Chapter 2: Experiment 1

1. Introduction

The aim of this first experiment was to show that both novices and experienced evaluators could make use of the author's set of evaluative principles, as then developed, to identify or uncover more usability problems than when using Nielsen's and Molich's ten heuristics (Molich & Nielsen 1990, Nielsen 1993, Nielsen 1994d) alone. The rationale was that heuristics such as Nielsen's seemed to be both very general in content and somewhat limited in scope, appearing to rely on evaluator expertise over content. A set of expanded and more focused principles might enable experienced evaluators to uncover a wider and more directed set of usability problems. It was thought that this might also be true of novices, if sufficient guidance was provided. Novices might thus represent a 'baseline' for the treatment of evaluation materials: if an effect for novices could be shown, such an effect for different experts would be easier to claim.

The focus in this experiment and its successors (Experiments 2 and 3) was on the materials used to generate problem accounts, rather than the procedures proposed for one or more evaluation methods. To this end, the method - intended to resemble Nielsen's descriptions of heuristic evaluation (Nielsen & Molich 1990, Nielsen 1993, Nielsen 1994d) - was held the same throughout each experiment. In the between-method comparisons summarised in Chapter 1, the main measure has usually been problem count. However, as Gray & Salzman (1998) point out, such comparisons are liable to criticism on considerations of construct validity if conclusions relating to whole methods are derived from a single dependent measure, be it problem count or any other. Any possible differences in the method of use of the evaluation materials compared in this thesis have therefore been left to future investigations. (However, as we shall see in Chapter 4, there is no guarantee that problem count is itself a reliable measure of the efficacy of a particular set of materials, let alone method.)

To this end, both novice and experienced subjects were asked to evaluate the same interface (here, a cut-down simulation of an existing system), some subjects using Nielsen's heuristics and some the principles set as then developed. Subjects were asked to 'think aloud' during their interaction with the system, their reactions and comments being recorded by the evaluator. A third, control, condition was also introduced, in which subjects were given no evaluation materials: it was hoped that this would represent a further 'baseline' for the materials, indicating the 'added value' which each might provide.

2. The Evaluation Materials

The author's full principles set and Nielsen's heuristics are described in Sections 3.3 and 3.5 of Chapter 1, respectively. Transcripts of the materials used in this experiment appear in Appendixes A and B.

2.1 Nielsen's Heuristics

(See Appendix A for a transcript as used in the experiment.)

In this and Experiments 2 and 3 the same version of Nielsen's heuristics was used. It was taken in full from Molich & Nielsen (1990), with the addition of heuristic 10, 'Help and Documentation', later included in Nielsen (1993) (p.20). Their titles[1] were as follows:

1. Simple and Natural Dialogue
2. Speak the User's Language
3. Minimize the User's Memory Load
4. Be Consistent
5. Provide Feedback
6. Provide Clearly Marked Exits
7. Provide Shortcuts
8. Provide Good Error Messages
9. Error Prevention
10. Help and Documentation

2.2 The Principles Set (Short Version)

(See Appendix B for a transcript as used in the experiment.)

In this experiment a short version of the author's then 26 principles set was used, covering just three pages (the version used in Experiment 2 ran to 31 pages and featured 23 principles). The derivation and format of the principles set are described in Chapter 1. The titles of the 26 principles (in normal typefont) are listed below. The seven groupings (in bold typefont) represent only one possible categorisation (but Chapter 4 presents a discussion of how problem *types* might later become conflated with problem *instances*). Title acronyms are as feature in the Results (Section 5). Principles marked * were those originally considered by the author to be most important on grounds such as have recently been characterised as effectiveness and efficiency (ISO 9241-11 (1998)).

Requirements and Functionality Principles

* Requirements match (RM)
* Functional Utility (FU)

[1] There are minor differences between the Molich & Nielsen (1990) and Nielsen (1993) titles. Those listed are as appear in the earlier paper.

User-System Principles

* Navigational Effort (NE)
  Memory Load (ML)
* Error Management (EM)
* Feedback (FE)
* Location and Navigation (LN)
  Choice Availability (CA)
  User Control (UC)
  System - User Match (SUM)

User Principles

* Modifiability (MO)
  Flexibility (FL)
  Appropriateness of Content (APC)
  Accuracy of Content (ACC)
  Salience (SA)

Comparative Principles

* Consistency (CON)
* Grouping and Linking (GL)

System Performance Principles

  Manipulability (MP)
* Responsiveness (RP)

Perceptual and Motor Principles

  Visio-perceptual Load (VL)
  Audio-perceptual Load (AL)
  Motor Load (MOL)
  Perceptual Clarity (PCL)
  Perceptual Contrast (PCO)

User Support Principles

  General Help (GH)
* Context-sensitive Help (CSH)

3. The Software

The software used for Experiment 1 was a cut-down simulation of the London National Gallery's public access hypermedia browser, the Micro Gallery. It was chosen because (a) the author already had available a near-complete version of the simulation, having developed it as a design exercise for the author's MSc in User Interface Design (Connell 1993), and (b) the Micro Gallery embodied aspects of the usability design issues which the author considers important and interesting (specifically, navigation and consistency).

The Micro Gallery System (hereafter referred to as the MGS) offers on-screen depictions of the whole of the National Gallery's collection of paintings and drawings, including historical and geographical information on the artists and works in the collection. The MGS is a touch-screen hypermedia system, allowing visitors to follow links and associations between different screen contents. The original MGS in-gallery installation was created by Cognitive Applications Ltd. and sponsored by the American Express Foundation. A CD-ROM version was subsequently created by Cognitive Applications and published by Microsoft® and the National Gallery under the title 'Microsoft Art Gallery'.

The MGS collection consists of a 'Painting Catalogue' of single-screen ('page') representations of the Gallery's paintings and drawings, plus secondary illustrations. Access to the Painting Catalogue is via four 'Top Level' categories or Sections (Artists, Historical Atlas, Picture Types and General Reference); from the top level the user may navigate 'downwards', through a different number of levels, to the paintings. Return to the Top Level Sections may be made at any time via an Index button. A large variety of navigational strategies are permitted, including 'horizontal' movements (within level), 'backward' movements (within, or back to the previous, level) and 'hyperjumps' (to related or associated pages).

The MGS simulation (hereafter referred to as the Simulation) used in Experiment 1 reproduced the main features of the National Gallery installation, but with very much reduced content (featuring in full just two artists). The remainder of the Simulation's pages offered a version of the MGS structure, duplicating the Index and Top Level Sections but incorporating only representative pages at lower levels. User input was via mouse and cursor rather than touch screen. Paintings, thumbnails and map illustrations were reproduced from 'Art Gallery'.

Figure 2.1 shows a representative page from the Simulation. It depicts the first of two pages in the Artists (A to Z) Section on the artist Degas. It offers biographical information and nine 'thumbnail' reproductions of the artist's paintings. Clicking on any of the shaded text items (here "media", "pastel" and "Impressionists") opens a 'pop-up' panel describing that term. Clicking on any of the thumbnails goes to the full-screen version of that painting in the Painting Catalogue. The icon (colour-coded) and label to the right of the title confirm that the page belongs to the Artists Section; the italicised text at bottom left of the page indicates that the next page (of these two pages on Degas), reached by clicking on the NEXT PAGE button, is about "Degas in Paris 1850-1900". Using the GO BACK button from that next page, or from the Painting Catalogue, would return to the page depicted; however, clicking on GO BACK from *this* (depicted) page would return to the source from which the page was reached: in this case likely to have been the page of the Artists A - Z listings which includes the entry for 'Degas'.

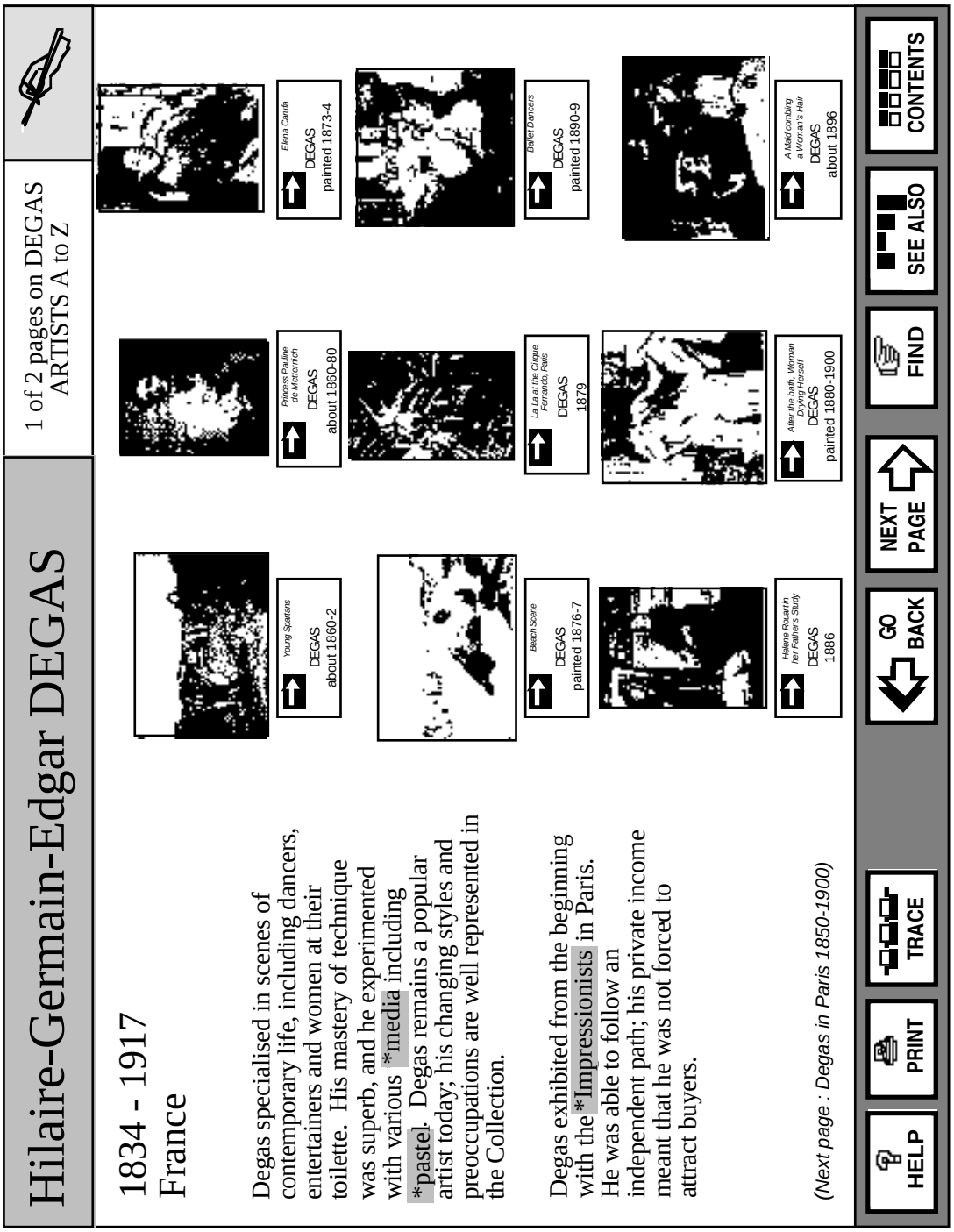


Figure 2.1. Experiment 1. Sample screen from the Micro Gallery System Simulation.

Other Simulation menu bar buttons include CONTENTS (a change of name from the MGS), allowing return to the Contents (Index) page (from which any of the four Top Level Sections can be reached); and SEE ALSO, which in the original MGS offers 'hyperjump' moves (via an intervening page) to any of the 16 pages (up to four per Section) which are related to the current page (the Simulation offered functional See Also moves from only three pages). In addition, the Simulation offered a new 'backtrace' or 'history' facility, by which users could jump up to eight moves back from the current page; this was accessed via the TRACE

button. A prototype HELP facility (with only indicative content) was included; printing (PRINT) and searching (FIND) were not implemented in the Simulation.

The Simulation was implemented in SuperCard® version 2.5, running on Apple Macintosh IIci under System 7.5. Window (SuperCard stack) size was reduced from the MGS's 19-inch full-screen to 240 x 190 mm.

In this experiment the intention had been to 'backwards engineer' a parallel version of the Simulation which included designed-in faults, and to compare usability problem counts for the 'poor' and 'improved' versions. To this end, two versions, A (improved) and B (poor) were produced. However, the time taken in recruiting experienced subjects meant that only version A was used. The consequences of this limitation for the conclusions which may be drawn from this experiment will be explored in the Discussion.

4. Method

4.1 Design

Experiment 1 was a 2 (group) x 3 (condition) between-subjects design. Subject groups consisted of novices (8 per condition) and experienced subjects (5 per condition); the three conditions were Control, Heuristic and Principle.

4.2 Subjects

Group status was assessed by a questionnaire, administered by the experimenter (the author) before the start of each experimental session. The questionnaire included items intended to measure both HCI knowledge and wider computing experience. Subjects scoring 0 to 20% inclusive on the questionnaire were deemed to be novices, while those scoring 30% and above were deemed to be experienced. The questionnaire was validated by adjusting the score weighting during the Experiment 1 Pilot study (see below), then refining to achieve 98% with a recognised HCI expert and 4% with a subject completely new to computers. See Appendix E for a transcript of the questionnaire.

Novices were mainly University of York Psychology undergraduates, others being recruited from Art History and Open University Summer School undergraduate courses. Most novice subjects were female and aged 18-21. Most were paid at hourly rates or received subject credit. Experiment 1 novice subjects' mean questionnaire score was 8.98% (median 8, standard deviation 3.93).

Experienced subjects were recruited from the University of York HCI research group. Most were male and aged 21-29. Some were paid at hourly rates. Experiment 1 experienced subjects' mean questionnaire score was 61.40% (median 64, standard deviation 19.98. The difference between novice and experienced scores was significant at the $p<0.0001$ level (one-tailed unrelated t-test).

In this experiment two post-pilot subjects' scores fell outside the novice experience range; these subject's results were not included in this or the succeeding analysis.

4.3 Procedure

Subjects were randomly assigned to either the Heuristic, Principle or Control conditions, according to the order of volunteering for the experiment. Heuristic and Principle subjects were asked to read the appropriate evaluation materials at the start of the session and were instructed to refer to the materials throughout the session. No time limits were imposed for the reading of the materials (Heuristic subjects took around one minute, Principle subjects up to five minutes). Subjects were informed about the simulated nature of the software, and were asked to act as visitors to the National Gallery who had volunteered to comment on an early prototype version of the MGS. They were asked to 'think aloud', that is, to verbalise their comments and reactions, and were told that they might be reminded to do so.

The procedure for all subjects was first to answer a non-trivial question regarding one of the paintings in the Simulation (requiring subjects to find and use the 'See Also' facility), and then to explore until all of the Simulation's substantive material had been encountered. This was intended to resemble Nielsen's recommendation for heuristic evaluation, namely that evaluators should first run through the whole of an interface in relatively brief fashion, and then proceed in more detail (Nielsen 1993 pp158-159, Nielsen 1994d p29). In this experiment, the experimenter (the author) sat with each subject and wrote down his or her comments while she or he verbalised. Comments (or user-system interactions) were not otherwise recorded.

Great care was taken not to prime or direct each subject, but unclear or incomplete comments were more fully elicited. Subjects were allowed to proceed at their own pace and to decide for themselves when they had finished exploring. On completion, Heuristic and Principle subjects were instructed to look again at the evaluation materials and to add any comments which they might have missed; Control subjects were merely asked to add any further comments. The experimenter's account was then read back to the subject and any required changes made. Any duplication in the subject's account, that is, where the same issue had been raised in more than one way, was agreed with the subject. Each subject was finally asked to rate each issue in terms of its severity, from 1 (trivial, might be ignored) to 7 (serious, must be addressed).

The principal dependent measure in this experiment (as in Experiments 2 and 3) was to be the number of problems which were reported by subjects. The source of this measure was the individual subject protocols which resulted from each session. As we shall see in Chapter 4, this process - the extraction of usability problems from subject protocols - is by no

means a trivial issue. The issue of problem extraction and reduction (from several subject accounts to a single set) will become a major component of the second theme of this thesis.

4.4 Pilot Study

A pilot study had been carried out in order to try out the procedures and to validate the prototype questionnaire. This involved five subjects, whose results were not included in the analysis. During the pilot, changes were made to the questionnaire format and content, but not to the evaluation materials.

5. Results

5.1 Problem Counts per Subject

See Figure 2.2.

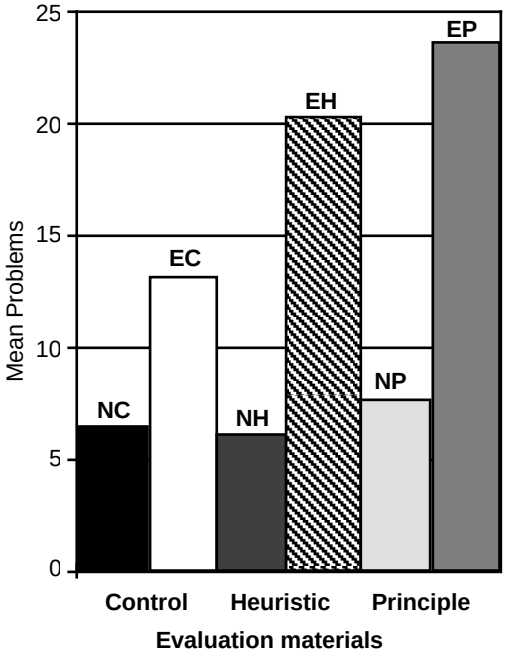
A two-way between-subject ANOVA of the complete problem count data (Table 2-1, Figure 2.2 (a)) showed a significant main effect of group (novices vs. experienced, $F[1,33]=71.42$, $p<0.001$). Further pair-wise (Newman-Keuls[2]) tests showed significant ($q_{0.05}$) differences between groups for the Heuristic ($W_2=8.42$) and Principle ($W_2=6.82$) conditions but not the Control condition ($W_2=7.85$). Thus significantly more usability problems (mean ratio 2.84) were uncovered by the experienced subjects than by the novices, both overall and in two out of three conditions.

Control		Heuristic		Principle	
Novice	Experienced	Novice	Experienced	Novice	Experienced
6.38 [2.83]	13.00 [6.44]	6.00 [2.07]	20.20 [6.57]	7.63 [3.70]	23.60 [5.50]

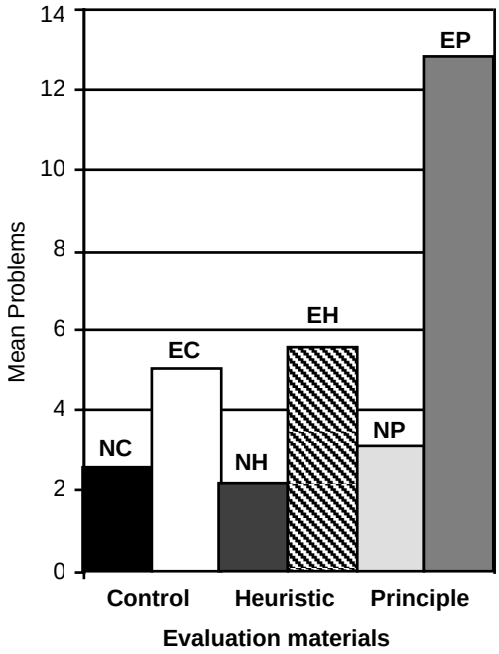
Table 2-1. Experiment 1. Mean problem counts per subject, all problems. Figures in [brackets] are standard deviations.

There was also a main effect of condition for the novice and experienced groups ($F[2,33]=5.60$, $p<0.01$), but no group x condition interaction ($F[2,33]=3.90$, $p=0.30$). Separate one-way ANOVAs showed no main effect of condition for novices ($F[2,21]=0.67$, $p=0.52$) but a marginally significant effect for experienced subjects ($F[2,12]=3.82$, $p=0.052$). A further post-hoc (Tukey HSD) test for experienced subjects showed significant ($q_{0.05}$) differences between only Principle and Control conditions. Thus while for novices the differences between evaluation materials had little effect on the numbers of problems identified, there appeared to be a limited (EP>EC) effect for experienced subjects alone.

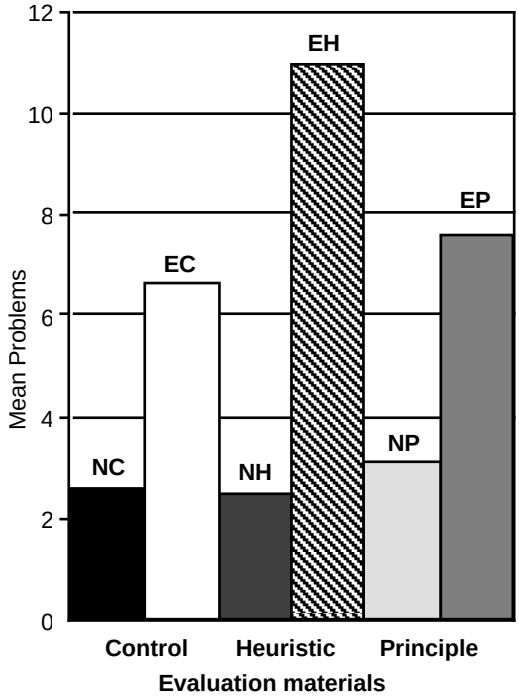
[2] Using the Games and Howell procedure for heterogeneous variances and unequal sample sizes (Howell 1997). Here and in subsequent analyses, this procedure will be used in place of the Tukey HSD test where there are wide differences in standard deviations.



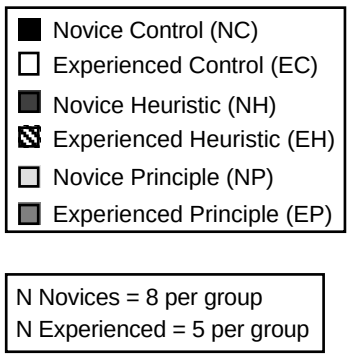
(a) Mean problem counts, all problems



(b) Mean problem counts, high severity problems



(c) Mean problem counts, low severity problems



N Novices = 8 per group
N Experienced = 5 per group

Figure 2.2. Experiment 1. Mean problems per subject.

However, when high-severity problems (rated by subjects as 5 to 7 inclusive on a 1 to 7 scale) alone were analysed, a more clear-cut view emerged (Table 2-2, Figure 2.2 (b)).

Control		Heuristic		Principle	
Novice	Experienced	Novice	Experienced	Novice	Experienced
2.50 [2.62]	5.00 [3.08]	2.13 [0.99]	5.60 [3.78]	3.13 [2.17]	12.80 [4.44]

Table 2-2. Experiment 1. Mean problem counts per subject, high severity problems. Figures in [brackets] are standard deviations.

A two-way between-subject ANOVA of high severity problems alone revealed a further significant main effect of group ($F[1,33]=31.67$, $p<0.001$). There was again a main effect of condition ($F[2,33]=8.94$, $p<0.01$), plus a group x condition interaction ($F[2,33]=5.88$, $p<0.01$). Separate one-way ANOVAs showed a main effect of condition for experienced subjects ($F[2,12]=6.50$, $p<0.05$), but not for novices ($F[2,21]=0.49$, $p=0.62$). A further post-hoc (Tukey HSD) test for experienced subjects showed significant ($q_{0.05}$) differences between Principle and Heuristic and between Principle and Control, but not between Heuristic and Control conditions (all $q_{HSD}[3,12]=3.78$, $W_3=6.44$). Thus experienced subjects had proved able to uncover significantly more high severity problems using the principles than the heuristics (and the Control), while novices remained unable to do so.

Similar analyses on the low severity (rated 1 to 3 inclusive) problems alone (Table 2-3, Figure 2.2 (c)) showed the expected main effect of group ($F[1,33]=28.34$, $p<0.001$), but not of condition ($F[2,33]=1.39$, $p=0.26$), with no group x condition interaction ($F[2,33]=1.82$, $p=0.18$). No further analysis is reported.

Control		Heuristic		Principle	
Novice	Experienced	Novice	Experienced	Novice	Experienced
2.63 [2.33]	6.60 [4.51]	2.50 [1.41]	11.00 [5.15]	3.13 [1.73]	7.60 [4.56]

Table 2-3. Experiment 1. Mean problem counts per subject, low severity problems. Figures in [brackets] are standard deviations.

This analysis has shown (a) a consistent effect of problem count for experienced subjects over novices, and (b) an effect of problem count for the principles set over Nielsen's heuristics (and the Control) for high-severity problems reported by experienced subjects.

5.2 Severity Ratings per Subject

Similar analyses were carried out on the mean severity ratings of problems (rated by subjects on a 1 to 7 scale). See Table 2-4. A 2-way ANOVA of all-severity problems showed no significant differences between either groups (novices vs. experienced, $F[1,33]=0.02$, $p=0.89$) or conditions ($F[2,33]=1.06$, $p=0.36$), with no group x condition interaction ($F[2,33]=0.30$, $p=0.74$).

Control		Heuristic		Principle	
Novice	Experienced	Novice	Experienced	Novice	Experienced
3.77 [1.26]	3.75 [1.17]	3.72 [0.82]	3.50 [0.87]	3.96 [0.72]	4.33 [0.66]

Table 2-4. Experiment 1. Mean severity ratings per subject, all problems. Figures in [brackets] are standard deviations.

When high-severity problems (rated by subjects as 5 to 7 inclusive) alone were analysed, a further 2-way ANOVA revealed a significant difference between groups ($F[1,28]=10.26$, $p<0.01$) but not between conditions ($F[2,28]=0.00$, $p=0.99$), with no group x condition interaction ($F[2,28]=0.89$, $p=0.42$). See Table 2-5. Similar analysis of low-severity problems (rated 1 to 3 inclusive) again showed no significant between-group ($F[1,32]=2.74$, $p=0.11$) or between-condition ($F[2,32]=1.06$, $p=0.59$) differences, nor a group x condition interaction ($F[2,32]=1.02$, $p=0.37$). See Table 2-6.

Control		Heuristic		Principle	
Novice	Experienced	Novice	Experienced	Novice	Experienced
5.38 [0.44]	5.81 [0.51]	5.19 [0.24]	6.04 [1.00]	5.44 [0.37]	5.76 [0.20]

Table 2-5. Experiment 1. Mean severity ratings per subject, high rated problems. Figures in [brackets] are standard deviations.

Control		Heuristic		Principle	
Novice	Experienced	Novice	Experienced	Novice	Experienced
2.20 [0.79]	2.26 [0.58]	2.69 [0.40]	2.18 [0.32]	2.54 [0.39]	2.18 [0.22]

Table 2-6. Experiment 1. Mean severity ratings per subject, low rated problems. Figures in [brackets] are standard deviations.

Thus while the differences between conditions had no overall effect on the severity ratings assigned to problems, experienced subjects appeared to be more critical than novices when assigning higher ratings. This result and the assignment of severity ratings will be taken up in the Discussion (Section 7.1) and later in Chapter 8.

5.3 Problem Tokens

5.3.1 Problem Distribution

The above analyses compared the total numbers of problems reported by each subject, regardless of the particular problems which were included in each subject's problem set, and of any overlap (shared problems) between subjects. While any duplications in individual subject accounts had been removed at the end of each experimental session (see Section 4.3), some problems were likely to be shared between different subjects while others were reported by a single subject alone. In this Section we consider the distribution of Unique Problem Tokens (UPTs) (Jacobsen et al. 1998a and 1998b) amongst both groups and conditions, examining the degree of overlap of UPTs between different subject populations. In any one comparison, a high degree of overlap between populations with differing problem counts would imply that any significant difference between them consists

mainly of UPTs not found by the group with the lower problem count (over and above those held by both); conversely, a low overlap implies that the higher and lower problem-count groups' UPTs are separate and distinct. (A model of the process by which a single set of UPTs, in which each unique problem or problem token appears only once, is generated out of the disparate problem sets arising from different subject protocols - the problem reduction process - will be presented in Chapter 4.)

Figures 2.3 (a) and (b) show the distributions of UPTs by group (Novices, Experienced) and condition (Control, Heuristic, Principle) for all and high severity ratings respectively. They were produced by determining the number of single-incidence (occurring only once in a group or condition) and shared-incidence (occurring in more than one group or condition) appearances of UPTs, between all groups and conditions and then between condition pairs. The amount of overlap in any case is indicated by the relative size of Both or All appearances compared to the single incidences, out of the total UPTs[3] in that case. Thus, for example (Figure 2.3 (a), By Group), 56.2% of the all-rated problems were reported by Experienced subjects only and 21.6% by Novices only, while 22.2% were reported by both groups.

It is clear that overall there was relatively little sharing of UPTs (maximum 22.2%, minimum 8.1%), between either groups or conditions. For example, we saw above that experienced subjects found on average 2.84 more problems than did novices, representing a maximum possible between-group overlap of 35.2%; thus while the experienced did find many of the same problems as the novices, there remained some 13% of their combined total UPTs which novices reported but experienced subjects did not.

Looking at both Figures 2.3 (a) and (b), there appear to be little differences between the overlap profiles by group, but some differences by condition. While the overall sharing between conditions for all and high-rated UPTs was similar at 8.1% (Figure 2.3 (a)) and 6.1% (Figure 2.3 (b)) respectively, as was Control only (by Condition) (22.3% and 21.5%), the relative difference between Principle only and Heuristic only appears higher for the high-rated UPTs than for all UPTs (48.5% to 23.9% compared with 38.7% to 30.9%). This difference appears particularly reflected in the higher proportion of Principle only to Heuristic only UPTs (59.5% to 26.4% compared with 45.4% to 33.6%), and (not illustrated) Experienced Principle only to Experienced Heuristic only UPTs (ratios of 2.37 to 1 compared with 1.2 to 1). Thus Principle subjects (particularly Experienced Principle subjects) appear to have reported a greater proportion of high-rated single-incidence problems (compared with Heuristic and Control subjects) than all-rated single-incidence problems.

[3] Counting shared problems once only.

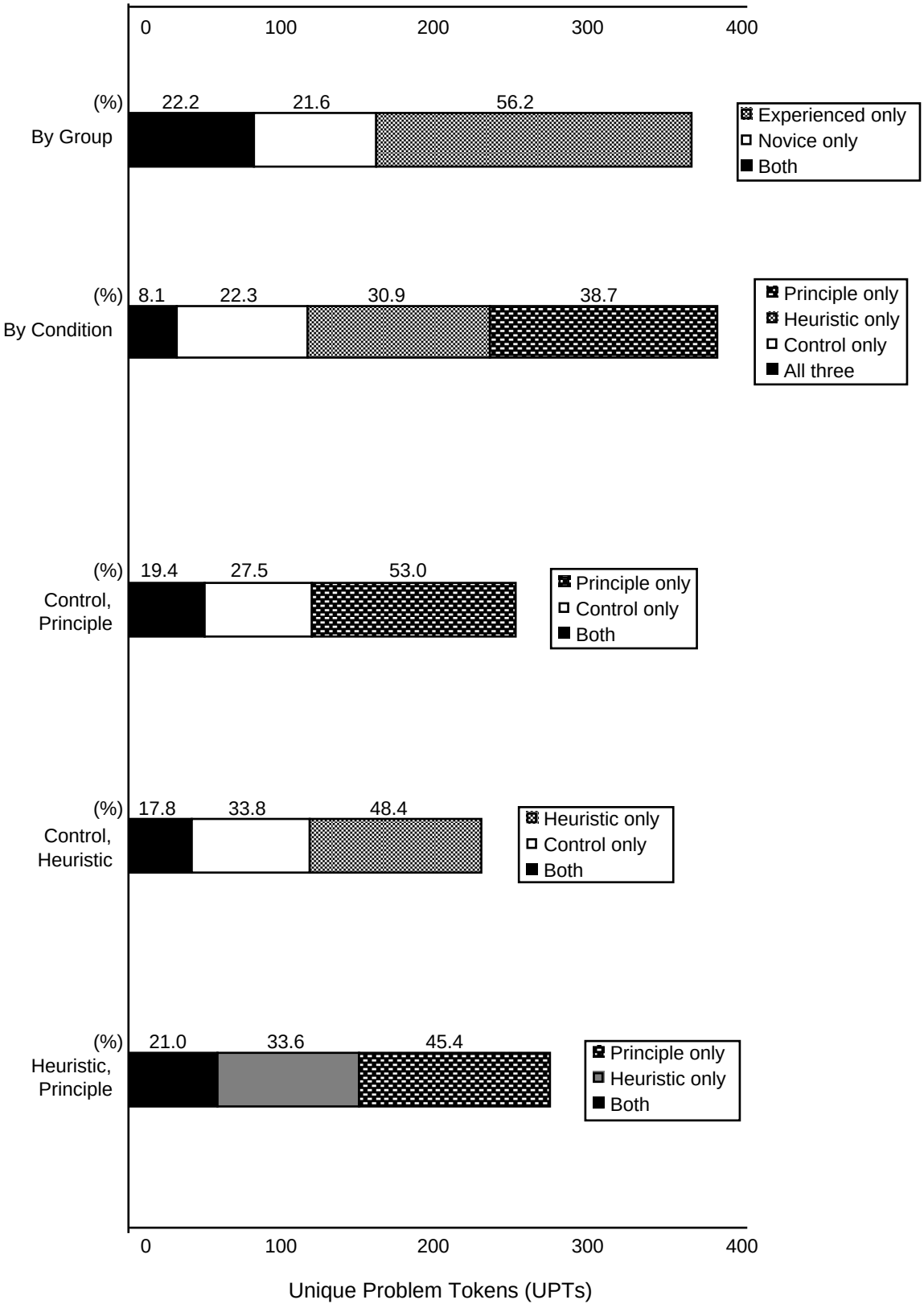


Figure 2.3(a). Experiment 1. Percentage problem (UPT) distributions for all problem ratings.

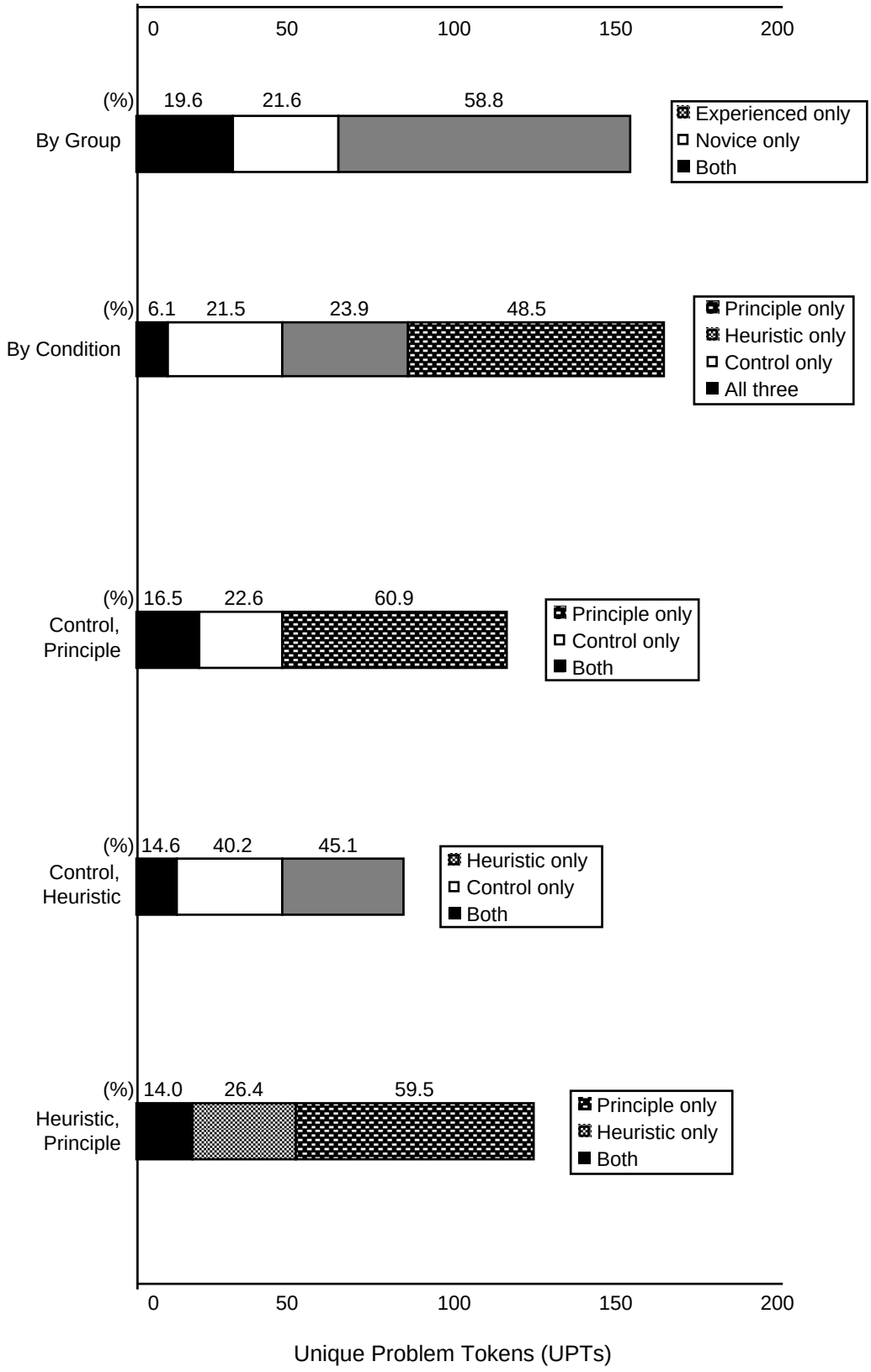


Figure 2.3(b). Experiment 1. Percentage problem (UPT) distributions for high severity ratings.

On *this* analysis, then, there appears to have been some qualitative difference between the Principle and Heuristic conditions (and Principle and Control) such that (a) most problems found by Principle subjects were different from those found by other condition subjects (in common with other between-condition profiles), and (b) these relative differences between

conditions were greater for high-rated problems. Thus while subjects using the principles were finding mainly different problems to those using the heuristics (and the Control subjects), the extent of that difference appears to be greater for high-rated problems and for experienced subjects.

5.3.2 Problem Listings

Tables 2-7 and 2-8 show the top ten percent of the full set of 230 UPTs, sorted by frequency of appearance and mean severity rating, respectively (see Appendix G for full listings). It will be seen that frequencies decline rapidly from a maximum of only 4.73% to 0.9%, while severity ratings fall little (and that all but one of the highest severity problems were reported only once). The skewed frequency distribution (70.4% of all UPTs appeared once, compared with 16.1% three times or more) is responsible for the low degree of overlap seen in the previous Section, while the severity distribution is reflected in the lack of significant difference between severity ratings in Section 5.2.

No.	Problem description	Freq.	%	Rtng.
64	Organisation [place within period] of History Atlas pages not understood	21	4.73	4.57
88	"Trace" not meaningful term	15	3.38	4.60
179	Slowness of system processes (page changes etc.)	15	3.38	4.15
211	Index pictures (and text) should be clickable (+ indicated)	15	3.38	3.17
82	"See Also" not meaningful term	12	2.70	4.25
156	Next Page behaviour inconsistent	10	2.25	4.45
201	Need for more obvious general help or introduction	8	1.80	5.33
219	History Atlas pages (in TLS): map names not clickable	7	1.58	3.87
30	Cannot return directly to source page of a set of pages, once further 'in'	6	1.35	5.50
86	"Top level", "Top Level Sections" not meaningful terms	6	1.35	3.44
89	Trace icon not meaningful/obvious	6	1.35	3.17
212	Pictures and captions / Index 'parts' could be (clickable) together	6	1.35	2.61
31	Next Page does not return to previously found 'next page'	5	1.13	5.00
32	Trace could be larger (more previous pages) : can't jump > 8 pages back	5	1.13	5
130	More information ('story') needed on (particular) works (etc.)	5	1.13	5
163	Cursor does not change over menu bar buttons	5	1.13	2.88
193	Italic text [on painting descriptions] hard to read	5	1.13	3.83
28	Go Back does not include Index or See Also pages	4	0.90	4.33
29	Can't go to/know of a previous page in a set unless just been there	4	0.90	5.00
62	Need 'already selected'/'visited' indicator (e.g. on Trace icons)	4	0.90	5
73	Need stronger/additional information on category (Section) changes	4	0.90	5
87	"Put away" [Index, See Also 'close' button] not meaningful term (or icon)	4	0.90	4.17
107	Not clear what 'shadow' on Go Back button indicates	4	0.90	3.67

Table 2-7. Experiment 1. Top ten per cent of full set UPTs, sorted by frequency of appearance. Freq. = frequency, Rtng. = mean severity rating.

No.	Problem description	Rtng.	Freq.
1	Not clear what the system is for, nor why users need it (!)	7	1
2	Too much art (for 'man in the street') (!)	7	1
6	Should be a map of whole gallery, organised by period	7	1
12	1st screen(s) should be more oriented to what (tasks) visitors want	7.0	1
14	No facility to search for terms/names	7	2
20	Prefer full-screen works, not painting plus text	7	1
22	Organisational layout too complex and specialised	7	1
50	General Reference (Top Level) should include selectable sub-types	7	1
58	"Busy" cursor should not still indicate after page changes	7	1
106	Next Page and Go Back icons not clear or adequate	7	1
112	Painting Catalogue should be part of system hierarchy, not separate	7	1
154	History Atlas navigation (time within place) not used elsewhere	7	1
155	History Atlas organisation [by time within place] not consistent with rest	7	1
168	Highlighting [enabled] effect on See Also button wrong way round	7	1
203	Should be able to use the system without help	7	1
217	Help Contents page : not clear what is 'active' [was only partly completed]	7	1
220	History Atlas search strategy not clear	7	1
5	Need for indication of where other works are (in other art galleries)	6	1
9	Need for 'suggested tour' of popular works in the gallery	6	1
15	General Reference should include names of works	6	1
19	First (Contents) page of Help not necessary (go direct to 'lower' pages)	6	1
34	Trace does not show all pages 'visited' (unwinds on Go Back)	6	1
37	Trace does not include Index/Contents page	6	1

Table 2-8. Experiment 1. Top ten per cent of full set UPTs, sorted by mean severity rating. Rtng. = mean severity rating, Freq. = frequency.

Table 2-9 shows the top ten percent of the same full set of UPTs, now sorted by the product of frequency and severity. (Note that in *this* case the severity distribution means that the product and frequency listings are little different.) Such a list might be used as a means of prioritising the usability problems found in a particular study, for the purposes of deciding which problems might be tackled first in any development programme. For the MGS (the simulation used in the experiment: see Section 3 and Figure 2.1) attention might be drawn to the Historical Atlas (problems 64, 221), terminology (problems 88, 82, 86, 87), system speed (problem 179), and, particularly, navigational issues (problems 156, 30, 31, 29, 28, 73, 62). The latter are in broad alignment with the detailed discussion of the CD-ROM version of the MGS ('Art Gallery') offered in Garzotto et al. (1995).

No.	Problem description	Prior.	Freq.	Rtng.
64	Organisation [place within period] of History Atlas pages not understood	96.0	21	4.57
88	"Trace" not meaningful term	68.7	15	4.60
179	Slowness of system processes (page changes etc.)	62.3	15	4.15
82	"See Also" not meaningful term	51.0	12	4.25
211	Index pictures (and text) should be clickable (+ indicated)	47.5	15	3.17
156	Next Page behaviour inconsistent	44.5	10	4.45
201	Need for more obvious general help or introduction	42.7	8	5.33
30	Cannot return directly to source page of a set of pages, once further 'in'	33.0	6	5.50
219	History Atlas pages (in TLS): map names not clickable	27.1	7	3.87
31	Next Page does not return to previously found 'next page'	25.0	5	5.00
130	More information ('story') needed on (particular) works (etc.)	25.0	5	5.00
221	History Atlas: featured places [described in text] not (all) on map	22.7	4	5.67
86	"Top level", "top level sections" not meaningful terms	20.7	6	3.44
29	Can't go to/know of a previous page in a set unless just been there	20.0	4	5.00
193	Italic text [on painting descriptions] hard to read	19.2	5	3.83
89	Trace icon not meaningful/obvious	19.0	6	3.17
28	Go back does not include Index or See Also pages	17.3	4	4.33
73	Need stronger/additional information on category (Section) changes	17.0	4	4.25
32	Trace could be larger (more previous pages) : can't jump > 8 pages back	16.7	5	3.33
87	"Put away" [Index, See Also ‘close’ button] not meaningful term (or icon)	16.7	4	4.17
212	Pictures and captions / Index 'parts' could be (clickable) together	15.7	6	2.61
62	Need 'already selected’/’visited' indicator (e.g. on Trace icons)	14.7	4	3.67
107	Not clear what 'shadow' on Go Back button indicates	14.7	4	3.67

Table 2-9. Experiment 1. Top ten per cent of prioritised UPTs, sorted on Frequency x Mean Severity Rating. Prior. = Priority, Freq. = frequency, Rtng. = mean severity rating.

5.4 Problem Types

After a single set of UPTs had been extracted from the combined subject protocols, each UPT was assigned by the experimenter to one of the 26 usability principles which made up the principles set as then developed (see Section 2.2 for a principles listing, and Appendix B for a transcript). Although subjects in the Control and Heuristic conditions were not exposed to the principles set, this categorisation by problem type was to enable comparisons to be made between conditions and groups according to the severity ratings which each broad problem type may have attracted (Section 5.4.1) and the issues which might have been involved in subjects’ problem reporting (Section 5.4.2). The validity of this approach, and problem categorisation generally, are taken up in the Discussion and later in Chapter 8.

5.4.1 Problem Severity and Problem Types

Some insight into the problem types (principles) which attracted higher severity ratings can be obtained by ranking the 26 types according to their mean severity rating by type (Table 2-10).

All		By Condition						By Group					
		Control		Heuristic		Principle		Novice		Experienced			
Type	Mean	Type	Mean	Type	Mean	Type	Mean	Type	Mean	Type	Mean	Principles	
CA	6	GH	6.00	***RM***	7	CA	6	***MO***	6	***RM***	6.25	***RM***	Requirements Match
MO	6	***RM***	5.50	FL	6	***MO***	6	GH	5.00	CA	6	***FU***	Functional Utility
FL	6	***FU***	5.05	MOL	6	***CSH***	6.00	***FU***	4.75	FL	6	***NE***	Navigational Effort
RM	5.50	***CON***	4.71	***GL***	5	GH	5.00	***CSH***	4.50	UC	5	***ML***	Memory Load
GH	5.00	***NE***	4.00	***RP***	4.63	***NE***	4.73	***CON***	4.43	GH	5.00	***EM***	Error Management
GL	4.75	UC	4	ACC	4.50	SA	4.69	***FE***	4.17	***GL***	4.75	***FE***	Feedback
FU	4.58	***RP***	4.00	GH	4.50	***FE***	4.63	APC	4.10	***RP***	4.56	***LN***	Location and Navigation
CSH	4.50	***LN***	3.88	***FE***	4.17	***FU***	4.61	***LN***	4.05	***FU***	4.41	CA	Choice Availability
SA	4.34	PCL	3.75	***LN***	4.14	***LN***	4.51	***RM***	4	SA	4.34	UC	User Control
CON	4.25	None	3.59	***FU***	4.08	***GL***	4.50	YL	4.00	***LN***	4.30	SUM	System-User Match
LN	4.18	MOL	3.33	SA	4	***CON***	4.39	***NE***	3.94	***NE***	4.11	***MO***	Modifiability
RP	4.15	APC	3.18	YL	4.00	SUM	4.30	SUM	3.76	***FE***	4.08	FL	Flexibility
FE	4.12	SUM	3.02	SUM	3.77	***RM***	4	***RP***	3.75	***CON***	4.07	APC	Appropriateness of Content
NE	4.03	**FE**	3	***CON***	3.65	UC	4.00	UC	3.50	MOL	3.94	ACC	Accuracy of Content
UC	4.00	MP	3	***NE***	3.35	***RP***	3.83	None	3.25	None	3.69	SA	Salience
SUM	3.70	PCO	2.50	PCL	3.25	None	3.83	PCL	3.13	SUM	3.63	***CON***	Consistency
None	3.47	YL	2	MP	3	YL	3.33	PCO	2.50	PCL	3.42	***GL***	Grouping and Linking
MOL	3.46	ACC	1.00	**CSH**	3	APC	2.92	MOL	2	YL	3.11	MP	Manipulability
YL	3.33	**ML**		None	3.00	PCL	2.88	ACC	1.00	MP	2.33	***RP***	Responsiveness
PCL	3.30	CA	2	ML	2	MOL	2.25	**ML**		ACC	2.17	YL	Visio-Perceputal Load
APC	2.84	***MO***		APC	2.00	PCO	2	CA		APC	2.00	AL	Audio-Perceptual Load
MP	2.33	FL	2	CA		**ML**	1	FL	2	PCO	2	MOL	Motor Load
PCO	2.33	SA		UC		ACC	1.00	SA		**ML**	1.50	PCL	Perceptual Clarity
ACC	1.70	***GL***		***MO***		MP	1	***GL***		**MO**		PCO	Perceptual Contrast
ML	1.50	***CSH***		PCO		FL		MP		**CSH**		GH	General Help
EM		***EM***		***EM***		***EM***		***EM***		**EM**		***CSH***	Context-Sensitive Help
AL		AL		AL		AL		AL		AL		None	No match
Mean	3.79	Mean	3.64	Mean	4.05	Mean	3.81	Mean	3.78	Mean	3.94		

See Appendix B for principle descriptions. None = problem(s) not matching any principle. Principles in ***bold italic*** are those deemed to be more important.
No decimal place = sole appearance of a problem of this type reported. No data = no problems of reported. Cut (thick cell border) at mean rating >= 4.

Table 2-10. Experiment 1. Problems sorted by mean severity rating per principle (overall, then by condition and group).

It can be seen that with the exception of Error Management (EM), which did not appear to match any of the problems reported[4], all problems rated 4 and above could be assigned to one of the 11 of the (then) 26 usability principles which had been previously deemed by the experimenter to be most important (see Section 2.2). This is encouraging news for the principles set. However, it should be pointed out that comparable rankings appear only in the Principle and Experienced lists, implying that the placement of the 10 principles is due to the higher[5] severity ratings assigned by only experienced subjects exposed to the principles. For example, the highest ranked principle overall, Choice Availability (CA), occupies that position because of a sole rating by an Experienced Principle (EP) subject.

Thus, on *this* analysis, the 'added value' offered by the principles over the heuristics appears to be bound up in the willingness of EP subjects to assign higher severity ratings to the problem types which they reported.

5.4.2 Problem Distributions by Type

Further insight into the problem types (principles) which featured most often in subjects' reports can be obtained by ranking the 26 types according to their frequency of appearance by type (Figure 2-11).

In contrast to the mean severity ratings (Section 5.2), it can be seen that distributions are skewed in favour of a small number of problem types, and that the most important principles are not confined to the higher incidences. Taking an upper threshold of 10%, we can see that overall the most frequently reported types were System-User Match (SUM) and Navigational Effort (NE), along with those problems which could not be assigned to any of the principles (None). Two other principles, namely Location and Navigation (LN) and Consistency (CON), were also found atop the rankings by condition and group (with EP subjects' problem types being more widely distributed). The preponderance of these four principles (plus None) can clearly be seen when plotting the problem incidences by type (Figure 2.4), from which little evidence of an EP effect (Experienced > Novice, Principle > Heuristic or Principle > Control) can be detected. Further, the relatively high number of problems (14%) to which no principle could be attributed (None) implies that subjects (even experienced subjects: 16.5%) were finding problems without such guidance.

---

4 This is probably because no software bugs were encountered in this experiment, and because of the nature of the task involved (free exploration).

5 Between-subject ANOVAs of problems rated 4 to 7 show a similar EP effect to that for ratings 5 to 7 (a two-way main effect of condition, $F(2,23]=6.20$, $p<0.01$, and a one-way main effect for Experienced, $F(2,12]=4.22$, $p<0.05$).

All		By Condition						By Group					
		Control		Heuristic		Principle		Novice		Experienced			
Type	%	Type	%	Type	%	Type	%	Type	%	Type	%	Principles	
SUM	21.6	SUM	22.4	SUM	24.2	SUM	19.0	SUM	28.1	SUM	18.0	***RM***	Requirements Match
None	14.0	None	14.7	***COM***	15.4	***ME***	15.6	***LM***	14.4	None	16.5	***FU***	Functional Utility
ME	11.5	***LM***	11.2	None	14.8	None	12.8	None	9.4	***ME***	14.1	***ME***	Navigational Effort
LM	9.5	***FU***	9.5	***ME***	11.4	***LM***	10.1	***ME***	6.9	***COM***	11.3	***ML***	Memory Load
COM	9.5	APC	9.5	***LM***	7.4	***FU***	6.1	***FU***	6.3	***LM***	6.7	***EM***	Error Management
FU	6.1	***COM***	6.9	***FE***	4.0	***COM***	6.1	APC	6.3	***FU***	6.0	***FE***	Feedback
APC	4.7	***ME***	5.2	***FU***	3.4	PCL	4.5	***COM***	6.3	APC	3.9	***LM***	Location and Navigation
PCL	3.6	***RP***	3.4	APC	3.4	***RP***	3.4	***RP***	5.6	PCL	3.5	CA	Choice Availability
RP	3.4	PCL	3.4	***RP***	3.4	***FE***	2.8	PCL	3.8	***FE***	2.8	UC	User Control
FE	2.7	MOL	2.6	PCL	2.7	APC	2.8	***FE***	2.5	SA	2.1	SUM	System-User Match
GH	2.3	GH	2.6	GH	2.0	SA	2.8	GH	2.5	***RP***	2.1	***MO***	Modifiability
ACC	1.6	***RM***	1.7	ACC	1.3	GH	2.2	**CSH**	1.9	MOL	2.1	FL	Flexibility
MOL	1.6	ACC	1.7	YL	1.3	ACC	1.7	UC	1.3	GH	2.1	APC	Appropriateness of Content
SA	1.4	PCO	1.7	***RM***	0.7	YL	1.7	ACC	1.3	ACC	1.8	ACC	Accuracy of Content
YL	1.4	***FE***	0.9	**ML**	0.7	MOL	1.7	PCO	1.3	YL	1.8	SA	Salience
RM	0.9	UC	0.9	FL	0.7	UC	1.1	***RM***	0.6	***RM***	1.1	***COM***	Consistency
UC	0.7	MP	0.9	SA	0.7	***GL***	1.1	***MO***	0.6	***GL***	1.1	***GL***	Grouping and Linking
GL	0.7	YL	0.9	***GL***	0.7	***CSH***	1.1	YL	0.6	MP	1.1	MP	Manipulability
MP	0.7	**ML**		MP	0.7	***RM***	0.6	MOL	0.6	**ML**	0.7	***RP***	Responsiveness
PCO	0.7	CA		MOL	0.7	**ML**	0.6	**ML**		CA	0.4	YL	Visio-Perceputal Load
CSH	0.7	***MO***		***CSH***	0.7	CA	0.6	CA		UC	0.4	AL	Audio-Perceptual Load
ML	0.5	FL		CA		***MO***	0.6	FL		FL	0.4	MOL	Motor Load
CA	0.2	SA		UC		MP	0.6	SA		PCO	0.4	PCL	Perceptual Clarity
MO	0.2	***GL***		***MO***		PCO	0.6	***GL***		***MO***		PCO	Perceptual Contrast
FL	0.2	***CSH***		PCO		FL		MP		***CSH***		GH	General Help
EM		***EM***		***EM***		***EM***		***EM***		***EM***		***CSH***	Context-Sensitive Help
AL		AL		AL		AL		AL		AL		None	No match
Total	100.0	Total	100.0	Total	100.0	Total	100.0	Total	100.0	Total	100.0		

See Appendix B for principle descriptions. None = problem(s) not matching any principle. Principles in ***bold*** are those deemed to be more important.
No data = no problems of this type reported. Cut (thick cell border) at occurrence >= 10 %

Table 2-11. Experiment 1. Problems types (principles) sorted by % occurrences (overall, and out of totals per condition and group).

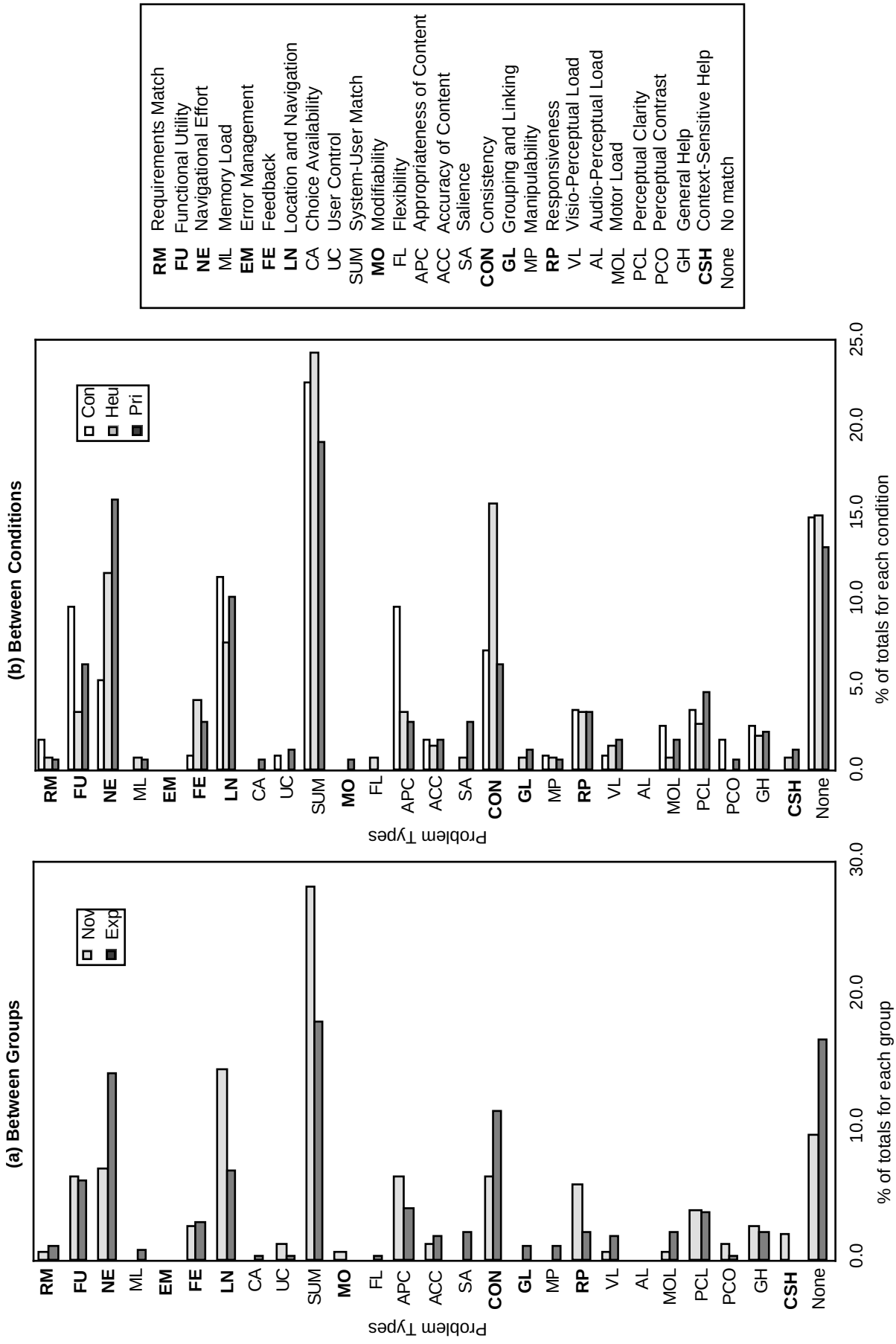


Figure 2.4. Experiment 1. Problem incidences (% occurrences) per problem type (principle): (a) between groups (Novice/Experienced), (b) between conditions (Control/Heuristic/Principle). See Appendix B for principle descriptions. Types in **bold** are those deemed to be more important.

Thus on *this* analysis the problem types deemed most important could not be considered to have had any effect on the likelihood of any one principle being featured in subjects' problem reports. Nor was the group effect (Experienced > Novice) found in Section 5.1 replicated in the type distributions. Hence any difference between principles and heuristics again appears to have manifested in the willingness of subjects to assign severity ratings (Table 2-10), rather than to raise any one problem type over another.

However, further indication of the 'added value' offered by the principles set might be seen in the distribution (shared and non-shared) of high-severity (rated 5 to 7) problem types amongst the Principles and Heuristic conditions (Table 2-12). We find that several problem types were reported using the principles only, particularly CSH (context-sensitive help) and FU (Functional Utility), with a preponderance of problems of type NE (Navigational Efficiency). Thus it might be those problem types which were particularly implicated in the principles-heuristics effect for severe problems.

Both	Heuristic only	Principle only		Principles	
CON	ACC	APC		**RM**	Requirements Match
CON	FL	CA		**FU**	Functional Utility
CON	**GL**	**CON**		**NE**	Navigational Effort
FE	MOL	**CON**		ML	Memory Load
FE	PCL	**CSH**		**EM**	Error Management
FE	**RM**	**CSH**		**FE**	Feedback
FU		**FU**		**LN**	Location and Navigation
FU		**FU**		CA	Choice Availability
GH		**FU**		UC	User Control
LN		**FU**		SUM	System-User Match
LN		**MO**		**MO**	Modifiability
LN		**NE**		FL	Flexibility
LN		**NE**		APC	Appropriateness of Content
NE		**NE**		ACC	Accuracy of Content
None		**NE**		SA	Salience
None		**NE**		**CON**	Consistency
SUM		**NE**		**GL**	Grouping and Linking
SUM		**NE**		MP	Manipulability
SUM		**NE**		**RP**	Responsiveness
SUM		**NE**		VL	Visio-Perceptual Load
SUM		**NE**		AL	Audio-Perceptual Load
SUM		None		MOL	Motor Load
		None		PCL	Perceptual Clarity
		None		PCO	Perceptual Contrast
		SUM		GH	General Help
		SUM		**CSH**	Context-Sensitive Help
		SUM		None	No match
		UC			
		VL			

Table 2-12. Experiment 1. Distribution and sharing of types (principles) of high-severity problems amongst Principle and Heuristic conditions. Types in **bold** are considered most important.

5.4.3 Problem Types per Subject

We have just seen that problems were not distributed evenly among problem types. Hence if we count problems by type rather than single instances, we should not expect to achieve the same between-condition effect as obtained in Section 5.1 (though a group effect might persist). However, since most of the higher-rated problems were of types also deemed most important (Table 2-10), it is worth attempting to replicate the above results for high-severity problems of the most important types. (It also has relevance for the later issue of problem instances versus problem types.) For comparison, we shall also do this for all types and all ratings.

Now counting by type[6] rather than problem instances, a two-way between-subject ANOVA of the complete problem type data (Table 2-13) showed a further significant main effect of group ($F[1,33]=45.53$, $p<0.001$), plus a main effect of condition ($F[2,33]=3.61$, $p<0.05$) but no group x condition interaction ($F[2,33]=0.82$, $p=0.50$). Separate one-way ANOVAs showed no main effects of condition for novices ($F[2,21]=1.42$, $p=0.26$) or for experienced subjects ($F[2,12]=2.25$, $p<=0.15$). Thus while experienced subjects again uncovered more usability problem types (as currently depicted) than did novices (mean ratio 2.02), there were no detected between-condition differences for either experienced or novice subjects in the number of types reported.

Control		Heuristic		Principle	
Novice	Experienced	Novice	Experienced	Novice	Experienced
4.50 [1.20]	8.00 [2.12]	4.00 [0.93]	9.20 [2.28]	5.63 [3.07]	11.20 [2.77]

Table 2-13. Experiment 1. Mean problem type counts per subject, all problem ratings. Figures in [brackets] are standard deviations.

Counting only types of high severity problems (rated 5 to 7) (Table 2-14), a two-way between-subject ANOVA showed a further significant main effect of group ($F[1,28]=14.62$, $p<0.01$) but no main effect of condition ($F[2,28]=1.54$, $p=0.232$) or group x condition interaction ($F[2,28]=2.79$, $p=0.08$). No further analysis is reported.

Control		Heuristic		Principle	
Novice	Experienced	Novice	Experienced	Novice	Experienced
1.83 [0.75]	2.40 [1.34]	1.57 0.53]	4.25 [1.26]	2.14 [0.90]	3.20 [1.64]

Table 2-14. Experiment 1. Mean problem type counts per subject, high severity problems only. Figures in [brackets] are standard deviations.

Counting only the eleven problem types (see Section 2.2) deemed to be most important (Table 2-15), a two-way between-subject ANOVA showed a further significant main effect of group ($F[1,33]=41.71$, $p<0.001$) but no significant main effect of condition ($F[2,33]=1.54$, $p=0.07$) or interaction ($F[2,33]=0.60$, $p=0.55$). No further analysis is reported.

[6] Including only those principles which featured in any one comparison, so that not all 26 principles (listed in Section 2.2) might be involved in any set. Also incorporating only one occurrence of each type in each set.

Control		Heuristic		Principle	
Novice	Experienced	Novice	Experienced	Novice	Experienced
2.00 [1.07]	5.60 [1.95]	2.88 [1.46]	6.80 [3.56]	3.13 [1.96]	8.40 [2.07]

Table 2-15. Experiment 1. Mean problem type counts per subject, most important types only. Figures in [brackets] are standard deviations.

Finally, counting only the most important types and only high severity problems, a further two-way between-subject ANOVA (Table 2-16) showed a further significant main effect of group ($F[1,33]=16.37$, $p<0.001$) plus a significant main effect of condition ($F[2,33]=4.38$, $p<0.05$) but no interaction ($F[2,33]=2.56$, $p=0.09$). Once again separate one-way ANOVAs showed no main effects of condition for novices ($F[2,21]=0.31$, $p=0.74$) or for experienced subjects ($F[2,12]=2.99$, $p=0.09$).

Control		Heuristic		Principle	
Novice	Experienced	Novice	Experienced	Novice	Experienced
1.13 [1.13]	3.20 [2.59]	1.00 [0.76]	2.00 [2.55]	1.50 [1.85]	5.80 [2.39]

Table 2-16. Experiment 1. Mean problem type counts per subject, most important types and high severity problems only. Figures in [brackets] are standard deviations.

Thus further analysis by problem type has failed to replicate the between-condition (EP > EH and EP > EC) result obtained above for problem instances, even for high-rated problems and the most important types. This is not unexpected, even given the relatively low means (and large standard deviations) involved; its implications will be further discussed in Chapter 4.

6. Summary of Results

1. Experienced subjects identified significantly more usability problems than did novices, overall and in two out of three conditions.

2. The short version of the author's principles set used in this experiment elicited from experienced subjects significantly more high severity (rated 5 to 7) problems than did Nielsen's heuristics alone.

3. Nielsen's heuristics did not elicit significantly more usability problems than did a control condition without such guidance. This was so for both experienced and novice subjects and for high severity problems.

4. Novice subjects were not able to make use of the short principles set to elicit a significantly higher number of usability problems than they did using Nielsen's heuristics.

5. There were no significant differences between either groups or conditions in the mean severity ratings assigned to problems by subjects. However, ratings in the high-severity range were significantly higher for experienced subjects than for novices.

6. There were more problems which were identified only once than were shared by more than one subject. The significant result in item 2 corresponded with the highest difference between numbers of problems reported only once in each condition.

7. All of the most often reported problems were within the top ten percent of UPTs when arranged by frequency of occurrence. Sorting on the product of frequency and mean severity rating is offered as one means of prioritising problems.

8. Most of the higher-rated (4 to 7) problems were of types deemed by the author to be most important (of the then 26 principles). This may have been a factor of the high severity ratings assigned by experienced subjects using the principles (item 5).

9. A small number of problem types (principles) also featured a disproportionately high number of times (though these were not the most important types).

10. Though experienced subjects again reported significantly more problem types (principles) than did novices, no one condition attracted a higher number of types (even for high-severity problems deemed most important). However, the distribution of high-severity problems between Principle and Heuristic subjects appeared to implicate a few problem types in the significant result in item 2.

7. Discussion

While the overall difference between experienced subjects and novices is to be expected, that between the principles and heuristics implies that the principles set did offer additional material sufficient to enable experienced subjects to uncover more high-severity usability issues than they would using the heuristics alone. The nature of the additional issues (the 'added value' which the principles set may offer) may be gleaned from the high-severity problem types which were reported using the principles rather than the heuristics (Table 2-12): this would seem to implicate CSH (context-sensitive help) and FU (Functional Utility) in particular, along with an emphasis on NE (Navigational Efficiency) and SUM (System-User Match).

However, we have seen that the severity ratings given to problems by subjects played an important part in the differences between problem numbers, and that the low degree of overlap between problem reports meant any between-subject differences are likely to have been made up of problems reported under one group or (especially) condition alone. The derivation of the problem types (categories based on the principles set) also requires comment.

7.1 Severity ratings

We have seen that there were no significant between-condition differences in problem numbers for either group (but a marginal Principle-Control effect for experienced subjects), until high-severity (rated by subjects 5 to 7) problems were partitioned out. We have also seen that the same problems which were most highly rated by the experienced subjects were responsible for the preponderance of the most important problem types in the ratings distributions, even though (a) there were no overall differences between ratings levels and (b) the type distributions did not reflect either these important types or the expected between-condition differences. (The only significant differences between mean severity ratings also proved to be for higher-rated (5 to 7) problems alone.)

Thus we may have to conclude that

(a) there was a particularly strong effect of the willingness of experienced subjects exposed to the principles to assign high severity ratings to more problems than they (or others of similar experience level) would have done using the heuristics or control materials alone;

and

(b) the strength of this effect was such that it swamped any other effect which could be detected (other than the difference between experienced and novice subjects), including any effect of problem type.

Thus the influence of the principles set seems to have been in encouraging subjects with sufficient experience in distinguishing salient usability issues to assign high importance to more issues than they otherwise would, while still not focusing on these issues more than any other (or identifying more examples of different issues).

This conclusion represents a difference of emphasis for the principles set - some influence on overall problem-selection which the heuristics do not exhibit, enabling a better focus on more important issues - from that which might have been anticipated. The fact that subjects using the principles were encouraged to identify more serious problems should be seen as encouraging, since it is those very problems which we would wish to focus upon in any set of improvements. The fact that these problems were seen as specially serious (and not just more of the same) says something about the ability of the principles set to pull out important issues (along with the trivia), not just that there were more than twice as many principles to run through as there were heuristics. And the lack of significant differences between the heuristics and control materials appears to support the proposition that short lists of heuristics such as Nielsen's rely on evaluator expertise over content alone (and in practice would need to be backed up by supporting material such as is found in Nielsen 1993).

As to the source of the severity ratings, it should be noted that in both this experiment and Experiment 2 it was subjects themselves who assigned severity ratings to their own usability problems. This is in contrast to the approach reported by e.g. Virzi (1990, 1992) and Jacobsen et al. (1998a and 1998b) where severity was assigned by other evaluators. The author's view was that to assign severity ratings to subjects' own problems reports may be to impose an experimenter's view on what subjects wish to say (i.e. to be part of a possible experimenter bias). However, the issue of severity assessment is an important one, and will be taken up in Chapter 8.

7.2 Problem Types

We saw in Section 5.4.3 that the effect of problem count was not replicated for problem types, even for high-severity problems and the most important types. It was also shown that the types and severity ratings distributions did not correspond. It should be pointed out, however, that the type categorisation was based on the 26 parts which made up the principles set as then developed; thus an attempt to apply these categories to the problems derived using the heuristics (or indeed the control) may be criticised on grounds of internal validity. For it was the very absence of the principles set which was being assessed in the other two conditions, and to retroactively assume some 'underlying' influence for the principles is dubious. (It would make more sense to also attempt this for the ten heuristics). However, the categorisation by principles is offered as a means of comparison at a higher level than individual UPTs, enabling a view of the common factors underlying any effects at lower levels. This issue will also be taken up in Chapter 8.

7.3 Problem Overlap

We have seen that there was a low degree of overlap between problem reports (individual subject protocols) in the problems found by different subjects. The differences between Principle and Heuristic (and Control) conditions are thus more likely to have been made up of problems found under each condition alone, rather than in more than one condition. Thus any attempt to conclude that the principles or heuristics tended to elicit certain problems and not others may be unfounded. However, the listing in Table 2-12 is offered as an indication of the types of problems to which the principles set alone (and not the heuristics) may have encouraged the assignment of high severity ratings.

The lack of problem-sharing was reflected in the large proportion of single-incidence UPTs and the fact that most of the higher-frequency problems were among the top ten percent of the sorted list of UPTs (Table 2-7). While variations in the incidence of unique problems for individual evaluators is a feature of usability evaluation studies (Jacobsen et al. 1998a, 1998b), the results from this experiment (70.4% of all UPTs appeared once, only 16.1% three times or more) represent an extreme version of that trend. In this case the relatively open-ended nature of both the software evaluated (a simulated hypermedia browser) and

the tasks involved (free exploration following a goal-driven task) may be responsible. Some explanations are offered.

(a) Firstly, as mentioned above, the intention had been to compare 'poor' and 'improved' versions of the software in order to focus on deliberately designed-in faults. Unfortunately, lack of experienced users meant that only the 'improved' version was used: even given the small size of the simulation, the large number of single-incidence UPTs represents a considerable scattering of focus, even amongst experienced subjects. (See Appendix G for a full problem listing.)

(b) The simulation was considerably reduced in size from that of original MGS, so as to emphasise the salient aspects of the interface over its content. This was explained to subjects at the start of each session, and seems to have been well understood.

(c) The very type of software was such that free exploration (following links and lines of interest from an initial goal) may be more representative of typical usage than prescribed tasks (e.g. find some information about a painter; locate an art-historical period for a painting). (This is even more likely to be the case with the current generation of internet sites and browsers than for CD-ROM or museum-based hypermedia).

(d) It is the author's belief that to stipulate user tasks *may* be to proscribe the very range of issues which may be identified, and that this *might* partly underline the much higher problem overlap (and lower problem counts) which are typical of published reports (e.g. Sears 1997, Virzi 1992). In an experimental setting, where software release deadlines and re-design programmes are likely to be less pressing, researchers have more leeway to include other (or all) aspects of an interface in a series of studies than is typical in commercial environments. It is the skill of a usability engineer to define the range of different user tasks which make up a typical usage scenario; nevertheless, to claim that a short series of prescribed tests can identify "the" (or even "most of the") problems "with" an interface is, in the author's view, to claim more for such results than may be justified.

While the inability to compare 'good' and 'poor' versions of the simulated MGS stands as the major failing of this experiment, it is the author's claim that both the large problem numbers and their wide distribution amongst subjects remain more representative of real usage than would be the result from more proscribed tasks. (The reader may inspect the full problem listing in Appendix G). This issue will be further explored Chapter 8.

However, it is remains possible that the particular nature of these results was due to something about the way that problem reports were (a) generated from subjects, or (b) extracted from subject reports, not to mention (c) the nature of the software and tasks. If (a), some bias of the experimental method may have been responsible; if (b), some validation of

the problem extraction procedure is necessary; if (c), some more restricted tasks, on more limited software, may still be attempted.

The issue of possible methodological bias is addressed in Experiment 2 (Chapter 3); more restricted tasks and software functionality will be examined in Experiment 3 (Chapter 5); and in Chapters 4 and 5 validation of part of the problem extraction (and reduction) process will be offered for experiments 1 and 3.

Summary of Chapter 2

In Experiment 1 it was found that experienced subjects were able to identify more high-severity usability problems using the author's then principles set than they did with Nielsen's heuristics alone. Novices proved unable to do this at any severity level. Experienced subjects also found consistently higher numbers of problems than novices. The first result was not replicated for problem types based on the principles set. It was also found that many more problems were reported by single subjects than by more than one subject, and that severity level rather than problem type played an important role in any difference(s) between principles and heuristics.