Chapter 5: Prototype Test 1 (GOfER)

This chapter, and the following one, describe the empirical testing of two of the ten graphical techniques described in the previous chapter. The process could, given enough time, be applied to any of the information graphics in that chapter.

The two graphics chosen for this final testing stage were selected by trying to balance the potential value of the graphical techniques with the effort required to develop them into a real data stage, the ease of reproducing them in currently used presentation media in the NICE appraisals process, and the ease of testing the graphics (in terms of having readily available comparators for testing, or the availability of experts to test them, for example).

The first prototype tested is the one presented in 4.2.1 and 4.3.1, which became known, first as COGS (Clinical Effectiveness overview Graphical Summary, and then as GOfER (Graphical Overview for Evidence Reviews). This information graphic offers a way of giving a graphical overview of the results of a systematic review, potentially addressing several of the issues raised in the information needs interviews in Chapter 3.2. The data used to construct the graphic comes from the PenTAG systematic review of the clinical effectiveness of cochlear implants. This was a complex review, mentioned by two people during the information needs interviews.

5.1 Introduction

5.1.1 Testing graphics

This chapter details one way of testing information graphics – with a comparative study. This forms one of the pathways from Prototype to deployment in the proposed design process model for information graphics in HTA (see Figure 5.1 – 1).

Information designers have long felt the need to validate their work through testing. A memorandum from The Otto and Marie Neurath Isotype Collection at Reading Uneversity, written by Marie Neurath suggests that: "The team should have some arrangement that they can test their products", because she felt that: "The teacher trainee should be able to get the whole message out of a chart by putting the right questions and thus making people look carefully and SEE the answers from the chart. If people can't see them, the charts are bad and have to be redesigned." (Marie Neurath 1955).

There is a long history of empirical testing in design. One recent proponent of such methods, David Sless, writes of not only a need to test newly designed presentations of information through task-based user testing, but also a need to establish the performance of existing presentation methods prior

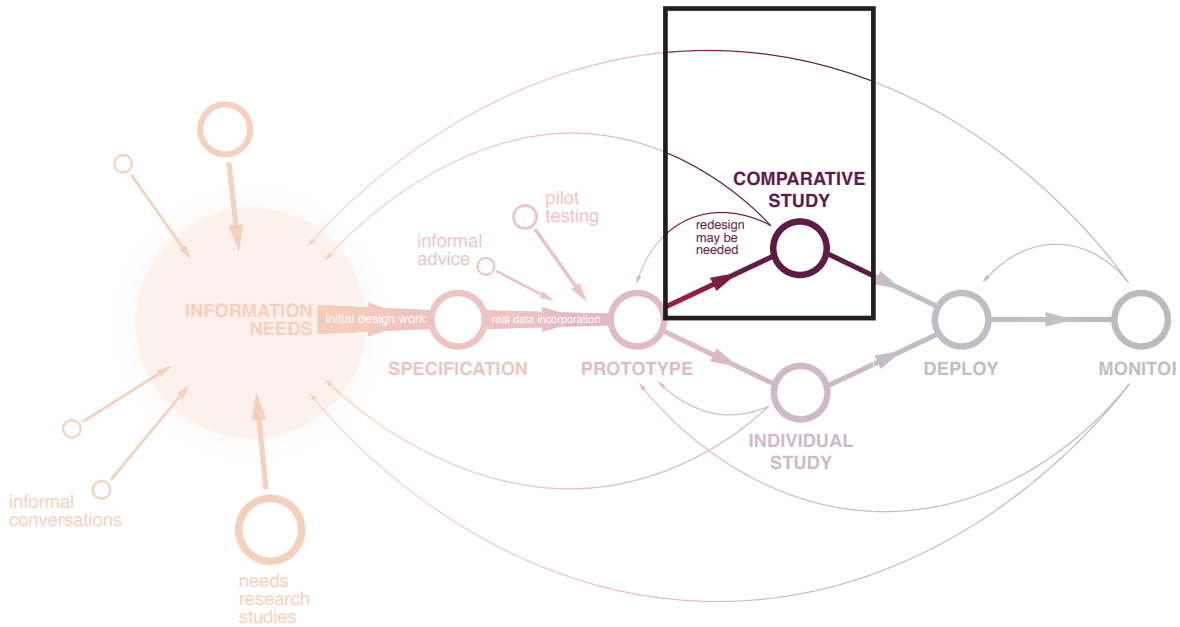

Figure 5.1 – 1
The phase of the proposed design process detailed in Chapter 5

to these tests (Sless 2008). The idea of such tests is to demonstrate relative effectiveness of new information presentations to the original ones. This can show an appreciable value of design to those without knowledge of the field. Given Easterby's views that designers are not good at providing any kind of validation of their designs beyond showing the designs themselves (Easterby & Zwaga 1984), this approach might be somewhat unusual. Not everyone has the experience provided by many years of design practice or training, however, and it is important to show the benefits of design in a language understood by those that might commission design and benefit from visual communication. In the scientific field of HTA, this is the language of empirical testing.

In HTA, the formal evaluation of different presentation techniques might serve to demonstrate the differences between numerical, textual, and graphical presentation of data, such as that produced from systematic reviews or mathematical models. It will be the researchers themselves that produce, and are tasked with disseminating, the data from systematic reviews and mathematical models in HTA, who will ultimately determine whether graphical techniques are used or not. An empirical study showing the strengths and weaknesses of these different presentation methods might be able to inform this research community, by being publishable in its own right in HTA journals.

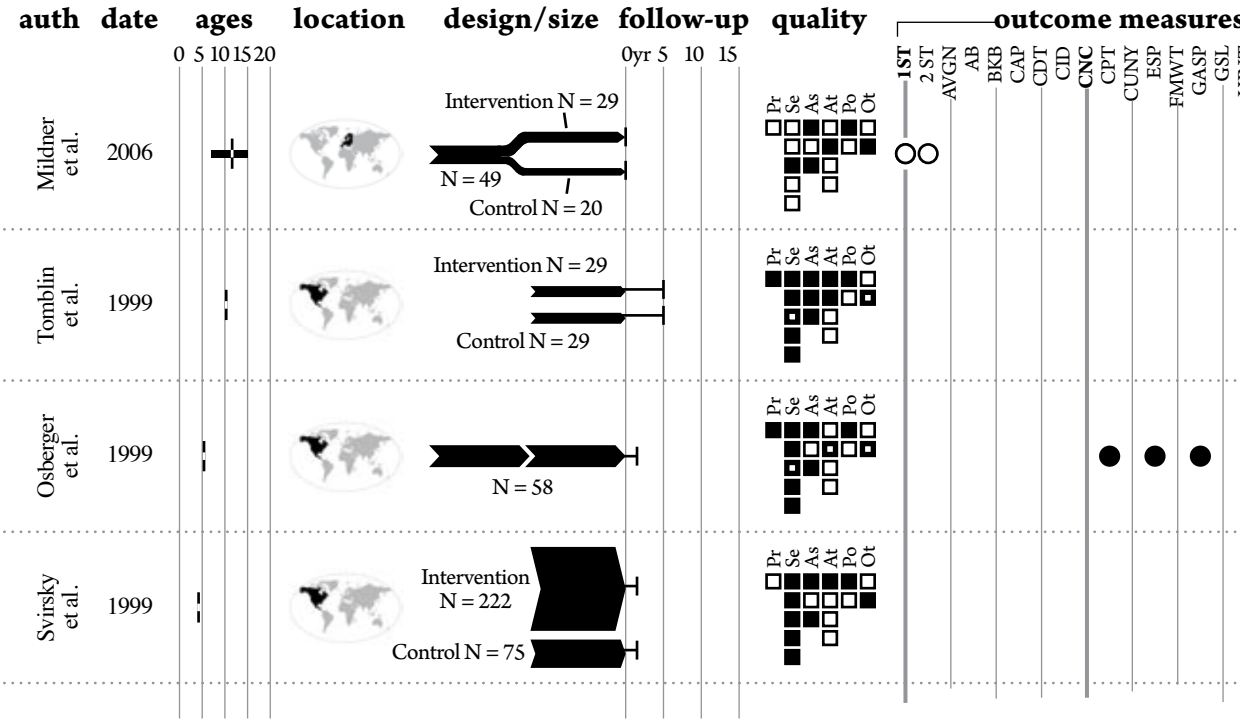

Figure 5.1 – 1
A small section of the tested GOfER display

Table 15 Summary of study characteristics unilateral cochlear implants vs. acoustic hea

Study ID	Design	Intervention group	Control group
Mildner et al[125] 2006 Croatia Length of follow up: n/a N = 49	Randomised, controlled	N = 29 Age yrs: mean 11.6 (7-15) Degree of deafness: profound >98 dBHL Mean age (yrs: months) at implant: 8.2 (2-12) Mean time (yrs) between deafness and implantation: Not reported	N = 20 Degree of deafness: profound > 98 dBHL
Tomblin et al[126] 1999 USA Length of follow up: 5 yrs N = 58	Non-randomised controlled Prospective trial with cross-overs allowed	N = 29 Age yrs mean (SD): 10 (2.9) Degree of deafness: profound Pre-lingually deaf Mean (SD) age (yrs) at implant: 4.76 (1.57, 2-13) Mean time (yrs) between deafness and implantation: NR	N = 29 Age yrs mean (SD): 9 (3.65) Degree of Pre-lingually deaf
		Participants	
Osberger et al[127] 1999 USA	Pre/Post prospective Repeated measures	N = 58 Age mean yrs: 5.4	

Figure 5.1 – 2
A small section of the data tables from the original report

5.1.2 Testing GOfER

The GOfER display, described in Chapters 4.2.1 and 4.3.1 was chosen for testing. A small section of this graphic is shown in Figure 5.1 – 1, which presents just four of the data on four of the 24 shown in the graphic. This full version of the graphic is included in Appendix C.

The GOfER graphic was chosen for further evaluation, for several reasons. Firstly, the members of the panel discussion detailed in Chapter 4.2.1 had a unanimously high opinion of its potential usefulness. Also, this graphic had a direct comparator available – the existing systematic review results section from the Cochlear Implants report (Bond et al. 2009) from which the data used to create the prototype was taken. This section contains all of the data that was used to produce the graphic, but presented in more traditional tables containing numbers and text (see Figure 5.1 – 2). This allows an experimental design in which participants' performance and experience with the presentation can be compared to that of others with a different presentation method.

The GOfER display presents information on a set of published effectiveness

studies, and how they were conducted. This information is often presented in technology assessment reports using tables of text and numbers. The graphical GOfER display is an alternative way of presenting this information, which has two main functions:

Firstly, it enables a larger amount of information about the trials included in the systematic review to be presented in the area available on a printed A4 page of a technology assessment report (the 'condensing' function from Chapter 1.3.2) (also see Tufte 2001). The condensing of information in a GOfER display is potentially useful, in that a more complete picture of a trial's characteristics can be presented in one place, enabling the viewer to compare multiple aspects of the trials together (therefore also fulfilling function 3 'comparison' from Chapter 1.3.3). For example, GOfER is designed to allow the viewer to see which studies had statistically significant outcomes in different measures at the same time as considering the size and quality of the trial. This information had to be split up over several separate tables using numerical and textual data presentations in the published report.

Secondly, the GOfER display is designed to facilitate a quick overview of important information (Function 4 from chapter 1.3.4). The information needs interviews in Chapter 3.2 established that decision-makers had limited time available to digest the information presented in TARs. The principles investigated in (Resnikoff 1989) suggest that the human visual processing system is a "high bandwidth channel". Our ability to quickly process, and judge the approximate sizes of objects is employed by the GOfER display. This might allow an HTA decision-maker to gain a quicker understanding of key characteristics of the studies included in a systematic review.

However, it is critically important that the graphic presentation technique does not sacrifice accuracy of understanding to provide these benefits. Therefore, the test will only be considered successful if the graphical technique enables the participants to find information more quickly, at the same level of accuracy (or more accurately in the same length of time).

As well as testing how long it takes participants to find information, the test should show how easily a participant can obtain a general overview of the reliability of the trials in the two information displays. This sense of considering the characteristics of studies is an important aspect of systematic reviewing. As noted in (Egger, Smith, & Altman 2001), the danger is that a traditional narrative review "ignores sample size, effect size and research design." They describe the systematic review as a way to consider many aspects

of the reliability of trials in an objective manner. The GOfER display should be able to support this key aspect of systematic reviewing, by giving an overview of the reliability of trials. This can be tested by asking participants about whether they have an overall sense of the reliability of the trials.

The GOfER test prototype was designed so that it could be included in an A4, black and white, printed report. The prototype for this test was made using the vector graphics software Illustrator, page layout software InDesign, and a spreadsheet for calculating the various sizes and positions of page elements. Most systematic reviewers in health do not have access to, or familiarity with, this software, and can not be expected to have the design training necessary to produce a clean, accessible layout for such a graphical presentation from scratch. Testing such a prototype is a way of assessing whether software to produce such graphical techniques would be valuable, without having to invest in designing sophisticated software to automatically produce them. It may, however, be more appropriate to design such graphics in exactly the way that this prototype was produced, making use of trained information designers.

5.1.3 Aims

The study detailed in this chapter has two main aims. They are related, but require very different information. Firstly, the study aims to:

1) Highlight the strengths and weaknesses of this graphical presentation method, in terms of accuracy and speed, in finding specific information on the trials included in the review.

This kind of question requires summative evaluation, and focusses on quantitative results to show the differences between information presentation methods.

Secondly, however, the study simultaneously aims to:

2) Capture suggestions for improving the graphic.

For this more formative question, a qualitative approach is required, and a more detailed level of information must be presented. It is not enough to know that an error has been made, but the *reasons* for this error must be understood, so that the presentations can be altered to increase their effectiveness in future. For this reason, the level of detail in the reporting of this study is quite high.

5.2 Methods

One-on-one interview is one of many techniques used to test information presentations, in areas such as interaction design usability testing (Preece, Rogers, & Sharp 2002). These techniques can be used to test paper-based presentations, with only a small amount of adaptation. One-on-one, task-based interviews were used to test the prototype version of the GOfER display, using data from a past review. The interview technique was based on that recommended by Preece, combined with ideas taken from the field of cognitive interviewing, a technique often used for validating and improving questionnaires (Drennan 2002;Willis 1999).

5.2.1 Population sampled

The graphic to be tested had a very specialist audience. Expert users, with a knowledge of the challenges of systematic reviewing and the HTA process in the UK, were required to test it.

One audience that would have had relevant experience would be committee members for the UK's NICE technology appraisal process. They have to understand systematic reviews of clinical effectiveness in a limited time. However, there were ethical constraints in asking them to participate in this test. They already give substantial amounts of time to the process, without any payment beyond travel expenses. Placing further burdens on their busy schedules just because they are prepared to give time already could be considered detrimental to the assessment process. There were also practical constraints. NICE appraisal committee members are spread out around the UK. Reaching, for example, each person on one committee, would require an unreasonably large travel budget, and a large amount of travel time, or some method of carrying out the interviews remotely.

One of the biggest challenges in remote interviewing, for a visual subject matter like using an information graphic, is controlling or recording what an interviewee sees or points to when answering questions (Birnbaum 1999). The graphic could be presented on screen, with a web-based display, either controlled by the interviewer to only show one page at a time, or navigable by the user but recording how long a participant spends on each screen of data. Mouse tracking could be used to see where a participant clicks or even

moves their mouse pointer. However, the graphic is designed to be included in printed reports. A typical screen might only display up to about 1200 by 1000 pixels, a far lower level of detail than a printed page. Zooming and panning would be required, and the experience would become much less like using the document as designed. It would be possible to send copies of the graphic in the post, and ask participants to use webcams or video conferencing systems and record the interviews in this way. It would be difficult to control for people opening the envelopes ahead of time and having an advantage in the tests. There would be less control over the quality of recordings and more variability in the setting in which the participants were performing the test. There might also be research bias in interviewing only people that have webcams, perhaps leading to more technologically able or well-equipped people being interviewed.

Instead of using NICE committee members, NICE staff, such as their technical advisors, interviewed in the study detailed in Chapter 3.2, could be asked to participate. They are grouped together, in two centres (Manchester and London), and were very willing to help in the previous study. However, five of the seven technical advisors were already interviewed for research that informed the design of the graphic. It was therefore thought preferable to search for different opinions to better evaluate the GOfER display.

The university-based research groups that provide technology assessment reports to NICE can equally be seen as possible users of the graphical technique. When producing one of these reports, part of their task is to communicate the results of a systematic review of clinical effectiveness. They also might be expected to have good understanding of the most important information to present, and a very good technical knowledge of the process of producing systematic reviews of clinical effectiveness. While their primary role is to produce such reports, rather than to make decisions based on them, they will still be very familiar with the experience of reading other reviews from any time that other systematic reviews have been identified as possible includes for their own reviews.

5.2.2 Sample size

Sless recommends around 6–10 users as an appropriate number to find the majority of faults in a design (Sless 2009). Preece also suggests of 5 to 12 users for interaction design testing (Preece, Rogers, & Sharp 2002). This is a fairly

common size for such studies.

Systematic reviewers and information scientists at two different TAR authoring centres were asked to participate in the tests (based in Sheffield and Exeter). A total of 9 systematic reviewers were interviewed (6 in Exeter, 3 in Sheffield).

5.2.3 Procedure

Willis gives two main variations of cognitive interviewing. The first is referred to as 'think aloud', in which the participant tries to talk through their thought processes as they use the document. The second is called 'probing', which has the interviewer asking participants to describe how they found answers to questions once they have found them.

The only available interviewer for this study was the designer of the graphical presentation, which has both advantages and disadvantages. The principle benefit is that the method gives direct feedback to the designer, who can shape the questions dynamically during the interviews, to receive the kind of response that might lead to a redesign. For example: "that's an interesting [i.e. incorrect] response... Which part of the diagram were you looking at exactly just then?" The two main disadvantages of the designer-interviewer are that, firstly, interviewees might be naturally and justifiably wary of being critical of someone's work to them in person, creating a bias in favour of the graphical presentation. Secondly, if a graphic like this was used for health decision-making, the designer of the graphical presentation is not likely to be on hand while reports are being read, to resolve confusion or misunderstanding of what is represented by different elements of a graphic. It might be difficult for the graphic's designer to refrain from assisting a participant asking for clarification, whereas the graphic technique should be understandable in and of itself. However, the first method of cognitive interviewing, the speak aloud test, can counteract biases like these, being more led by the interviewee (Willis 1999).

The speak aloud method is also appropriate due to the fact that the interviewer available had relatively little previous interview experience. This method is more demanding on the interviewee than the interviewer (Willis 1999).

One possible disadvantage of using the think aloud method in this case is that some people might be naturally more or less talkative than others. Since a time measure is used in the study, care had to be taken that this did not affect the

analysis unduly. This 'verbosity bias' was accounted for by using randomisation (see Chapter 5.2.4), and a proportional time measure in the analysis (see the beginning of Chapter 5.3.2)

5.2.4 Experimental design

Participants were shown both the new, graphical, GOfER display, and the original report section from which the data came (in numerical/textual, tabulated format). To account for possible inter-subject variability, each participant was shown both the existing presentation method from the published TAR, and the new graphical method. To account for carryover effects, where seeing one of the presentations first might influence the results for the other, a sequential, randomised experimental design was used.

Each participant was given twelve tasks, in the same order, asking them to find specific data relating to the trials used in review. For the first four questions, the participants were given either the graphical summary or the clinical effectiveness section of the report, selected randomly. After this initial set of questions, the first presentation method was removed, and replaced with the other for four more questions. Finally, four more questions were asked in which the participant could choose which of the two presentation methods to use (see Figure 5.2 – 1).

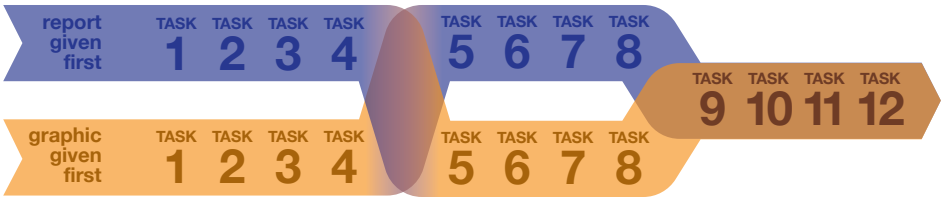

Figure 5.2 – 1
The experimental design. Blue represents report, and orange the graphic. The two groups are represented by the arrows, which cross after task 4 and merge after task 8

The randomisation of participants was conducted with a toss of an ordinary coin, until five participants had received one of the two presentation methods first. The other method was then given exclusively, to provide roughly equal numbers receiving each presentation method first. The randomisation of participants was important to help to control the effects of the 'verbosity bias', where some participants speak more about their thought processes as per the speak aloud method of cognitive interviewing. Randomisation is often relied

upon to control for characteristics of participants that have been identified as possible confounders, but can also help to control for other, unforeseen biases (Schulz 1997).

5.2.5 Outcomes

Interviews were conducted one-on-one, and video recorded. This allowed three things to be recorded at once. The time taken by participants to complete the tasks, their accuracy, and perhaps most importantly, qualitative data about their experiences in using the presentation methods.

The main focus of the study is to get qualitative information from participants on their experience with the graphics. This gives more context, explaining any effects on time and accuracy. It also addresses the other two aims of the study, to record suggested improvements, and to give an idea of how the systematic reviewers think such graphics should be produced and used, as well as who they might be useful for.

It was important to have an appropriate questioning strategy for the study. Certain questions are likely to be intrinsically easier with the GOfER display than the report, such as comparing two pieces of information that are presented separately in the report. Conversely, some questions will be easier to answer with a report, for example where the reviewer happens to have prepared a textual summary of a particular area.

Designing a questioning strategy for the study was particularly challenging. Some information in the main body of the report in data tables was not presented in the GOfER display, such as the degree of deafness of the participants, and the brand of the implants tested. A pilot study with one participant included some questions in the combination format part of the test (the last four questions) which asked for information that was not presented in the GOfER display. These questions were removed from the final test, as they failed to say anything relevant about the effectiveness of the graphical presentation. The decision was made to stick largely to simple search and find tasks with definite answers for the twelve timed questions. These would require the participants to find information that was available in both the report section, and also the graphic. Broader questions, which asked participants what they thought about the general reliability of the trials in the review, were also used, to assess the function of the graphic as a summary

should.

The questions asked are summarised below:

First presentation (randomised between report and graphic):

- Task 1: Which trial has the largest N? (ie. sample size)
- Task 2: How many of the trials were conducted in the UK?
- Task 3: Which trials used the Lexical Neighbourhood Test (LNT)?
- Task 4: Can you tell me about selection bias in the Peters et al. (2007) trial please?
- Do you have any overall impressions of how reliable the trials seem in general?

Other presentation:

- 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?
- Has your opinion of the overall reliability of the trials changed at all, since you've been working with this other presentation method?

Combination format presentation:

- Task 9: How many trials used a cross-sectional study design?
- Task 10: Which outcome measures were used by Nikolopoulos et al. in their 1999 trial?
- Task 11: Which trial (or trials) have the lowest mean age? (of those that report this). How old is this?
- Task 12: Which trial seems to have the highest quality, according to the checklist used in the report?
- Probe: Why did you select your chosen presentation format(s) to answer those questions?

General questions:

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

A full version of the script used in these interviews is available in Appendix D.

5.2.6 Analysis methods

A shorter time to find the answer to the questions given could indicate a more efficient information display, which was noted as potentially important in the information needs interviews in Chapter 3.2. This could not be a main focus of the study however. A minor issue is the 'chatty' bias already identified, which potentially affects this measure, even after randomisation, especially with interview techniques where only small numbers of people are interviewed. More importantly, however, the GOfER display, while providing a comprehensive overview of the characteristics of the trials, will never be able to present every relevant piece of information on the trials, even on the longest dimension of an A4 page. Therefore, searching through a report for information will necessarily take longer.

The time needed to find information in the displays must also be balanced by the participants' accuracy in the tasks, which can also be seen in the video recordings. A shorter time to find information, but with less accurate answers,

would be highly undesirable. The accuracy and strategies used to come to answers is also useful to qualify qualitative statements. If an interviewee was to say how much they liked a representation, for example, they would also need to be able to use it correctly to complete their tasks for this view to be taken into account.

Quantitative data was analysed using statistical tests, to compare the time and accuracy of the participants using different presentation methods for the same tasks.

The main focus of the study, however, was the qualitative data. To prepare this data for analysis, the video recordings of the interviews were gist transcribed, including making a record of actions such as turning pages and those gestures which added meaning to what the participants were saying. A sample transcript is available in Appendix E. These transcripts were analysed with a framework approach (Ritchie & Spencer 1994)*. The thematic framework was developed largely from participants responses, and can be seen in Table 5.2 – 1.

* also see Chapter 3.2.2 for more information on the framework approach

cat. no.	category	new code	subcategory	description
1	**potential functions**	**1.1**	complexity	where the graphic is noted to contain a large amount of data
		1.2	condensing	data is compressed into a smaller space using graphical techniques
		1.3	comparison	two or more things are presented in such a way that they can be compared. Horizontal and vertical comparison (between trials and within trials) Pattern-finding.
		1.4	limited time	where participants mention that it is quick (or quicker) to find information, or that time would be limited in the decision context
		1.5	selective focussing	bringing certain information to the fore to the exclusion of other information.
2	**organisation**	**2.1**	treatment comparisons	different comparisons (sections of review from report). Unfair advantage of collating all comparisons together in graphic.
		2.2	ordering	ordering of trials in presentations, ordering of outcome measures
		2.3	trial repetition	same studies repeated
		2.4	consistency	repeating headers, changing layout of tables
		2.5	appendices	what belongs in appendices?
3	**trial characteristics**	**3.1**	size	trial size
		3.2	design	trial designs
		3.3	outcome abbreviations	meaning of outcome measure abbreviations
		3.4	outcome categorisation	types of outcome (speech perception, production, etc)
		3.5	findings	lack of negative findings, lack of significant findings
		3.6	effect size	effect size of outcomes / standardised mean difference
		3.7	confidence	confidence intervals, sampling error, measuring uncertainty
		3.8	non-graphical	trial characteristics that can't be represented graphically, qualitative data
		3.9	non-reporting	trials not reporting data
4	**interpretation**	**4.1**	uncertainty	confusion - participants' uncertainty in meaning of representation. Asks interviewer for clarification. self doubt, double checking
		4.2	tie resolution	difficulty in judging difference in close numbers
		4.3	alignment	difficulties with judging which objects fall on which line or area
		4.4	overview	getting an overall sense / summary of data / report structure. "Eyeballing", scanning with eyes
		4.5	cognitive load	something takes a lot of mental effort. "painful" tasks.
		4.6	sizing	difficulty seeing small things, comments about relative sizes of page elements
		4.7	learning	novelty, unfamiliarity with graphic, learning effects, novelty affects preference?
5	**judgement**	**5.1**	reliability	reliability of evidence
		5.2	quality weighting	relative importance of quality checks
		5.3	narrative	importance of textual narrative / disadvantages of textual narrative
6	**strategies**	**6.1**	navigation	with indexing / contents / bibliography
		6.2	focussing aids	using finger or other aid to focus on data presentation, for example, tracing a line with a finger, or marking a place on the page
		6.3	counting aids	use of counting aids (paper and pen, fingers)
		6.4	memory	recalls something that helps. carryover effects
		6.5	peering	participant has to bend or hold up document to look closely
		6.6	bookmarking	using fingers or other objects as page markers
7	**preference**	**7.1**	graphic preference	preferring graphic, reasons for
		7.2	report preference	preferring report, reasons for
		7.3	no preference	similarity of report and graphic. Not sure which one would be preferred.
8	**design**	**8.1**	production	should graphic be designed on an individual basis, or produced with software (design vs automation)
		8.2	locking up	availability of data for further analysis (vs 'locking up data' in graphics), having to measure bar charts
		8.3	review variability	differing importance of data for different reviews / universally required information
		8.4	detail	level of detail, needing extra information
		8.5	media	display media (paper / screen)
		8.6	suggestions	design ideas and suggestions
9	**application**	**9.1**	context	lack of familiarity with context of decision, not having read whole report, outside participant's specialism
		9.2	decision	informing decision, using the systematic review, interpreting, decision question
		9.3	use	how graphical summary would be used in reports
		9.4	information responsibility	who decides which information is used for a COGS display?
		9.5	audience	who would use the graphics?

Table 5.2 – 1
Thematic framework

5.3 Quantitative results

All nine participants were able to complete the majority of the tasks that they were asked to perform. The interviews lasted from 36 to 78 minutes, with time differences largely due to some participants being more talkative than others in their speak aloud tasks.

5.3.1 Characteristics

Several questions about participants' background and experience were asked before the displays were shown.

5.3.1.1 HTA familiarity

After an initial introduction, each participant was asked how long they had been working with systematic reviewing, and in HTA. Experience ranged from only 6 months to over 10 years. Two of the participants (nos 4 and 6) were information scientists. Participant 8 was a public health systematic reviewer, and participant 9 worked mainly with qualitative research. The other five were involved for the majority of their working time with systematic reviews of clinical effectiveness for NICE appraisals, such as the one that provided the data

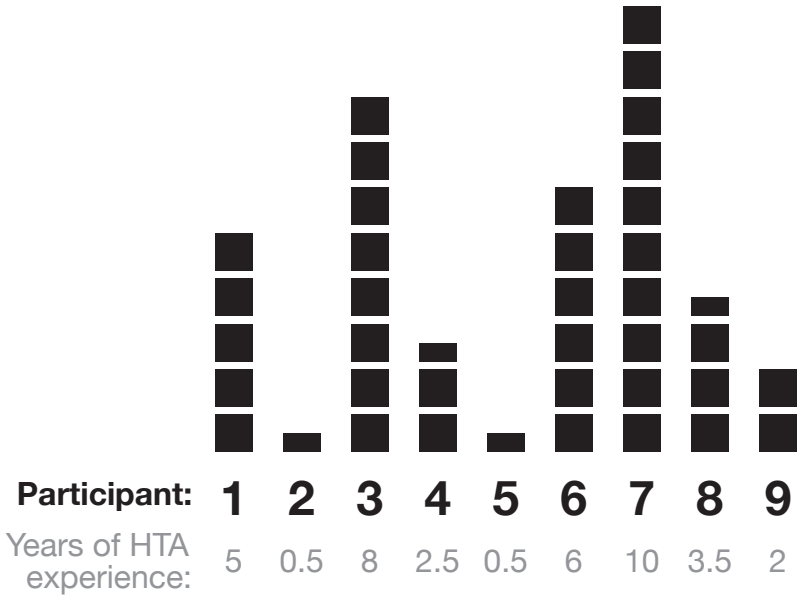

Figure 5.3 – 1
Experience of participants

used in the experiment. Figure 5.3 – 1 shows the results of the question.

With only 9 participants, it is difficult to make statistical inferences with confidence on the effect of experience on the quantitative data collected. The participants with 6 and 8 years' of experience in HTA were the only participants who were generally less accurate in their answers with the GOfER display than with the report. However they were also the two that took proportionally longest with the report, compared to how long they spent answering with the GOfER display (256% and 393% of the time taken with the GOfER display respectively across all 8 comparative tasks). The participant with 10 years of reviewing experience was more accurate using the GOfER display than the report, taking a longer time with the report also (215% of the time taken with the GOfER display).

5.3.1.2 Familiarity with cochlear implants report

None of the participants had read the report before. Participant 3 had shared an office with the person conducting the systematic review at the time, and so knew some of the challenges that they had faced. No participant was familiar with the structure of the report or the data presented therein.

5.3.1.3 Learning preferences questionnaire

Each participant completed a learning style questionnaire called VARK (Fleming 2010). This tool gives people a numerical score on four different learning styles, or preferences: visual, audio, read-write and kinesthetic. A person with a higher score in one of these categories is said to prefer that learning style. While statistical significance is not likely given the sample size, results are presented here to show whether this approach may be suitable for future quantitative studies. A summary of the results is presented in

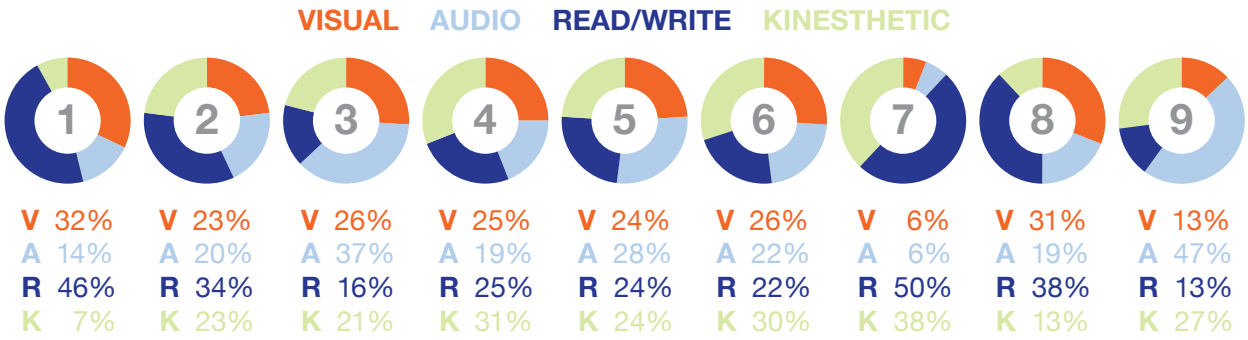

Figure 5.3 – 2
Learning preferences of participants

Figure 5.3 – 2.

In theory, people scoring highly on the 'visual' style might be expected to favour the graphic over the report. Similarly, 'read/write' learners might be expected to perform better with the report.

In practice, the two people that made more accurate decisions with the report than the graphic (participants 3 and 6) both had relatively high 'visual' scores, and low 'read/write' scores. Of two people that spent longer with the graphic than the report, participant 5 had medium scores on both 'visual' and 'read/write', and participant 9 had quite low scores with both.

One person (participant 7), mentioned that they had previously done a learning styles questionnaire, and remembered greatly preferring reading styles to visual styles. This tool seemed to confirm this, with this person scoring the highest of the nine participants on the 'read/write' style (50%) and the lowest on the 'visual' style (6%). However, this participant had much higher accuracy and task completion speed while using the GOfER display.

5.3.2 Task performance

Quantitative results such as time taken and accuracy of participants can give a brief overview of the relative effectiveness of the graphical presentation vs the original report. However, to allow for more detailed qualitative analysis a small sample size was used, and therefore statistical tests can not give conclusive evidence of the effectiveness of the technique. Having said that, the quantitative results are presented here, to show that larger, quantitative assessments would be a feasible way of showing effectiveness.

To control for the 'verbosity bias' mentioned in Chapter 5.2.3, the task completion times in the following sections are given as a proportional measure. The times taken in each task for a participant were summed, to provide a 'total task time' value for each person. The results are presented as the time taken to answer, as a percentage of each participant's total task time.

5.3.2.1 Overall results

Overall, those using the GOfER graphic were able to find information more quickly than those using the report. Those using the report in the first eight tasks took a mean of 12.8% of their total time for each task, while those using

GOfER used an average of 7.1% (two-sample $t(69) = 4.4$. $p < 0.001$.) These results are presented in Figure 5.3 – 3.

Despite spending less time finding their answers, accuracy was also higher for

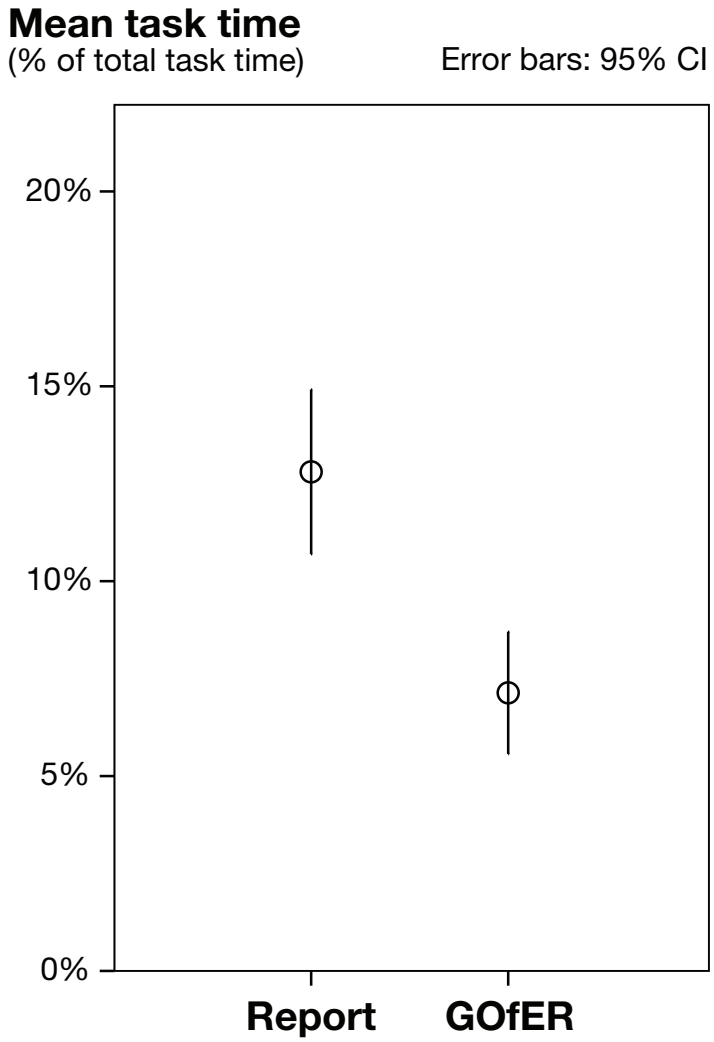

Figure 5.3 – 3
Average time taken by presentation used

those using GOfER. Those using the report answered 46.4% of their questions correctly. Those using GOfER answered 74.3% correctly ($c2\ (1, N = 63) = 5.12$, $p = 0.024$).

5.3.2.2 Results by task

Due to the small sample size, looking at each task individually is likely to be representative of the true population means. However, some general trends can still be observed. The quantitative results sections relating to the twelve tasks performed are detailed in Table 5.3 – 1. More detailed results of each task

Report

GOfER graphic

statistically significant (p<0.05)

Task	N (number of testers)	Task time (% of total) mean & 95% CI (0 … 30)	t	p	Accuracy (% of testers responding correctly)
1			3.16	**0.016**	50%
2			2.22	0.62	75%
3			0.93	0.384	0%
4			0.22	0.834	75%
5			3.81	**0.009**	75%
6			1.78	0.119	20%
7			0.51	0.623	0%
8			3.16	**0.016**	20%
9					
10					
11					
12					

Table 5.3 – 1
Quantitative results from tasks

are also presented in Appendix F.

The time measure for the first 8 tasks had two independent samples, which would allow statistical comparisons in a larger sample. Even with these results, three of the tasks (1, 5 and 8) would show significantly less time with the graphical presentation, reporting at $p<0.05$. However, as a Bonferroni correction would be required for carrying out 8 tests, this would not be considered relevant in this case. However, this trend is also observed across all other tasks, which indicates that the GOfER technique was largely successful at presenting much of the data, and participants seemed to be able to scan the graphical presentation more quickly than the tables of the report. It might be argued that this finding could be confounded in this instance by the larger amount of data in the report that needed to be searched through, particularly as most tasks required searching through four intervention subgroups, which were presented separately in the report, but together in the graphic. However, it may equally be argued that given the small amount of time available for decision-makers to read reports (see Chapter 3.2) that simpler presentations like this could provide benefits in a real context of use, as long as the detailed raw data was still available (perhaps in appendices).

The small sample size meant that accuracy could not meaningfully be tested for statistical significance. For a larger group, a chi-squared test might be used for this purpose. However, the trend is again for higher accuracy with the graphical presentation in most cases.

For the last four tasks, participants were allowed to choose which presentation to use. No participant chose to use the report for any task during this section of the study. While this may be explained in part by the fact that the GOfER display was novel, it still suggests that participants felt confident enough to rely solely on the graphical technique in this section. However, accuracy was extremely low for task 9, which asked participants to count the number of trials with a particular design. The example on the key was an unusual trial with a very large intervention arm, and a much smaller control group. Seven of the participants didn't realise that there were also four smaller trials included in the review that had the same study design, perhaps because they looked so different to the larger trial with very different-sized arms.

5.3.3 Randomisation effects

Several participants reported that they felt the GOfER display allowed them to get an overall sense, or overview of the data. This is reflected in the fact that participants who were given the GOfER display first tended to take significantly less time to carry out their tasks with the report in the second part of the test (those given GOfER first took a mean of 99.5% of their GOfER task time with the report, those given the report first took a mean of 268.5% of their GOfER task time with the report, two-sample $t(7) = 4.0$, $p = 0.005$).

This may just be an indication that the second set of tasks happened to be relatively quicker to perform with the report than the graphic. However, looking especially at the substantially quicker task times for GOfER in tasks 5 and 8, this does not seem to be the case. Also, given the participants' general comments about feeling that they had more of an overall sense of the evidence after seeing the GOfER display, it may be that the graphical presentation gave those who were given it first an overview of how the report was structured, enabling them to complete their tasks more quickly with the report in the second part of the interviews.

5.3.4 General questions

After all of the tasks were completed, participants were asked some general questions about their experiences with using the two displays.

5.3.6.1 Usefulness of GOfER for presenting data from cochlear implants review

All nine participants thought that the graphical summary provided by the GOfER display would have been a useful addition to the cochlear implants review. Four people mentioned that it gave a faster overview of the trials used, and could be used to initially familiarise oneself with the evidence before going into more detail as necessary. Three people mentioned that more information was available in the GOfER display in one place than in the tables of the report, enabling the viewer to get an overall sense of the key aspects of each trial.

Participant 2 mentioned that as they weren't familiar with the research area, and didn't know about the outcomes, it was difficult to see what the text was saying about the trials. They felt more able to concentrate on the specifics of the studies with the GOfER display. Participant 6 felt more confident about

making a judgement of the quality of the trials with the GOfER display, in a way which they weren't with the report.

Participant 9 was the only person that said anything negative about the GOfER display when asked this question. While they said that they liked the data display, they thought that "the key needed some more work".

5.3.6.2 Potential usefulness of graphical summary in other reviews

Eight of the participants were asked whether they thought a GOfER display could be used in other reviews. If so, they were then asked if different data would need to be presented, and how much this would change from review to review.

Every participant thought that the GOfER display could be used in other reviews. The amount of change necessary varied substantially, however.

Participants 2 and 5 thought that a GOfER display like this would need exactly the same information for every review. However, these were the two people with the shortest length of experience in HTA, at six months each. More experienced researchers tended to feel that the display would need from minor to substantial revision in response to the individual complexities of different reviews. Ages, locations and study designs were some of the data types that participants thought might be more or less important in different reviews.

One participant thought that genetic markers might become increasingly important in the future. One thought that the display could be adapted to a public health context, including data such as who delivered the intervention, and how much contact people had.

5.3.6.3 Task validation

Seven of the participants were asked if the tasks that they had been asked to do were representative of what they would do to understand a systematic review. This is a key question, as the results may be largely dependent on the task design.

All seven said that the tasks were representative of what they would do to understand a systematic review. However, participants 2 and 4 (0.5 and 2.5 years' HTA experience) mentioned that it was also important to look for patterns in the data on the trials.

Participants 1 and 3 (5 and 8 years' HTA experience) also said that getting an

overall sense of the data was useful. However, they thought that interpreting a systematic review like this was a case of looking at 'the answer', and then working back to see how much that answer should be believed. Participant 3 mentioned that they would commonly start by looking at which studies had reported significant outcomes, and working back from there, trying to account for the differences.

5.3.6.4 Missing Information / extraneous information

The written questions included a section asking whether participants thought that anything was missing from the reports, or if anything was unnecessary. These questions had generally already been addressed by the point at which it was planned to ask them, particularly while talking about how GOfER might be adapted to different reviews.

The interviewer only felt the need to ask whether anything was missing from the GOfER display as a separate question four times, and two of the people asked felt that they didn't know the decision context well enough to answer. The other two didn't have anything to add to what had already been said.

The interviewer only asked once whether there was any extraneous information, to participant 1. They responded that: "the outcome measures presented are bewildering". They couldn't see why these measures had not been brought together more in the review, perhaps using a standardised mean difference metric.

5.3.6.5 Potential Usefulness of Interactive Version

Eight of the participants were asked if an interactive version of the GOfER display would be useful, that could be sorted by the various different kinds of data in the graphic. While this was universally thought to be potentially useful, several people thought that they would have to see this proposed interactive tool in action to make an informed decision.

Specific examples of how a sortable, interactive tool might be used included:

- looking at the spread of ages within trials of a certain design (participant 1).
- looking at how many UK studies there were regardless of comparator (participant 3).
- sorting by study design (participant 5).
- sorting by study size (participant 6).
- sorting by each of the quality checks, to detect patterns of other characteristics that might be affected in trials that were in some way found to

be flawed (participant 8).

Three people (participants 4, 6 and 8) mentioned that an interactive tool like this would only be of use to the analyst. They thought that a fixed snapshot would be needed for a report, to present to decision-makers.

5.4 Qualitative results

Qualitative results are presented in sections corresponding to the categories of the thematic framework developed during the analysis of the interview transcripts (see Table 5.2 – 1). Each theme has several sub-themes, results of which are summarised in tables throughout this section.

5.4.1 Thematic category 1: Potential functions

The 'potential functions' theme was the only one that was established before beginning the analysis of the interview transcripts. The five sub-themes used here correspond to the potential functions of information graphics established in Chapter 1.3 – potential functions of information graphics in HTA.

The GOfER display seems to have incorporated all five potential functions to some degree. As the display was able to condense information into a small space, comparisons both between and within trials could be more quickly and easily made.

A more detailed summary of each sub-theme for the 'potential functions' category is included in Table 5.4 – 1.

Sub-theme	Notes
Complexity	Mentioned most often while looking through the report or considering the overall reliability of all the trials. Both report and graphic were described as complex, or containing a large amount of data, particularly in relation to the large number of outcome measures used by the studies reviewed.
Condensing	After being given the second presentation, most of the participants mentioned that they were impressed by how much more condensed the information was in the GOfER display.
Comparison	The GOfER display enabled people to more easily compare the magnitude of certain variables, such as the ages of participants and the size of the trials, to those in other trials in the review. This sometimes enabled them to make connections between the data that were not made in the report, due to the information being spread across multiple tables.
Limited time	All nine of the participants mentioned that they could find information more quickly in the GOfER display during the general discussion after the tasks were all complete. Two of them mentioned that they would use it to find information quickly, and only go into the tables and text of the report to find detail if necessary.
Selective focusing	This theme was mentioned by three people when asked about the overall reliability of the trials in the review. They thought that the GOfER display allowed them to pick out information of interest.

Table 5.4 – 1
Sub-themes for the 'potential functions' category

5.4.2 Thematic category 2: Organisation

Several participants talked about how the report and graphic were ordered and layed out. The splitting of the report into treatment comparison sections seems to have caused difficulties in some tasks. This may indicate that the questions asked were biassed in favour of the GOfER display in these tasks.

The way that both GOfER and the report were split into treatment comparison sections may also have led to errors in double-counting trials that appeared in more than one section. The lack of consistency in the 'summary of results' tables may explain why so many errors were made by those using the report for task 8.

A more detailed summary of each sub-theme for the 'organisation' category is included in Table 5.4 – 2.

Sub-theme	Notes
Treatment comparisons	Both the report and GOfER graphic were separated into five different sections for different treatment comparisons (such as acoustic hearing aids vs cochlear implants, or single implants vs two implants). This meant that, in the report, the data tables needed to complete the tasks were separated by large quantities of text and tabulated data that did not appear in the graphic. Participants therefore struggled when using the report to complete tasks 1, 2, 3, 5, 6 and 7, which required data from each trial in the review across all treatment comparisons. This could be seen as giving an unfair advantage to the GOfER display in much of the evaluation. Several participants also mentioned that it affected their decision to use the GOfER display in tasks 9, 11 and 12.
Ordering	Two people mentioned that they found the alphabetical ordering of outcome measures in the GOfER display useful. Two people asked about the ordering of the trials, one with each display. Both were presented in reverse publication order, which took a short time for them to work out.
Trial repetition	Data on two of the trials were given in two different treatment comparison sections. This was often noticed in questions beginning "which trials…", as the participants were reporting author names. However, in questions beginning "how many trials", participants did not tend to look at author names, and they were frequently double counted in both displays.
Consistency	During task 8, participants using the report struggled to identify significant outcome measures in the 'summary of results' tables in the report (see Figure 5.4 – 1). These tables were presented over multiple pages, but the second row of headers was not repeated as it should have been. This caused confusion, especially as the column widths varied between pages (see Figure 5.4 – 2).
Appendices	Two people (one using each display) noted that they would expect to find more information on quality checks in appendices. One participant assumed that the characteristics table in the report was part of an appendix, which might suggest that such large volumes of raw numerical data is more suitable for appendices than the main body of the report.

Table 5.4 – 2
Sub-themes for the 'organisation' category

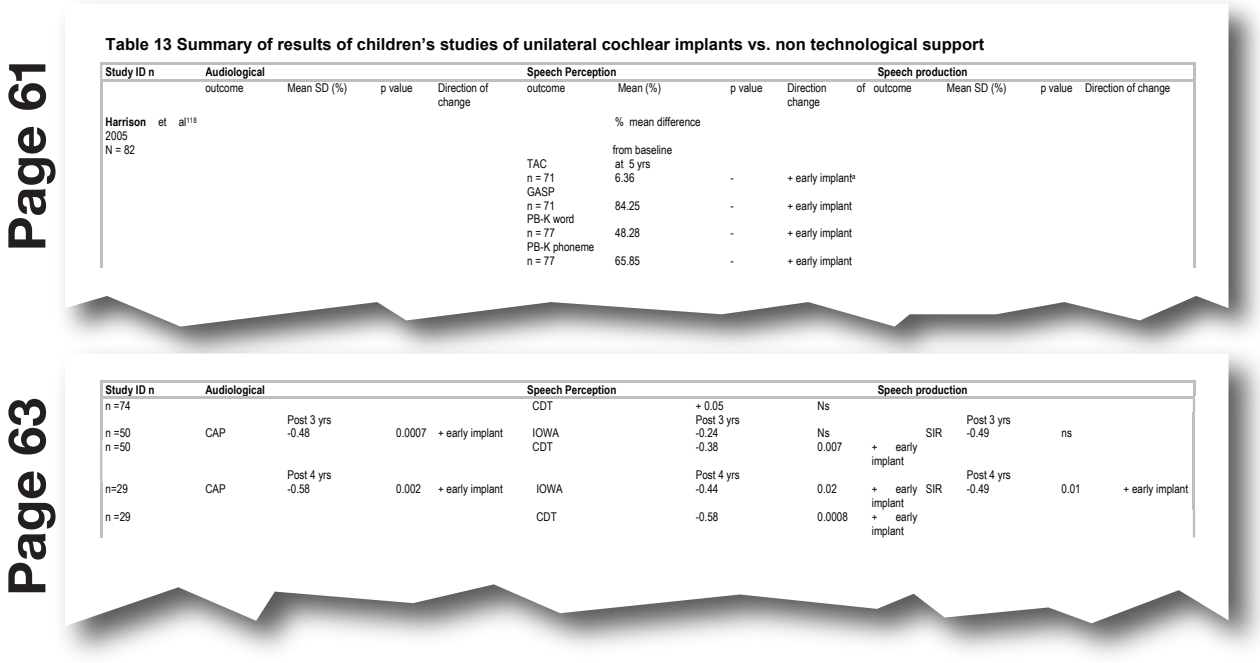

Figure 5.4 – 1
The top of two pages in the report, showing parts of the same data table

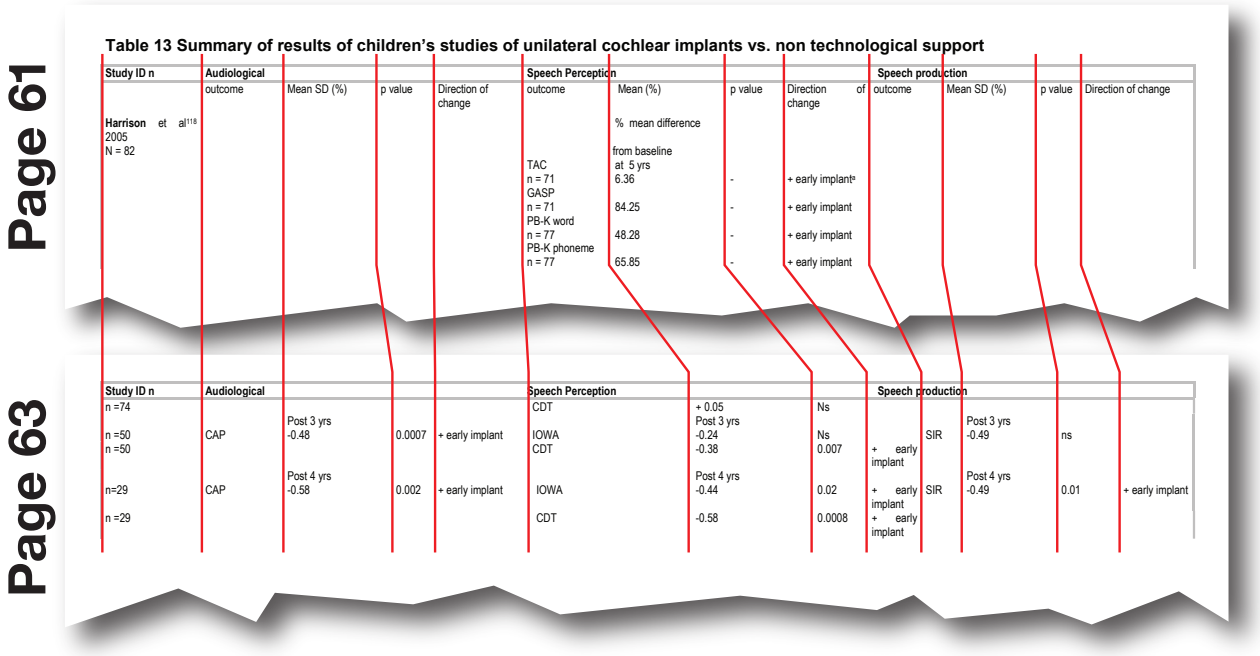

Figure 5.4 – 2
The same two pages, showing the misalignment of columns

5.4.3 Thematic category 3: Trial characteristics

The characteristics of the individual trials in the review were presented in the report in four different kinds of data tables. These included far more data than the GOfER display. In the tasks used in the experiment, the volume of data presented often made it a more laborious task for participants to find the information that they needed in the report for their tasks. If decision-makers have a limited time to digest the reports, this may mean that a quick visual summary like GOfER is able to inform them better than a very detailed set of data tables. However, the full detail should still be made available in appendices in this case.

A more detailed summary of each sub-theme for the 'characteristics' category is included in Table 5.4 – 3. This shows how the authors of the report and the designer of GOfER had chosen to represent various different aspects of the trials, and how the participants were able to understand the patterns and ranges of this data in the review.

Sub-theme	Representation in report	Representation in GOfER	Notes
Trial size (ie. number of participants in study)	Numerical value presented in each of the following tables: 'Summary of study characteristics' (see Figure 5.1 – 2). 'Summary of study results' (see Figure 5.4 – 1). 'Visual summary results table' (see Figure 5.4 – 3).	Shown with the width of the lines in the Sankey diagrams, supported by numerical value shown beneath.	The size of trials was considered a key factor in determining the reliability of trials by almost all participants. The size of the trials became quite obvious to participants using the GOfER display, and four people using GOfER mentioned trial size when considering the reliability of the trials after the first four tasks. In contrast, only one of the people using the report said they considered trial size. Participant 1 mentioned that they would be cautious about believing the results of a large trial with a pre-post design, which shows the importance of the way that GOfER presents key data in one place so that it can be considered in conjunction with other factors.
Study design	Short textual description in 'Summary of study characteristics' tables. Also shown with a two- to five-letter abbreviation in the 'Visual summary results tables'.	The shape of the Sankey diagrams indicated, for example, whether the people in the trials were randomized to receive cochlear implants, or if the same people were tested before and after being given the implants.	This was also used as an indicator of the trials' reliability by three of the people using GOfER, but not any of those using the report. However, all participants seemed to have a poor understanding of how the different study designs were represented in GOfER (see Appendix F – 5.2). Only if the study designs could be successfully communicated, perhaps with better description in the key, would this be an advantage for the GOfER display.

Table 5.4 – 3
Sub-themes for the 'characteristics' category
(continued on next page)

Sub-theme	Representation in report	Representation in GOfER	Notes
Outcome abbreviations	Outcome measures were presented in 'Summary of study results' and 'Visual summary results' tables. Both used abbreviated forms of the outcome measure names. The abbreviations were explained in a separate table towards the beginning of the report.	Outcome measures were also abbreviated in the GOfER display. To fit the information into a small space, some outcome measures were abbreviated more heavily than in the report. A similar table to interpret these abbreviations would be necessary if GOfER were used in an HTA report.	Several participants felt the need to look up the meaning of the outcome abbreviations, although none of the tasks specifically asked for this level of detail. Both report and graphic would need an explanatory table. It might be helpful to provide a page number for reference when such abbreviations are used.
Statistical significance	For each study in the review, the 'Summary of study results' tables presented a numerical 'p-value' (measuring statistical significance of the result). The 'Visual summary results' tables also presented the statistical significance and direction of change using a matrix with shaded squares.	The GOfER display showed the statistical significance using the circles on the outcome measures grid. Black circles indicated significant results, and white non-significant.	The summary of results table in the report caused confusion in the participants that used it, none of which were able to find the correct values for their tasks. Those using the visual summary tables in the report or GOfER were more able to pick out significant results. The visual summary tables are themselves an unusual thing to find in reports, however, and should be encouraged for complex reviews such as this.
Direction of change	The 'Summary of study results' tables showed the 'direction of change' (showing which treatment the test favoured), in a separate column, with a + or – symbol.	The direction of change was also shown in GOfER using the circles on the outcome measures grid. Diagonal lines across these circles would have indicated a change in favor of the control in each comparison if this had appeared in the data set.	Every trial in this review reported an effect in favor of cochlear implants, therefore the display was not tested in its ability to show this data. In other reviews, this may be a more important consideration.
Effect size	Data from each outcome measure used was presented in 'Summary of study results' tables, in numerical form, although standardised scores were not used so that the large number of different outcome measures used were not easily comparable.	Not shown.	Many reviews use statistical techniques to compare the magnitude of an effect. This can even be shown across a range of different outcome measures (using meta-analysis techniques). This was not possible in the cochlear implants review, due to the variety in study designs used. Two of the participants thought that it might be useful to display effect size in a GOfER display in other reviews.
Confidence	The 'summary of study results' tables presented the standard deviation or 95% confidence interval of the mean results for each outcome measure, in numerical form. This data was rarely reported by the trials, however.	Not shown.	The confidence interval is another important statistical tool, used to show how certain a result is. One participant mentioned that they would like to have seen the confidence intervals of the outcome measures. This might require substantial adaptation of the display, but could be valuable in other reviews.

Table 5.4 – 3
Sub-themes for the 'characteristics' category
(continued from last page)

Table 39 Visual summary results: unilateral cochlear implants vs. non technological support – adults

Speech perception outcomes

Study design (follow up, months)	Author year	N	Speech perception outcome tests – summary results for cochlear implant condition	cochlear implant condition											
				BKB	AVGN	AB monosyllable words	CUNY words	CUNY sentences	CUNY sentences in noise	HINT sentences	MAC vowels	MAC consonants	CID sentences	NU-6 monosyllabic word test	Everyday telephone sentences
PP (P) (9)	UK Cochlear Implant Study Group 2004	316		■	■										
PP (R) (>18)	Mawman et al, 2004	214		▒		▒									
PP (P) (3)	Parkinson et al, 2002	216					■	■	■	■					
PP (P) (24)	Kessler et al, 1997	238						▒			▒	▒	▒	▒	▒

Quality of life outcomes

Study design (follow up, months)	Author year	N	cochlear implant condition		
			HUII3	GHSI	GBI
PP (P) (9)	UK Cochlear Implant Study Group, 2004	316	■	▒	▒

Key: ■ = positive significant outcome p <0.05; ▒ = positive outcome (not significant or no significance reported p <0.05); //// = negative outcome (not significant or no significance reported p <0.05); //// = negative significant outcome p <0.05; NRC (P) = non randomized controlled trial (prospective); NRC (R) = non randomized controlled trial (retrospective); PP (P) = pre/ post (prospective), PP (R) = pre/ post (retrospective), XSOC= cross sectional own control

Figure 5.4 – 3

Example of 'Visual summary results' table from report

5.4.4 Thematic category 4: Interpretation

GOfER seems to be able to give an overview of trial characteristics, particularly in the quality grids. While the quality grids did not affect time or accuracy in tasks 4 or 7, the qualitative analysis shows that participants found it easier to take in the visual grid as a whole than the tables in the report.

However, participants were uncertain about the maps, quality grids and study designs in GOfER, at least to start with. While the maps and quality grids were generally understood despite the participants' uncertainties, the study designs were more problematic (as shown in Chapter 5.3.2.2). As in many scientific diagrams, some learning overhead is likely to be necessary before such diagrams are widely understood.

A more detailed summary of each sub-theme for the 'interpretation' category is included in Table 5.4 – 4.

Sub-theme	Triggers for categorisation	Notes
Uncertainty	This sub-theme was marked where participants stated that they were uncertain about their response, or where they asked the interviewer what something meant (these questions were not answered at the time, but noted down).	Participants tended to display more uncertainty with the GOfER display, as might be expected with an unfamiliar graphic (see Figure 5.4 – 4 on the page after next). GOfER uncertainty was largely displayed in understanding the representation of data, and the meaning of graphical elements. Report uncertainties tended to be around understanding what the values in tables were. GOfER uncertainty particularly related to the maps, the quality grids and the study designs. Uncertainties about maps and quality grids did not lead to task errors, whereas the uncertainties about study design did, and were therefore more serious. See Appendix F – 5.2 for more information.
Tie resolution	Difficulties in judging the larger or smaller of two very close values were categorized under this sub-theme.	This theme appeared while participants were using GOfER in task 5 (judging longest follow-up) and task 11 (judging the lowest mean age). Two people felt that it would be important for decision-makers to know the exact numerical values for these quantities, but one other person thought that the graphic emphasizing how close the values were was useful.

Table 5.4 – 4
Sub-themes for the 'interpretation' category
(continued on next page)

Sub-theme	Triggers for categorisation	Notes
Alignment	Whenever a participant had difficulty seeing which of a series of objects fell on a particular line or space, this sub-theme was marked.	One participant misjudged which page number belonged to a heading in the report contents. One other person mistook which line a circle fell on in the GOfER outcome measure grid.
Overview	This theme represents participants trying to get an overall sense or summary of data, or structure. The term 'eyeballing' was often used, to describe the process of looking quickly for an overview of a particular set of data.	Three people mentioned that it was easy to see an overview of the quality of the trials from the quality grids. Participant 3 tried to get an overview of the quality of trials from the quality tables in the report, stating that there were "quite a lot of yeses" in the quality table for the first comparison, and "more nos" in the third comparison. In fact, the first comparison table had an almost identical proportion of both yeses and nos to the third comparison. Most people thought that the GOfER display gave an overview of the trials, as might be expected since this is a fundamental aim of the graphic.
Cognitive load	This theme showed where great mental effort was needed. Participants frequently displayed they were experiencing cognitive load by saying that doing a task was "painful".	Most often displayed while using the report, particularly when trying to find the right table for a task, or go through a large amount of numerical data to find an answer.
Sizing	This theme represents people saying that something was too small, or other comments about the relative size of page elements.	Two people thought the GOfER maps were too small, one person thought the quality grid should be bigger, and one person thought the key should be bigger.

Table 5.4 – 4
Sub-themes for the 'interpretation' category
(continued from last page)

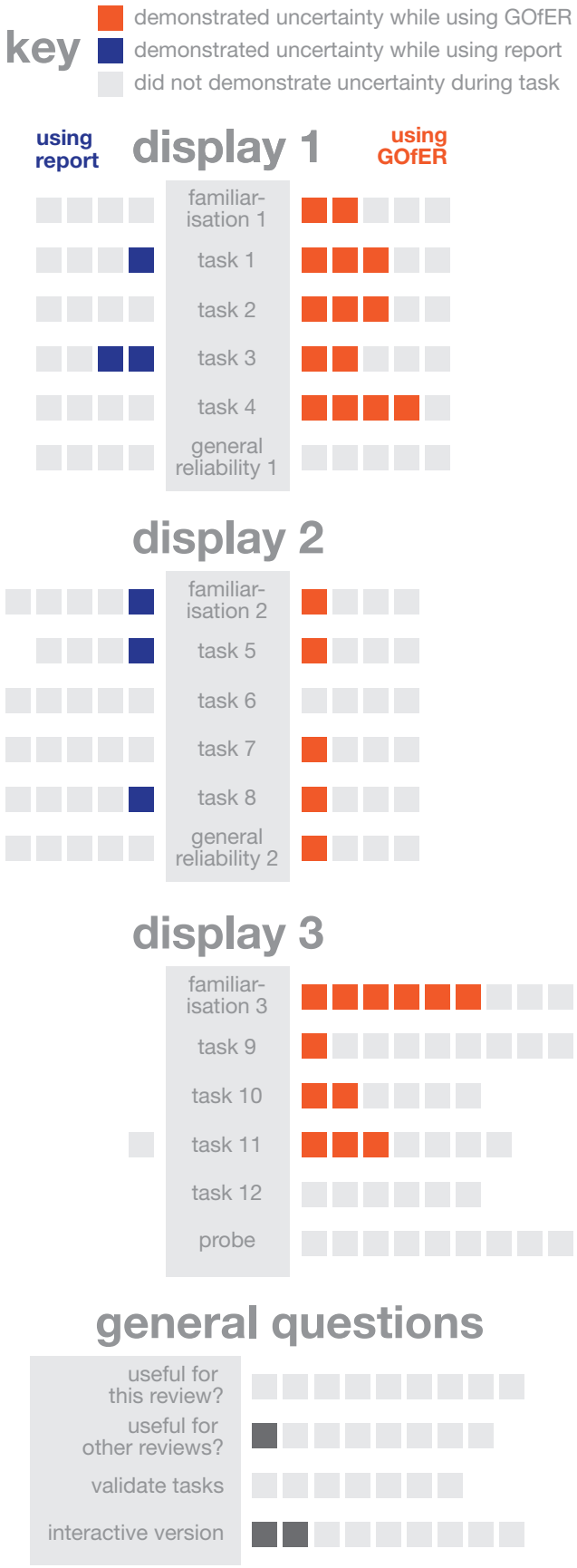

Figure 5.4 – 4
People demonstrating uncertainty during tasks

5.4.5 Thematic category 5: Judgement

Making judgements about the quality and reliability of trials is a complex and challenging task. Participants mentioned many different characteristics that can affect the reliability of trials, many of which must be considered together. For example, different quality checks are more or less important for trials with different study designs, and these must be considered alongside other factors such as the size, location, and follow-up periods of studies. The GOfER display can give an overview of the most important characteristics in one place, which could be supported by textual narrative, and raw data in appendices.

A more detailed summary of each sub-theme for the 'judgement' category is included in Table 5.4 – 5.

Sub-theme	Notes
Reliability	Many different aspects of the studies were considered important to judge the reliability of a trial: • Quality checklist for study • Size of study • Length of Follow-up • Statistical significance of outcome measures • Location of study • Date of study It was commonly mentioned that multiple factors such as these should be considered together to determine the reliability of trials. One participant, for example, said that they would consider the design of a study to know which of the quality checks were most important. Another thought it would be important to see if positive findings were only found in studies with weaker quality checks.
Quality weighting	Several participants noted that not all quality checks in the quality assessment tool used would be equally important. A common strategy used with GOfER was to count the number of black squares in the quality grid, which would not have accounted for this. However, some participants felt they did not have enough experience to know which quality checks were more important. Participant 1, using GOfER, was able to look at the two squares that they had learned would be important in pre/post studies.
Narrative	While there were differences of opinion on the importance of narrative, the general consensus was that the GOfER display could be used alongside textual description to explain the overall results of the trials, in the same way that data tables are currently. Those with a longer experience in HTA seemed to be more comfortable with the idea of reading a large volume of text, even though this may not be possible for time-limited decision-makers.

Table 5.4 – 5
Sub-themes for the 'judgement' category

5.4.6 Thematic category 6: Observation

Some of the participants' actions revealed how they were using the presentations. The large report section document required extensive use of navigational aids, such as contents and indexing. Participants also marked pages to return to with fingers and other 'bookmarks'. The GOfER outcome measures display required the use of focussing aids, such as fingers or pens, to trace from the results to authors' names and the outcome abbreviations at the top of the page. Also, several participants had to look closely at small page elements, suggesting that some of them were approaching the limits of clear perception.

A more detailed summary of each sub-theme for the 'observation' category is included in Table 5.4 – 6.

Sub-theme	Notes
Navigation	Those people using the report needed to use indexing strategies such as contents, indexes and bibliographies as paths to information. Three people looked for a list of tables, which might have helped them to find the information they needed. Three people voiced dislike for the contents page, but mostly without qualifying this with a reason. Better use of type size, font weight and spacing could have helped establish hierarchy in the contents.
Focusing aids	All participants used a finger or pen to mark places on a page, often using them to trace from one place to another. Those using the report used aids to trace from headings in the contents to the page numbers, which were quite widely separated. Those using GOfER often traced from the circles on the large outcome measures grid to either the outcome abbreviations at the top of the page or the author names on the left. In either the report contents or the GOfER outcome measures display, a small expenditure of space might be worthwhile to separate the lines into sections.
Memorising	Some participants remembered something from earlier in the interviews that helped later on in the tests. This largely tended to concern remembering something from the familiarisation with report or GOfER, suggesting that carryover effects between the presentation methods were minimal in this study.
Peering	Participants occasionally had to look closely at a page, particularly while using GOfER. People peered at: • Location maps (2 people) • Outcome measure titles (2 people) • Quality grids (3 people) • Length of follow-up (1 person) However, only a few people specifically mentioned that any of these elements should be bigger (See table 5.4 – 5: Sizing). This may indicate that elements of GOfER are very close to, but not quite at the limits of clear perception.
Bookmarking	Most people using the report used some kind of bookmark, either sticky tabs provided by the interviewer, or fingers as temporary bookmarks. As might be expected, this was not so necessary with GOfER as the document was so much shorter.

Table 5.4 – 6
Sub-themes for the 'observation' category

5.4.7 Thematic category 7: Preference

Participants generally seemed to prefer GOfER to the data tables in the report. Figure 5.4 – 5 shows the number of times the participants voiced a preference for GOfER or the report's numerical displays. Participants 2, 3, 4, 5, 6 and 7 gave a general preference for GOfER, often while being asked about the general reliability of their second presentation (whether given GOfER first or second).

Specific aspects of GOfER that participants preferred were: (participant numbers are in brackets)

- The outcomes display (5, 7)
- The quality grid (3, 8)
- Follow up display (7)
- Being able to compare characteristics, quality and outcomes together (3)
- Being able to compare characteristics between studies (2)
- The study design symbols (2, 9, 5)
- Age display (3)

It is interesting to see that three people stated they preferred the graphical presentation of the studies' experimental design using Sankey-style flow diagrams, even though all participants made mistakes interpreting this representation. It suggests that with a better key, this might still be a useful presentation method for study designs.

The one person that stated a preference for a numerical display at any time was referring to having a spreadsheet with

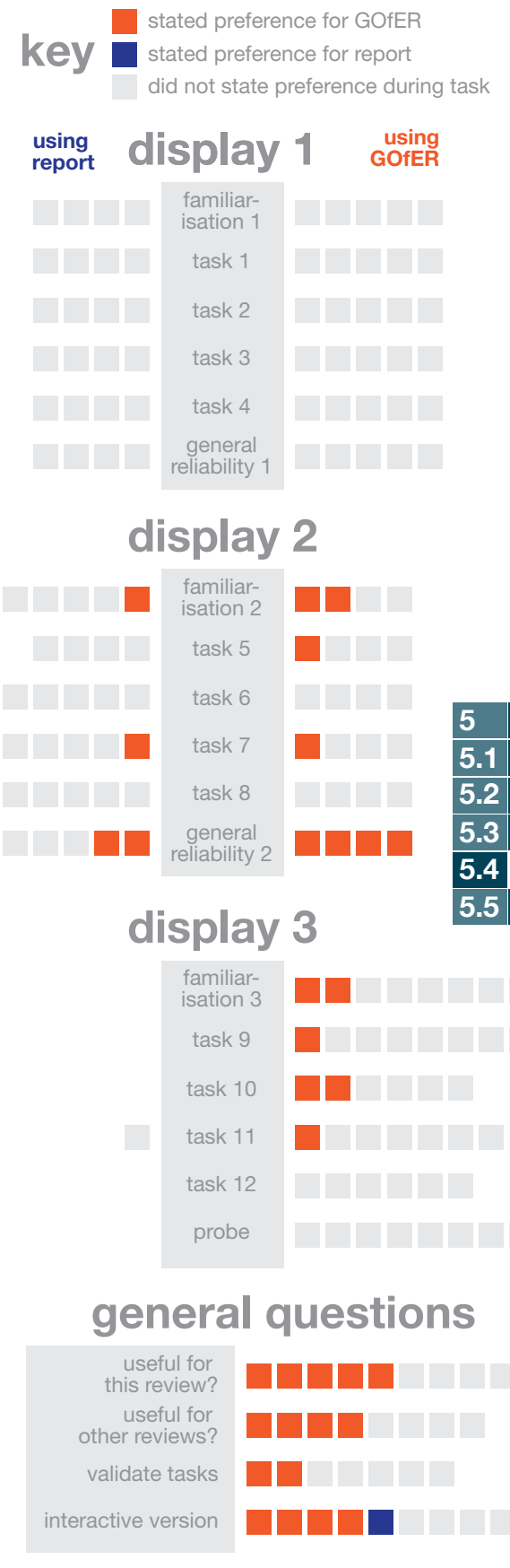

Figure 5.4 – 5
People stating a preference during tasks

numbers as opposed to a sortable graphical display in the last question of the interview.

5.4.8 Thematic category 8: Design

There were very few specific references to how a graphic should be produced, or what media should be used. Participants generally felt that the GOfER display gave an overview of the information, but more detail would be needed as well. During the interviews, both participants and interviewer offered several suggestions for improving the graphic in this and other reviews. A summary of these suggestions, along with a more detailed summary of each sub-theme for the 'design' category, is included in Table 5.4 – 7.

Sub-theme	Notes
Production	One participant assumed that the GOfER display would be produced with an automated tool, with "preset fields that you can click on" to assign in which area of the world the study was.
Detail	The general opinion of the participants was that more detail might be needed than was presented in GOfER. However, they seemed to think that GOfER was useful to provide an overview of the evidence before looking at the detail. Task 11 showed this view particularly clearly, with 4 people choosing to use the GOfER display for the task, and then saying that if they wanted to know exact number, they would look it up in the tables of the report.
Media	Three participants mentioned that a version presented on a computer could link to more detailed presentation of data on request, perhaps by clicking on an element to see a data table. However, a limited version of this functionality could be provided in a static, printed version by providing table or page references within the GOfER display. Two people thought that an interactive version would only be of use to a reviewer, and a fixed, static version would always be needed to include in reports.

Table 5.4 – 7
Sub-themes for the 'design' category

5.4.9 Thematic category 9: Application

Participants envisaged the GOfER display as a way of summarising key points about the trials, alongside the textual narrative of the reports. They thought that decision-makers could use them to quickly see what kind of evidence the economic modelling in a NICE technology assessment report was based on.

A more detailed summary of each sub-theme for the 'application' category is included in Table 5.4 – 8.

Sub-theme	Notes
Context	Several participants noted that they lacked knowledge of the specific intervention under consideration, which impacted on their ability to judge how reliable the trials were. Three of the five people given the report second thought that they would have to have read the whole report to be able to judge trial reliability.
Use	Participants generally thought that the GOfER display would be included in the report, surrounded with textual information, as the data tables are currently, rather than how it was presented in the tests, as a kind of 'visual executive summary'.
Decision	Participants 1 and 3 thought that decision-makers would be "starting with the answer and working backwards". They would use the systematic review to judge how much they believed the outputs of the economic analyses in the report.
Information responsibility	While the GOfER display condenses a large amount of information onto a single page, there are limits to what it can show. One participant mentioned that it might either be the responsibility of the reviewer to decide what was presented in a graphical summary like this, or that the decision-making body might have some input into what was shown.
Audience	While the decision-maker was the most commonly mentioned audience for the GOfER display, two people thought that it might aid the reviewer, while writing the report or when returning to it at a later date to remind themselves of the trials. One person thought that the modelers doing economic analysis might find it useful as well.

Table 5.4 – 8
Sub-themes for the 'application' category

5.5 Conclusions

5.5.1 Statement of findings

Overall, the GOfER display seemed to reduce the time taken by participants to find information, particularly in tasks 1, 2, 5 and 8. these tasks involved finding large studies, finding UK studies, looking at length of follow-up, and finding outcome measures used. GOfER also seemed to increase accuracy in task 3, another task using the outcome measure grid. However, these results should be treated with caution, even when presented as statistically significant, as the sample size is very small. Also, the larger volume of information in the report, particularly textual narrative, put it at a disadvantage for most of the tasks used (except task 8). This is not surprising, as GOfER is a summary of the report, but in this case, it might have been more relevant to focus on individual treatment comparisons, as in task 8, rather than expecting participants to find an overall answer across all subgroups.

However, the largest advantage shown for GOfER in the randomised section of the interviews was in the ability of participants to get a quick overview of the reliability of the trials. Participants were able to consider a much wider range of study characteristics simultaneously with the GOfER display than with the report in the time available to them. As the interviews in chapter 3.2 showed that decision-makers are likely to have limited time to read reports before the meetings at which they must make decisions based on them, this is an important advantage of the GOfER display over the tables used in the report.

The most serious problem with GOfER highlighted by the quantitative results was the general inaccurate interpretation of the visual display of the experimental designs used by the studies in the review. This confusion may partly stem from inadequacies in the key, but also a could be a result of a more fundamental flaw introduced in the design of this graphic by lack of understanding on the part of the designer about the differences between the study designs used. For example, the cross-sectional design seemed to one participant to represent people being studied over time, when this was not, in fact, the case. This is perhaps inevitable for collaborations between content experts and expert communicators, and is exactly the kind of flaw that evaluations such as these are designed to highlight. Suggestions for improving understanding in this area are made in Chapter 5.5.5.

The qualitative results suggest that participants favoured the graphic more than its modest performance gains in the specific tasks would explain (see Chapter 5.4.7 – Thematic category 7: Preference). For example, one of the things that participants liked was the GOfER quality grid, even though it did not seem to improve quantitative performance greatly. However, the qualitative results suggest that the quality grids helped them to get an overall sense of the quality of the trials more quickly (see Table 5.4 – 4: 'overview'). The tasks that asked them to find specific information in the quality grids were perhaps not very well suited to the way that the GOfER display was designed.

Several people noted that the graphic was novel to them (see Table 5.4 – 4: 'learning'). They said that it was a refreshing change to textual and numerical information presentation, which may have had a positive effect on their preference.

It is interesting to note that some people seem to feel more confident about a numerical value, feeling that they have to 'transform' a visual representation into a number to understand it. However, numbers themselves can mislead just as much – or even more so – than a graphical representation. This is a fact that shopkeepers have long exploited, pricing goods at £2.99 rather than £3.00. A graphical representation of these figures would truly show how close these values were, but the effect of beginning the numerical representation with a 2 rather than a 3 has a disproportional effect on the hapless shopper. This suggests that both the graphical and numerical presentation must be transformed into another form in the brain to be understood. This is in line with Resnikoff's suggestion that the visual processing system can be used to 'pre-process' data, enabling its understanding more quickly than having to interpret abstract words and numerical symbols (Resnikoff 1989). These interviews suggest, however, that there is still a general misapprehension that quantities must be presented numerically, and that graphical presentations must be somehow transformed back into numbers in the brain. Numerical values may indeed be needed for further analysis by other reviewers, and so all numerical values should be presented. However, to present data to time-pressured policy decision-makers, this study suggests that it would be more appropriate for the raw data to appear in other tables or appendices. A graphical presentation supported by text and key outputs is all that is needed for them to understand as much about the data as possible in the time available, and to easily identify *relative* similarities and differences between studies.

In terms of production and use, this graphic was designed for a particular

problem with proliferation of outcome measures in the cochlear implants review. The people that have tested it suggested that other reviews might need a larger map, or something else completely different to be displayed. The large number of outcome measures in the Cochlear Implants review may not be a problem in other reviews, leaving space for other information to be presented within the display. In areas like HTA, where each review has individual complexities, it will always be preferable to have a trained information designer to present the results of the review in the most appropriate manner. However, the necessary skills, experience and resources for this work may not often, if ever, be available.

While the GOfER display was not initially conceived as an automated tool that could be used for different reviews, the idea might have some merit. This could take the form of a spreadsheet–like program with preset columns and rows for different trial characteristics. The reviewer (perhaps with input from the decision-maker commissioning the review) could select certain important information, an output size (likely to be A4), and the software would produce something similar to the GOfER display from the raw data. The size of different elements could then be adjusted for clarity, the characteristics in the display could be used to order the trials in different ways while the reviewer analyses the data, and then a 'fixed' version could be outputted as a pdf file, or as separate image files, to be included in a report. Such software could be produced as a complete package, or perhaps as an add-on to existing software, such as an Excel plug-in or Open Office extension. Alternatively, it could be a web-based tool, which also might allow the sharing of results more easily.

The results raise a question concerning information responsibility (see Table 4.5 – 8: Information responsibility). A GOfER display can present a large amount of information, but not everything shown in the data tables of the report is used. In the case of the cochlear implants review, the designer omitted the deafness of participants, and the brands of implants used, for example. For a graphic like this to be used in a live report, the decision would have to be made which information should be presented in this display. While the reviewer will almost certainly have a fairly clear idea of what should be presented, the decision-maker might also have input. Also, in the case of having an information designer to produce graphics like this, they may be able to present several less important pieces of information in the space of one thing that had individual importance, but less importance than the sum of all the other things that could be presented in that space. For example, the outcome measures table in the cochlear implants GOfER display took over half the available space, as much as the ages, design, size, follow-up and quality

together, which might be considered more important to give a summary of. Such design issues will have to be carefully negotiated between designers and users of information graphics in HTA.

5.5.2 Discussion

Returning to the main aims of these interviews, it seems that GOfER would be a valuable way of giving a quick overview of key information on the characteristics, quality and outcomes of trials included in a systematic review. Its particular strength is that it condenses the data, allowing multiple aspects of the trials to be presented together on the page, giving a more holistic view of each study (allowing study characteristics, quality and results to be more quickly and easily considered together).

However, the GOfER display may be challenging to produce exactly with the current skills and software available for use in technology assessment report authoring groups. While displays like this, designed on an individual basis by trained information designers is an ideal, resource constraints may make this an unrealistic option. A software tool that could produce a GOfER display automatically would be less costly, but would mean sacrificing flexibility, and a commercial case would have to be made for it to be developed,

In terms of production and use in HTA, researchers should be encouraged to present visual summaries of systematic reviews using whatever means is available to them. Therefore, the name GOfER should not be used to refer to a presentation of this exact set of data, but as a graphical summary that could include different data for different reviews.

When it comes to inclusion in reports, whatever method a GOfER display is produced with, it seems that such graphics should be placed with the text in the main report sections, so that they can be used along with the textual narrative to explain the evidence to the reader. The raw numerical data must also be presented, in case exact values are needed for further analysis at a later date. This raw data can mostly be presented in appendices, and the graphic could be used to guide the reader to the relevant information, with page numbers or section headers included in the display.

5.5.3 Strengths and limitations of methods

The speak aloud method of cognitive interviewing worked well in most cases, but some people were reluctant to share thoughts, or only half-vocalised thoughts, which may have been too quick to capture in this way. It also affected people's task times, but this was fairly straightforward to account for in the analysis, using proportional time measurements instead of absolute. A more experienced, and/or less biased interviewer could try the probing method instead (Willis 1999), especially where time to complete tasks is important.

The main flaw in the experimental design was the fact that the GOfER display was presented in a separate document, not bound together with text as it would be in a report. It also presented less information than the data tables that were used in the report, which made it much quicker and easier for participants with GOfER to find the right place to look for the information. Tasks like task 8, which asked for data about a single comparison section might have been a better choice. Alternatively, a fairer comparator to test GOfER against might be a set of tables which only presented the data used in the graphic, but showing it in numerical/textual form. However, this would not be representative of the way reports are actually used in the decision context.

It might be possible to test the original report against a version of this same report with the GOfER display included, perhaps replacing the 'visual summary' tables (which were an unusual feature of the cochlear implants report). The decision would have to be made then whether to move some or all of the raw data tables to appendices in this report with integrated GOfER displays. Doing so might be instructive in answering the question of whether this is appropriate or not, and more closely represent how a GOfER display might be used in practice (as a replacement for, or a supplement to, more conventional presentations of information in the main body of reports).

The participants indicated that the tasks they performed were representative of how they might interpret a systematic review. However, two of the most experienced reviewers mentioned that they were likely to look at the results of the review first, and then work back to the characteristics of individual studies to judge how much they believed this answer. The tests may therefore give an accurate sense of how a decision-maker might interpret the results tables or the GOfER alternative, but not account for the context of looking at them with the outcome of the whole review in mind.

If this GOfER display were to be evaluated again, it would be better to test it 'in

situ' within the report itself. It would also be best to test with the whole report, including appendices, rather than just using a single section. This would make the experimental condition closer to what a decision-maker would be expected to deal with.

The VARK learning styles questionnaire did not seem to be a good predictor of the opinions of participants. However, the length of experience in HTA seemed to affect responses more. Those with more familiarity with the field had some very valuable comments. In future studies, a useful method of sampling could be to ask more people to take part than is practical, and interview only those with the most experience of working in HTA (or the relevant field of application).

Some research on information responsibility might be valuable to the further development of GOfER. It would be helpful to know how valuable a decision-maker would find the various elements of the display, so that this could be taken into consideration when deciding which data should be included in a graphical summary. The designer of a graphic like GOfER would then be able to balance the space taken by each element with how valuable it would be to a decision-maker.

There might also be the possibility of testing the GOfER display with different audiences. As this experiment has shown that GOfER has some value, it might be possible to include it in a report which will be used to make a decision at a NICE appraisal committee. While a detailed, controlled, task-based experiment would not be possible, a questionnaire or short telephone interviews with decision-makers might be instructive. Such research would show how the display is working in practice, and if anything should be changed from their perspective.

The GOfER display is already the product of 'iterative design' (Nielsen 1993), in that it has been shown to HTA experts and redesigned a number of times. The results of this experiment also suggest some fairly significant further improvements that might be made to GOfER. It could be tested again after a major redesign, perhaps with initial informal testing before another formal empirical study was undertaken.

5.5.4 Implications of findings

The most significant failing of the GOfER display was in task 9, where none of the 9 participants in the evaluation managed to identify all of the cross-sectional designs used in the studies included in the review. They seemed not to pick up on the smaller studies with this experimental design. This could be at least partially addressed by altering the key, showing both large and small studies with each design, as shown in Figure 5.5 – 1.

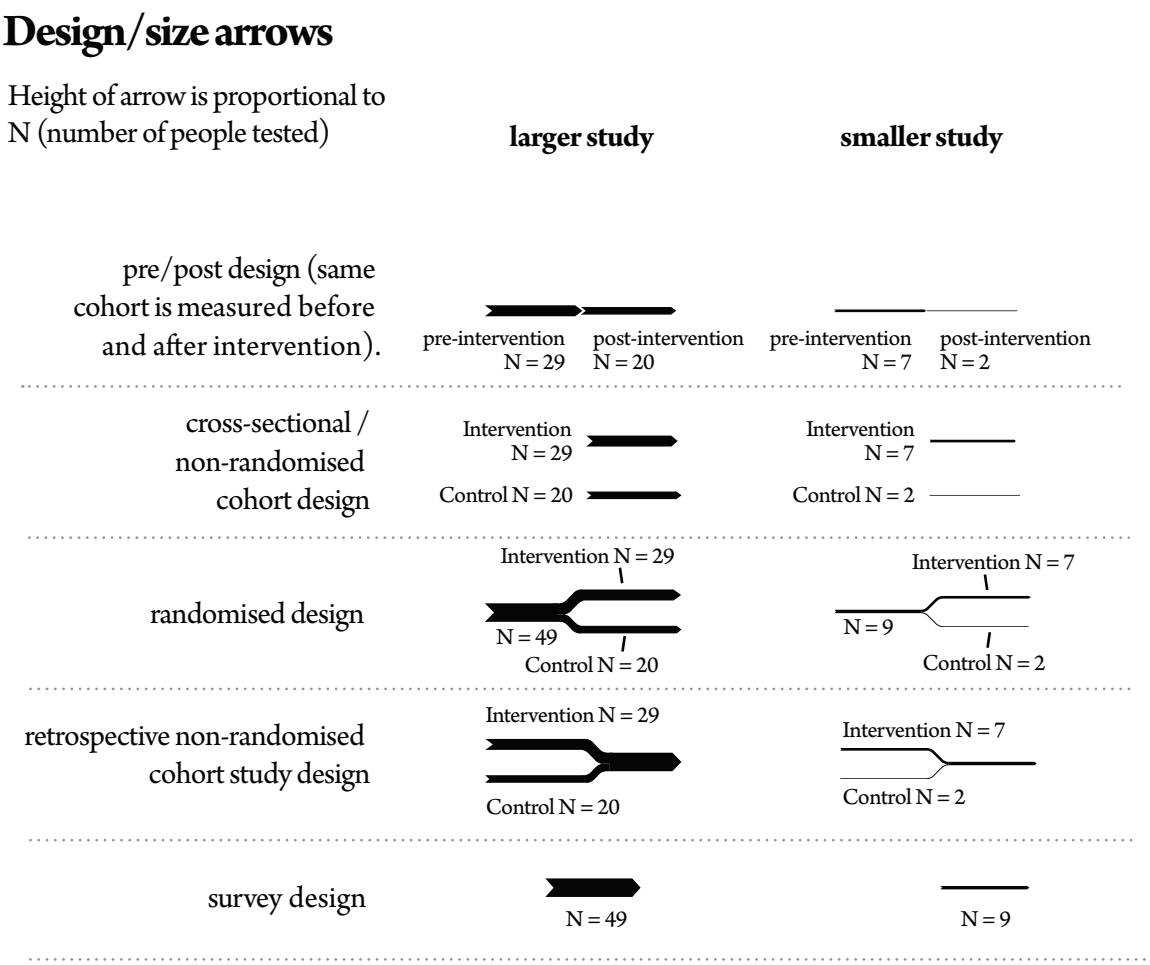

Figure 5.5 – 1
Revision to key, showing both large and small studies

Also, one participant mentioned that the cross-sectional design looked to them like time was passing. It might actually be possible to combine the length of follow-up of the studies with the design/size arrows, as in Figure 5.5 – 2. This might also allow outcomes that were measured at several different time points to be shown alongside the graphic, so that it could be seen at which point the

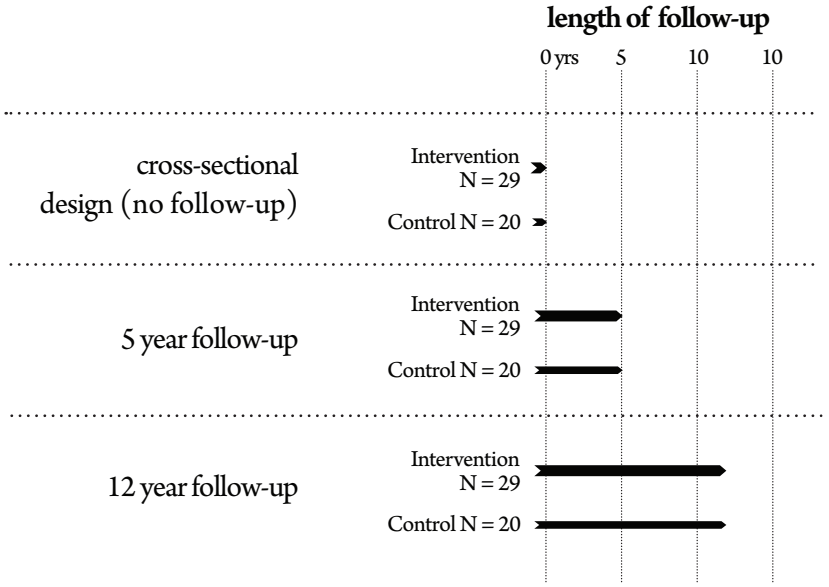

Figure 5.5 – 2
Design/size arrow used to show length of follow-up

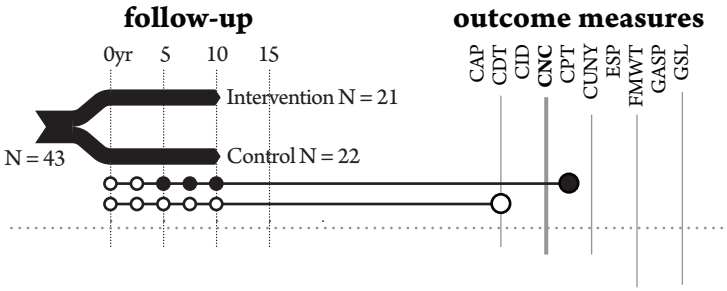

Figure 5.5 – 3
Design/size arrow used to show length of follow-up and the time at which an outcome measure became statistically significant

test became significant (see Figure 5.5 – 3). This might need a slightly larger space for each trial, and therefore fewer than 8 per page. For reviews in which the number of people tested at each stage was reported, it might be possible to incorporate dropout into the diagram, especially if it was larger. In Sankey diagrams used in engineering, losses in energy flows are often represented with arrows branching from the main flow. Something similar could be used here.

Also, concerning design/size arrows, if sophisticated page layout software was not available, it might be possible to simplify this element of the GOfER display to just show the N number and length of follow-up with straightforward solid coloured bars. The length of these would represent follow-up and the height sample size.

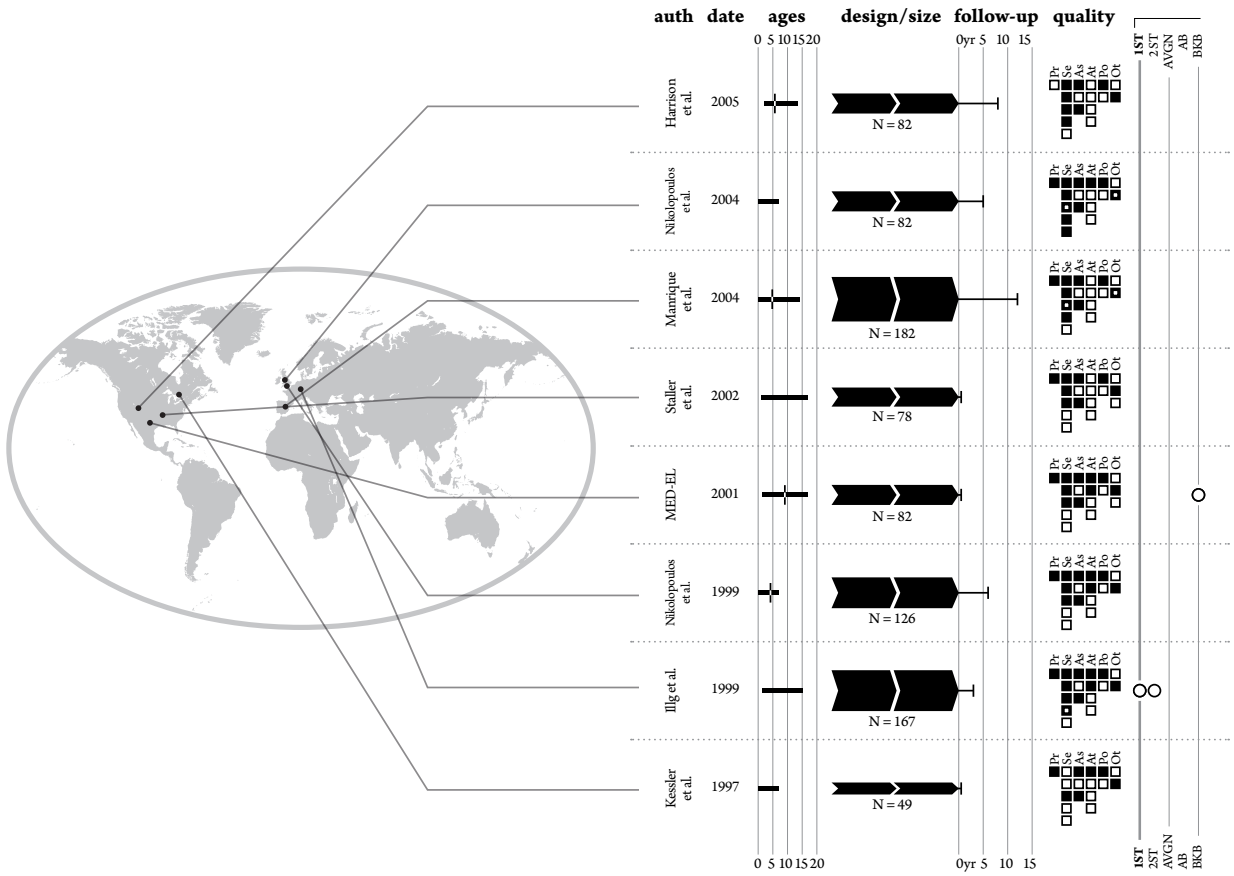

Figure 5.5 – 4
Display for heightened importance of the area of the world

However, different interventions might require very different displays. For example, in the case of the area of the world being very important, a graphic like that shown in Figure 5.5 – 4 might be more suitable, or even that in Figure 5.5 – 5, which just shows location and size of trial, represented by the area of the circles. Interestingly, this is no longer recognisable as a 'GOfER display' as such, supporting the conclusion that the term itself should not be used to mean a particular combination of visual techniques, but any kind of visual overview of information from a systematic review of clinical effectiveness.

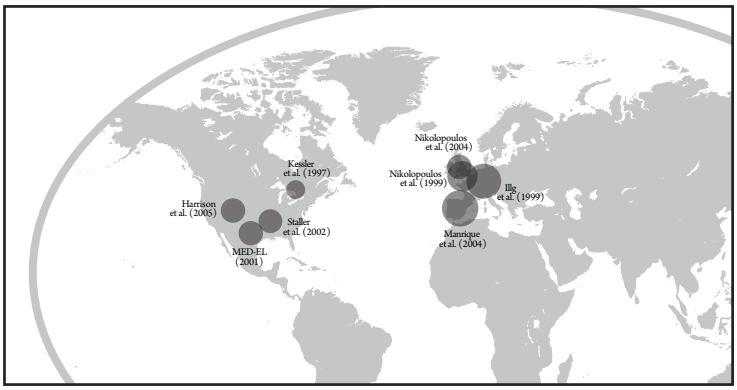

Figure 5.5 – 5
Location and size display

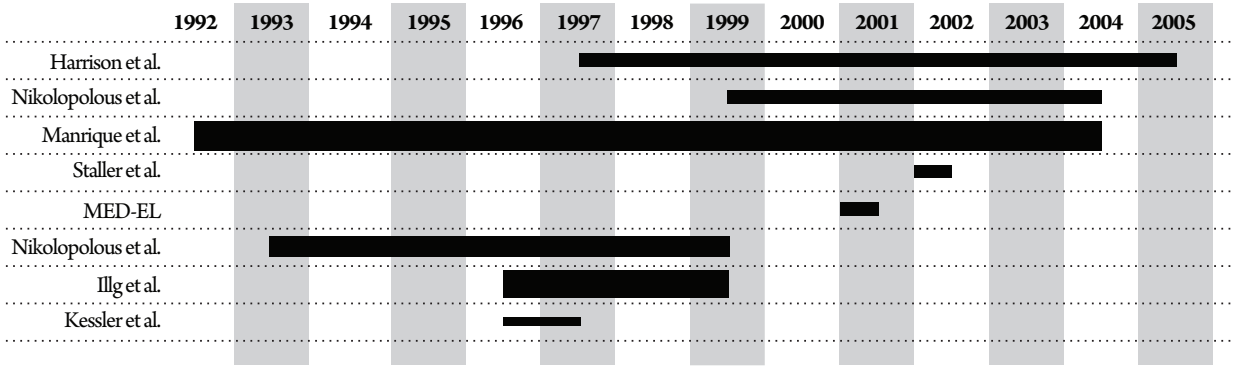

Figure 5.5 – 6
Display for heightened importance of timing of study

Another example of different information importance which might substantially change the design of the graphic display was if the time at which the study was being conducted was particularly important (as it might well be with a medical device such as an implant, due to technological improvements). In this case, the display in Figure 5.5 – 6 might be more appropriate. The right-hand end of the bars shows the date at which the study was reported. The left-hand end of the bars shows the date the trial would have started given the follow-up. The vertical height of the bars show the sizes of the trials. Again, this looks very different to what started as parts of GOfER displayed in a different way. However, other elements such as quality or outcomes could be built back in, especially if this display was condensed in width.

Lastly, a number of people mentioned that the outcome dots, while useful, didn't give much information on what was reported beyond whether it was significant. In other reviews, it would be quite possible to incorporate a display of the mean and confidence interval of each outcome measure, which would begin to look similar to a forest plot (Lewis & Clarke 2001). However, this would probably mean that the display would have to be separated into groups depending on which outcome measure was used.

Even in cases where a meta-analysis was not possible, it might be possible to give an indication of effect size using similar circles to those used in the GOfER display tested. A dark circle could represent a large effect, a medium grey a medium effect, and so on. The width of the outer ring could represent significance of statistical tests (see Figure 5.5 – 7).

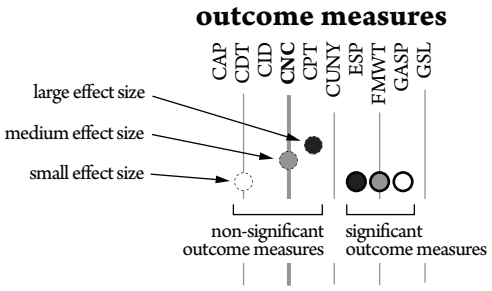

Figure 5.5 – 7
Effect size display