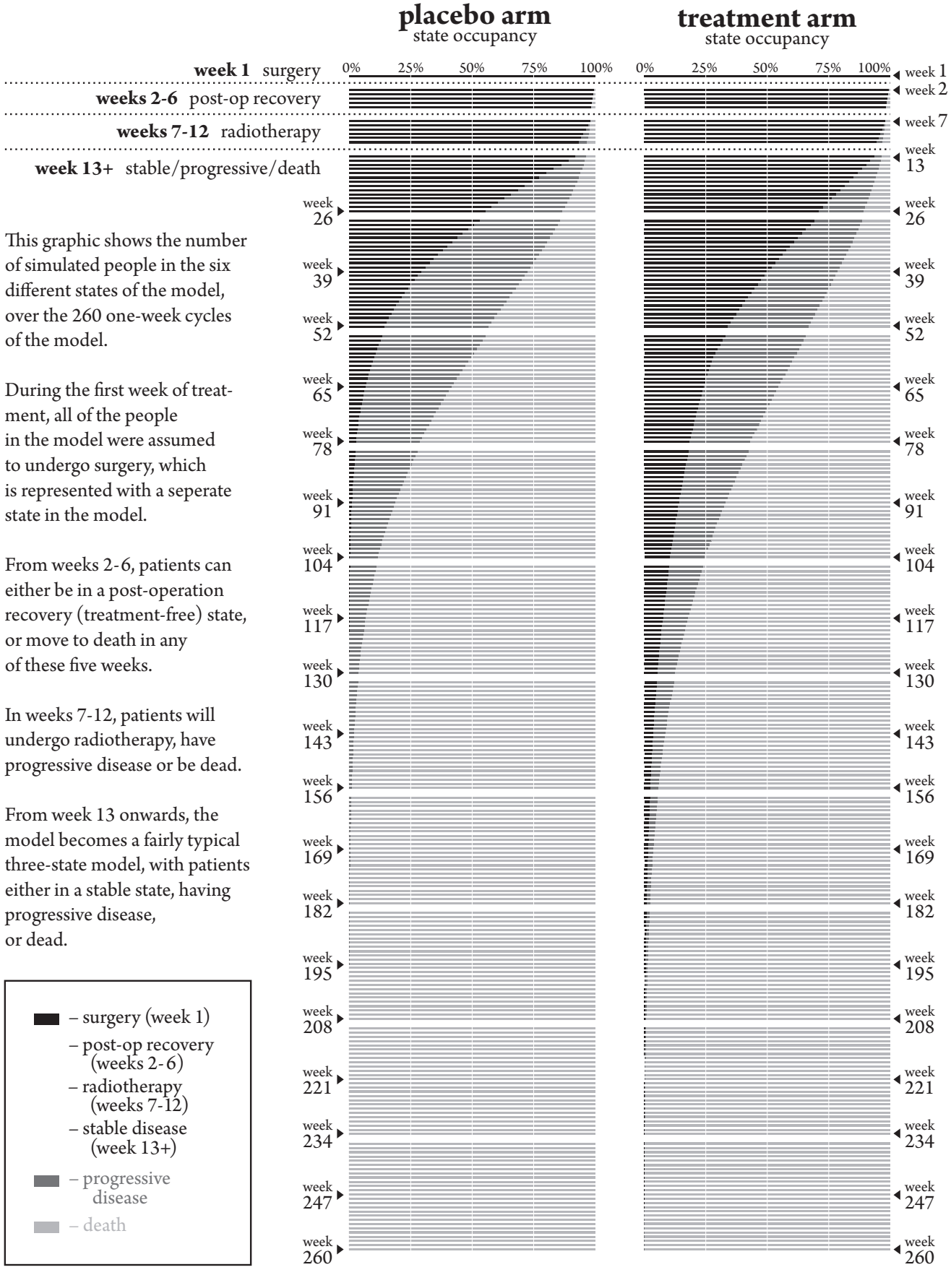
The version of the soc graphic that was tested in Chapter 6 is presented over the next seven pages. The graphic has been scaled to 90% of the tested size, to fit within the required document margins for this thesis.

State Occupancy Charts (SOCs)

temozolomide vs placebo
for the treatment of newly
diagnosed high-grade glioma

State Occupancy Chart

1



This graphic shows the number of simulated people in the six different states of the model, over the 260 one-week cycles of the model.

During the first week of treatment, all of the people in the model were assumed to undergo surgery, which is represented with a seperate state in the model.

From weeks 2-6, patients can either be in a post-operation recovery (treatment-free) state, or move to death in any of these five weeks.

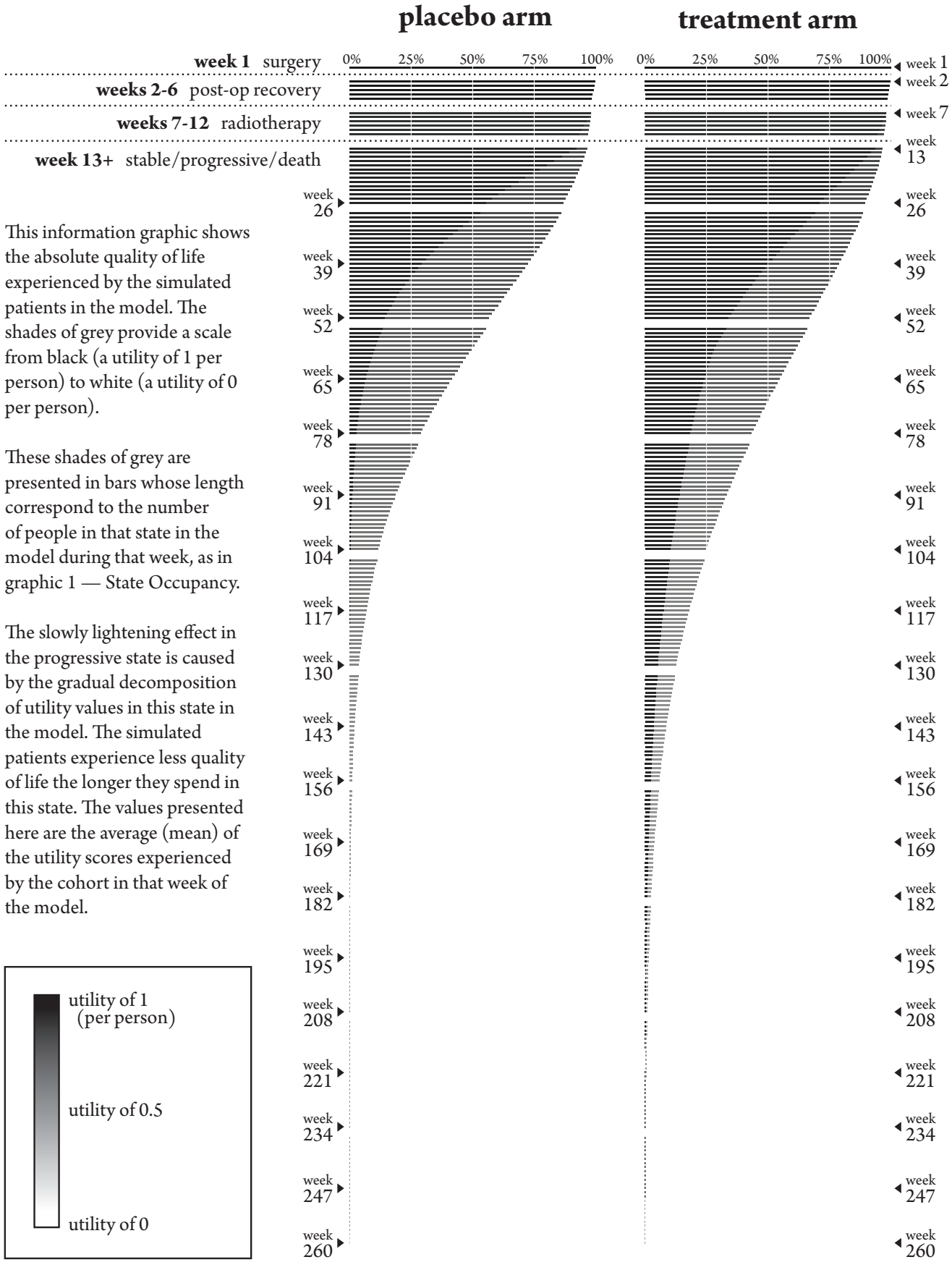
In weeks 7-12, patients will undergo radiotherapy, have progressive disease or be dead.

From week 13 onwards, the model becomes a fairly typical three-state model, with patients either in a stable state, having progressive disease, or dead.

– surgery (week 1)
– post-op recovery (weeks 2-6)
– radiotherapy (weeks 7-12)
– stable disease (week 13+)
– progressive disease
– death

State Occupancy & Absolute Quality of Life

2



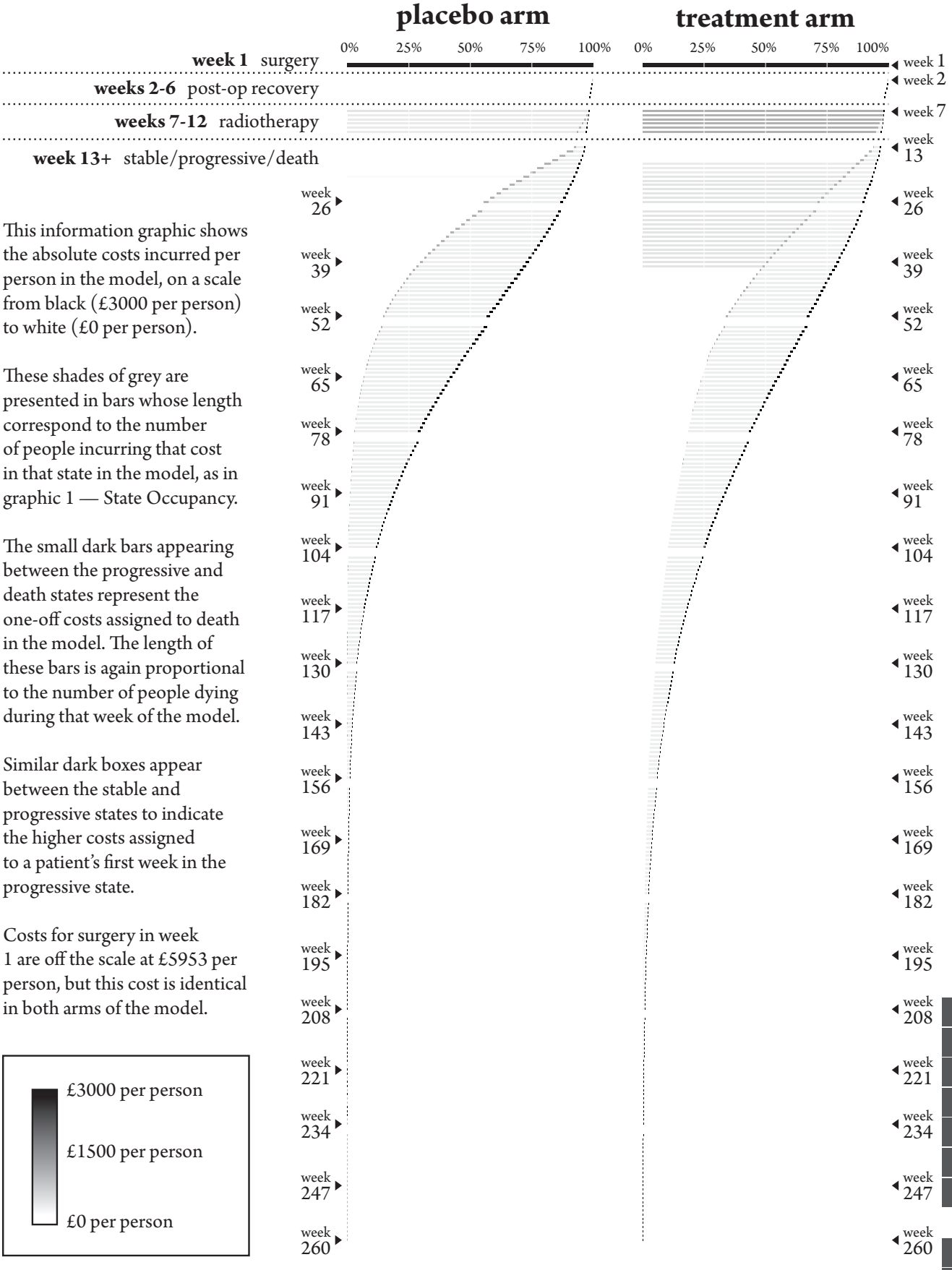
This information graphic shows the absolute quality of life experienced by the simulated patients in the model. The shades of grey provide a scale from black (a utility of 1 per person) to white (a utility of 0 per person).

These shades of grey are presented in bars whose length correspond to the number of people in that state in the model during that week, as in graphic 1 — State Occupancy.

The slowly lightening effect in the progressive state is caused by the gradual decomposition of utility values in this state in the model. The simulated patients experience less quality of life the longer they spend in this state. The values presented here are the average (mean) of the utility scores experienced by the cohort in that week of the model.

State Occupancy & Absolute Costs Per Person

3



This information graphic shows the absolute costs incurred per person in the model, on a scale from black (£3000 per person) to white (£0 per person).

These shades of grey are presented in bars whose length correspond to the number of people incurring that cost in that state in the model, as in graphic 1 — State Occupancy.

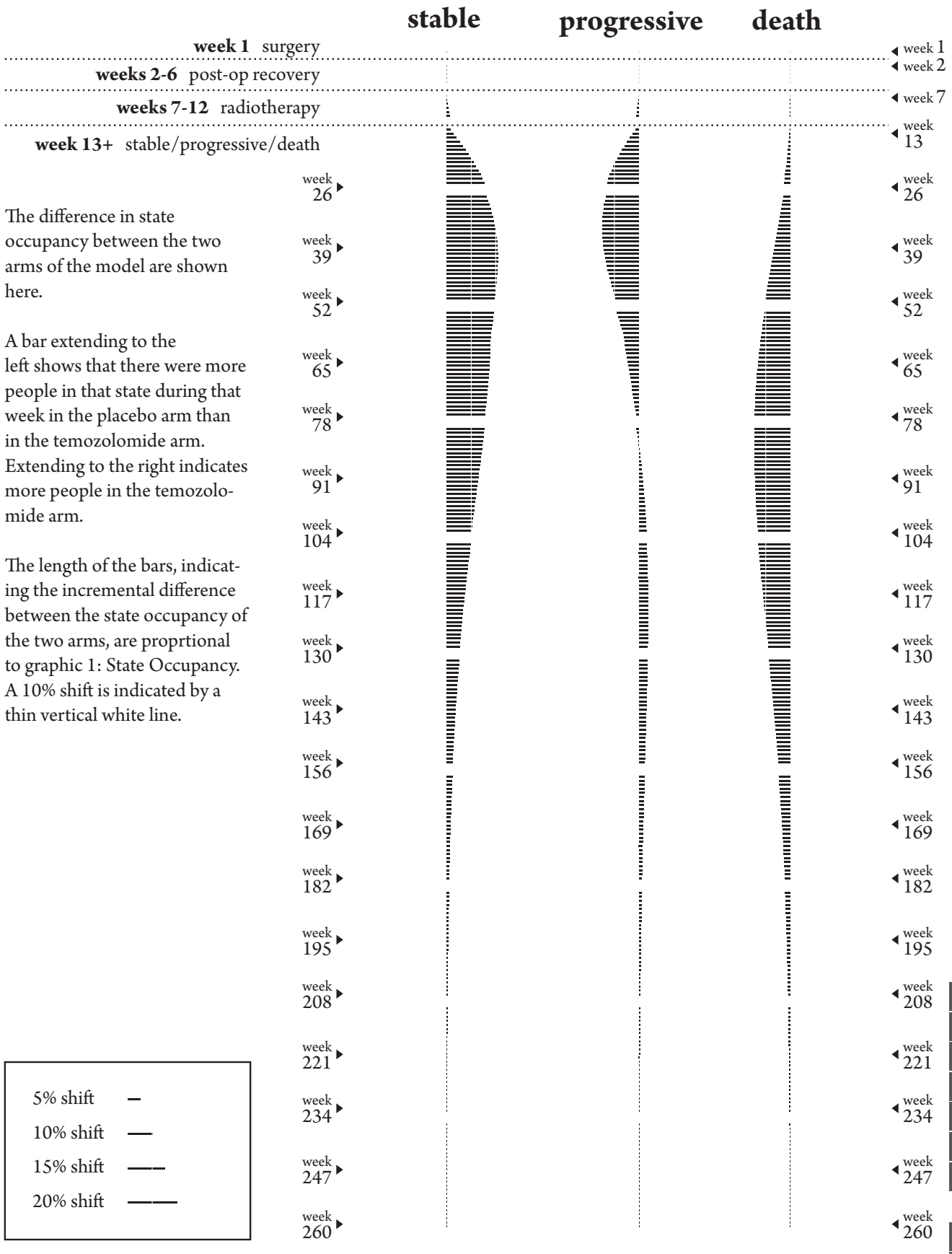
The small dark bars appearing between the progressive and death states represent the one-off costs assigned to death in the model. The length of these bars is again proportional to the number of people dying during that week of the model.

Similar dark boxes appear between the stable and progressive states to indicate the higher costs assigned to a patient's first week in the progressive state.

Costs for surgery in week 1 are off the scale at £5953 per person, but this cost is identical in both arms of the model.

Incremental State Occupancy

4



The difference in state occupancy between the two arms of the model are shown here.

A bar extending to the left shows that there were more people in that state during that week in the placebo arm than in the temozolomide arm. Extending to the right indicates more people in the temozolomide arm.

The length of the bars, indicating the incremental difference between the state occupancy of the two arms, are proprtional to graphic 1: State Occupancy. A 10% shift is indicated by a thin vertical white line.

5% shift –
10% shift —
15% shift ——
20% shift ———

Incremental QALYs

5

This graphic shows the differences between the two arms of the model in terms of the quality adjusted life years (QALYs) that would be experienced by a simulated cohort of 1000 people.

A bar extending to the left shows that, during that week, more QALYs were experienced by the people in the placebo arm than the temololomide arm. A bar extending to the right represents more QALYs experienced in the temozolomide arm.

Thin vertical white lines show the number of QALYs that would be experienced in a cohort of 1000 simulated patients. A bar that reaches one line represents one QALY.

The death state is not shown, as no QALYs are experienced in that state, as it was assigned a utility of 0.

The "total" column on the far right shows a sum of the values from the other two states.

incremental QALYs

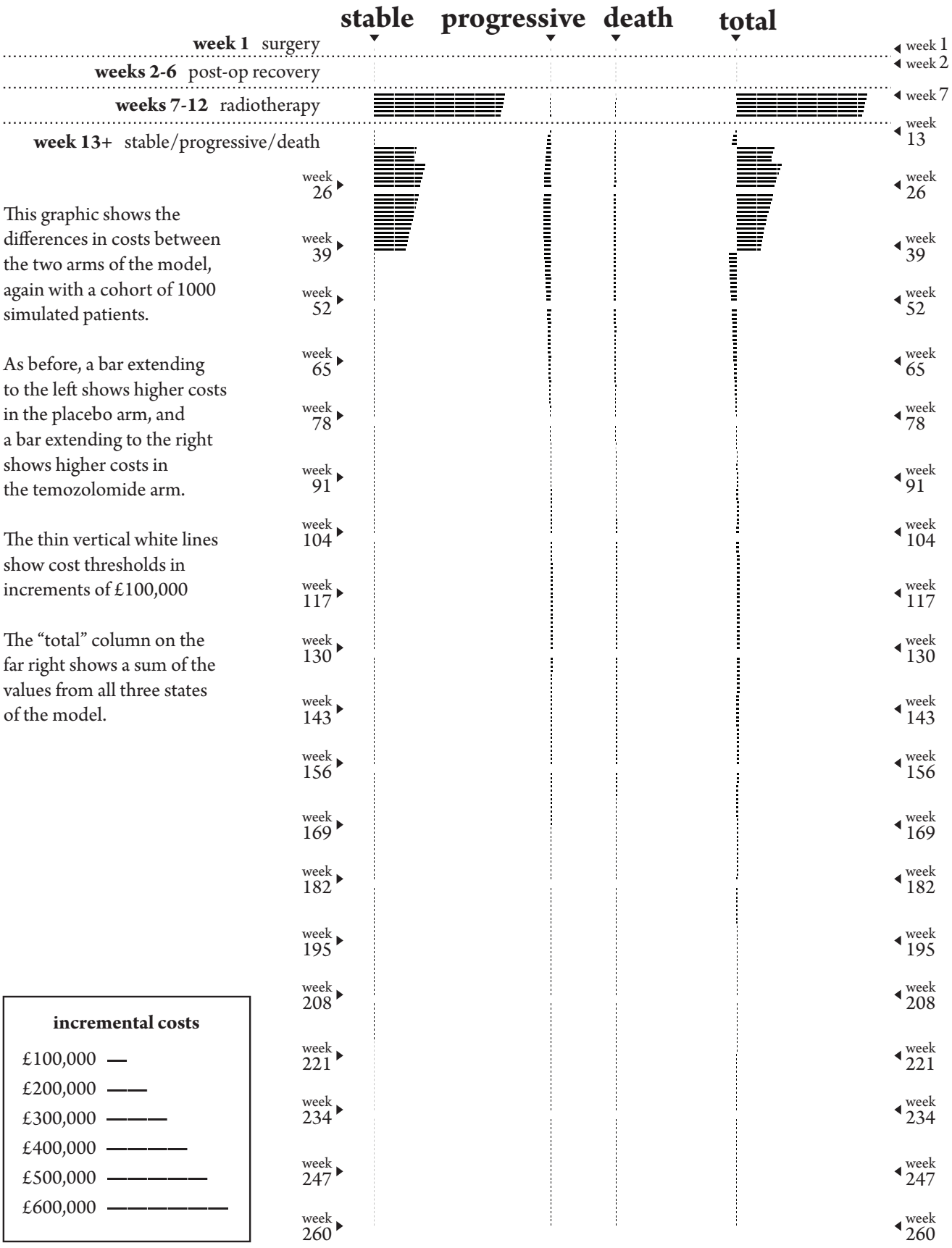
1 —

2 ——

3 ———

Incremental Costs

6



This graphic shows the differences in costs between the two arms of the model, again with a cohort of 1000 simulated patients.

As before, a bar extending to the left shows higher costs in the placebo arm, and a bar extending to the right shows higher costs in the temozolomide arm.

The thin vertical white lines show cost thresholds in increments of £100,000

The "total" column on the far right shows a sum of the values from all three states of the model.

incremental costs

£100,000 —
£200,000 ——
£300,000 ———
£400,000 ————
£500,000 —————
£600,000 ——————

Appendix H

Script for SOC test interviews

22nd July 2010

---intro---

Thank participant.

Presentation format that we are considering using.

Your answers help us decide how to present HTA data at PenTAG.

What we plan to do with research.

analysed results & anonymised transcripts in PhD.

Care will be taken that you cannot be identified.

analysed results may also be used in paper for IJTAHC

will not give the recordings or transcripts to any third party, or use them for any other purpose.

happy to be video recorded?

enable me to concentrate on tests, I will not have to make extensive notes.

---context---

considering use of graphical presentation technique in HTA work.

a way of showing state occupancy in cycles of Markov model

relates this to costs and QALYs

data from PenTAG model - Temzolomide vs placebo - treatment of newly diagnosed high-grade glioma.

I'll be giving you 6 tasks,

requiring you to pull information out of the graphics for me.

don't worry if difficulty finding information.

presentation method being tested, not you.

welcome to ask questions

I may choose not to answer if I feel that the presentation method should be able to answer.

process of finding answers without my help more representative of how someone would use graphic if in a report.

---personal information collection---

Before we begin, may I collect some personal information about you?

welcome to answer only questions feel comfortable with.

- Tell me about your experience with modelling / simulation modelling.
- What is your experience with HTA?
- Are you familiar with the carmustine implants / temozolomide for high-grade glioma report?
- Do you have the learning styles questionnaire I asked you to complete?

- Have you ever felt the need to show where the costs and QALYs come from in a markov model?
- In your experience, are decision-makers content to just be presented with an ICER?

Okay, we'll begin the test.

Here is the information graphic.

---SOC graphic given---

Please take a few minutes to familiarise yourself with information it contains.

---5 mins to look through---

I'll now give you tasks to perform.

We'll have a chance for general discussion after we finish,

but feel free to ask questions or make suggestions as we go through.

1)a How many people have progressive disease in week 52 of the temozolomide arm of the model?

1)b Is this more or less than in the placebo arm?

1)c How did you find this information?

2)a Where do the costs tend to come from in the model?

2)b Does this information help you to understand this model?

2)c Why is this?

3)a What would you say the average utility value of those in the

progressive state was in week 78 of the model, for the temozolomide arm?

3)b How does this compare with those in the placebo arm?

4)a Is there a particular model state in which the simulated patients experience more QALYs in the placebo arm than in the temozolomide arm?

4)b Why do you think this is so?

5)a At what point are more people in the progressive state in the temozolomide arm than in the placebo arm?

5)b **If graphic 4 used** -- Would you have been able to judge this without using the incremental state occupancy display?

If graphic 1 used -- Why did you prefer to use graphic 1 over graphic 4?

6)a Where does the greatest difference between the costs of the two arms lie?

---general questions---

- Was there any information that you thought was missing from the presentation that would be useful to show in an HTA report?
- Do you feel that any of the information is unnecessary for an HTA report?
- Do you think this kind of display would be useful or confusing to a decision-maker at NICE?

- IF YES - what might it show them that they do not already know?
- IF NO - is there anything that might be improved to make it more helpful?

- Do you feel it would be useful to anyone else involved in the HTA process?
- How easy do you think it would be to compare this graphic to one produced for a different appraisal?
- Do you think this graphic would be more useful for STAs or MTAs?
- Do you have any questions for me?

---outro---

Thank participant

Mention talk - 3pm tuesday

Appendix I

Sample interview transcript from SOC tests

interview transcript 6

Interview date: 27th July 2010

Interviewer

Participant

Timings

Questions/tasks

Participant's actions

<6s> - shows where participant is quietly reading or gathering information. In this case, for 6 seconds.

p32 / p vi is used as shorthand for the participant turning to a particular page.

question is used for shorthand to indicate that the participant is starting to look at the currently visible question card.

three dots (...) indicate an unintelligible syllable.

a hyphen at the end of a word (word-) indicates an unfinished sentence.

And I'll hide it away so it doesn't prove enormously distracting to us so just ignore it. Just concentrate on asking you the questions. So, oh and I should have said you won't be identified.

So I am considering a graphical presentation technique for use in HTA-type TAR reports. and it's a way of showing the state occupancy in different cycles in a markov model and relating those to the costs and QALYs that are incurred in different cycles of the model. So we've used data from this technology

assessment report which is high-grade glioma. A treatment called Temozolomide which is the drug that's given at the same time as radiotherapy.

Okay.

It's a sort of very typical end of life drug and it's just data comparing that to a placebo so it's a very simple situation I guess for this one.

1:00

and what I'll do is I'll give you a few moments to familiarise yourself with it and then I'll give you six tasks to perform just pulling information out of it. Don't worry if you find, if you have difficulty in finding any of the information for the tasks it's the presentation that I am testing not yourself so it's probably just that it's hard to find it if that's the case. Ask questions as you go through feel free especially anything that would be in the report that you would know if you were reading it and I am quite happy to tell you about the structure of the model, or anything like that that would be in the rest of the report. I might choose not to answer questions if I feel that the presentation method should be able to answer it itself because the process of you finding the answers to your own questions without my help would be more representative of how someone might use the report

The actual new one or the old one?

Sorry, the new one. you won't really have to look at the old one particularly.

Okay.

So

2:00

before we begin can I just collect a bit of general information about yourself? Can you tell me something about your experience with modelling particularly simulation modelling?

Okay I have been at ScHARR since '96 so however many years that is and then.

About fourteen years is that?

Yeah and since then, since the start of that I've been doing modelling, mathematical modelling and simulation in healthcare.

Extensive I would say.

Yeah okay.

And very relevant and was that your first experience of HTA in '96

when you joined ScHARR?

Yes more or less yes.

And has it all been for HTA work or- was it NICE work or-

Well NICE- I wouldn't include NICE as HTA. It depends it's a question is it just straightforward pharmaceutical drug A versus drug B work model or, because we get, I haven't, I've done a bit of capacity planning right at the start,

3:00

and I've also done more interesting work on infectious diseases and transmissions that go beyond the standard Markov so I am not really sure how long I think you want me to go on for about that.

Yeah just any indication that you could give me of how your work relates to what we do at PenTAG which is producing mostly producing technology assessment reports for NICE.

well it's totally-

But it's totally different to that?

Okay well it's, it's totally similar

Excellent

It's totally similar on the- I'm effectively- I am trying to think- _____ runs your group, doesn't ___?

Yes.

I'm effectively _____ [same name as before]. Except that I do far more modelling than _____ [same name] does. So I am on the committee as well as you know.

So you've got a real familiarity with the committee, that's brilliant.

Yes so I'm on committee C which is one of the Manchester committees.

Okay.

I've done fifteen of those

Points to report.

Yeah.

I guess, I've just had to update my CV for some of them which is why I know, I think I've done ten first authors or something like that

4:00

so I mean I know it very, very well.

So it's more management of the team that does the report.

You would hope that that would be the case but no I've rescued enough people and quite often I'm involved. We've just finished one last week where I've led it, done the modelling so, as well as supervising everyone else, I quite often do it.

Thank you, that gives me a really good idea of your role in it and you're familiarity with the committee will be very helpful. The thing I am interested in really is how to communicate the structure of the model to people at NICE and other decision makers. People that need to have an understanding of the model, but haven't actually got to grips with it themselves so that's going to be quite interesting. Great so are you familiar with that report at all?

No I've never seen it.

Great that's fantastic. You've given me your learning styles questionnaire sorry it was slightly irritating. I need to find a better one, but it has useful categories, so-

5:00

So if you were trying to present a Markov to somebody would you normally be content to just give them the kind of you know the outputs, the ICER value or do you need to sort of show them something about what happens in the model to produce that value?

To a committee, yes?

To a committee particularly, yes.

The strategy that I would apply- I think the problem is if it goes to a committee then it's a lead team member are you aware of this?

Yes.

And then they decide what they present so you get some influence, but, even if I was doing it when I do my lead committee ones I would be happy to give the summary values but I would know for sure where, what the key drivers were, what the utilities were in each state so that if anyone asks the question I could answer it.

That's interesting so it's often the lead team presentation- the person who is giving that the lead member who will know inside and pout the structure of the model, and present just key results from it, but obviously if it becomes crucial

6:00

they will have the information.

Yes and no at the committee what you've got is you'll have two or three people from the Institution.

Yes.

And they will be there but because for reasons of consistency probably style, grammar and knowing all the ins and outs properly the lead team- the assessment group's never allowed to present.

Yes.

So they are there and what you'll find is somebody will present their results and then as soon as the nitty gritty questions start coming the lead team member backs away and then it's the assessment team

then it's the TAR team

Yes.

Yes.

So, so the question, there's two questions (a) what you present in that

points to the report

and (b) what do you, what is presented to the committee.

But all committee members do get to see the report?

Yes

almost laughs

this is an in- the committee's made up of diverse members and there's probably four

7:00

designated health economists and the problem is a health economist is about, you get people who know all about the SF36 the EQ5D utilities and who know mathematical modellers but they are lumped together and they are vastly different.

So you could end up with four that are more into the utilities side of things.

You could do, and that would be a disaster from my perspective.

But obviously, but you try and balance those, I am sure sure they

do-

Yes they do, they do try and do that but outside of those and the people who are really interested I am unsure how many people read it. To me there's probably about ten people driving the decision from each committee. So those are the ten that really matter.

That's really interesting.

And so then it's going to be interesting how you, those should be the ones who as we saw in the email that went back just when me and _____ were having an in joke, I am sure most- I wouldn't be surprised if a third of the committee members couldn't tell you what a CEAC actually was.

8:00

Yes, sure.

And they then misinterpret it.

Yes.

So, but whether or not you want to get the information across to them- obviously the more information you have the better, but whether or not they choose to use it is a different issue.

Of course and you can never really predict what happens when you give someone the information. How long they have to get through it all, the conditions that they are studying it in, are they doing it at a weekend and on a train or-

To give you an indication for the last committee I went we got 1200 pages come through to read in five working days.

Good grief.

Before the committee.

Yeah it just seems insane to actually expect-

Well I had a chat with, not _____ who's the other, who's your other modeller?

_____ _____.

No not _____, _____ works because he has been thinking about applying and I think I probably put him off not intentionally but he asked about workload so I actually gave him an honest answer.

At that's a really interesting situation there, and I guess anything that speeds up the process of them getting information is good and therefore over-complicating things probably not.

It depends though.

9:00

I was teaching yesterday for the NICE Association of British Pharmaceutical Industry as soon as, I am also worried that people understand things a little and then make totally wrong conclusions because they don't understand it fully. So it's not, I am not sure how much you can simplify it apart from having people in who can do it in their sleep.

And I guess a committee is a selection of individuals, each with their own individual strengths, so those who do really understand it will be to guide it to some extent.

And the others don't which is may be what I was saying in a long-winded way.

Yes, no, no that's fine, that's fine it's good to say things in different ways to know that I am slightly sort of slightly understanding, and listening to the tape will be good as well. Okay then I think I've got a fair idea about that so what I'll do is I will give you the graphic presentation that we are testing here.

Yes.

And I will give you about five minutes to familiarise yourself with it and then I'll give out to you six tasks to pull out the information

10:00

and then we can have a bit more of a general discussion after that.

Yep, okay.

Cool so that's the graphic there

takes graphic from interviewer.

and if you put it there yeah that will be great.

reads cover <5s>

Oh if you wanted to see the structure of the model, that's

interviewer opens report, places it next to graphic

the state transition diagram there. It is fairly simple kind of end of life thing. There are a few sort of pretty much tunnel states at the start, and then it goes into stable, progressive or death. A simple end of life drug.

All right.

p1 <19s> report <2s> p1

11:00

<20s> p2 <54s>

12:00

p3 <64s>

13:00

p4 <41s> p5

14:00

<74s>

15:00

p6 <32s>

Okay?

Okay. Do you want me to make any comment at all?

Well we'll have a chance general discussion afterwards.

Okay I think one of your descriptions is slightly wrong down the left.

Probably is, yes. I'm not an expert in these things. My job is to produce the graphical things,

16:00

and I rely on the guidance of people for- so that would be really good to know, later on.

Okay.

But if I can ask you do the six tasks

Yep.

and then we'll have a chance to bring those things out when we go. Task number one.

interviewer gives question 1 [16:11]

reads question 1 <4s> p1 <6s>

You actually want the number for this, or-

Approximately yes that will be fine.

Okay.

bends to look closely at page

That's all a graphical thing can ever really give you.

points finger around week 52 on treatment panel

Okay. I was just checking. Hm.

key, question

Treatment arm week 52

p1

progressive disease

moves finger to right of progressive disease bar in week 52

65 minus

moves finger to left of progressive disease bar in week 52

30 about 35%.

completes response [16:51]

Excellent thanks very much. Do you think that is more or less in the placebo arm during that week?

bends over to look at page

17:00

That is less.

Great and you found that information by-

I did it expressly- I did my method for you on video.

Yes.

I just working out the length of the progressive disease arm.

Yes it's not a massively good quality video but it's just a sort of medium grey bar you are looking at.

Oh yes medium grey bar

And it was just above the white line where it says week 52, I guess.

Yeah, yeah.

Great that's perfect thank you very much. Shall we move onto the next one.

interviewer gives question 2 [17:35]

More of a judgement based question I suppose.

reads question <3s>

It could be just: tell me about the costs in the model.

Well if the model's been done correctly you've got a state costs for each of these ignoring your one off costs right at the start, you've got a state cost presumably per week or per whatever time cycle you are using for progressive disease

18:00

for- you'll have a one-off cost of death I think you said, and for stable disease. And just the sum product of those two.

Yes it's slightly have a look at the second page there.

p2

Okay.

Sorry the third page.

Third page all right.

p3

It's slightly different because, well, tell me if you can see from there that they might be different.

That might be different, okay.

turns page landscape

apart from assuming zero costs for stable disease.

Yes it's not quite zero, but it's so close to zero that it doesn't show up.

Right, okay.

<2s> rotates graphic portrait <2s> p5

I just want to check on the incremental costs to see whether or not

p3

those two

points to surgery in placebo arm, p4

are actually the same because obviously you've got the drug costs on there as well is that the bit you want me to

Is that for surgery or-

p3

No I am just saying on the treatment arm I didn't

19:00

point out the fact that whoever is on treatment will also be getting drug costs.

p6

Sorry, that was- I did it without saying. So,

<11s>

yeah which explains why at some point, across just individual time weeks, you get different, the costs in the treatment arm is actually

<2s>

more which would be due to the drug costs.

Yes, yes that's where the drug's administered. Okay great. So this one.

interviewer gives question 3 [19:45]

reads question <5s>

okay, so

p4, p5

incremental QALYs-

p6, p5, p4, p3, p2

20:00

here we are then

question, turns graphic sideways

I'm interested that you're naturally turning it on its side. That's a more familiar way of-

Well it is- for this one it just was, the other ones I can do it by I think I was mentally turning it on its side for the other ones. So for here

question

progressive

p2

state, you are in the treatment arm, you're at week 78. Progressive- that's stable progressive. Yeah.

looks at key twice

it's going to be around 0.5 mark.

completes response [20:41]

Okay great and how does that relate to the value in the placebo arm?

It should be the same unless you are getting toxic effects from the drug. The trouble is, I'm used to doing these so I'd be looking- if it wasn't very near the same I would want to know why it wasn't

bends closely to page

21:00

but let me just check. Umm no my eyes aren't good enough it looks about the same.

Yeah I think it was something to do with the quality of the printout as well. But no I think it is almost exactly the same, actually.

But if you are getting side effects it might be 0.01 less which you aren't going to pick up.

points to graphic.

Well exactly but then you might argue that it's not necessary to pick it up so much, as there isn't make such a difference.

It depends on what the outcome is I guess.

Very true, very true. Yes I think in this model there is a slight, the progressive state is done in a slightly complicated where it slowly degrades over time, so the longer someone's spent in it, the less utility value-

I suppose the point I was making there, I didn't go into explaining it in full the point I was making was if you have two identical drugs, bar one gave you stomach ache, that would come across on your mean cost per QALY table, but not on here.

Ah yes, yes.

So that's what I was saying.

Yes.

22:00

You wouldn't discount it because no one would take the second drug, it would be dominated.

Yeah sure, sure. Okay then.

interviewer gives question 4 [22:11]

reads question <8s>

Okay.

p2 <4s> question <6s>

And that question means particular model state.

Yes,

okay

sorry, it's a bit of a long question

No, no, no I was making sure that-

turns document portrait

unless I am reading this wrong let me just check what else you have given me,

p1

I'll give you my initial gut answer as it will hopefully teach you the way people- or give you an insight into the way people think.

p2

23:00

My gut feeling is that it's got just be on this graph, and I'm comparing the line colours

puts a finger on each arm of the graphic, around week 80, question

whether or not

p1

there's anything more,

p3

if there's a hidden thing I am falling into

p4

but immediately

p1

because it's per state you only want it to compare

p2

the three separate things.

points to placebo then treatment arm

Yes I see what you mean.

turns document landscape

Unless I am missing something let me just have a look,

holds graphic up closer to eyes <5s>

from that

question

apart from the side effects that we wouldn't pick up and also potentially a one off side effects of surgery

p2, turns document portrait.

which I am not sure you can see

I'm not sure that was reported

They should have done.

question

The answer is no

not sure.

Yeah well my answer's no.

or there isn't a particular state, okay.

Yeah.

Try and look at page five

p3

that might be a bit easier to see

p4

from there.

p5

24:00

Yeah I know but this isn't per state this is adjusted per by people in state isn't it.

That's a good point.

That was why I was asking what the question was.

Yeah, yeah, yeah.

What I'm saying is, say there was- the utility was 0.6 in both arms.

Yes.

But you have more people in the placebo arm.

Yes, yes.

That to me is a no but were you thinking that was a yes.

Yes I was yes.

Okay that was why I checked on the wording.

I understand.

Okay so I can add-

So you're really more interested in the absolute value- in whether they are actually produced because there are more of them because that's something that you've come across before?

<2s>

I immediately would have put per patient on the end of that.

points to question.

Yes.

Which there isn't so that

Yes, correct, I understand exactly what you are saying now. I was asking-

To me I would be more, it's obvious if you've got more people in there you would get more that's so obvious you don't really need it but what you do need to know is per patient to see if there's trickery going on

25:00

because it might be that the side effects are so bad that it's really driving the model.

Yes.

So in which case to answer that to answer that question then yes there will be because more people progress therefore you've got more in the progression [stage end?].

completes response [25:16]

The follow up question was why do you think that is, and obviously that was in your mind before even answering, so-

Yeah it would be interesting to see how many people you get- obviously for my sins, I've been doing this for too long.

That's really, really helpful thank you and I think that that kind of information might show me what would be more useful to present in different versions as well.

Although why and I'll ask you at the end I was going to ask you why here

points to negative QALYs during radiotherapy and early stable state for the stable display

unless it's a side effect of the actual-

Good point I am not actually sure.

I think that would be a side effect of the- maybe it's- they shouldn't both be that way obviously so it's got to be an offset. This is good, because it will indicate to me that, in the model, hold on, something is happening in these weeks that I need to know about.

26:00

Is it that this drug has a really nasty nausea for the first two weeks, or whatever it is.

You are pointing at the stable state in week 13.

Yeah, yeah.

There's two weeks where there are-

In week 13 and 14 the incremental QALYs are negative for both stable and progressive whereas we would suspect that more people would be as happens in week 39 for instance that there will be more for the stable in the treatment arm and less in the progressive. So something that would be good at indicating it.

Okay, sure, excellent thank you very good.

interviewer gives question 5 [26:37]

reads question <6s>

Okay.

closes document, p3 <4s> p2, p1

... ...

turns document landscape, question

... ... people in the progressive state

p1

27:00

places two adjacent index fingers below placebo panel, around weeks 91 and 104

so for your tape I'm looking to see when the progressive bars are roughly equal.

Yes.

Which isn't going to be-

makes strained noise

hang on

turns head to view portrait

for the sake of this about 68 I guess about

turns head back to view landscape

oh no more 78

turns head to view portrait

about 80

completes response [27:18]

places fingers on progressive state in both arm panels, around week 80.

I was just trying to look and see where the bars were identical.

That's pretty perfect, thank you for saying that out loud. So is there any particular reason that you prefer to use that one, this question could be got from that or from page 4 and particularly

I think it probably be because I prefer working with solid bars.

p4

Yes.

Well you can't get it from four if there's going to be a different utility effect can you, hold on oh this is your incremental state yeah, no you can do so you could have been looking for when that hits zero.

Yeah.

Yeah okay well I wasn't far off then. I think the main reason I went to that one is

28:00

'cos I saw it first and knew I could get the answer for the rough degree of accuracy I wanted but you are right for more accuracy you should have gone for four.

But it might also be the function of having seen quite a lot of state occupancy diagrams like that before. I mean normally you would use lines I suppose, but that might be more familiar.

Possibly but this is, I mean I actually like these graphs they are useful but they are not difficult to interpret so it's it is a better way there you didn't have to do the eyeballing two lines this is a far better way of doing it.

And the last one.

interviewer gives question 6 [28:35]

reads question <3s>

Ooh.

p5

that's a bit more of a

p6

yes, a judgemental one.

Hold on I need to be sort of looking at two at the same time

holds hand in p6, turns back to p3

to see which one is the better one.

<3s> p2, still holding p6,

29:00

p1, p2, p3 <3s> p6, question <2s> p6 <3s>

Okay I would be saying there's clearly something going on with

radiotherapy costs in weeks 7 and 12 but then there's also the cost presumably of the drug costs of stable patients and just general treatment within weeks- I don't know if that's 15 or 13, down to about 39.

Yes would you say either of those had a greater effect?

um. ts ts tsss.

<3s>

It would be interesting to see them added up. They probably look about equal although this one may be the 39 might shave it but four lines at double

30:00

is going to be the same as eight lines at half I know that isn't quite right but it's hard to tell.

No that is really interesting I think they are almost exactly equal so it's interesting. It seems that people can measure that fairly naturally. I was wondering if one would outweigh the other, but that doesn't seem to be the case. Great, well that's the end of the questions thank you very much indeed.

Can I tell you that thing

p5

just while I am think of it.

Yes do yes.

I think yeah let me just check. I knew when I read it about three times just to check.

Sorry.

points towards bottom of text area <14s>

I just don't think I had it right earlier in my head. It doesn't show the number of QALYs that will be experienced, it's obviously the incremental QALYs

Ah.

QALYs.

31:00

Yes so it's the incremental QALYs that would be experienced.

It's as soon as you were saying the cohort, well I don't think the cohort size, okay the cohort of 1000

Yes.

It, in number four when you pulled it out it slightly changed the sentence and made it better I think. You said the incremental difference between the state occupancy- oop, typo

Which one, proportional?

yes. A ten percent shift is indicated- okay so that's just ten percent so that one's fine I think it was just a bit strange that it just said the number of QALYs from a thousand.

Yeah, yeah the number of, yeah you are absolutely right that should be much clearer. I am slightly hampered because I don't actually have a background in-

No, no, no it's just in case 'cos I knew that one and I read it through three times just making sure I understood what you meant.

Yeah, yeah that's fair enough yeah that will be helpful luckily this isn't going into a live report.

32:00

It's only minor anyway.

That's great no thank you I can see what you mean. Right well if it's okay we, if I can go on to asking you a few general questions about it if that's fine we've talked about half an hour now so if we can go on for another ten minutes of so.

No I'm fine.

Brilliant thank you very much. First of all was there any information that you thought was missing from these graphics that should be added in that would help to understand the model?

Are you taking it for red

p1

that the mean costs and QALYs are already upfront in the table?

Yeah they should be presented numerically elsewhere and I am pretty sure that this was never obviously be able to replace the numerical-

Yeah, yeah. No it was very intuitive and useful. It would only- it would be of most use for detecting errors within the model.

Interesting.

Interesting for that one I picked up and said: well, why is that

happening. So it would be good for eliciting or making explicit assumptions made within the model.

So it might be useful in the actual modeller

33:00

themself to look at as well.

Yes they should be doing that anyway I would say but yeah the modeller should always understand the model he's but yeah if he didn't mean

p5, points to negative QALYs at top of stable state

that to happen.

Yeah.

Then, yeah if you didn't mean the two bits on the left of week 13-ish to happen then obviously there's a problem in the model but if you did mean it to happen you could say to the committee this is due to reason a, b, c.

Yes okay.

So no

p6

I mean

p5

it was intuitive enough to follow it pretty much easily.

If you had to take one of those six pages out which would it be?

p6, p5, p4

Okay

p3, p2

amusingly going back to

p1

my previous answer

p4

4 you can take,

34:00

4 might well be the candidate

p1

because you can calculate that yourself.

Looking at 1.

Yeah which is the longer way that I did it rather than the shortcut.

Yes.

So.

p2

If there was a real premium for space in ....

Well yeah okay if there was a real premium for space two could go and you could also have on the top

p1

of here or somewhere on here a chart that says

points near top of text, p1

utility state and treatment, non treatment and you could just do that.

p2

I mean that's a handy way of doing it but you could always- I suppose where I am going to is probably most of it- there's two things, if you are really worried about space you could probably kill almost all of this. If you are wanting for informative stuff that will help decision then it is useful.

Cool.

But I, you know you could extract almost all of this from here points to report but not as quickly.

p3

Yeah okay.

So yeah.

p1

35:00

Which would be the most useful

p4

for that, I suppose, is the question.

p5

Five immediately showed generally QALYs drive it, drives the decisions particularly over a long term disease. So seeing how they were built up was good.

p6

ditto with the cost if you had to just,

p4

if you only had to run with two and the rest were in tables

p3

five and six would be the ones you would keep

p1

alongside one.

Great thank you for that that's great. Okay then do you think this kind of display would be useful opportunity to a decision maker at NICE?

laughs

It comes back to my first question. If you only assume that fifty percent of people understand the mathematical modelling that goes on then it will be useful to fifty percent who will be the ones making the decision the other fifty percent probably just glance at it and allow other people to make the decision.

That's interesting.

36:00

But yeah I mean if this was in, this was in a report, I'd think: Yeah, that's good.

Okay and we've already talked about other people in the HTA process. Do you think there's anybody else apart from modellers and committee members that it might be useful for?

Any interested person it will be useful to but it depends who they are. It might be drug competitors working out how the model was put together, it would certainly be the drug company wanting to know if they believed your model.

Yes.

So it would always be useful, anyone who wants to validate the model it would be handy a handy tool.

Good that's good and if I produced a different one of these for another intervention how easy do you think it would be to compare the two?

Different disease areas or same disease area?

Yeah different disease area

37:00

and different treatments.

The question would be why do you want to compare them unless you're running a cost equality matrix in which case the table up front is your one.

Yeah that's true, that's true so I guess I'm saying because there are some fairly arbitrary values like the maximum cost, the blackness is 3,000,

Oh okay.

in another treatment it might be 6,000. Do you think that's a weakness?

No because it would be making all these within a certain disease area and that if it was an STA it would be confined to that one drug and its comparators within that disease area and within the MTAs it would be within that disease area.

So sort of break it down into smaller comparisons.

More interesting would be and we will come on to in the end is what you do when you don't want to use a Markov model or when you use a complicated Markov model that has multiple states.

That's true yeah like 16 or 18 states.

Yeah if you are putting patient history in there that would be an interesting one.

Yeah.

38:00

There was a small about of that there, but yeah. It might, yes that's an interesting point and actually, may as well say something about that now. I was thinking that perhaps if it was something that people were quite familiar that that degree of blackness equals the amount of costs produced, it might be a bit more difficult to interpret which stage is which because you certainly see at what point in the model the costs were coming out in which states especially with good labelling.

Yeah you should be able to. Mine is more for instance if you had an MI but then break your wrist you don't want to lose you know the wrist utility is something like 0.99. You don't want to lose the fact that you've had an MI and you are actually at point A. So you have to be containing within the model, you might need an MI and wrist state. So all of a sudden you are now at- it's going to be a lot harder to- it depends.

39:00

85% of models that we see aren't difficult and would benefit from this. The other fifteen might be harder to deal with

sure.

there's an osteo report I did that was an individual patient model that I think had something stupid- it was in the- if I had done it as a decision tree or as a Markov if would have needed something like- as a decision tree it would need to be something over like two million states.

Good grief.

But then you've got four different fracture types, MI, CHD and you've got to be keeping a history of them. So

Wouldn't you do something like discrete event simulation then?

Yeah which is, that was back in 1999 when I did that work now I am doing it all on DES.

Yes.

So the interesting, I do a lot of the DES the interesting thing would be how this would come across in a DES model.

A very good point. It's designed for Markov, just because

Yes I am not knocking it, for a Markov it looks really good and I am just throwing out- because I do a lot of DES work.

But would it work with DES?

It would be harder to do

40:00

but that doesn't stop you saying this is a good idea. I don't want to really put a damper on it.

No, no that's fine and- the only reason is that there are a lot of Markovs used at the moment.

Well mainly due to NICE's rules on software

oh, right

but at the minute they are only letting through models- oh, well it's not only, there are exceptions if you notify NICE in advance and the assessment group accept.

Yes.

but it's Excel, Triage and WinBugs there may be another free one that's in there oh what's it called-

But you're not actually- unless you've got specific permission, you're not going to be using Simul8, or-

No well I love that program that's the one I use all the time. I've done an assessment group in Simul8 and got my knuckles wrapped. One's come through in Simul8 and I had to go and teach _____ [another TAR team] how to use Simul8 so they could do it, but no, not allowed which is a crying shame but yes.

41:00

The reasons for that 'cos I'm not just saying- this the reasons for that is that if you use Simul8 them what about Arena and and what about the other twenty competitors but that's a much different question I suppose. I'm veering off, so I'll

No that's fine, that's fine. We are very nearly finished actually. Oh I just wanted to ask you do you think this would be more useful for STAs or MTAs?

Probably about the same the only thing that could be interesting on an MTA is that you could do, you could obviously make this, it would be nice to get then all on one page, but if there was more than three it might be problematic.

I guess you could probably do four and

Yeah you would ideally only have only do one on the efficient frontier or the CEAF.

Oh yes.

Are you happy with the cost effectiveness acceptability frontier?

42:00

I think so.

It's the ones that aren't dominated or extendedly dominated, so you take them out, and then you just get

I see the- so you have an incremental gain in each case.

Yeah, yeah so that would be, if I was doing it or wanted to see

it for an MTA then that would be the ones that would be the obvious candidates but if you ordered them in the right order it wouldn't be a problem anyway and saying that people can always print out a second page and just put them beside it

That's quite a good point actually yes.

That's what I'd do.

I know that seems simple, but that does- I mean, this is stapled because I know that people will get them in a bound report and won't be able to look at them side by side and so I've been testing everything on that basis.

Well you would expect that because of this cost cutting in NICE means that we might actually have to print out everything ourselves so the bound report is- yeah, I think that's going to cause a few arguments but the bound report may be a thing of the past.

Yes, distribute them in PDF format,

43:00

and let them read it on screen.

But when you've got 1200 pages and people are already doing it for free and their work is then getting charged for it.

Yes.

We'll see. it helps me and it helps people like _____ [director of another TAR team] because it helps us run, and see what's going on, and make better decisions.

Yes.

But for the people-

You have people like me saying you should be printing them in colour because you get more dimensions of data in the graphic, but I don't think that's going to happen.

No you can tell from this more or less anyway.

That's true yes

And this wouldn't replace anything

no.

this would just be a really handy go to sheet in the committee when you've already read it and you've already read the numbers in the table but this is quickly let me just test my theory again.

Yes I guess so. Yeah, I think that's how I see it being used. have you ever seen anything like this done before?

Surprisingly no 'cos it is simple and easy.

Yes.

It's good that you are, that you are putting this together and it will come out.

44:00

You may have hit on something nice and easy that no one has done before, which is a good thing to do.

I mean presumably you must have seen state occupancy diagrams like that

turns p1 landscape

with a line, curve, graph-

turns p1 portrait

Yeah I mean that one is fairly standard. I don't know whether it's laziness, time pressures why we don't normally do them. I guess it's a bit of both.

Well I had to use some slightly unusual software to get the value- the grey scale values on there I didn't do that in Excel that was an open source piece of software called-

The interesting thing would be, sorry,

p2

it's something that I was thinking about 'cos you couldn't pick it up off it but was this a really chronic disease as that went for thirty years it wasn't end of life? Presumably by the time you got to year 30, almost everything would be close to white.

Yes.

Because of the discounting. and I think that was what you meant with your-

points to top of third paragraph on p2, talking about the lightening effect in the progressive state.

Yes, and this is un-discounted as well actually, but yeah we-

Oh okay

45:00

I was assuming what's the gradual decomposition thing then?

Oh sorry that's not discounting that's,

just age then

that's the wording that was used that was used in the report but I don't know what the reasons for it are I am afraid but it just, the utility value slowly-

And it isn't discounting, because it screams discounting to me but

It could be actually that's a good point the costs aren't discounted and maybe I've done something horrible and discounted the QALYs and not the costs. Because the discounted costs value which I didn't use.

Okay it's just that, yeah, each year- well there's two things going on with the QALYs and it could be one is each year you get older you get less utility so there's that part one and then there's the discounting so maybe it's that first one I will be surprised if they didn't do discounting though.

Oh, they did yes, I just didn't use it for this particular-

But the thing is no decision is ever made not discounted.

Yes.

So you would need that and so

It could be done.

I guess where I am going to is once you get beyond year 15 this graph will, this colour scheme won't be particularly useful.

46:00

Yes.

Unless you are using it, unless you make it damn clear that it's not discounted in which case

p4

these

p3

numbers

p5

here though

points to p5

would need, these have to have discounted so you couldn't tie up the two anymore but that might be a worthwhile trade I am not sure so to be able to see what's going on.

Thank you. That's good and do you think it would be useful to display scenario analysis with these, you know, alter a value and then display more of them?

Yes but then you are coming to a common sense decision over how many

pages you're willing to take up with them.

Yeah if you only did it for a small number of scenarios that really were key and driving it then it would bring home why the values changed or why the decision changed.

Which ones would you want to see different-

47:00

Oh I don't know for this- not here, I am just saying that if you decided that actually

Oh- Yes I see-

If the treatment had a ten percent chance of killing you.

Yes.

Rather than a one percent chance just say there was new data that came in from observational study and the committee wanted to explore this higher death rate, that could totally flip the decision and then it would be better to-

points to p5, makes disappointed/frustrated sound.

death would be harder to see. You could pick it up on your

p4

state occupancy, and your incremental and there

circles graphic with finger

but- and also on your first one, but things like that. Or if you suddenly found that your utility differed, that progressive actually was nowhere near as bad as you thought it was is now not 0.5, it's 0.8

p5

then showing how this

points to around week 39 in progressive state, indicates moving out to the left

would bulge out more then this

points to total column around week 39, indicates moving to the left

would come more that side. So the trouble is the only thing, the people who I suppose it's useful and it is good to make sure we understand it

48:00

but the people who will be driving the decision will already be telling me what's going to happen like I did then. I would be, it's good because if it didn't match my expectations I would want to know why.

Very true.

So it's a validity- if the utility for progressive went up and this graph didn't bulge out more here

points to around week 39 in progressive state, indicates moving out to the left

and like that

indicates leftward bulge moving further down in time

and bit more like that way

indicates rightward bulge after week 91 extending more to the right

and then that

points to total column

didn't alter accordingly I would want to know why. It would show that something in the model that isn't right.

That's really interesting so you might be able to just run these on various things, maybe not necessarily include them all in the report, but just say: this is an interesting finding, this goes into the report.

Hopefully if it's an interesting finding they will have fixed it

before it gets to that point

yes, or explained why so it might be useful for them looking at it. I firmly believe that every modeller should be: (a) pre-guessing every result that he runs (b) checking that the results match his expectations if they don't, prove to himself why he was wrong or fix the model.

49:00

So it could be useful more like that than rather for a committee member but-

Yes and it's quite useful if we're producing graphics, as well, if we're saying that should look like that from those values, look at the shape of it and thinking to yourself-

We deal with numbers at the minute only because we don't have- well, probably laziness and lack of experience in how to do it, lack of the knowledge.

I know that everything is sort of, you are quite limited software available as well.

I've certainly drawn these in Excel for my own reasons to check, just for validation so that they are used I suppose then, but everyone builds their model differently so

Well that's great I think that's my last question do you have any questions for me at all or

I don't think so.

Is there anything else you want to know about it?

No, no what's the, so is this part of your PhD or

Yes well I've designed about ten different presentations to be used in different areas of TAR reports-

interview ends [50:00]