Chapter 8: Conclusions and Discussion

1. Introduction

The substantive results from the preceding Chapters may be summed up under the two main themes of the thesis, as follows.

1.1 Principles versus Heuristics

1. In Experiments 1 and 2 neither novices nor experienced subjects reported significantly more all-severity usability problems using the author's principles set (short or full versions, respectively) than did similar subjects using Nielsen's heuristics. However, significant differences were revealed for high-severity (rated 5 to 7 out of 7) problems reported by the experienced subjects in Experiment 1. In spite of no overall difference in severity ratings, these subjects' ratings in that range were significantly higher than the novices'. They also reported more (mean ratio 2.8) all-severity problems than did novices.

2. In Experiment 3 there were significant differences between Principle and Heuristic subjects at the Knowledge task level. At this but not the Skill or Rule levels, novice subjects using a subset of the principles materials showed greater accuracy and lower redundancy than those using a corresponding subset of the heuristics. (There was an apparent trade-off between these differences and the higher hit rates of both groups at the Rule level.) There were also significant differences in accuracy and redundancy, but not hit rate, for high-severity problems at the Knowledge level.

3. In Experiments 1 and 2 there were no differences between Principle and Heuristic subjects in the number of problem types (principle set attributes) which were reported. This was so even for high-severity problems deemed most important (Experiment 1), and for interface-specific problems (Experiment 2). Nor did subjects in these experiments identify either more, or different, problem types using the heuristics (or the principles) than the control. However, in Experiment 1 most of the problems assigned higher severity (4 to 7 out of 7) by subjects were of types deemed most important by the author.

4. In neither Experiment 1 nor Experiment 2 were there significant differences between the numbers of problems reported by subjects using Nielsen's heuristics and under control conditions without such guidance. This was true for both experienced and novice subjects in Experiment 1, and for novices in Experiment 2 (who also failed to report significantly more problems using the full principles set). However, Heuristic subjects in Experiment 2 did report more high-severity problems than Control subjects.

5. Attempt was made to match the changes which followed evaluations of the first two versions of the VP-Lab search interface to the materials used in Experiment 1. There were more candidate matches with the short principles set than with the heuristics.

6. Sorting unique problem tokens (UPTs) on the product of frequency and mean severity is offered as one means of problem prioritisation. In Experiments 1 and 2 this allowed a focus on the most salient usability issues out of the large numbers of reported problems.

1.2 Cumulative Problem Curves

7. The cumulative problem curves from both Experiments 1 and 2 exhibited patterns different from those employed by Nielsen and others in support of the '3 to 5' claim. The numbers of experienced subjects required to uncover 75% of the relevant totals of both all and high-severity problems were much greater than 5. The corresponding numbers of novice subjects were outside the limits claimed for novices in Experiment 1 (but within limits in Experiment 2). These discrepancies were shown to correspond to lower detection rates than the 33% predicted by Nielsen (or the 25% required by the model).

8. It was shown that in Experiments 1 and 2 cumulative curves of the form predicted by Nielsen and others could be generated by assigning problems to categories derived from the principles set. This led to the speculation that other researchers may have allowed implicit categorisations to influence the early stages of the problem reduction process. It was suggested that alternative interpretations of the problem descriptions in two published papers might allow for increases in problem counts sufficient to lower detection rates below the '3 to 5' threshold.

9. The cumulative curves from Experiment 3 showed that only on the very contrived Rule level tasks did novices' predictive performance approach that required for a '3 to 5'. On the more realistic Skill and Knowledge level tasks, neither Heuristic nor Principle subjects managed to reach target numbers of problems observed in the Test condition. Only Heuristic subjects correctly predicted 75% of (all-severity) target problems at the Skill level and (high-severity) problems at the Knowledge level, within limits claimed for novices.

10. Problem distributions in both Experiment 1 and Experiment 2 revealed similar patterns, even though subjects in Experiment 2 wrote their problem descriptions directly into a prepared spreadsheet. In both experiments, there were more problems which were reported only once (including those responsible for the significant result in Experiment 1) than were shared by more than one subject. In that experiment, all of the most frequently occurring problems were within the top ten percent of UPTs.

11. In Experiment 3, problem distributions at the Knowledge level were more like those in Experiments 1 and 2 than those at the other two levels. Higher proportions of both predicted and observed problems featured only once, and a lower proportion of predicted problems were shared between Heuristic and Principle conditions, than at the Skill or Rule levels.

12. Inter-rater reliability tests on the data from both Experiments 1 and 3 showed good correlation between the author and independent raters on the 'between subjects' stage of the problem reduction process (Experiment 1), and the match between predicted and observed problems (Experiment 3).

13. In Experiment 3, comparison of predicted and observed problems revealed that subjects at all task levels and in both conditions tended to over-predict. The differentiation made between prediction rate (all predictions to observed UPTs) and hit rate (correct predictions to observed UPTs) allowed comparison with the curves for detection rate (predicted UPTs per population) from Experiments 1 and 2. It was seen that these latter curves exhibited a pattern similar to that for prediction rate in Experiment 3.

14. The earlier ticket vending machines data was re-analysed in the light of the results from Experiments 1 to 3. Differences in the experimenter's hit rates and accuracy on the three machines were shown to be consistent with the later findings. It was also shown that the detection rate on only the smaller and simpler FFM had been high enough to comply with a '3 to 5' for observed user errors. In this light it was suggested that the Experiment 3 observed problem totals might have been greater (and hit rates still lower) had additional subjects been observed.

2. Discussion

Discussion is structured around the same two main themes.

2.1 Principles versus Heuristics

This thesis has demonstrated that both novices and experienced evaluators could make use of evaluation principles such as the author's to either report or more accurately predict more usability problems than similar subjects using short heuristics such as Nielsen's. However, the novices were only able to do so for a relatively 'closed' and non-contrived task (at the Experiment 3 Knowledge level), and the experienced subjects (on realistic but 'open-ended' tasks and systems in Experiment 1) only for problems to which they themselves had attributed high severity. Novices' heuristic-based performance was better on only simple and constrained tasks (at the Experiment 3 Skill and Rule levels), though on an open-ended task (in Experiment 2) they did manage to report more high-severity problems using the heuristics than a control.

The VP-Lab study in Chapter 7 implies that it might be relatively simple to demonstrate some results for principles over heuristics. In volume terms alone, there should be more matches, more to be uncovered, with an expanded principles set than with a shorter set shorn of its background material. However, a principles-heuristics effect proved to be more elusive than first envisaged. In Experiment 1, it was demonstrated only for experienced subjects and for high-severity problems; in Experiment 3, only for subsets of both principles and heuristics, and after controlling for task type; and in Experiment 2, not at all. The likely reasons for the limited success of the first main theme of this thesis will now be explored. They can be summed up under two headings, namely internal and external validity.

2.1.1 Internal Validity

Internal validity issues include the use of problem count as dependent measure, the method of severity assessment, the means of problem recording, the subject experience questionnaire, and the assignment of problems to types.

2.1.1.1 Problem Count

Even if usability problem count were not an unreliable measure of evaluator performance (as Chapter 4 onwards has demonstrated), it would still be a crude yardstick by which to assess the quality of an evaluation method. Yet it is the main dependent measure in most UEM studies, and much attention has been directed at comparisons of the numbers of problems extracted from various subjects-as-evaluators (Chapter 1). As Gray & Salzman (1998) point out, studies which rely on a single dependent measure (be it problem count or any other) in comparisons of whole methods are liable to criticism on grounds of construct validity. For that reason, this thesis has focused on materials rather than methods, holding the latter the same within experiments while varying the former.

Attempt has been made to explore more than just problem count. The prioritised problem listings in Appendixes G to I are included in order to flesh out these sometimes conflicting results. The extracts featured in earlier Chapters represent one way of cutting through the noise which such large problem sets represent. The thesis has also gone beyond problem count to differentiate between measures of predictive power (hit rate, accuracy and redundancy) and cumulative performance (detection rate and prediction rate). In the terms used by ISO 9241 Part 11 (1998), predictive power can be considered as a measure of the effectiveness (accuracy and completeness) of an inspection method, while cumulative performance assesses the efficiency (resource expenditure) of groups of evaluators.

However, problem count *is* the main dependent variable in these experiments. As has been said, the author believes that it is more important to establish some internal validity for this measure than to worry about whether or not the procedures used do, or do not, qualify as heuristic evaluation (or any other method). For that reason, some evidence was offered in Chapters 4 and 5 for the inter-rater reliability of the most important ('between subjects') stage

of the problem reduction process in Experiment 1 and the predicted-observed problem matches in Experiment 3. Nevertheless, in that the procedures used *do* resemble Nielsen's prescriptions for heuristic evaluation (Chapter 1, Section 5.3), it is believed that the conclusions concerning that particular method can be drawn with more confidence from these experiments than from studies whose interpretations of the method are less explicit.

2.1.1.2 Severity Assessment

Two of the grounds on which usability problems may be prioritised are severity rating and problem frequency. In Experiments 1 and 2 severity (ranked 1 to 7, low to high) was assigned to problems by subjects themselves, using criteria established by the experimenter (the author). Subjects were asked to consider only the impact of each problem (from 'trivial, might be ignored' to 'serious, must be addressed'). In Experiment 3, severity was assessed by the experimenter. Three forms of frequency were computed, namely frequency of reporting (Experiments 1 and 2), frequency of prediction (Experiment 3) and frequency of occurrence (Experiment 3). Thus while severity and frequency were measured separately (Lewis 1994), the thesis made use of only two of Nielsen's (1994d) four criteria (frequency [of occurrence], 'user impact', 'persistence' and 'market impact') for severity assessment.

The author's view is that problem persistence is, or should be, bound up user impact (which includes the difficulty users will have overcoming a problem). Market impact might best be assessed outside of inspection or testing (e.g. in focus groups). As far as assigning priority is concerned, it is more important to be clear about which form of frequency is being used. The tactic of basing priority on the product of frequency and severity seems sound, bearing in mind that these are ordinal rather than interval measures. (A further factor might be estimated time-to-fix individual problems.)

Severity is important for this thesis. The significant result for experienced subjects in Experiment 1 was closely bound up with severity ratings, the single effect in Experiment 2 was for Heuristic subjects' high-severity problems, and the only within-limits cumulative performance on non-trivial tasks by novices in any of these experiments was for high-severity Knowledge level problems in Experiment 3. The fact that no independent assessment of severity was obtained may detract from the validity of these results, particularly the first. However, the cumulative curves for both experienced subjects in Experiment 1 (Chapter 4, Figure 4.4) and novices in Experiments 1 (not shown for high-severity problems) and 3 (Chapter 5, Figures 5.5 and 5.6) did confirm the general finding from studies such as Virzi (1992) and Jacobsen et al. (1998a, 1998b) that high-severity problems are reported more frequently than low-severity ones. (Experiment 2 high-severity problems were not further analysed.) This implies that these subjects' *relative* severity assessments were reliable (and *not* that their choice of severe problems would have been the same as the that of the author or other raters).

In spite of no overall differences in severity ratings, the only principles-heuristics effect in Experiment 1 was for predominantly single-incidence high-severity problems reported by experienced subjects. The same subjects were also more critical than novices in their assignment of high-severity ratings. This suggests that the effect of the principles was to encourage higher severity assignment to a largely different (and bigger) set of problems than those reported by heuristics users. The higher ratings of experienced subjects at this level *may* be due to a 'risky shift' effect, the very identification of a larger number of 'serious' problems causing a shift towards the top end of the (high) severity scale. If that is so, then a principles-heuristic effect would involve not only the identification of additional usability problems but an assignment of disproportionate importance to those same problems. In that case, the issue is not whether evaluators are more likely to pick out severe problems than minor ones (they probably do), but whether those problems are severe in the first place.

Most UEM studies (an exception being Virzi 1992) do not offer independent confirmation of 'actual' severity. While this is also true of Experiments 1 and 2, Experiment 3 set out to assess the predictive ability of Heuristic and Principle subjects against a set of 'real' errors in the Test condition. Unfortunately, these subjects were not asked to assess severity (this was done by the experimenter); however, in Experiment 3 it was possible to compare predicted and observed frequencies, there being a higher correlation at the Skill and Rule levels than at the Knowledge level (see Chapter 5, Figure 5.2 for problem distributions). If the relationship between frequency and severity holds for this experiment, then it is reasonable to conclude that severe problems were predicted more often at the first two levels. In the event, however, the experimenter did not attribute high severity to any of the Skill or Rule problems (errors), preventing further speculation. A recent comparison of estimated and 'actual' severity levels (Hassenzahl 2000) found no correlation between the two measures, implying that severity, like detection rate, is unreliable even when rated by more than one evaluator (Jabobsen et al. 1998a, 1998b).

2.1.1.3 Problem Recording

In contrast to most UEM studies, in these experiments no additional means of problem recording were used beyond subject or experimenter transcription. In Experiments 1 and 3, the experimenter (the author) sat with subjects while they 'thought aloud', writing down their comments and reactions on paper. In Experiment 2, subjects entered their own problem reports online (into a prepared spreadsheet) without experimenter intervention. While the strong similarity between the problem distributions from Experiments 1 and 2 (see Appendixes G and H) supports the view that the experimenter presence in Experiment 1 did not unduly bias the subjects' reporting (and that those transcripts were accurate), nevertheless the lack of opportunity for post-session analysis and independent assessment of the subject protocols is the first major omission from these experiments.

In hindsight, it would have been preferable to have recorded at least a sample of subject sessions in some form, such as audio, video or automatic logging. Not only would the opportunity then exist for inter-rater comparisons, but the protocols would be available for re-analysis. The three stages of the problem reduction process depicted in Figure 4.5 of Chapter 4 represent only some of the points of potential unreliability in problem extraction, the very first being the veridical recording of a 'subject session'. Even without the doubts which have been raised concerning the veracity of think-aloud protocols, to filter verbal reports through another level of interpretation - the observer in a heuristic evaluation or guideline review (Nielsen 1993, 1994d), the evaluator in user testing (Jacobsen et al. 1998a, 1998b) may be to invite distortion.

The inter-rater reliability tests carried out on the results from Experiment 1 showed that the reductions performed by the author between the initial (low-level) problem listings and the final (top-level) problem sets in Appendix G were not unrepresentative of other experimenters. However, it has been pointed out that the wording of the problem descriptions used in this and the Experiment 3 inter-rater test were as prepared by the author. The problem reports had thus already been subject to a certain degree of filtering (from voicings to written record) and then had been edited at the 'within subjects' stage. Ideally, then, the early stages of the problem reduction process would also receive independent verification. Under constraints of time and resources, the inter-rater tests focused on the most important parts of the two processes, namely the 'between subjects' stage of Experiment 1 and the predicted-observed matchings in Experiment 3. This is far from ideal, but it is more than many UEM studies have offered.

In running these experiments the author learned a great deal about the mechanics of problem recording. The Experiment 1 report forms were rudimentary, allowing only for additional comments. The Experiment 3 forms included the likely causes and consequences of each (predicted or observed) error, as well as recommendations for improvements. Experiment 2 subjects were asked to record the principle or heuristic to which each problem related, as well as remedies and other comments (see Appendix J for a sample.) The low-level problem sets from these Experiments are a result of a particular view of how usability issues might be extracted and recorded. That this view largely coincides with that proposed by Cockton & Lavery (1999) in their Structured Usability Problem Extraction Framework, or SUPEX, goes some way to offset the failure to make additional recordings of subject sessions.

2.1.1.4 Subject Experience Questionnaire

The questionnaire used to measure subject experience contained questions designed to assess both computing (including interface design) experience and HCI knowledge. (See Appendix E for a transcript of the Experiment 1 version[1].) Assessment was both objective

[1] The questionnaire content was the same for three experiments in all but the first question.

(e.g. software used) and subjective (e.g. awareness of usability issues). The scoring was adjusted during the Experiment 1 pilot study to achieve 98% for a recognised HCI expert and 4% for a subject with no computing experience. Novices were deemed to be those whose scores ranged from 0% to 20% inclusive, while experienced subjects were those scoring 30% and above. Any subject whose scores fell outside these ranges (one in Experiment 1, two in Experiment 3) were omitted from subsequent analysis. Mean novice scores from Experiments 1 to 3 were 10.43%; mean experienced scores from Experiment 1 were 61.40% (median 64, standard deviation 19.98). In Experiments 1 and 3 the questionnaire was administered by the experimenter (the author) at the start of each subject session. In Experiment 2, it was administered by e-mail following subject recruitment (before the start of first, training, session).

It is clear that experienced subjects were allowed a much wider range of scores than novices. Thus although there was a distinct gap between the two scoring bands, it would be unwise to claim that the questionnaire was measuring "HCI expertise" in more than rudimentary fashion. However, most of the expected novices' scores did fall within the narrower range. As for the experienced subjects in Experiment 1, it is uncontroversial that these subjects did (and still do) know a great deal more about user interface design and evaluation than first-year Psychology undergraduates. Nevertheless, it remains unproven that this knowledge was solely responsible for the single problem count difference on which the claim for an Experiment 1 principles-heuristics effect depends.

2.1.1.5 Problem Types

It was demonstrated in Chapter 4 that different conclusions could be derived from the same problem data according to the level at which problems were categorised. The approach used in this thesis has been to resist problem typing ('same-as' matching) until the grounds for such categorisation have been established. In the terms of the model introduced in Chapter 4 (Figure 4.5), problem grouping is left to the last, 'between types', stage of the problem reduction process.

The reliability of the analyses on which such categorisations are based will depend on the both the accuracy of the problem matchings and the validity of the groupings under which matchings are made. The former is the subject of the inter-rater reliability tests performed on the data from Experiments 1 and 3 (discussed in Section 2.1.1.3). The latter will be addressed in this Section.

The tactic adopted in Experiments 1 and 2 was to base categories on the versions of the author's principles set used in each experiment. Thus the 230 UPTs resulting from the Experiment 1 problem reduction (Appendix G) were grouped according to the (then) 26 principles in the short set (leaving 14% unmatched), while the 114 UPTs from Experiment 2 (Appendix H) were matched with the 30 attributes in the full set (leaving only 7%). (Experiment 3 problems were not categorised beyond the three designated task levels.)

Not all principles or attributes were deemed to be represented in each set, Experiment 1 matching 24 of 26 principles while Experiment 2 matched only 14 of 30 attributes. Though the novices-experienced effect persisted in Experiment 1, neither this nor Experiment 2 exhibited significant differences between Principle and Heuristic subjects when counting by problem type.

There are several problems with problem types. The first is that they are arbitrary. Even if it can be demonstrated that two or more evaluators would match a problem set in the same way (and this is what the inter-rater tests do, broadly, demonstrate), there is no guarantee that the resulting groupings are representative of a real effect for one problem type over another. Thus, for example, although the two independent raters in Experiment 1 agreed with the author about the similarity of problems originally labelled NE (Navigational Effort), it does not mean that they agreed with author about what those problems represented (i.e. they might have considered the descriptions to mean something different). The assignment of problems to 'correct' categories (e.g. evaluation criteria) remains to be demonstrated for these experiments. In experiment 2 attempt was made to have novices attribute the source of problems to themselves or the evaluation materials; not surprisingly, no between-condition differences[2] were demonstrated for these guesses. Karat et al. (1992) reported that experienced evaluators performing a heuristics-based walkthrough tended to misattribute the source of problems compared to those observing user testing.

The second problem with comparisons by problem type is that it is invalid to apply categories derived from one set of evaluation materials to the problems arrived at using another set. As has been said, to retroactively assume some 'underlying' influence for the principles set for heuristics (and especially control) users is very dubious. Thus the lack of between-condition problem type differences in Experiments 1 and 2 may be criticised for its attempt to compare unlike with unlike. In Experiment 2, it is tempting to suppose that the want of attributions to the principles set categories was responsible for the lack of results at the problem type level. While this is a reasonable assumption, confirmation would require one-to-one matching of types between heuristics and principles. Such comparisons are probably best restricted to the mainly qualitative analyses attempted in Chapter 7.

The third problem with problem types is that they offer a distorted view of UEM effectiveness. In one important sense it does not matter which categories are applied to a problem set as long as it is done consistently. It can be shown that *any* equivalent grouping can reproduce the effect for problem types demonstrated in Chapter 4 (that high-level categorisation can raise detection rates sufficiently to comply with a '3 to 5'). In that sense, findings based on such categorisations (no matter how unwitting) *may* mask a different

[2] Principle-Heuristic t-tests (of proportion of problems attributed to materials rather than self): all severity p=0.34, high severity p=0.44, both two-tailed.

version of the underlying (low-level) usability issues. This issue is further discussed in Section 2.2.

2.1.2 External Validity

External validity issues focus around the extent to which the conclusions from these experiments can be extrapolated to other varieties of evaluation materials, other evaluators, other systems and other user tasks, plus the role of empirical support for principles and heuristics.

2.1.2.1 Evaluation Materials

As has been said several times, these experiments did not set out to compare evaluation methods, nor even competing versions of the same method. The focus on principles versus heuristics was designed to achieve some closure on just one aspect of one form of inspection method, namely the use of different evaluation materials in guideline review. Nevertheless, insofar as the methods used have resembled heuristic evaluation as prescribed by its originator, some limited conclusions can be drawn for that method regarding the value of materials other than heuristics.

The evidence from this thesis is that evaluators in a heuristic evaluation can make use of principle-based material such as introduced in these experiments to identify a wider range of usability issues than when using heuristics such as Nielsen's. In Experiment 1, this was demonstrated for experienced subjects using a short version of the full principles set. In Experiment 3, it was shown that even novices could make use of a subset of the full set on even a non-trivial task. However, in Experiment 2 a similar group of novices were unable to make better use of the full principles set than they did the heuristics. Some evidence was also presented that Nielsen's heuristics were no better than control materials in enabling both subject groups to identify usability problems. However, in Experiment 2 novices using the heuristics did report more high-severity problems than those using the control.

These conclusions do not deny the validity of attempts to compare different forms of UEM including heuristic evaluation. However, the implication of the review in Chapter 1 is that unambiguous conclusions regarding such comparisons are difficult to draw due to the lack of control which has been a feature of many such studies (Gray & Salzman 1998). If the aspirations of this thesis seem restricted, it precisely because of these considerations that the 'soft option' of principles versus heuristics was chosen in preference to comparisons of different principles. In the event, the difficulty experienced in demonstrating even this effect shows that it was probably wise not to attempt more.

To return to evaluation materials, the reason why Experiment 2 failed to show a principles-heuristics effect is probably the relative size and scope of the two sets of materials. This conclusion is supported by the relative (but nevertheless partial) success of Experiment 3, which limited both sets of materials so as to be directly relevant to the tasks and scope of the

experiment, and by Experiment 1, which used a short version of the full principles set (and made use of a higher proportion of this short set than was evident in Experiment 2). Though it is doubtful that these novices could have found their way around the full principles set even with shorter and more focused tasks (for example, chosen to relate to clusters of two or three principles apiece), at this point it is not possible to draw conclusions concerning the precise size and scope of a candidate 'set of full principles' (as opposed to the full principles set used in Experiment 2).

Moreover, even if it were possible to show superiority for one large set over another, without more careful control on tasks and materials than was exercised in Experiments 1 and 2 it would be difficult to know what it was about the winning (and losing) set which made it successful. There would be no guarantee that apparently similar sets would perform the same way, even on identical tasks and subjects. Thus even the apparently simple goal of a principles-heuristics effect turns out to be more difficult than first envisaged.

Nevertheless, some conclusions can be drawn concerning the differences between principles and heuristics. First, it is clear that the wider scope of the principles set(s) did enhance even novices' ability to identify usability problems, compared with the ten short heuristics. In this respect Experiments 1 and 3 were successful whereas Experiment 2 was a failure. Second, there are strong indications of a common set of core issues (e.g. consistency, feedback, error management) which are addressed by both sorts of materials. Beyond that core, differences between materials would be manifest in their respective abilities to attract issues of different varieties under the sorts of controlled conditions which were not attempted here. Third, the implication is that the heuristics are no better than control materials in eliciting problems (but may still be better than nothing when other materials fail).

While Experiment 3 did go some way towards a more careful approach, having brought in such controls (on systems, tasks and materials) all at once it is not possible to say for sure which of them was responsible for the resulting differences. First steps in any further studies would thus be to establish a core set for restricted tasks, and then to expand both sorts of materials as different tasks and/or systems are introduced. If heuristics prove incapable of predicting problems with more complex tasks where principles succeed, then that may be evidence that heuristics, do, as suspected, require supporting background material. The nature of that background might be established by adding and removing identical principle-based material until the lack of an effect for the heuristics is mirrored in the opposing set. An established core set would also represent a validated control against which individual heuristics or principles might be measured.

While both heuristics and principles are undoubtedly preferable to the huge collections of guidelines (of which the Smith & Mosier 1986 are the best and most influential) outlined in

Chapter 1, it remains to be established that such materials manage to codify HCI guidance in a way that can be used by designers and evaluators alike. Little attention has been given in this thesis (as elsewhere) to the differences between guidelines as design advice and evaluation material. In-house style guides such as the Microsoft (1995) Windows™ guidelines are intended as design guidance, whereas the most profitable use of heuristics and principles may be as aids to evaluation. While guideline collections have largely failed in their purpose (of guiding designers), it remains probable that the corpus of guidance which is represented by the HCI literature (and held by practitioners) is capable of being codified in more manageable form than has so far been achieved. The degree to which such materials may be craft-based (from experience and practice) or truly principle-based (from empirical evidence including cognitive psychology) will be discussed in Section 2.1.2.4.

2.1.2.2 Evaluators

The second major failing of the three main experiments is their use of novice subjects. Only Experiment 1 compared the performance of experienced and novice evaluators, the rest relying on a pool of undergraduates. Thus the conclusions that can be drawn regarding the differences between materials, tasks and methods are limited by what can be assumed about such novices. The issue is the ecological validity of results drawn from mainly female Psychology undergraduates at a British university.

The reasons for the predominance of novices are resource-based. The pool of available subjects in this group changes each year, with (at York) a supply of up to 100 fresh[3] individuals. By contrast, the much smaller pool of experienced researchers (even including postgraduates) changes little. Further, the practicalities of recruitment and session-running are such that in Experiment 1 it took perhaps 3 times longer to complete the same number of sessions with experienced subjects than with novices. In addition, Experiment 2 required subjects for two consecutive sessions, while Experiment 3 needed to be completed in as short a time as possible. These constraints are not unique to HCI research. However, the nature of these experiments meant that apart from the two inter-rater tests, recruitment of experienced subjects was not attempted beyond Experiment 1. In the absence of experienced subjects, it was hoped that the novices would provide a 'base-line' for evaluator performance.

In the event, the results from Experiment 3 imply that even novices can make use of suitably targeted principles when performing non-contrived tasks. The Experiment 1 results showed that experienced evaluators did much better than novices on open-ended tasks. However, little can be concluded from Experiment 2 except that those novices failed to make use of a welter of principles material. Unfortunately, the differences between the three main experiments mean that extrapolation from the first to the second and third experiments

[3] Not that they stay that way for long. Even when paid, the no-show rate for Psychology undergraduates is alleged to be around 60%.

cannot easily be made. Thus at this point it is impossible to infer precisely how much better experienced subjects would have performed in Experiment 2 (in particular) and Experiment 3 (other than that gender and age are not anticipated to be factors). Clearly, such conclusions necessitate a replication of both studies with experienced subjects (here defined as those whose questionnaire scores exceed 30%). The performance of other relative novice populations (such as Computer Science undergraduates) might also be assessed, particularly on more general tasks such as those in Experiment 3. It may also be possible to establish the role of prior exposure to the Experiment 2 materials, by running sessions without training for additional subjects from both groups.

Since evaluator expertise plays a strong role in heuristic evaluation (Nielsen's 1992 projections for combined novice performance being much lower than those for both 'single' and 'double specialists': Chapter 4, Figure 4.2), the conclusions that can be drawn from these experiments concerning expert performance in heuristic evaluation are limited. However, Experiments 1 and 3 showed that even novices could make use of principle-based materials. On the above 'base-line' assumptions, these experiments can be considered as having established a minimal role for principles in heuristic evaluation in particular, and guideline review in general. Experiment 1 showed that both subject groups shared a common core of popular (high-frequency) problems, but that experienced subjects reported between two and three times more problems than novices overall. In an early study, Hammond et al. (1984) compared expert inspections with novice user testing. While both groups identified a set of low-level problems (or 'procedural and conceptual difficulties') with the same system, the experts provided a more 'integrated' view, including hypotheses concerning the sources of problems. Such a view is akin to what we might expect of expert use of evaluation materials. The issue for this thesis, then, is whether principles offer more to such experts than do heuristics. The implication from Experiment 1 is that they do.

2.1.2.3 Systems and Tasks

Experiment 3 showed that different results could be obtained by manipulating task scope. In that experiment, the systems chosen (Word and Excel) were also more 'closed', that is, more easily admitted of goal-directed tasks, than the 'open-ended' varieties (gallery browser, teaching software) used in Experiments 1 and 2. The issue for this Section is what can be concluded about the use of principles and heuristics with different systems and different task types. (For the purposes of this discussion, an 'open-ended' system is one for whom user goals and outcomes may be difficult for even designers to predict; a 'closed' system is one for whom outcomes may be more easily specified. Databases and web browsers thus qualify as open-ended systems, while word processors and spreadsheets are more closed.)

It has been established that the Experiments 1 and 2 problem distributions and detection rates were very similar, but that those in Experiment 3 were comparable only at the more complex Knowledge level. Experiment 3 novices' detection rates were higher at the simpler

and more contrived Skill and Rule levels (at the Rule level, high enough to comply with a '3 to 5' for the 'double specialists' in Nielsen 1992). In contrast, only at the Knowledge level was there a (partial) principles-heuristics effect for predicted problems compared with observed problems. The implication drawn in Chapter 5 was that the application of short heuristics may be limited to simple tasks or sub-systems, whereas longer principles may be best applied to complete or more complex tasks. However, in Experiment 3 both tasks (including the Knowledge level task) and materials were much reduced in scope compared to those in the first two experiments. Thus the validity of this conclusion depends on supposed similarity between the Experiment 3 tasks and those in other studies, compared to the 'open-ended' tasks in Experiments 1 and 2.

On the face of it, the comparison seems valid. Most UEM studies have used defined tasks or sub-tasks, even on 'open-ended' systems such as database front ends. Most of the Experiment 3 tasks appear comparable to those in other studies (the exception being the very contrived Rule level tasks). Thus one difference between heuristics and principles may be that the 'added value' of principles works best on more substantial tasks. This is good news for principles subsets. However, it remains to be seen how much validity can be claimed for the Experiment 1 and 2 results, in terms of the low detection rates there exhibited (this will be taken up in Section 2.2.2.2). And on the very Knowledge level task on which this proposition depends, the combined performance of Heuristic subjects was better than Principle subjects in their predictions of high-severity problems. The difference between this result and that for Experienced Principle subjects in Experiment 1 is that the latter were allowed to assign their own severity levels, whereas in Experiment 3 severity was established by the author. However, the Experiment 3 result involved novices and only six high-severity problems, so this comparison may be in invalid.

It appears, then, that the manipulation of task and system types in Experiment 3 produced an (albeit limited) principles-heuristics effect for novices at the more realistic and complex task level. This is in contrast to Experiments 1 and 2, where such an effect was not exhibited for open-ended tasks. The fact that it required tasks, materials and systems to be closely controlled in order to demonstrate such an effect makes it difficult to make more general claims about these factors (but this is a difficulty for all controlled experimentation on real-world systems). The failure of novice performance in Experiment 2 (and the poor performance in Experiment 1) may mask potential results for tasks, levels, materials, which cannot now be revealed. As mentioned in Chapter 2, the original intention in Experiment 1 was to compare 'poor' (B) and 'improved' (A) versions of the gallery browser software. Had the author known how long it would take to run and analyse even one of these, the preference would have been for version B.

Why, then, the choice of open-ended tasks and systems in Experiments 1 and 2 ? The approach was that usability evaluation is about more than having users perform set tasks.

The author believes that unguided interaction is a valid inspection and test procedure. Indeed, Nielsen's (1993, 1994d) prescription for heuristic evaluation (Chapter 1, Section 5.3) is specific in recommending that evaluators first run through the whole of an interface before only then proceeding in more detail (by simulating 'tasks'). Subjects in Experiments 1 and 2 were allowed to explore the whole of the (already reduced-scope) software used and to report freely. The fact that this produced results at variance with much of the literature (discussed in Section 2.2.2.2) led to the reinstatement of the more constrained tasks used in Experiment 3.

This view of task scope has some support. Lee (1998) proposed that the focus of user testing should be widened to encompass the goal formation and action execution stages of Norman's (1986) theory of action (as well as just the action specification on which testing appeared to concentrate). Karat (1994) supported the use of both prescribed tasks and self-guided exploration in inspection methods. Desurvire & Thomas (1993) were able to enhance the predictive performance of evaluators (compared to that reported in Desurvire et al. 1992) by having them adopt different 'perspectives' (e.g. "self", "sociologist", "spoilt child"). The author's view is that both closed and open-ended tasks have a place in inspections, but that the nature of many recent systems (e.g. web browsers, multimedia teaching packages) is such that it may be difficult to define a set of closed tasks when the precise outcome of an exploratory session may not be known in advance.

2.1.2.4 Empirical Support for Evaluation Materials

It cannot be credible that HCI research in general and cognitive psychology in particular has nothing to contribute towards collected guidance for interface design and evaluation. The role for empirically based findings from cognitive psychology and other fields must be wider than Landauer's (1991) pessimistic view (Chapter 1, Section 5.5).

The author's principles set represents one attempt to combine a wide range of guidance material into more compact form than the large collections mentioned in Chapter 1. As made clear in that Chapter, the inspiration for this process was the set of "sensitive dimensions" devised by Marshall et al. (1987). In turn, the source for that set were the collected guidelines from cognitive psychology contained in that book (Gardiner & Christie 1987). As was also made clear, the sources for the author's principles set have widened to include a range of principles and heuristics including Nielsen (1993), Scapin & Bastien (1997) and ISO 9241 Part 10 (1996). A common feature of these and the other sources is that they offer sufficient rationale to allow an evaluator to weigh up and generalise their guidance rather than apply it by rote.

The issue for this Section is the extent to which the results from these experiments are evidence that evaluation materials derived from such sources as the above have genuine explanatory and predictive power. The question is not whether principles are better than

heuristics because they offer more background rationale, but whether that rationale is sound in the first place. In short, the issue is whether the principles have a solid empirical base.

A test of the 'general rules which can be applied to a range of interface styles and design issues' (Chapter 1) would imply assessment of the explanatory and predictive power of at least the core principles from the author's set against some known user-system results. This might mean the tracking of such a core set of issues against problems raised in practice over a long period. This is clearly beyond the scope of these experiments, including Experiment 3. The John & Mashyna (1997) and John & Marks (1997) studies remain the single example known to the author of an attempt to assess the effectiveness of changes made as a result of different evaluations.

Even though most of the sources cited in the author's principles set (see Chapter 1, Table 1.5) are based on solid empirical grounds, these are still secondary sources, and the width of the trawl is no guarantee of the quality of the catch. The principles set shares with heuristics such as Nielsen's a certain amount of 'backwards rationality', that is, the search for justification in the literature for material already presented. However, the tying of every assertion in the 30 attributes (sub-principles) to an empirical base would require more work than the author is presently able to undertake. Thus the principles set is not yet a proper exercise in cognitive ergonomics as described in Chapter 1 (Section 5.5).

2.2 Cumulative Problem Curves

Chapters 4 onwards have demonstrated a discrepancy between the cumulative performance of the evaluators used in Experiments 1 to 3 and those in much of the UEM literature. On all but the very contrived Rule level tasks in Experiment 3, both detection rates and hit rates were lower than those found elsewhere. The numbers of experienced subjects in Experiment 1 required to uncover 75% of both all-and high-severity totals were much greater than the 5 claimed. The corresponding numbers of novice subjects in both Experiment 1 and Experiment 3 Skill and Rule tasks were outside the limits claimed for novices. However, novices in Experiment 2 were within these limits, as were Experiment 3 Heuristic subjects at the Skill level and (on high-severity problems) at the Knowledge level. It was shown that in Experiments 1 and 2 curves of the form claimed could be generated by counting by problem categories (types) rather than UPTs (instances). This had the effect of reducing problem totals sufficient to push up detection rates to within the necessary 33% (or the 25% required by the model on which the '3 to 5' claim is based).

When differentiation was made in Experiment 3 between prediction rate and hit rate, it was revealed that detection rate alone is an unreliable measure of predictive ability, and that the collective performance of subjects in Experiments 1 and 2 may have been even lower than previously envisaged. However, it was also shown that Experiment 3 subjects tended to over-predict relative to correct predictions, and that the curves from the first two experiments

may have masked a considerable number of false positives. The similarity between problem distributions at the less contrived Experiment 3 Knowledge level with those from Experiments 1 and 2 implied that this performance in the first two experiments (on 'open-ended' tasks) was not unrepresentative of what might be seen on 'closed' tasks. Re-analysis of the earlier ticket vending machines study revealed that only on the smaller and simpler machine (the FFM) were detection rates for observed errors comparable to those claimed elsewhere (i.e. high enough for a '3 to 5'). In this light it was suggested that the Experiment 3 Test condition total (of observed problems) might have higher (and hit rates for Heuristic and Principle subjects still lower) had additional Test subjects been used.

2.2.1 Internal Validity

Internal validity issues concern the three measures (hit rate, accuracy and redundancy) used to assess predictions against observed problems, and the differentiation made between hit rate, prediction rate and detection rate as measures of collective performance

2.2.1.1 Hit rate, Accuracy and Redundancy

The distinction between predicted and observed problems in Experiment 3 allowed differentiation between hits (correct predictions), misses (problems observed but not predicted) and false positives or FPs (problems predicted but not observed). This offered a wide variety of potential ratio (percentage) measures, including hits to misses, FPs to observed and FPs to hits. The chosen measures were hit rate (% hits to observed problems), accuracy (% hits to predicted problems) and redundancy (% FPs to predicted problems). The distinction between hit rate and accuracy allowed assessment of the difference between correct prediction and the effort expended in making those predictions. In terms of the usability criteria in ISO 9241 Part 11 (1998), hit rate and accuracy may be seen as, respectively, measures of the effectiveness and efficiency of the predictions arising from an inspection. Redundancy was chosen for its ability to assess over-prediction in relation to total predictions. However, it later became apparent that accuracy and redundancy are not independent, and that the one can be derived from the other (accuracy=1-redundancy). So redundancy is, well, redundant.

An alternative measure of over-prediction might be 'wasted effort' (%FPs to observed problems). However, it is not clear to this author whether this is also derivable from the other measures. The good news is that hit rate and accuracy do discriminate between effectiveness and efficiency in the predictive process. This author is more confident in offering these as valid dimensions, though there may well be others. The second inter-rater reliability test in Experiment 3 showed good correlation between the author's predicted-observed matchings and those of two independent raters. Sears (1997) used similar measures, respectively 'thoroughness' and 'validity' for hit rate and accuracy, plus a third measure, 'reliability' (standard deviation of hits / mean hits). To the author's knowledge, the third measure has not been employed in other UEM studies.

2.2.1.2 Hit Rate, Prediction Rate and Detection Rate

In this thesis differentiation has been made between hit rate (% correct predictions vs. observed problems), prediction rate (% total predictions vs. observed problems) and detection rate (% problem rate per population). Plotting these as cumulative measures enabled comparisons to be made between the combined performance of the subjects in each experiment. This revealed that detection rate as used in Experiments 1 and 2 may have masked a considerable tendency to over-predict, more reliable measures being prediction rate and hit rate. It was also shown that with the exception of the very contrived Rule level task, subjects in Experiment 3 were not reaching targets for correct problems as revealed in the Test condition (though at the Skill level Heuristic subjects did manage to reach 75% of predicted problems within novice limits).

The distinction between prediction rate and detection rate has shown that different results can be achieved according to the choice of denominator in such ratio measures. As will be shown in the next Section, many studies in this domain have offered detection rate as the yardstick of UEM effectiveness. Whether or not evaluators do over-predict (and the implication from Experiment 1 is that this is true of experienced subjects as well as novices), it is clear that detection rate is seriously flawed as an effectiveness measure. The next Section will also show that there has been considerable variation in this reported detection rates between studies, implying that this measure is both invalid and unreliable.

Figure 5.5 of Chapter 5 showed that cumulative curves of prediction rate ran consistently outside of their theoretical (probabilistic) equivalents (as determined from the height of the y-intercepts), while those for hit rate ran inside. The latter is the form predicted by the model used by Nielsen & Landauer (1993) as the basis for the '3 to 5' claim, replicated by Virzi (1990, 1992). The reasons why hit rate should conform better to the model than prediction rate are, for the moment, unclear. The author's hypothesis is that the uneven distribution of UPTs which is typical of total predictions (as seen in Experiments 1 and 2) is more likely to produce lines which exhibit little curvature (thus running outside of the ideal), while the distribution of hits is more likely to be evenly spread. This hypothesis is so far untested.

However, it *is* clear from Experiments 1 and 2 that the simple expedient of counting by problem type rather than instance enabled those curves to be brought inside of their probabilistic equivalents (see Chapter 4, Figure 4.6), thus resembling the curves for hit rate. Figure 4.4 of the same Chapter implies that the correspondence between actual (permuted) and ideal (probability) curves on which the '3 to 5' is based begins to break down when detection rates (and p values) are smaller than the minimum of 25% required by the theoretical model. Figure 4.4 suggests that this is so for detection rates of around 10%, though the actual limits remain to be established. The discussion of probability theory in Chapter 4 also suggested that the factors determining the shape of cumulative curves include the available number of evaluators (out of which total UPT figures are forged). Such

cases as the permuted curves in Figure 4.4 then represent *truncated* versions of those which would result from (very) much larger subject numbers.

If the above argument is correct, it appears that the curve fitting on which the theoretical model introduced in Nielsen & Landauer (1993) was based depended on the very same assumptions that are required for the '3 to 5' itself; that is, that both detection rates and UPT totals have to be *already* relatively high in order for empirical curves to fit the model. Once more, then, the low detection rates (and large problem totals) exhibited by Experiments 1 and 2 have served to identify some limits to the correspondence between actual and theoretical curves on which the '3 to 5' depends. The differences between these curves now comes be seen as a consequence of many factors, only one of which is the distribution of UPTs among evaluators. If this is so, then the above hypothesis is much more complex than first envisaged.

2.2.2 External Validity

External validity issues concern observations as a datum for problem prediction, and detection rate as measured in other UEM studies.

2.2.2.1 Observed Problems

It was shown in Experiment 3 that novice subjects consistently over-predicted compared to correct predictions (hits), and that distributions of predicted problems were similar to those from Experiments 1 and 2. It is reasonable to conclude, therefore, that subjects in all three experiments were predicting or reporting usability issues which did not and would not be observed in practice. If this is a general tendency of subjects in UEM studies, or more widely, of evaluators in usability inspections, then it deserves comment.

As was said in Chapter 5, while subjects might be expected to attempt to be as accurate (make as few misses) as they can, the reasons for over-predictions (false positives) are more difficult to identify. More particularly, if FPs are an artefact of the experimental process, but occur to a lesser extent in the 'real' world (of product inspection and testing), then this is bad news for laboratory-based studies of inspection methods.

The differentiation between hit rate and prediction rate is one means of assessing the degree to which subjects over-predict. In Experiment 3 this was assessed in relation to observed problems in the Test condition. However, comparison with the ticket vending machines results in Chapter 6 implied that even such a large number of observations as this might not exhaust the variety of problems to be found with simple interfaces and highly constrained tasks, let alone the complexity of desktop GUIs. Thus the use of observed problems as a datum for problem prediction may also be invalid unless large numbers of test subjects are involved.

Without such a datum it will be difficult to assess the true degree of over-prediction inherent in usability inspections. Yet Nielsen (1994b) once more estimated the number of non-specialist think-aloud test users required to report 75% of problems to be around 4 or 5, and it is rare for user tests to run more than 7 or 8 (Rubin 1994 (p93) recommends 4 to 5 for "less formal tests", and 10 to 12 for "true experimental design"). Jacobsen et al. (1998a, 1998b) have shown that the 'evaluator effect' applies as much to multiple evaluators (of user tests) as it does to multiple users (as test subjects). Thus it is possible that numbers of both test subjects and test observers required to identify 'most problems' have been underestimated in the same way as for evaluators in usability inspections.

The evidence from studies that have measured it (e.g. Bailey et al. 1992, Mack & Montaniz 1994, Cuomo & Bowen 1994, Sears 1997) is that even experienced subjects over-predict. The above discussion implies that it may be difficult to verify the degree of over-prediction, in this thesis as elsewhere. As to whether false positives should be accepted as valid usability data, the author's view is that all reported problems should be recorded as valid until demonstrated to be noise[4]. In the absence of validated test data, some means of prioritisation (such as frequency x severity) may be used. Virzi (1997) (p709) says that both observed (test) problems and FPs are valid data, and should be considered in re-design. Nielsen's (1994d) view of problems which would be *impossible* to find by user testing (thesis author's italics) is that they are "still usability problems" (p46). The richness of the subject protocols from Experiments 1 to 3 (see Appendix J for a sample) implies that even novice users have much to contribute when not limited by task constraints.

2.2.2.2 Detection Rates

It has been shown that detection rates are an unreliable indicator of the effectiveness of inspection methods in that they mask figures for hit rate and prediction rate. It is surprising, then, that most of the UEM studies such as reviewed in Chapter 1 have not attempted to measure hit rate, relying on the false assumption that the combined number of problems reported or predicted by evaluators represents the sum total of a method's effectiveness. Where hit rates have been assessed, they (like detection rate) are difficult to verify without a sufficiently detailed account of the problem reduction and matching procedures on which they are based. (Section 1 of Chapter 5 outlines some studies which have measured hit rate).

Table 8-1 presents a summary of the detection rates and hit rates from the studies in this thesis, including the ticket vending machines (TVMs). In the latter case, detection rates refer to the observed problems (errors) per subject population, while hit rates are based on predictions made by the author in an error analysis.

4 Noise in Appendix H (Experiment 2 problem listing) included categories for "useless" and "picky".

	Experiment 1			Expt. 2	Experiment 3 (Novice)			TVMs		
	Nov.	Exp.	All	Nov.	Skill	Rule	Knowledge	FFM	MFM	QF
Detection rate (%)	7.7	10.1	5.6	8.4	26.7	32.9	24.7	20.0	9.7	10.4
Hit rate (%)	-	-	-	-	19.2	63.1	11.2	25.0	56.3	-

Table 8-1. Detection rates and hit rates from Experiments 1, 2 and 3, all problems, and the ticket vending machines study. Experiment 3 figures are means from the Heuristic and Principle conditions. Nov = Novice, Exp. = Experienced, Expt. = Experiment, TVM = Ticket Vending Machine, FFM = Few Fare Machine, MFM = Multi-Fare Machine, QF = QuickFare machine.

If hit rates are a true representation of predictive performance, it is clear that the novices in Experiment 3 were managing to conform to the requirements for a '3 to 5' only on the very contrived Rule level tasks (while failing to reach predictions for novices at all but the Skill level). In particular, performance at the more realistic Knowledge level was well below novice predictions. It is equally clear that detection rates alone would have offered a distorted view of novice performance at all but the Rule level. Since problem distributions in Experiments 1 and 2 were similar to that at the Knowledge level, it is likely that detection rates in these experiments mask lower figures for hit rate. The low detection rates on the larger of the three vending machines (MFM and QF) are a consequence of the larger number of observed errors on these machines compared to the uniformly low error incidences on all three machines (see Table 6-2 of Chapter 6). The higher hit rates on the FFM and MFM were derived from a single evaluation session and have not been independently assessed.

Chapter 4 demonstrated that the detection rates from Experiments 1 and 2 could be raised by the expedient of counting problem types rather than instances. This suggests that studies that have achieved higher detection rates on tasks and systems of comparable scope to those in Experiments 1 and 2 were using an approach to problem identification and reduction which is at variance with that of the author. In order to assess this would require access to the sort of low-level data and procedural details which are not generally available. However, just such a comparison has been undertaken by Morten Hertzum and Niels Jacobsen in a paper (Hertzum & Jacobsen 2001) published subsequent to time of writing. In this paper, Hertzum & Jacobsen extend previous work by Jacobsen et al. (1998a, 1998b). In the latter two papers, Jacobsen et al. demonstrated that evaluators of user test sessions showed substantial disagreement in both their problems identifications and their nominations of severe problems. This disagreement was confounded by the differences between test users in problems exhibited. Jacobsen et al. characterised the resulting 'evaluator effect' (of user test sessions) as an interaction between evaluators and users. For the later (2001) paper, Hertzum & Jacobsen have compared the detection rates from 11 published UEM studies including Jacobsen et al. (1998a) and Connell & Hammond (1999) (which summarised Experiments 1 and 2). The data collected is reproduced (by permission of Morten Hertzum) in Table 8-2, which here includes the results from Experiment 3.

Reference	UEM	Evaluated system	Task scen-arios	Total UPTs	Evaluators	[Exp. or Nov]	Detn. rate, all probs. [%]	Detn. rate, sev. probs. [%]	Any-two agree-ment [%]
Lewis et al. 1990	CW	Electronic mail	Yes	20	4 (3 CW developers and 1 CE+ novice)	Exp	65	-	-
Dutt et al. 1994	CW	Personnel recruitment	Yes	32	3 (2 CS grad. students and 1 HCI researcher)	Nov	73	-	65
Hertzum & Jacobsen '99	CW	Web-based library	Yes	33	11 CS grad. students	Nov	18	21	17
Jacobsen & John 2000	CW	Multimedia authoring	No	46	2 CS grad. students	Nov	53	-	6
Nielsen & Molich 1990	HE	Savings	No	48	34 CS students	Nov	26	32	26
		Transport	No	34	34 CS students	Nov	20	32	-
		Teledata	No	52	37 CS students	Nov	51	49	-
		Mantel	No	30	77 computer professionals	Exp	38	44	45
Nielsen 1992	HE	Banking	No	16	31 novices (CS students)	Nov	22	29	-
					19 usability specialists	Exp	41	46	33
					14 double specialists	Exp	60	61	-
Nielsen 1994d	HE	Integrating	Yes	40	11 usability specialists	Exp	29	46	-
Connell & Hammond '99	HE	Hypermedia browser *	No	33 [11]	8 undergrads.	Nov	18	19	9
				84 [26]	5 HCI researchers	Exp	24	22	5
		Interactive teaching **	No	57 [34]	8 psychology undergrads.	Nov	20	16	7
Expt. 3 of this thesis	GR	Word and Excel	Yes (Skill)	N/A	8 psychology undergrads.	Nov	27	-	16
			(Rule)	N/A			33	-	37
		Excel	(Knowl-edge)	N/A			25	-	29
Jacobsen et al. 1998a	TA	Multimedia authoring	Yes	93	4 HCI researchers	Exp	52	72	42
Molich et al. 1998	TA	Electronic calendar	No	141	3 commercial usability labs	Exp	37	-	6
Molich et al. 1999	TA	Web-based email	No	186	6 usability labs	Exp	22	43	7
		* Expt.1 ** Expt. 2		[ ] severe problems		Mean	35.10	38.00	22.45

Table 8-2. Table 1 of Hertzum & Jacobsen (2001), reproduced with permission. Data from Connell & Hammond (1999) refer to Heuristic condition subjects from Experiments 1 and 2 (labelled HE in the paper); detection rates and any-two agreements for these subjects are calculated from population totals, and may differ slightly from the table when published. Column 7 does not feature in the original table. Data from Experiment 3 are averages of Heuristic and Principle conditions. Cited sources feature in the References (data from Dutt et al. 1994, Connell & Hammond 1999 and Molich et al. 1999 made available to paper authors). CW = cognitive walkthrough, HE = heuristic evaluation, TA = think-aloud. GR = guideline review. Expt. = experiment, grad. = graduate. Exp = experienced, Nov = novice. Detn. rate = detection rate. Probs. = problems. Sev = severe.

Hertzum & Jacobsen have made a significant contribution to measuring evaluator agreement in their introduction of **any-two agreement**. Any-two agreement is the number of UPTs two evaluators have in common divided by the number of UPTs they collectively detect, averaged over all possible pairs of evaluators. Thus it is a measure of the amount of agreement between evaluators in a population, regardless of the total number of unique problems detected by that population. It encapsulates in a single figure the 'problem overlap' which distributions such as Figures 2.3(a) and 2.3(b) of Chapter 2 attempted to depict. Any-two agreement might also be a better measure of the collective performance of a group of evaluators than the ratio measures (detection rate, hit rate and prediction rate) used in this thesis. However, the relationship (if any) between these measures and any-two remains to be determined. For the moment, any-two and detection rate can both be used to compare the results from Experiments 1 to 3 with the rest of Table 8-2.

Clearly, detection rates from the smaller subject pools (Heuristic condition subjects) in Experiments 1 and 2 ('Hypermedia browser' and 'Interactive teaching') are nearer to the other studies than those in Table 8-1 (if still insufficient for a '3 to 5', particularly for high-severity problems). This treatment is consistent with the rest of the table except for the three subject groups in Nielsen (1992). In contrast, the any-two means for these populations are among the lowest in the table. The author's interpretation is that the any-two figures better reflect the distributions of problems (and lack of agreement between subjects) in these experiments than does detection rate alone. As for Experiment 3, it is equally clear that both detection rates and any-two are more in line with the other studies. In particular, the low any-two for predictions at the Skill level compared with the Knowledge level is interesting in view of the better cumulative hit rates at the Skill level (Chapter 5, Figure 5.5).

Though Hertzum & Jacobsen caution against between-study analyses, some interesting comparisons emerge when data are partitioned by (presumed) experience level and whether or not task scenarios were used. As for experience level (Nov or Exp), detection rates for novices in Experiments 1 and 2 and experienced in Experiment 1 are almost but not quite the lowest, whereas those for Experiment 3 novices are not far from the mean (of 32%). Of the studies that used defined tasks, the Experiment 3 detection rates are not the lowest, while those for Experiments 1 and 2 (with open-ended tasks) are the lowest but for the Experiment 1 Experienced Heuristic subjects. These comparisons suggest that the effect of using defined tasks in Experiment 3 was to raise novice detection rates (but not, as we have seen, hit rates) uniformly, while even the detection rates for the smaller populations in Experiments 1 and 2 were lower than other studies without task scenarios. Experienced subjects in Experiment 1 still managed higher detection rates than the novices, even with very low any-two agreement.

Hertzum & Jacobsen's view is that the collected data is evidence of a substantial evaluator effect, even for more formal procedures (c.f. heuristic evaluation) such as cognitive

walkthrough and user testing (with think-aloud). This author's view is that detection rates of less than 50% need to be seen in context of the 'mere' 33% required for a '3 to 5'. In that respect, Experiments 1 to 3 are not as far removed from the other studies as they might appear, though the effect of introducing defined tasks was to make a considerable difference to novice performance. What is clear is that the evaluator effect is real, and that studies which rely on either inspection (by evaluators) or testing (of users) alone will fail to capture the full range of usability issues.

3. Final Remarks

The claim that 75% of problems can be found with only 5 evaluators was shown to have been supported for (a) limited tasks and functional scope, and (b) higher-level problem categories, thus setting some boundaries on what can be claimed for heuristic evaluation and guideline review. The implication from this thesis that this sort of inspection may be a more expensive and lengthy process than has been portrayed, and (as might have been expected, and as Nielsen himself has said) needs to be combined with other empirical approaches such as user testing. Alternatively, open-ended inspections might be combined with other more clearly task-based methods such as cognitive walkthrough. However, the effect of task constraint may be to limit the scope of the issues to be uncovered.

The author's view is that restricting tasks and system functionality in evaluation may be to restrict the range of issues that will be uncovered. Of course this is necessary when testing for specific aims or specific functions (for example, new or changed), but to make such claims as "there were n problems with this interface" is to underestimate the potential for what evaluators, even novices, can be encouraged to identify. This goes much wider than mere 'tasks', towards the issue of utility versus usability introduced in Chapter 1. The underlying rationale for this thesis is that it should be possible, with suitably judged principle-based material, to encourage evaluators appropriate to the software to identify more usability issues than they would using simpler guides such as heuristics. So far this has proved to be possible on both open-ended and constrained tasks, but not with both experienced and novice evaluators (and, for the latter group, only partially so). Novices, it seems, do not have the confidence (or maybe just the experience) to go beyond defined tasks and materials subsets.

Yet the richness of even the novice reporting in these experiments belies the belief that it is not worth the attempt to assess both usability and utility. In Experiment 1, one of the problems reported with the software evaluated (the art gallery simulation) were that "It is not clear what the system is for, or why users need it" (Appendix G). At that time (1996) a CD-ROM version (published by Microsoft and called "Art Gallery") was available. The CD-ROM package was usable, but it has been withdrawn. No doubt it was extensively tested: could

an average user find a painting by a particular artist, or learn about types of painting ? - yes, it had a search engine (not in the original) and a simple 4-category index; could a user find his or her way around the historical/geographical atlas ? - no, but it had a search engine and a simple 4-category index; could a user look at the many other paintings by featured artists which are not in this particular collection (The National Gallery, London) ? - no (but there are lots of art books).

In one of the rejoinders to Gray & Salzman's (1998) critique of UEM studies, Andrew Monk wrote that "Experiments should be employed to answer small questions, and HCI researchers should not be embarrassed about the limited scope of the conclusions they draw" (Monk 1998 p302). This thesis started out to achieve some closure on a single small question, namely the scope of evaluation materials in guideline review. That the answers should have touched on so many larger questions (usability vs. utility, inspection vs. observation, cumulative problem curves, probability theory, problem prediction vs. problem matching) bears out the difficulty in HCI (as in applied psychology) of finding issues which are small enough to be addressable yet still worth undertaking. Some of the embarrassments - the lack of audio and video, the doubtful severity assessment, the reliance on novices, the redundancy of redundancy, the omission of Experiment 1 version B - may, however, be offset by the insights - the thought experiment (lottery balls), the effect of problem types on cumulative curves, the task levels manipulation, hit rate vs. detection rate, Hertzum & Jacobsen's any-two agreement, even the principles set.

Two limited propositions arise from this thesis. Like all propositions, they require falsification.

1. Expanded principles materials such as the author's can be used by both experienced and novice evaluators to identify more observable usability problems than shorter heuristics such as Nielsen's.

2. In usability inspections using guidelines or heuristics, even experienced evaluators will correctly identify smaller proportions of observable problems, and will disagree more among themselves as to the problems identified, than have been claimed.