Using laddering and on-line self-report to elicit design rationale for software

Gary Sherlock & Gordon Rugg

Abstract

This paper discusses issues relating to elicitation of design rationale for software, and describes how we used two techniques in combination to elicit information about design rationale for software (in this case, Web pages). We found that students and professional designers did have more knowledge, and more richly structured knowledge, than members of the public; however, this knowledge made little explicit reference to design guidelines from the academic literature. Using two techniques in combination allowed us to identify and clarify issues which would probably have been missed by using each technique independently, and we recommend that more systematic use should be made of combinations of techniques for work in this area.

Keywords: Web page design; design rationale; laddering, online self-report

Introduction

There are numerous well established techniques such as QOC and IBIS for representing design rationale. This article describes the use of two elicitation techniques in a way intended to complement the use of design rationale notations. The study described here involves design of Web pages, but the same approach could be applied to design of any class of software. Using this domain allowed direct comparison of two trained groups with untrained members of the public, for reasons described below.

Although design rationale techniques offer obvious advantages, such as explicit representation of reasoning, there are equally obvious grounds for unease about the validity of the knowledge which they elicit. For instance, if designers are using these techniques to represent their own design decisions, this will be more likely to produce a “front” version than a “back” version (i.e. one for public consumption rather than one reflecting the reality), in Goffman’s (1959) dramaturgical metaphor. A more interesting set of problems arises from the cognitive routes used in the design: although designers have introspective mental access to a large set of explicit guidelines, this does not necessarily mean that their designs are accomplished using an explicit route with a high degree of verbalization – the literature on expertise consistently reports a high degree of pattern-matching in expert performance across a range of domains.

We therefore chose techniques on the basis of Maiden & Rugg’s (1996) framework, which takes explicit account of selecting appropriate techniques for elicitation of tacit, semi-tacit and explicit knowledge. Use of the appropriate techniques would then allow a reduction of problem space. The techniques chosen were on-line self-report, to catch contents of short term memory, and laddering, to unpack rationales and explanations recursively. Laddering also makes it possible to measure depth and breadth of explanation, as described in Rugg & McGeorge (1995), though for brevity the results from this are not described here.

Whatever problems the Web page designer may face, lack of guidelines is not one of them. In addition to numerous published guidelines, there is also a plethora of commercial courses and university modules dealing with this topic. Many of the specific guidelines are grounded in extensive bodies of established research by both academics and practitioners – for instance, the guidelines on use of color, or on selection of font, both based on extensive sets of findings from usability research. There is, however, one nagging question about this apparently idyllic situation: does anyone actually follow the guidelines?

The motivation for this research came from experience of seeing undergraduate students' responses to usability courses in general and to Web page design courses in particular. A common trend was for students to learn some of the course content, but apparently to view it as a pointless academic exercise: in examinations, and when asked to design Web pages, they ignored large portions of what they had been taught, and instead followed their own opinions about what constituted good design. This raised uneasy questions about which viewpoint was out of step with reality.

Our response was to investigate the design guidelines which were actually used by three sets of respondents, in evaluating good and bad design practice on commercial Web pages. One group consisted of students who had been taught Web page design; a second consisted of Web page designers; the third was a control group which consisted of ordinary members of the public. Our aim was to see what differences, if any, there were between the three groups. This would then allow us to assess the impact (if any) of the Web page design courses which the students had attended, with a view to feeding our findings back into academic practice. The same principle could be applied to other areas of software design.

There are well-established techniques for representing design rationale, such as QOC and IBIS. Although these are very useful during design of a product, they were less suited to our needs, since we wanted to be able to trace the respondents' guidelines back to their source. The method we used instead involved a combination of on-line self-report and laddering. This allowed us to elicit the design features which the respondents considered important, and then to unpack precisely why these features were considered important, and what the sources of evidence were for our respondents' beliefs. The following sections describe our procedure in more detail, and report our findings.

On-line self-report

Short term memory (STM) is an important and well-known issue in interface design; STM is an ephemeral type of human memory, with a typical capacity of about seven items and a typical duration of a few seconds (Miller, 1956). For interface design, it is often important to know which of the available on-screen information is being used by the user, and which is being ignored, so that information can be prioritized. There are similar issues in general product design, where it is often important to know which design features are significant for the users, and which are viewed as peripheral.

For these purposes, on-line self-report provides access to the contents of a respondent's short term memory. This method involves simply asking the respondent to think aloud while performing the relevant task, such as using an interface, or viewing a new product (Ericsson &Simon, 1993). The method is also known by other names, such as think-aloud protocols and concurrent verbalization. The "on-line" part of the name refers to the "live" nature of the session (as opposed to off-line self-report, in which the respondent performs the task and then reports afterwards on what they have done); it does not mean that the task has to involve a computer.

Not all tasks are suitable for on-line self-report. An obvious instance is safety-critical tasks, where the verbalization can be a dangerous distraction. Another obvious example is tasks involving language, such as simultaneous translation, where the respondent cannot perform the task and think aloud at the same time. Less obviously, many skills become so habitualized that the respondent no longer has valid introspective access into how they perform these skills. Also, respondents in many tasks involving spatial problem-solving encounter difficulties or fail to solve the problems if they think aloud during the task It should, however, be noted that there is also evidence that concurrent verbalization may help with some types of problem solving; this is a complex issue which goes beyond the scope of this article.

There are other varieties of report which get round these problems to varying extents. One is off-line self-report, where the verbalization occurs after the task. This can be useful for explaining decisions which could not be explained at the time (for instance, a driver's reasoning about what to do during a driving emergency), but it does not offer access to STM. One way of tackling such cases is on-line report of others, in which one respondent gives a running commentary on what another respondent is doing. It is also possible to perform an off-line report of others, in which one respondent views another performing a task, and comments on it afterwards. This may be used, for instance, in assessing the performance of trainees in a complex task which can be tackled in more than one way, and where the trainee's reasoning only becomes clear with time.

Laddering

Laddering is a useful technique, derived from Personal Construct Theory (Kelly, 1955) for eliciting hierarchically organized knowledge. It was originally developed by Hinkle (1965) and has since been further developed within areas such as knowledge acquisition (e.g. Boose, Shema, & Bradshaw, 1989; Rugg & McGeorge, 2002). Examples include goals and values, classification, and explanation of technical and of subjective terms. A very limited set of probes is used to move around the respondent’s knowledge. For instance, the probes may be used to elicit progressively higher-level goals and values relating to a design choice, or to unpack subjective terms to the point where they become reasonably objective. Each probe derives in a standardized way from something previously said by the respondent; this means that the responses fit into a tightly specified knowledge representation, but also have very wide scope for representing the respondent’s knowledge in their own terms.

A typical example of laddering on goals and values will start with a choice between two randomly chosen domain items, such as screen dumps of two Web pages, and the question: “Could you tell me which of these you would prefer, and why?” After the respondent chooses one and states a reason, the next question would be: “Could you tell me why you would prefer that?” This question would be used repeatedly in reaction to each response until the respondent could not give any further higher-level goals (for ethical reasons, it is usually advisable to stop when the responses become too personal). The process usually reaches top-level goals surprisingly swiftly (three to five levels), and responses usually fan-in towards a small number of high-level goals. In the domain of IT products, respondents often cluster either around instrumental high-level goals (i.e. goals involving getting the job done more swiftly and efficiently) or around expressive high-level goals (i.e. goals involving demonstrating to others what sort of person one is). This has obvious implications for marketing and for product design – some users will want a feature for very functional reasons, while others will want it simply because owning a device with that feature will be a sign of peer-group status.

Laddering downwards on explanations is similar in concept, though with different probes (Rugg & McGeorge, 2002; Honikman, 1977). The probe will typically start with a concept mentioned by the respondent elsewhere in the session, and will take the form: “You mentioned X earlier in the session. Could you tell me how you could tell that something was X?” For instance, “X” might be “cluttered”. The probe: “How could you tell that something was cluttered?” might produce a response such as “too many items on the screen”; this would be used as the seed for the next probe, “How could you tell that there were too many items on the screen?” and produce a response such as “There are more than about half a dozen”. “About half a dozen” is sufficiently specific to be used in design decisions, unlike “cluttered”. This “bottoming out” typically occurs fairly soon, depending on the domain and the expertise of the respondent – between one and five levels of explanation is the usual range.

Laddering can also be used on classes, with the initial probe: “Can you tell me some types of X?” and then subsequent probes to unpack the types, sub-types, sub-sub-types, etc. This can be useful in clarifying the extent of respondents’ expertise; for instance, when one of us (Rugg) applied this approach to computer science students’ knowledge of sources of information, it swiftly identified areas where the students were unaware of useful information sources. The results are being used by the University College Northampton library to improve provision to students.

Once mastered, laddering is a swift, powerful and flexible technique, and a very useful complement to other techniques such as on-line self-report and card sorts (Upchurch, Rugg & Kitchenham, 2001), particularly for clarifying the meaning of technical and subjective terms.

The case study

The case study used three groups. One group consisted of four professionally qualified Web designers; the second group consisted of six students who had been taught Web page design; the third was a control group consisting of six people who had experience of using the Web, but who had not been taught Web page design. All groups contained members from a range of ages and ethnic backgrounds.

The materials used were screen dumps of six Web pages from a range of commercial sites, deliberately selected to show a variety of color schemes, layouts, etc, so as to give the maximum opportunity for respondents to identify design features which they considered good or bad. Each Web page was reproduced twice, once as a laminated, full-color reproduction of the page, and once as a monochrome, unlaminated paper copy. The laminated full-color version was used to show respondents exactly what the page looked like; the monochrome paper version was used as a recording device, on which we and the respondents could write comments, etc, as the session progressed.

The procedure used a combination of techniques, following guidelines on choice of elicitation techniques in Maiden & Rugg (1996) and in Rugg, McGeorge & Maiden (2000). Using on-line self-report allowed us to identify features which the respondents recognised, as well as those they recalled; it also gave access to features being processed in the respondents' short term memories. Laddering allowed us to investigate the respondents' knowledge systematically, with recursive unpacking of their design rationales. The procedure used was to show respondents each screen dump in turn, and ask them to identify examples of good and bad design on the screen dump. This part of the procedure was very similar to normal on-line self-report, in which the respondent provides a running commentary on their actions and thoughts while performing a task (in this case, the task of critiquing the Web page design). After this, we used laddering to unpack each of these responses. For this domain, we found it useful to use customized prompts, developed during pilot sessions, as suggested in Rugg & McGeorge (2002). The prompts which we found most useful were:

- what problems are associated with X?
- what are the reasons, do you think, for designers using X?
- why do you think that this is an indicator that this site has been well/poorly designed?

The response to each initial question was written directly on the monochrome hard copy, making it clear which design feature was being addressed each time.

It was then possible to follow up the response to each question with further questions, until the reasons had been tracked back to their source. For instance, if an initial response had been: “This site is cluttered”, this might lead to a more specific reason for avoiding clutter, namely: “People have trouble handling more than about seven items in a list” and this in turn might lead back to design guidelines based specifically on research into the limitations of working memory, derived from Miller’s classic paper on the topic (Miller, 1956). Alternatively, they might lead back to a response such as: “That’s what I think”, or: “I find that users prefer this”, with no reference to the literature.

We varied the order of presentation of the screen dumps, to allow for boredom effects, and counterbalanced the order in which respondents from each group were used, to allow for practice effects with this customised method. The sessions lasted between fifteen and seventy-five minutes, and were audio recorded.

Results

There were differences in the duration of the sessions, with the mean duration of the student sessions being 43 minutes, compared to 38 minutes for the designers and 23 minutes for the control group. Interestingly, there were no substantial differences in the number of initial responses obtained from each Web page, with a range from 79 responses (site C) to 94 responses (site D). There was also no substantial difference between the number of initial responses given by respondents to the first page they critiqued (total = 88) and the number given to the last page they critiqued (total = 92). This suggests that the respondents were neither showing boredom nor diminishing returns.

The initial on-line self-report produced clear but small differences between the three groups, with the students giving a mean of 36.3 responses and the Web designers giving a mean of 33.35. The control group gave a mean of 27.0 initial responses.

A different pattern occurred with the laddering results, where the designers gave most responses – a mean of 81.0, compared to 70.5 from the students and 52.2 from the control group.

The next set of results relate to content analysis of the responses.

A useful first step is to tabulate instances of verbatim agreement, where two or more respondents use identical wording to each other in responses. This frequently occurs when the respondents are using well-codified knowledge (such as knowledge which they have been explicitly taught, as opposed to knowledge which they have learned implicitly via experiential learning). In this study, we found hardly any verbatim agreement within groups: only 9 phrases, each used by only two respondents. The phrases involved were:

- color
- small text

- wishy washy green
- what are they?
- gloomy
- what are the numbers?
- dislike color
- open
- waste of space

The first seven of these were generated by the student designers; the last two by the control group. The Web designers had no verbatim agreement. When responses are compared across groups, two more instances of verbatim agreement occur: one involving “font”, and the other involving “navigation”.

The next step was to group the responses in terms of gist agreement, where different respondents were using different wording, but with the same underlying meaning. This was done by an independent judge. This is useful for assessing agreement between respondents. At this level of analysis, there was considerably more agreement within groups, and also more agreement between groups. The areas of gist agreement for each Web page were as follows.

Page A: intuitive navigation, color differentiation, font size, excessive information, use of logos. (All these were mentioned by student designer group and by the control group, but not by Web designer group.)

Page B: additional navigation. (Mentioned by student designer group and Web designer group.)

Page C: uninteresting. (Mentioned by student designer group and by Web designer group.)

Page D: unclear icons. (Mentioned by all groups.)

Page E: logical navigation (mentioned by all groups); poor use of space (mentioned by student designer group and by control group).

Page F: clear branding (mentioned by Web designer group and by control group).

The next stage would normally be a further aggregation of responses, bringing together those which were semantically related, but which were not necessarily equivalent in meaning to each other. (For instance, “links” and “menu” are semantically related, since they both involve navigation, but they are not synonyms.) This is useful for identifying themes in the responses, as opposed to agreement about the values of the responses. In this case, the themes for all three groups were close to recurrent themes in the design literature: navigation; color; font; site control information; user interaction; user interaction; grammar; site information; layout, style and design; graphics.

Discussion

The results indicate that student Web page designers do know more than members of the lay public about Web page design, and that professional designers know more about Web page design than student Web page designers do. This expertise, however, is largely "behind the scenes" and emerged only when laddering was used to unpack the initial responses.

This suggests that students do learn something during Web page design courses, rather than resorting to unsupported opinion at the same level as an untrained member of the public. This finding is encouraging for the usability and design communities. However, the very low incidence of verbatim agreement within the student designer group and the Web designer group suggests that any formal education and training in this area has had very little impact on the terminology used by these groups. This in turn raises the suspicion that the lack of shared terminology reflects further-reaching conceptual differences within and between groups. The pattern of responses here is quite different from that in well-codified domains which we have studied in the past, where there tends to be a high degree of verbatim agreement even when visual pattern-matching is involved.

One thing which was significantly absent from the results for all groups was mention of any specific sources for design rationale. References to design rationale issues such as color blindness and color sensitivity of the human eye were fairly common, but were stated as facts, rather than being spontaneously backed up by sources. There are various possible explanations for this. One is that the respondents simply could not remember the sources; another is that the respondents had internalised the underlying principle by a process of skill compilation (Anderson, 1990), and had not found it necessary to internalise the source of the principles at the same time. A similar possibility is that the skills had moved from episodic memory to semantic memory, with the terminology being treated as incidental detail rather than part of the deep structure. In academic writing, use of referencing is an important issue rather than a superficial one, and academic researchers will typically remember key details of significant publications, such as the authors' names, the title and the date of publication; this is not such an issue for practitioners. Another possible explanation, that the respondents treated the sources as taken for granted knowledge (Grice, 1975) and did not bother to mention them explicitly, is unlikely, since laddering is normally a good way of eliciting taken for granted knowledge. The implication is, then, that the principles used by designers may have become uncoupled from the source literature, which raises interesting questions both for practitioners and for those producing design courses.

Conclusions

One conclusion from this study is that choice of elicitation method is an important issue. The on-line self-report was good for identifying the design features which the respondents considered important, but was not good for investigating knowledge as opposed to opinion. Laddering was good for unpacking the respondents' knowledge, and showing how much knowledge lay under the surface of their initial responses. It was also good for eliciting specific design issues, such as unpacking "dislike vertical side text" into the component reasons ("awkward font" and "awkward orientation").

Another conclusion was that students do appear to gain something from being on a Web page design course, and (in our sample at least) respond more like designers than like members of the lay public. Quite what they gain, however, is an interesting question. Our study was designed to investigate how much, and what, the three groups had to say about Web page design. Investigating the accuracy of the respondents' assertions(for instance, that a futuristic design will attract young customers) is a topic for a separate study. It would also be interesting to see how these results would compare with those from studying design rationale among, say, Java programmers.

In conclusion, these techniques appear to provide a useful addition to the toolbox of researchers investigating design rationale, and we recommend that they be more widely used.

References

Anderson, J. R. The Adaptive Character of Thought. Erlbaum, Hillsdale N.J., 1990

Boose, J.H., Shema, D.B. and Bradshaw, J.M. Recent progress in AQUINAS: A knowledge acquisition workbench. *Knowledge Acquisition*, (1): 185-214 (1989).

Ericsson, K. A., & Simon, H. A.. Protocol analysis: Verbal reports as data (rev. ed.). Cambridge, MA.: MIT Press, 1993

Goffmann, E. The Presentation of Self in Everyday Life. Doubleday, New York 1959

Grice, H.P. Logic and Conversation. In: P. Cole & J.L. Morgan, (Eds.) Syntax and Semantics 3. Academic Press, New York, 1975

Hinkle, D. The change of personal constructs from the viewpoint of a theory of construct implications. Unpublished Ph.D Thesis, Ohio State University, 1965
Cited in: Bannister, D. & Fransella, F. (1980). *Inquiring Man*. Penguin, Harmondsworth, UK.

Honikman, B. (1977) Construct theory as an approach to architectural and environmental design.

In: Slater, P. (ed.) (1975). The Measurement of Interpersonal Space by Grid Technique: Volume 2: Dimensions of Interpersonal Space, John Wiley and Sons, London, UK.

Kelly, G.A. (1955). The Psychology of Personal Constructs, vols 1 and 2. Norton, New York.

Maiden, N.A.M. & Rugg, G. ACRE: a framework for acquisition of requirements. Software Engineering Journal, pp. 183-192, 1996

Miller, G.A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, **63**, 81-93, 1956

Rugg, G. and McGeorge, P. Laddering. *Expert Systems*, **12**(4), pp. 339-346, 1995

Rugg, G., McGeorge, P. & Maiden, N.A.M. (2000) Method Fragments. Expert Systems **17**(5), pp. 248-257

Rugg, G. & McGeorge, P. (2002) Eliciting Hierarchical Knowledge Structures: Laddering. Encyclopedia of Microcomputers, vol. 28, supplement 7, pp. 69-110. Marcel Dekker, Inc, New York

Upchurch, L., Rugg G. & Kitchenham, B.

Using Card Sorts to Elicit Web Page Quality Attributes

IEEE Software 18(4), July/August 2001, pp 84-89