Chapter 3: Experiment 2

1. Introduction

In Experiment 1 (Chapter 2) it was shown that experienced evaluators could make use of the author's set of (then) usability principles, in short form, to identify more high-severity problems than using Nielsen's heuristics alone. However, the novices in Experiment 1 (mainly Psychology undergraduates) were unable to do so, at any severity level.

In this experiment we shall see if a similar group of novices could do any better, this time given prior training on the full version of the principles set. In this case the software (teaching material for Psychology students) was chosen to be directly relevant to the subject population (Psychology undergraduates). One source of possible bias in Experiment 1 was that the experimenter sat with subjects and wrote down her or his comments; in Experiment 2 subjects worked throughout the online material on their own and entered problem reports directly into a prepared spreadsheet.

The rationale for again using novices (apart from their relative availability) was that they offer a potential 'baseline' for the treatment of evaluation materials, since any effect for experienced subjects (such as that achieved in Experiment 1) may be due in large part to evaluator experience rather than the materials themselves. (If novices can do it, we have little difficulty claiming that experts can; if only experts can, it does not follow that moderately experienced subjects, or even other experts, will.) The issue of evaluator experience will be taken up in Chapter 8.

In this experiment the fully worked, 31-page version of the author's principles set described in Chapter 1 was compared with the same set of ten short heuristics (Molich & Nielsen 1990, Nielsen 1993) as was used in Experiment 1. A pilot study implied that even with prior exposure, the sheer bulk of the expanded material inhibited its use, so in the experimental sessions Principle subjects were additionally provided with a summary of the full set. Once again a control condition was included in which subjects did not receive either the heuristics or principles. All subjects were given prior introduction to both the materials and the evaluation procedures, in the form of an initial training session. It was hoped that this time the 'added value' of the 31-page principles set would be sufficient to enable this group of novices to identify more problems than they would with a shorter set such as was used in Experiment 1.

2. The Evaluation Materials

The 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, C and D. (Appendix C shows the preface material as given to Principle subjects.)

In this experiment the same general material was used to preface both the heuristics and principles sets. This took the form of an introduction to usability evaluation heuristics or principles (as appropriate), the general aims of the evaluation process, a guide as to what constitutes a usability problem, and how problem severity might be judged.

Control subjects were given only the preface material, without the introduction to heuristics or principles. This constituted only the first paragraph and second page (minus its second and third paragraphs) of Appendix C.

Heuristic subjects had the preface material, including introduction (to heuristics rather than principles), plus the ten heuristics. This constituted most of Appendix C, but minus its second paragraph.

Principle subjects had the preface material, including introduction (Appendix C), plus the 31-page principles set (Appendix D). In order to facilitate their use of the material, in the experimental sessions they were provided with an additional 8-page summary of the principles set, and were instructed to refer to the full version as necessary. The summary constituted merely the 'Explanation' parts of each principle or attribute.

2.1 Nielsen's Heuristics

(See Appendix A for a transcript.)

In both this experiment and Experiment 1 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 (Full Version)

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

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

In this experiment the full version of the now 23 principles[2] was used, now covering 31 pages including the introductory material. The same 7 overall groupings were retained as in Experiment 1, but some principles were split into sub-principles or 'attributes' (if not split, the attribute has the same name as the principle). There were thus a total of 30 attributes, numbered below, which were to be the focus of interest. Acronyms are as feature in the Results. The derivation, format and contents of the full principles set were described in Chapter 1.

Requirements and Functionality Principles

Requirements match

1. Functional Needs (FN)
2. Requirements Needs (RN)

Functional Utility

3. Functional Organisation (FO)
4. Functional Provision (FP)

User-System Principles

Navigational Effort

5. Minimum Steps (MS)
6. Minimum Retraction (MR)

Memory Load

7. Memory Load (ML)

Error Management

8. Error Management (EM)

Feedback

9. Feedback (FE)

Location and Navigation

10. Locational Information (LI)
11. Locational Modes (LM)

Choice Availability

12. Choice Availability (CA)

User Match

13. Terminology and Language Style (TLS)
14. Visual Metaphor (VM)

User Principles

Modifiability

15. Functional Modification (FM)
16. Step Modification (SM)

[2] Following Experiment 1, the content of three of the original 26 principles (User Control (UC), Appropriateness of Content (ACC) and Grouping & Linking (GL)) were merged into others (becoming new attributes), and one was changed by name (from System-User Match or SUM to User Match or UM).

Flexibility

17. Multiple Initiation (MIN)
18. Multiple Inputs (MIP)

Accuracy of Content

19. Accuracy of Content (ACC)

Salience

20. Salience (SA)

Comparative Principles

Consistency

21. Consistency (CON)

System Performance Principles

Manipulability

22. Manipulability (MP)

Responsiveness

23. Responsiveness (RP)

Perceptual and Motor Principles

Visio-perceptual Load

24. Visio-perceptual Load (VL)

Audio-perceptual Load

25. Audio-perceptual Load (AL)

Motor Load

26. Motor Load (MOL)

Perceptual Clarity

27. Perceptual Clarity (PCL)

Perceptual Contrast

28. Perceptual Contrast (PCO)

User Support Principles

General Help

29. General Help (GH)

Context-sensitive Help

30. Context-sensitive Help (CSH)

Apart from its length, the full principles set covered a wider scope than the version used in Experiment 1. It also included more examples, and indications of exceptions and potential trade-offs. (See Chapter 1, Section 3.3.)

3. The Software

The software used for Experiment 2 was part of the PsyCLE™ teaching material for undergraduate psychology students. It was chosen because (a) the content was directly

relevant to the subject population (psychology undergraduates) available to the author, (b) the software had been developed and produced locally (University of York Department of Psychology) and (c) the software could be run in parallel using available facilities.

PsyCLE offers multimedia teaching material based around seven modules concerning different aspects of psychological research, namely Language, Sound, Vision, Memory, Experimental Design, Developmental Psychology and Social Psychology. (Experimental Design was later re-designed as a separate package called SPEED™.) PsyCLE was created by a consortium of UK university Psychology departments, co-ordinated at University of York. The latest version (Version 2.0) was produced by Paul Askins and published by Blackwell Publishers. A student workbook (Hammond & McKendree 1998) accompanies the package.

This experiment made use of two of the PsyCLE modules, namely 'Developmental' and 'Social' (originally developed at Warwick and Kent Universities respectively). They were chosen because (a) they were relatively short compared to other modules, (b) they contained roughly equivalent amounts of interactive and experimental material, (c) they were relevant to undergraduate courses recently or then running, and (d) the author is more familiar with their content than with some other modules. Like other modules, both Developmental and Social Psychology consist of two parts, or half-modules. For this experiment two separate half-modules, respectively 'Piaget's Tasks' and 'Attribution Theory', were used: the former for the first, training, session and the latter for the second, experimental, session.

PsyCLE is based on a hierarchical menu structure organised around the seven modules and half-modules. Users navigate between the parts of each module or half-module using Next Page, Back and Return buttons. The content of a module can be a combination of simulated experiments, interactive demonstrations, question and answer sessions, animations, videos, and text. Most modules also contain bibliographies and/or further reading, plus a glossary linked to text items.

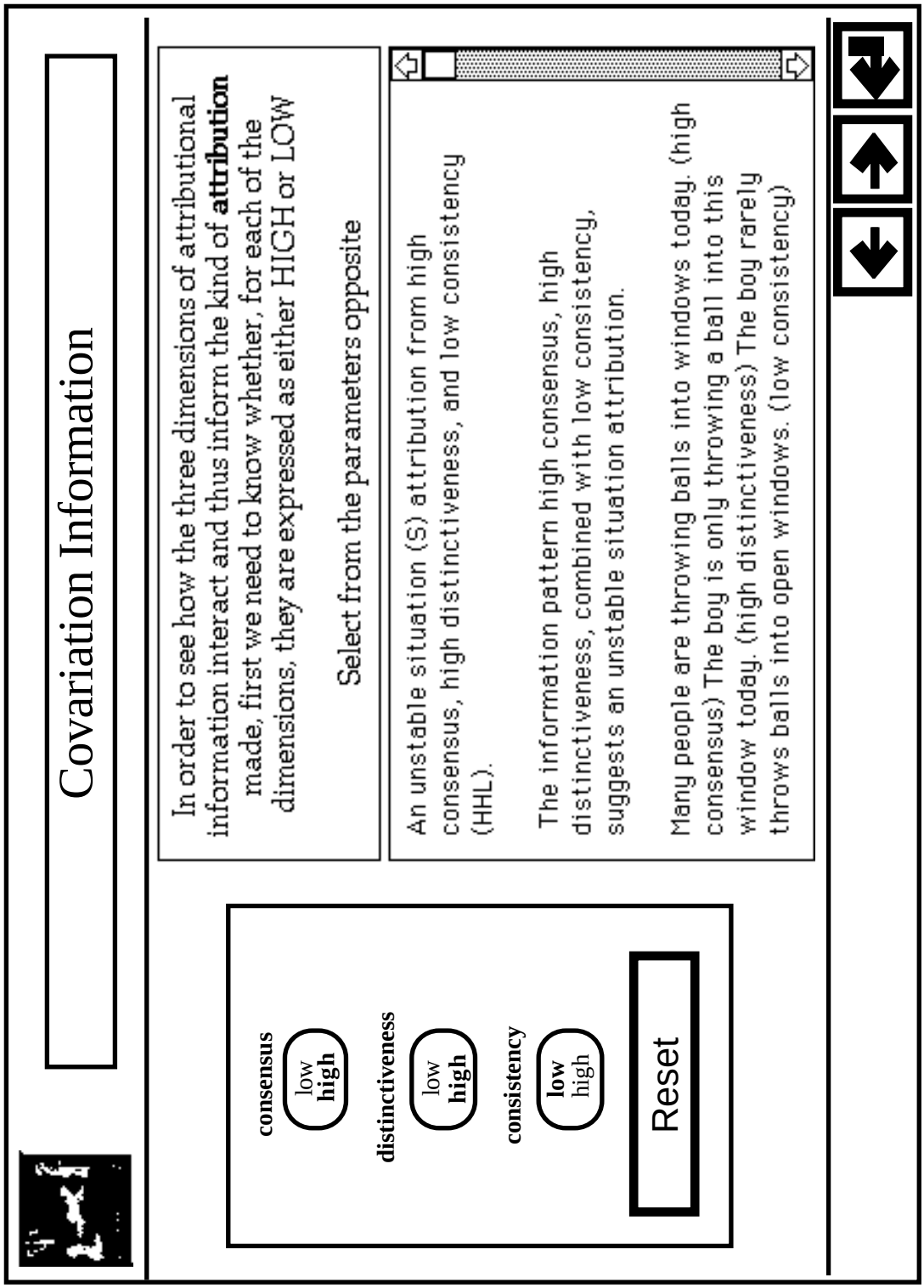


Figure 3.1. Experiment 2. Sample screen from PsyCLE half-module 'Attribution Theory'.

Figure 3.1 illustrates a sample screen from the 'Attribution Theory' half-module used in the experimental sessions. It shows the second screen in a sequence intended to demonstrate the principle of covariation in situational attribution. From the buttons on the left, the user can select any combination of high and low values for the three dimensions of attributional information, namely consensus, distinctiveness and consistency; a matching explanation (using situations from previously seen videos) then appears in the bottom right panel. (It

may be necessary to scroll the panel to read succeeding text.) In order to select a new combination of dimensions the user may reset all three values or simply click on different parameters. She or he may then proceed to the next screen in the sequence, return to the previous screen, or go back to the first (contents) page of the half-module.

PsyCLE is CD-ROM based, running on both PC (Windows® 95 and later) and Macintosh (System 7.1 and above) platforms. It was implemented in Macromind Director®, incorporating Quicktime® and Moviemaker®. Window size is 230 x 175 mm. The version (2.0) used by subjects in this experiment ran on Macintosh Quadras under System 7.1.

4. Method

4.1 Design

Experiment 2 was a three condition between-subjects design, in two sessions (training and experimental). Only one subject group, novices (different individuals from those in experiment 1), was involved; the three conditions were Control (N=7), Heuristic (N=8) and Principle (N=8).

4.2 Subjects

Novice status was assessed by the same questionnaire as used in Experiment 1 (with changes to the first question regarding previous use of the relevant software), this time administered by e-mail prior to the first session. (See Appendix E.) All subjects were University of York Psychology undergraduates. Most were female and aged 18-21. All were paid at hourly rates or received subject credit; all were additionally paid to read the appropriate experimental material prior to the first session.

Experiment 2 subjects' mean questionnaire score was 11.78% (median 10, standard deviation 5.81). A one-way between-subjects ANOVA (Table 3-1) showed no significant difference between conditions ($F[2,22]=2.86$, $p=0.08$), though scores for Principle subjects were lower than those for the other two conditions.

Control	Heuristic	Principle
14.00	13.50	8.13

Table 3-1. Experiment 2. Mean questionnaire scores per subject, by condition.

4.3 Procedure

Each subject was required to participate in two sessions, namely training (Session 1) and experimental (Session 2), held one week apart. Subjects were assigned to conditions as they volunteered, in order Heuristic, Principle, Control, so as to achieve a minimum of 2 participants per session. All subjects were sent the appropriate evaluation material (described in Section 2) a few days before the first session, and were instructed to bring the material to both sessions. Subjects were asked to read the material before the first session. Heuristic subjects received the preface material and one-page heuristics, while Principle subjects received the preface and the full principles set plus summary. Control subjects received only the preface material.

In both sessions (training and experimental) subjects were seated at alternately spaced computers and asked not to confer with other participants. The procedures were explained at the start of each session. (In Session 2 only a brief reminder proved to be necessary.) The session procedure instructions were designed to lead subjects through the steps required to perform their evaluations and complete the on-screen spreadsheet (and how to switch between the two tasks). These materials included description of the intended users of the software (the subjects themselves) and the primary purpose of the experiment, namely that subjects should "critically evaluate the module as a learning tool for psychology students" (the secondary purpose being for them to learn about the module contents). In Session 1 subjects were to work through a set of questions based on the teaching software (these responses were later returned to subjects and were not used in analysis).

The procedure in both sessions was the same: first, to run through the whole half-module relatively briefly (without running the experiments or videos), and then to work through the material in detail. As in Experiment 1, this was intended to resemble Nielsen's recommendation for heuristic evaluation, namely that evaluators should first go through the whole of an interface in relatively brief fashion, and then proceed in more detail (Nielsen 1993 p158, 1994d p29). Session 1 was designed to introduce subjects to the evaluation materials and procedures, so that familiarity with both could be assumed in Session 2. (Though Control subjects did not need such full exposure, this aspect of the procedure was held the same.) To that end, in Session 1 subjects were encouraged to seek clarification about any aspect of the procedures and (especially) use of the materials, while in Session 2 no such assistance was given (so as to avoid any possible biasing of subjects' problem identifications). For this reason, subsequent analysis will be confined to data from only the second session.

In both sessions subjects entered their evaluations directly into a prepared spreadsheet. In Session 2 this was without experimenter intervention, and even in Session 1 great care was taken not to influence the content of subject responses. Though subjects were instructed to use the appropriate experimental material as their guide, it was also emphasised that they

could identify additional usability problems not directly deriving from the materials (and that they should not attempt to 'backwards justify' any such problems).

Against each usability problem identified, all subjects were asked to enter the following.

- A severity rating, from 1 (low, trivial) to 7 (high, must be addressed)
- Any suggested remedy or remedies
- Any other comments

Heuristic and Principle subjects were additionally asked to enter the following.

- The heuristic(s) or principle(s) (attribute(s)) to which the problem related (if relevant)
- A judgement of the source of the problem, from 1 (wholly themselves) to 7 (wholly the evaluation materials).

(See Appendix J for a sample response sheet.)

In both sessions subjects were allowed to proceed at their own pace. In Session 1 a limit of one hour was imposed (but no pressure was applied to complete the half module). In Session 2 there was no time limit, though most subjects finished inside one hour. All subjects were asked to read through their responses before finishing and to add any further comments.

Other than minor editing for spelling and clarity, subject responses were used unchanged as the basis for subsequent analysis. The principal dependent measure (as in Experiments 2 and 3) was to be the number of problems which were reported by each subject. In order to extract this measure, any duplication in each account had first to be assessed; this was done by the experimenter (the author) following the experimental sessions. The attendant difficulties involved in this problem reduction process form part of the second main theme of this thesis, and will be taken up in full in Chapter 4.

4.4 Pilot Study

A pilot study (using the Session 1 half-module) had been carried out in order to try out the procedures, session instructions and evaluation materials (also given out in advance). This involved three volunteers, each using the principles set; their responses were not included in the analysis. As a result of the pilot it was realised that the size of the principles set necessitated additional summary material, which was provided in Session 1.

5. Results

The results to be reported refer only to the quantitative analysis performed on data from the second, experimental, session, thus relating to the 'Attribution Theory' half-module.

5.1 Problem Counts per Subject

See Figure 3.2.

A one-way between-subject ANOVA of the complete problem count data from Session 2 (Table 3-2, Figure 3.2 (a)) showed no significant effect of condition ($F[2,20]=0.78$, $p=0.47$). A further post-hoc (Tukey HSD) test confirmed no significant ($q_{0.05}$) between-condition differences ($q_{HSD}[3,20]=3.58$, $W_3=5.36$). Thus in this experiment the differences between principles and heuristics (and the control) did not enable Principle subjects to identify more usability problems than did similar novices using the heuristics. Nor did the heuristics enable these novices to find more problems than did Control subjects.

Control	Heuristic	Principle
8.71 [4.42]	11.13 [4.16]	9.00 [3.89]

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

When high-severity problems (rated by subjects as 5 to 7 inclusive on a 1 to 7 scale) alone were analysed, a similar view emerged (Table 3-3, Figure 3.2 (b)). A one-way between-subject ANOVA of high-severity problems alone revealed no significant effect of condition ($F[2,20]=0.42$, $p=0.27$). Thus the differences between the materials did not enable Principle subjects to find more high-severity problems than did Heuristic or Control subjects. However, further pair-wise (Newman-Keuls[3]) tests revealed a significant ($q_{0.05}$) difference between Heuristic and Control conditions ($W_2=2.79$) (and not between Control and Principle ($W_2=2.62$) or Heuristic and Principle ($W_2=3.35$)), implying that Heuristic subjects alone found more high-severity problems than Control subjects.

Control	Heuristic	Principle
2.43 [1.27]	5.50 [3.21]	3.63 [2.97]

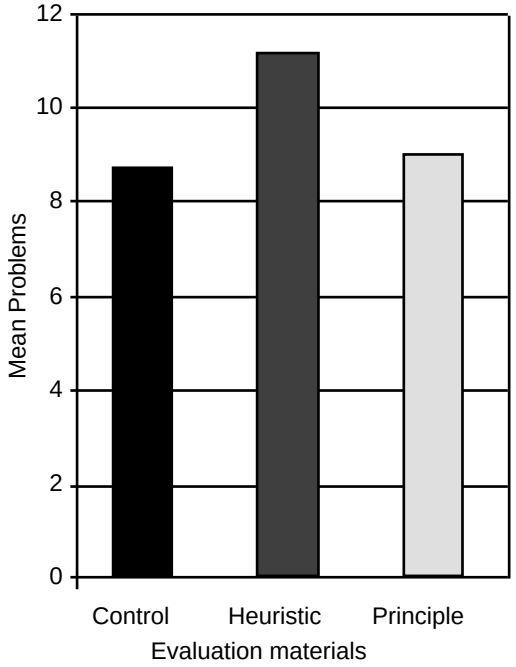
Table 3-3. Experiment 2. Mean problem counts per subject, high severity problems. Figures in [brackets] are standard deviations.

Looking at low-severity (rated 1 to 3 inclusive) problems (Table 3-4, Figure 3.2 (c)), a further one-way between-subject ANOVA of low-severity problems alone revealed no significant effect of condition ($F[2,20]=0.12$, $p=0.89$). Further pair-wise (Newman-Keuls) tests confirmed no significant between-condition differences.

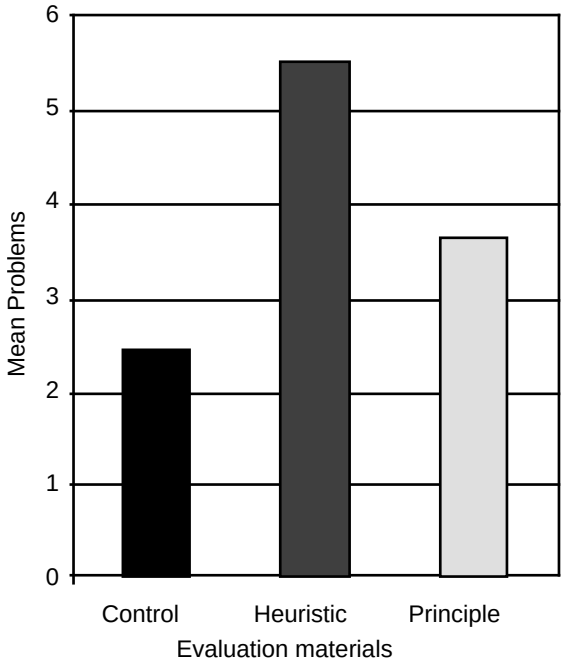
Control	Heuristic	Principle
3.71 [3.50]	3.38 [1.85]	3.25 [1.39]

Table 3-4. Experiment 2. Mean problem counts per subject, low severity problems. Figures in [brackets] are standard deviations.

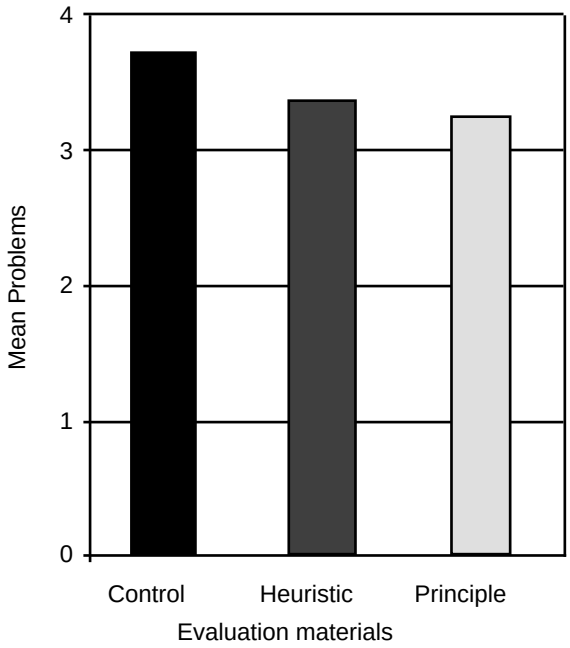
[3] Using the Games and Howell procedure for heterogeneous variances (Howell 1997).



(a) Mean problem counts, all problems



(b) Mean problem counts, high severity problems



(b) Mean problem counts, low severity problems

N Control = 7 per group
N Heuristic = 8 per group
N Principle = 8 per group

Figure 3.2. Experiment 2. Mean problems per subject.

This analysis has revealed no effect of problem count for novices using the full principles set over Nielsen's heuristics, at any severity level.

5.2 Problem Tokens

5.2.1 Problem Distribution

Distributions of unique problem tokens (UPTs) amongst conditions were computed in the same manner as in Experiment 1, that is, by determining the proportions of single and shared appearances of UPTs. Figure 3.3 shows the percentage of overlap (problem sharing) between all three conditions, and between each pair of conditions, for all severity ratings.

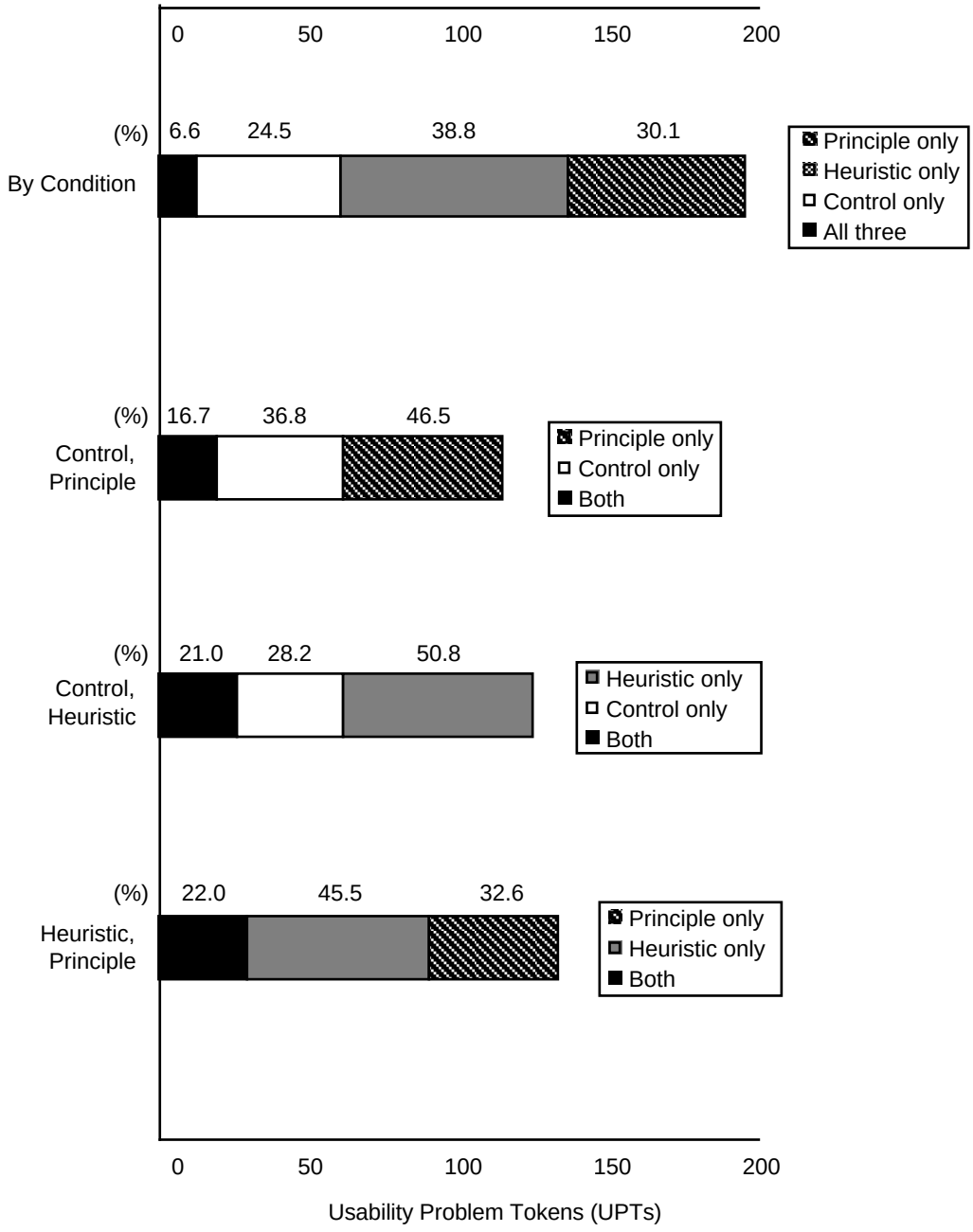


Figure 3.3. Experiment 2. Problem (UPT) distributions for all problem ratings.

It can be seen that once again the overall amount of problem sharing was low (6.6%), with between-condition distributions reflecting the all-ratings problem counts analysed in the previous Section. Comparison with the equivalent distributions from Experiment 1 (Chapter 2, Figure 2.3(a)) reveals an almost identical profile of shared to non-shared problems (but

involving half the total numbers of UPTs than in Experiment 1). Thus though any experimenter influence in the number and type of problems identified was restricted to the introductions which prefaced the evaluation materials, these novices showed the same lack of overall agreement in their problem reporting as was evident in Experiment 1. Once again, Principle subjects were finding mainly different problems to those using the heuristics (and the control), though this time there were no between-condition differences in the totals found.

5.2.2 Problem Listings

Problem (UPT) listings from this experiment show a similar distribution pattern to those from Experiment 1: of the 114 UPTs, 57.9% were reported once, with only 10.5% found 4 times or more. (See Appendix H for a full listing.) Tables 3-5 and 3-6 show the top ten percent of the full set, sorted by frequency of appearance and frequency x mean severity rating, respectively. It will be seen that the two lists are little different, showing that once again severity ratings were evenly distributed among UPTs.

From Table 3-6 we might conclude that in the first instance attention should be drawn to the lack of evidence of the glossary (problem 25), the organisation of the Further Reading section (2), the covariation principle exercise illustrated in Figure 3.1 (34, 35, 55, 46), the lack of a general help facility (21), the videos (85, 87, 53), and the lack of a search facility (77). These results indicate that subjects (Psychology students) were experiencing difficulties with an important aspect of the Attribution Theory half-module (the glossary, help and search issues being common to other PsyCLE modules). (See Appendix J for a sample completed report form.) The richness of the subject reports, including both criticism and constructive suggestions, offset the failure to achieve significant quantitative results in this experiment.

No.	Problem description	Freq.	%	Rtng.
25	Existence of glossary not obvious (other simpler things are explained)	11	9.65	4.3
2	Further Reading long, boring, tedious (too much text, poor English)	8	7.02	4.6
85	Video clips start automatically on opening and returning to page	7	6.14	3.7
34	Covariation explanations arising from video clip(s) unclear, unrelated to rest of material, boring, inadequate, repetitive	6	5.26	4.8
35	Covariation example terminology (A, S) inadequately explained	6	5.26	4.7
87	Video sound quality poor (background noise)	6	5.26	3.5
21	No general help provided	5	4.39	5.4
55	[Covariation exercise:] Insufficient or no feedback given on user input	5	4.39	4.4
66	Difficult to tell when scroll bar is activated [more text below]	5	4.39	1.8
53	Answers to questions on videos not given	4	3.51	4.3
68	Return to start [exit and return to top level] not clear or indicated	4	3.51	3.5
84	Video controls not clear, not noticed: a bit small and complex	4	3.51	3.0

Table 3-5. Experiment 2. Top ten per cent of full set UPTs, sorted by frequency of appearance. Freq. = frequency, Rtng. = mean severity rating.

No.	Problem description	Prior.	Freq.	Rtng.
25	Existence of glossary not obvious (other simpler things are explained)	47.3	11	4.3
2	Further Reading long, boring, tedious (too much text, poor English)	36.8	8	4.6
34	Covariation explanations arising from video clip(s) unclear, unrelated to rest of material, boring, inadequate, repetitive	28.8	6	4.8
35	Covariation example terminology (A, S) inadequately explained	28.2	6	4.7
21	No general help provided	27.0	5	5.4
85	Video clips start automatically on opening and returning to page	25.9	7	3.7
55	[Covariation exercise:] Insufficient or no feedback given on user input	22.0	5	4.4
87	Video sound quality poor (background noise)	21.0	6	3.5
77	Provide a finder (e.g. search by keywords, topic index)	18.0	3	6.0
53	Answers to questions on videos not given	17.2	4	4.3
46	Covariation Information (principle dimensions): examples inadequate	15.9	3	5.3
95	Covariation Information : ‘S’ not meaningful to Psychology students (too technical); "yields consistency information" not meaningful	15.9	3	5.3

Table 3-6. Experiment 2. Top ten per cent of full set UPTs, sorted on frequency x mean severity rating. Prior. = priority, Freq. = frequency, Rtng. = mean severity rating.

5.3 Problem Types

5.3.1 Problem Distributions by Type

Insight into the types of problems which featured most often in subjects’ reports can be obtained by ranking problem types according to their frequency of appearance (Table 3-7). This time problem categories were based on attributes from the full principles set, using only those 14 (of the 30) attributes which were deemed to have featured in any subject’s reports. (See Chapter 1, Section 3.3 for the 30 attributes and Appendix D for a full transcript.)

All		By Condition							
		Control		Heuristic		Principle			
Type	%	Type	%	Type	%	Type	%	Attributes	
TLS	23.9	TLS	24.6	TLS	22.5	TLS	25.0	FO	Functional Organisation
FP	10.4	FP	14.8	MS	13.5	FO	16.7	FP	Functional Provision
FO	9.9	None	13.1	FP	11.2	None	8.3	MS	Minimum Steps
MS	9.0	LI	11.5	CSH	9.0	CSH	6.9	MR	Minimum Retraction
LI	8.1	GH	8.2	GH	7.9	GH	6.9	EM	Error Management
GH	7.7	MS	6.6	MOL	6.7	LI	6.9	FE	Feedback
None	7.2	FO	6.6	LI	6.7	FP	5.6	LI	Locational Information
CSH	6.3	ACC	6.6	FO	6.7	MS	5.6	TLS	Terminology and Language Style
ACC	5.9	PCL	4.9	ACC	6.7	PCL	5.6	ACC	Accuracy of Content
PCL	5.0	EM	1.6	PCL	4.5	ACC	4.2	MOL	Motor Load
MOL	4.1	CSH	1.6	None	2.2	MOL	4.2	PCL	Perceptual Clarity
MR	0.9	MOL		MR	1.1	EM	1.4	CSH	Context-sensitive Help
FE	0.9	MR		FE	1.1	FE	1.4	GH	General Help
EM	0.9	FE		EM		MR	1.4	None	No match
Total	100.0	Total	100.0	Total	100.0	Total	100.0		

Table 3-7. Experiment 2. Problems types (attributes) sorted by % occurrences (overall, and out of totals per condition). See Appendix D for attribute descriptions. None = problem(s) not matching any attribute. No data = no problems of this type reported Cut (thick cell border) at occurrence >= 10.

Once again we see that the distribution is skewed in favour of a few types, though less markedly so than for Experiment 1. Overall the most frequently reported types were Terminology and Language Style (TLS), Functional Provision (FP), Functional Organisation (FO) and Minimum Steps (MS). The Principle distribution appeared to be more skewed than that for either the Control or the Heuristic, with None (problems which could not be assigned to any attribute) being third ranked (like that for the Control). This suggests that Principle subjects were not gaining much additional benefit from the principles set over the control (let alone the heuristics). The implications of this finding will be taken up in the Discussion.

5.3.2 Problem Types per Subject

We have just seen that problems were not evenly distributed among problem types, and that those for Principle subjects appeared most skewed. Hence if we count by problem type rather than instance, we might still find some differences between conditions, perhaps even in favour of the principles.

However, a one-way between-subject ANOVA of the mean problem types per subject, based on the 14 attributes listed in the previous Section (Table 3-8), showed no significant main effect of condition ($F[2,20]=0.14$, $p=0.87$); a further post-hoc (Tukey HSD) test confirmed no between-condition differences ($W_3=2.87$). Thus neither Heuristic nor Principle subjects were able to make use of the evaluation materials to identify any more problem types (as here analysed) than did Control subjects.

Control	Heuristic	Principle
5.71 [2.21]	6.25 [2.43]	5.65 [1.98]

Table 3-8. Experiment 2. Mean problem type counts per subject, all problems. Figures in [brackets] are standard deviations.

Problem types will feature in the discussion of types versus instances in Chapter 4.

5.3.3 Pedagogic/Non-Pedagogic Problems per Subject

It was noticed from the problems listings that a large proportion of problems were concerned with the purely teaching-related (pedagogic) content of the PsyCLE modules, as opposed to what might be called the human factors or HCI aspects. For example, we have seen that the most frequently reported types of problem with this half-module concerned terminology and language style, which turned out to be mainly related to the Covariation Principle exercise illustrated in Figure 3.1. Thus it seemed worth extracting from the data those problems which were not pedagogic in nature (leaving those concerned with navigational issues, minimal steps between related pages, etc.). In order to compare conditions, the (%) ratios of these problems to condition totals[4] were used. Again, however, this analysis revealed no significant main effects of condition for either all or high-severity problems (one-way between-subject ANOVAs, $F[2,20]=2.46$, $p=0.11$ and $F[2,16]=0.45$, $p=0.45$, respectively). See Tables 3-9(a) and 3-9(b). Further post-hoc tests for all problems (Tukey HSD: $W_3=22.56$) and high-severity problems (Newman-Keuls[5]: Control/Heuristic $W_2=43.81$, Heuristic/Principle $W_2=41.17$) also failed to confirm suspected between-condition differences in favour of the heuristics rather than principles.

Control	Heuristic	Principle
43.67 [17.00]	60.81 [17.18]	44.02 [18.07]

Table 3-9(a). Experiment 2. % ratios of non-pedagogic problems to total problems for each condition, all problems. Figures in [brackets] are standard deviations.

Control	Heuristic	Principle
25.00 [29.34]	48.44 [41.27]	36.57 [22.93]

Table 3-9(b). Experiment 2. % ratios of non-pedagogic problems to total problems for each condition, high severity problems. Figures in [brackets] are standard deviations.

Thus it does appear that even considering the non-teaching-content aspects of the PsyCLE material alone, the principles did not enable these subjects to elicit more of such sorts of problems than did the heuristics or Control materials. Nor was the previous Heuristic > Control effect replicated for such problems.

4 Based on UPTs rather than problem types.
5 Using the Games and Howell procedure for heterogeneous variances (Howell 1997).

6. Summary of Results

1. Overall, the full principles set used in this experiment failed to elicit from novice subjects a significantly higher number of usability problems than did either Nielsen's heuristics or a control condition without such guidance.

2. However, the heuristics did appear to have enabled the novices in this experiment to identify more high-severity problems than did Control subjects[6].

3. More problems were reported only once than were shared by more than one subject. Though these subjects were free from experimenter influence during problem reporting, their problem distributions were similar to those for Experiment 1.

4. Sorting on the product of frequency and mean severity is again offered as one means of prioritising problems. The top ten percent of prioritised UPTs reflects the particular difficulty which subjects reported with one part of the software evaluated in this experiment.

5. Comparisons by problem type (principle attributes) showed that no one condition elicited a significantly higher number of types (of those which featured in subjects' problem reports).

6. The problem types distribution suggests that Principle subjects were not identifying types different from those of Control (let alone Heuristic) subjects.

7. Even considering problems not primarily concerned with the teaching content of the software, no significant differences were found between the principle and heuristic (and control) materials in the numbers of such problems reported.

7. Discussion

7.1 Problem Counts

Once again, novices in this experiment proved unable to make use of even the full principles set to identify more usability problems than they did using the heuristics or control material. This was so given prior introduction to the materials and a training session. In fact the number of both problems and problem types found using the single-page heuristics was higher than that for the 31-page principles, though not significantly so (the only significant difference being between high-severity Heuristic and Control problems). Further, the analysis by types suggested that subjects using the principles were not gaining much additional benefit from the extended material, even over control subjects. The lack of a significant principles-heuristics effect also persisted for problems not directly related to the module contents.

---

6 This finding was unintentionally omitted from Connell & Hammond (1999).

The most likely reason for this (non-) result is that the sheer size of the full principles set (especially compared to the heuristics) 'swamped' any additional benefit that its content might have offered. This view is supported by feedback that Pilot volunteers were spending as much time looking through the principles materials as interacting with the software. (Though these relatively experienced[7] evaluators worked for only one session, they too were given the materials in advance.)

Following the Pilot, Principle subjects were provided with an additional 8-page summary and instructed to refer to the full set as necessary. Since even this appeared to be too much for these novices, one alternative strategy may be to reduce the relative sizes of both sets of materials, by excluding heuristics and/or principles not relevant to the software under evaluation. This is the approach adopted in Experiment 3 (Chapter 5), where three heuristics were compared with just two principles, this time using closely defined tasks and more restricted software content.

Another possible reason is that these novices were too inexperienced to be able to make sufficient use of the full principles set, even in summary form. This implies something about either the subjects' experience level, or the complexity of the principles materials, or both. As to the first, it is interesting that the mean questionnaire score for these subjects (Section 4.2) was higher than that for the novices in Experiment 1 (11.78% compared with 8.98%, unrelated t-test $p<0.05$), though that for the Principle subjects was lower at 8.13% (but not significantly so). The increase may well be an effect of the rapidly increased exposure of this undergraduate population (as others) to desktop software in the intervening two years. However, any effect of this increase was probably lost in the unfortunate (even if not significant) drop in Principle subjects' scores (condition order was varied as volunteers came forward).

As to the second possible reason, the reader may judge whether the content of the full principles (reproduced in Appendix D) are pitched at a level appropriate for Psychology undergraduates, even when using software specifically designed for themselves. The underlying rationale for this thesis is that with suitably judged principle-based material, it should be possible to encourage even novice evaluators to identify more usability issues than they would using simpler guides such as heuristics. Thus far, the judgement of what is an appropriate level appeared to have failed, at least as far as novices are concerned.

Perhaps the setting - the same computer laboratory in which these students had undergone many teaching sessions, and the one which they used for their own work - was such that they viewed it as just another tutorial exercise. In that case, using a different, more clearly 'experimental', environment might help to shift the experience away from the familiar.

[7] Pilot subjects were not sent the questionnaire.

However, at this distance from the experiment it is impossible to assess the effect of such manipulations, or of introducing novices from different sources (such as computer science undergraduates). The limitations of relying on mainly Psychology students as even a novice resource are acknowledged, and will be taken up in Chapter 8.

A more pragmatic reason for the lack of a result for principles over heuristics might be that subjects simply did not spend enough time reading the principles set beforehand (even following the training session), later relying mainly on their own judgement. (This even after being paid - but not much - to familiarise themselves with the materials.) It remains to be seen, therefore, how much training would be necessary to succeed with such novices; there is likely to be a trade-off between the amount of material which unmotivated volunteers[8] could be expected to assimilate, even when paid, and any benefits which might ensue.

One positive (if minimal) result was that, unlike in Experiment 1, the heuristics did manage to elicit more high-severity problems than the control materials. This can be seen as a partial success for the experiment, perhaps representing a 'baseline' for Nielsen's ten heuristics. On *this* single result, it seems that novices can make use of such short heuristics in identifying more important usability issues than they would otherwise, *provided* that (a) they are introduced to the materials beforehand, (b) the software evaluated is relevant to themselves, and (c) they are left to get on with it. It is the author's view that with brief materials the software (and task) orientation will be more important than training (see below for a discussion of the problem reporting method). In Experiment 3 (Chapter 5) we will examine the effect of reducing both software and task scope on novices' ability to make use of both heuristics and principles.

Another positive outcome was the richness of many of the subjects' problem reports, as illustrated in the sample protocol in Appendix J. Though of course not all subjects were as thorough, this form of reporting generated a set of substantive issues which could be used by future developers.

7.2 Problem Overlap

The crucial difference between Experiments 1 and 2 is that in Experiment 1 the experimenter (the author) sat with subjects while they 'thought aloud', and the experimenter wrote down subjects' comments and reactions. In Experiment 2, however, subjects were left free to enter usability problem reports on their own, without experimenter intervention. The possible results of that difference in procedures will now be explored.

Although in Experiment 1 every attempt was made to avoid influencing subjects' responses, a criticism of that experiment might be (a) that the proximity of the experimenter may have

[8] Two hours' subject time per term is compulsory for first-year Psychology undergraduates.

influenced what the subjects were able to report, and (b) that what was recorded was couched in the experimenter's, rather than the subject's, terms. In Experiment 2, both of these possible biasing factors were removed, potential experimenter influence being confined to the introductory parts of the evaluation materials. (See Appendix C for a transcript.) It is significant, therefore, that the distributions of reported usability problems for Experiment 2 subjects were very much like those for Experiment 1, with similar proportions of problems being reported only once. Though this profile is typical of many UEM studies (Jacobsen et al. 1998a, 1998b), in these experiments usability problems have generally been extracted from problem reports in the form of video and/or audio protocols, rather than (just) subjects' own written responses. The fact that the degree of problem-sharing (overlap) between subjects in Experiments 1 and 2 was so similar thus implies something about a tendency of subjects to identify separate usability issues, rather than variations in the method of problem recording. (It also implies something about the verisimilitude of written, as opposed to voiced, protocols.) These issues will be taken up in Chapter 8.

7.3 Problem Extraction

Even if the reliability of subject accounts can be asserted, there still remains the lengthy and complex process by which individual usability problems are extracted from subject protocols once recorded. That is, the reliability of the whole problem reduction process, from separate subject protocols to a single set of unique UPTs, has yet to be established. In all the experiments reported in this thesis, problem reduction has been carried out by the same experimenter (the author); it would not be too surprising, then, if a consistent pattern of problem distribution emerges. For that reason, attempt has been made to validate at least part of this process for Experiment 1 (described in Chapter 4) and Experiment 3 (described in Chapter 5). Chapter 4 also contains a detailed discussion of the problem reduction process. Though Experiment 2 failed to produce significant results for the principles over heuristics, the insights gained into the issues of problem extraction and reduction, plus the later generation of cumulative problems curves for all three experiments, are positive outcomes for this experiment.

7.4 Problem Prediction

One further issue. Even having asserted the reliability of subject accounts as faithful reports of a 'dialogue' with a system, and after claiming some validity for the process by which these accounts are processed, it still remains to be demonstrated that the usability problems which have been identified do, in fact, 'exist'. That is, just because a majority of even experienced evaluators claim the same issue to be a 'problem', it does not mean that any, let alone most, of a given user population would encounter that same problem in actual use. This view - that a proportion of the issues identified in think-aloud usability sessions are an artefact of the process itself, rather than having an externally verifiable basis - is an important one for this thesis, and will also be taken up in Chapter 8. For now, we look forward to Experiment 3

(Chapter 5), where attempt will be made to assess the ability of both heuristics and principles to elicit accurate predictions of observable usability problems.

Summary of Chapter 3

The full principles set used in Experiment 2 failed to elicit from further novice subjects significantly more usability problems than did either Nielsen's heuristics or control materials without such guidance. This was also so for problem types based on the principles and for non-pedagogic problems. However, subjects using the heuristics did manage to find more high-severity problems than did similar novices using the control. It was again found that more problems were reported by single subjects than by more than one subject, the pattern of problem distribution being very similar to that for Experiment 1. The lack of experimenter intervention lead to a discussion of the reliability of subject accounts and the processes of problem extraction and reduction.