Chapter 5: Experiment 3

1. Introduction

In Experiment 1 (Chapter 2) it was shown that experienced but not novice evaluators could make use of the short (three-page) version of the principles set to identify more high severity problems than using Nielsen's heuristics alone. In Experiment 2 (Chapter 3) similar novices to those in Experiment 1 also proved unable to identify more problems, even when given prior training on the 31-page version of the principles set. It was later shown (in Chapter 4) that the detection rates from both experiments were lower than the 25% to 33% required to satisfy the Nielsen prediction for problem-sharing (that 75% of problems could be uncovered by 3 to 5 experienced evaluators). It was speculated that the reason for the discrepancy between these and other published results may be that other researchers have allowed implicit categorisations to influence the early stages of problem reduction process. It was also suggested that the failure of novices to make differential use of the principles set was due to the relative size and complexity of these materials compared with the single-page heuristics.

In this experiment we shall see if a further group of novices could improve on their predecessors, this time guided by only two attributes (or sub-principles) from the full set used in Experiment 2. Comparison was made with the corresponding three of Nielsen's and Molich's ten heuristics. The tasks to be evaluated were designed so as to allow subjects to explore as much as possible of these two (sub-)principles or three heuristics, making use of an error model to differentiate between task levels. In contrast to the previous two experiments, where detection rates were measured against the problems found by all subjects in a population, subjects' predictive performance was measured against a set of observed problems exhibited by test subjects on the same tasks. This allowed assessment of not only the predictive power of each set of evaluation materials, but also any tendency to 'over-predict' problems which did not appear in the test sessions.

This experiment, then, made the distinction between problems which are predicted (that is, reported in a usability inspection as having a potential to impact users) and observed (either previously or concurrently, in observations or user tests). In this case, the comparison was between problems reported in two conditions (Principle and Heuristic) in which subjects were specifically asked to predict the nature of the difficulties which other users might experience, and a third, Test condition. This enabled differentiation between correct and incorrect predictions (hits and misses), as well as over-predictions (false positives). It also allowed comparison between detection rate as previously assessed and two new measures of predictive performance, namely hit rate and prediction rate.

There have been surprisingly few UEM studies which have compared predicted and observed problems. The results from those which have done so are as variable as those

elsewhere. Nielsen (1994d) reported a mean rate of correct predictions (hit rate) of 29% for heuristic evaluation, while Desurvire et al. (1992) had mean hit rates of 34% and 12% for expert and novice users of the same method (and 19% and 6% on cognitive walkthrough). Cuomo & Bowen (1994) reported respective hit rates of 22%, 46% and 58% for guidelines, heuristic evaluation and cognitive walkthrough users. Most of these would easily qualify for a '3 to 5'. However, without being able to inspect the problem sets and predicted-observed matchings it is difficult to assess their reliability. John & Mashyna (1997) pointed out that the Cuomo & Bowen results rely on types rather than instances. In John & Mashyna's own (ibid.) case study of a cognitive walkthrough (which included detailed problem listings), only 5% of test problems were predicted precisely (and other 5% vaguely) by a researcher new to the method. It appears that the effectiveness of inspection methods in making accurate problem predictions remains to be established.

A further difference between this and the previous experiments lies in the tasks and software used. In contrast to Experiments 1 and 2, where open-ended, exploratory tasks were chosen, in this experiment tasks were selected to focus on specific aspects of an error model (Reason's Skill-Rule-Knowledge, or S-R-K, model, citations below). This was partly to allow comparisons between different levels of the model, but also to reduce any tendency towards the low overlaps and low detection rates which were a feature of both previous experiments. For while it was demonstrated that the discrepancy between the Experiments 1 and 2 problem profiles and those required by Nielsen's '3 to 5' model can be eliminated by problem categorisation, it remains to be shown that the differences were not due to the nature of the tasks and/or system types. To this end, in Experiment 3 both tasks (text editing, data filtering) and software (Microsoft® Word and Excel) were selected to embody closed rather than open-ended procedures, enabling assessment of success, failure and error-making in goal-directed action.

The aim, then, was to achieve success with novices for the principles over the heuristics (if only for two principles), and to see if the previous lack of compliance with the '3 to 5' would hold even with reduced-scope tasks and sub-systems. Comparison of predicted with observed problems would enable a more realistic assessment of both success and compliance than was afforded by the first two experiments. It was hoped that the reduced size and scope of both tasks and evaluation materials would enable this new batch of novices to focus more closely on the relevant usability issues (here, in the form of error prediction) than in the previous two experiments. In what is believed to be a novel contribution, the differentiation between Skill, Rule and Knowledge tasks would also allow assessment of the effect of task level on the ability of such materials to enable accurate prediction of observed problems.

2. Reason's S-R-K (GEMS) Model

Reason's (1987b, 1990) general model of human error (GEMS) attempts to encapsulate the whole range of individual human error making into a single three-level model. It situates the distinction between mistakes (an error in the intention to act) and slips (errors in carrying out the intention) within skill-based, rule-based and knowledge-based task levels. Slips remain at the skill level, arising from a failure of monitoring of known tasks or sub-tasks. Mistakes, however, can occur due to problem-solving failures at either of the other two levels: at the rule level, they can involve the incorrect transfer of "strong but wrong" rules (which work in other, related, tasks) into a new or different domain, while at the knowledge level they can be associated with a range of error-reduction strategies including trial and error. According to the model, rule-based error-making is characterised by a reluctance to look for solutions at the knowledge level (when faced with a new task, we try what we know, or think we know, before attempting other solutions whose outcome is uncertain). Only when forced to do so, by failure at a lower level, do we 'dip into' the level above; but the achievement of some new technique or strategy at the knowledge level may allow a return to the rule level (which can in turn lead to more slips at the skill level).

GEMS is an elaboration of earlier models including Norman's theory of action (Norman 1981, 1986), Norman's slip-mistake dichotomy (Norman 1983), Rasmussen's original (1982, 1986) three-level model, and Reason's own (1987a) error taxonomy. For the purposes of this Chapter, these complexities will be simplified by reducing the model to its essentials, bound up in the acronym S-R-K (Skill-Rule-Knowledge). The next Section will describe the seven tasks, designed around the three levels, on which Experiment 3 was based. Before doing so we shall look briefly at some of the other HCI research which has made use of these models.

Examples of the use of the three-level model in error analysis include Zapf et al. (1992), Wright et al. (1993) and Salminen & Tallberg (1996). Zapf et al. (ibid.) (and Prümper et al. 1992, Brodbeck et al. 1993) claimed to have integrated the model into a single error taxonomy based on action theory (Norman 1981, 1986), adding a fourth level dubbed the "knowledge base for regulation" (Zapf et al. 1992). Their validation of the taxonomy involved both novice and expert computer users. Wright et al. (1993) used the GEMS model to classify errors made on a graphics package into skill-, rule- and knowledge-based. Salminen & Tallberg (1996) classified fatal and serious occupational accidents according to the original Rasmussen (1982, 1986) taxonomy. However, to the author's knowledge there have been no attempts to relate the S-R-K or GEMS models to the predictive ability of inspection methods. Specifically, studies of heuristic evaluation do not appear to have used the three-level model to differentiate between evaluator tasks. In this respect, Experiment 3 may represent a novel contribution to the UEM literature.

3. Software and Tasks

3.1 The Software

The software used for Experiment 3 comprised two of the most heavily used Microsoft® Office tools, namely Word (version 5.1a for Macintosh) and Excel (version 5.0 for Macintosh). They were chosen because (a) they and the hardware on which they ran was familiar to the available novice population (mainly Psychology undergraduates), (b) they embodied the sort of closed, goal-directed tasks which were required for this experiment, and (c) the possible range of such tasks varied from very simple to complex, from familiar to new. The recruitment of subjects was couched in terms of "learn new computer skills", a side-effect of the experimental sessions being subjects' acquisition of both general (keyboard techniques) and specific (data analysis) skills.

The software ran on a Macintosh IIci platform under System 7. Screen size was 240 x 190 mm. All tasks were enclosed within a single window (document or worksheet), and except in one case no task required window scrolling.

3.2 The Tasks

There were seven experimental tasks (a further pair were designed but not used, thus serving as an additional 'learning experience' for those subjects who completed early). As described above, there were three task levels, namely Skill, Rule and Knowledge. Tasks 1 and 2 were Skill tasks, with tasks 3 and 4, 5 and 6 forming Skill-Rule pairs. Task 7 was the single Knowledge task. Subjects attempted the tasks in sequential order; no subject failed to attempt all seven.

All tasks were intended to be representative of real use (of a word processor or spreadsheet), with clearly defined objectives. The type and length of tasks were intended to present ascending difficulty, from Skill to Knowledge levels. The Skill-Rule pairings of tasks 3 and 4, 5 and 6 were designed to manipulate the tendency to rely on known (Rule-based) techniques rather than explore novel (Knowledge-based) alternatives (Reason 1987b, 1990). Thus the acquisition of a Skill technique using one tool (Word) would be followed by a corresponding task in another tool (Excel) for which the identical technique does not work in the release versions used.

Each task was presented entirely on-screen, in separate Word documents or Excel worksheets. The task objective appeared at the top of the screen. The strategy used to decide whether or not to provide instructions or assistance in achieving that objective was as follows. Procedures had been designed to manipulate task level difficulty, rather than to represent a 'real' test of the software. Thus, though a goal for subjects was to enhance computer use, it was not necessary to have them flounder and fail (like real users would) unless that was part of the experimental manipulation. To this end, some degree of assistance was given in the form of on-screen 'step by step' instructions and verbal

assistance, but otherwise subjects were to try (and fail) for themselves. Help was not offered until subjects failed or made the same error repeatedly. Test condition subjects were given no assistance of this sort until they had completely failed.

3.2.1 Skill Tasks (Tasks 1 and 2)

Tasks 1 and 2 taught the use of drag-and-drop to move either text or spreadsheet cells to new locations. They were Skill tasks in that little rule-following was involved, operating at the hand-eye coordination rather than problem-solving level. (Both tasks did involve decisions as to the order in which text or cell ranges would be moved.) The main difficulty was in the selection of the source text or cells and the drop at the desired location.

Task 1 (Word: Move sentences) involved re-arranging the sentences in a 138-word paragraph by use of drag-and-drop alone (i.e. using the mouse and cursor rather than cut and paste). This required the selection, movement and drop of successive sentences. The goal (how the re-arranged paragraph should look when completed) appeared below the task text. Subjects were told (but not shown unless failing) how to do text drag-and-drop: "Select the text you want to move, then hold down the mouse button and drag". Other than this, subjects were free to choose the order in which sentences were selected.

Task 2 (Excel: Move cells) involved the extraction and relocation of the incorrect cells from sets of numerical data, by use of drag-and-drop alone (i.e. using the mouse and cursor rather than cut and paste). The incorrect, unordered data was mixed with ordered data in two columns headed **10..19** and **20..29**; the task was to extract this data and form the correct sequences in two new columns headed **30..39** and **40..49**. The goal (how the columns should look when completed) appeared below the task cells. Subjects were told (but not shown unless failing) how to do cell movement: "Select the cells you want to move, hold the cursor over a cell edge until it changes to a left-pointing arrow, then hold down the mouse button and drag". Other than this, subjects were free to choose the order in which cells or cell ranges were moved.

3.2.2 Skill-Rule Task Pairs (Tasks 3 and 4, 5 and 6)

The next two pairs of tasks were designed to manipulate the Rule-level tendency to use known solutions ("strong but wrong rules") rather than search for novel techniques. (According to the S-R-K model, these would be found at the Knowledge level). To this end, each of tasks 3 and 5 set up an expectation in the user of being able to use a certain keyboard technique in Word which did not work in Excel (the two releases of Word and Excel used in the experiment are inconsistent in this regard). In tasks 4 and 6 the subject was to find the correct technique on her/his own (all failed). Apart from this manipulation, tasks 3 and 5 were like tasks 1 and 2 in operating mainly at the hand-eye coordination level.

3.2.2.1 Tasks 3 and 4

Task 3 (Word: Emphasise words) involved the emboldening of single words and a set of short text, using only the appropriate keyboard shortcut (⌘ and B together). This required the selection of the required text and the manipulation of the Command (⌘) and B keys in the correct order. There were 10 single words in the first (three-paragraph) text; 4 short sentences (arranged list fashion) formed the second text. To aid location, the words to be emphasised were in capitals. Subjects were told (but not shown unless failing) how to do the keyboard shortcut: "Hold down the ⌘ key and press B".

Task 4 (Excel: Emphasise column headings) involved the emboldening of all the column headings (also in capitals) of a set of (six) columns of data, "using *only* an appropriate keyboard shortcut". This required the selection of the required cell or cells and the manipulation of the Command (⌘), Shift and B keys (not just the ⌘ and B keys) in the correct order (⌘+Shift+B or Shift+⌘+B). Subjects were not told in advance what the shortcut was.

3.2.2.2 Tasks 5 and 6

Task 5 (Word: Correct spelling) involved the use of the Shift and arrow keys (<-- -->) to select pairs of characters from incorrectly spelt words, prior to correcting the characters. This required (1) placing the cursor at the front or end of the characters to be corrected, (2) holding down the Shift key while moving the cursor to the right or left, and (3) re-typing. There were three such words, in separate paragraphs. To aid location, the words were in bold type. Subjects were told (but not shown unless failing) how to do the keyboard manipulation, in the form of on-screen step-by-step instructions.

Task 6 (Excel: Correct spelling) involved the selection of adjacent characters from incorrectly spelt words (in three column headings), prior to correcting the characters. This required the use of the mouse and cursor only, Shift+arrow key combinations merely extending focus to the adjacent cell. The steps involved were (1) select the required cell, (2) place the cursor at the front or end of the characters to be corrected, (3) hold down the mouse button while moving the cursor to right or left, (4) re-type. (Step (2) could be performed in-cell or in the formula bar). On-screen instructions were to "use the keyboard keys to select the characters to be changed".

3.2.3 Knowledge Task (Task 7)

Task 7 (Excel: Filter a table) involved the extraction and relocation of three groups of cells from a column of mixed data, using Excel filtering (Advanced Filter). This was the 'automated' equivalent of task 2, in which filtering was performed using drag-and-drop alone. In this case the filtered data was to be copied rather than moved, thus preserving the original data. (The default filter mode, AutoFilter, filters in-place, thus changing the data's appearance.) The goal (how the cells should look when completed) appeared below the task cells. Subjects were told (but not shown until failing) how to do the whole operation, in

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Task 7: Filter a table**											
2												
3	By using **Advanced Filter ...**, copy the scores from groups A, B and C into the table on the right.											
4	Use the far right columns K and L as temporary placeholders. There are 6 subjects per group.											
5												
6												
7	**ie. make this ...**											
8												
9		**Subject**	**Score**	**Group**			**Subject**	**Group A**	**Group B**	**Group C**	**Group**	**Score**
10		1	23.2	A			1					
11		2	34.6	C			2					
12		3	35.1	C			3					
13		4	34.8	C			4					
14		5	24.5	A			5					
15		6	23.5	A			6					
16		7	11.4	B								
17		8	36.2	C								
18		9	12.5	B								
19		10	13.0	B								
20		11	35.5	C								
21		12	23.0	A								
22		13	34.0	C								
23		14	22.4	A								
24		15	13.5	B								
25		16	11.9	B								
26		17	22.5	A								
27		18	12.6	B								
28												
29												
30												
31												
32												
33												
34	**look like this ....**											
35												
36		**Subject**	**Score**	**Group**								
37		1	23.2	C								
38		2	34.6									
39		3	35.1									
40		4	34.8									

The procedure for each group (A, B or C) is as follows:
(Make use of any dialogue box settings remaining from the previous attempt).

1. Enter the Group name (A, B or C) to be filtered into cell D10.
2. From the menus, select **Data --> Filter --> Advanced Filter ...**
   In the dialogue box:
3. Select **Copy to Another Location**.
4. Specify the **List Range** to be filtered: cells B9 to C27 inclusive.
5. Specify the filter **Criteria Range**: cells D9 and D10.
6. Specify the first row of the range of cells to **Copy to**: K9 and L9
7. Click **OK.**
8. Move the resulting filtered scores into the appropriate column of the middle table.

Figure 5.1. Experiment 3. Task 7 (Excel filtering). The task required the extraction of the three sets of scores from columns B and C, by filtering on a label (filter criterion) in cell D10. The temporary destination for each filtered set was columns K and L. Rows 34 to 55 show the result after filtering the final set (Group C).

the form of on-screen step-by-step instructions. Figure 5.1 shows the screen layout at the start of the task.

The task objective was to copy the three sets of six scores labelled A, B and C from columns B and C, finishing with the scores correctly placed in the table in columns G to I. This required the use of a filter criteria range (cells D9 and D10), D10 being filled with one group label at a time; the filter operation had thus to be performed three times in succession. Since the procedure risks (apparent) corruption of the source data, the temporary destination for each filtered group was to be columns K and L; the final step in each sequence was to move the just-extracted group into the appropriate column of the final table (using the drag-and-drop technique learnt in task 2).

This complex task was intended to be representative of real data manipulation (Psychology students are likely to have to separate mixed data into tables prior to analysis, and it is a good idea to preserve raw data). However, it (along with much else in Excel) was unlikely to have been used by undergraduates (none had), thus representing a suitable Knowledge-level challenge. Though such precise instructions as were provided would not normally be available (Excel's on-line help for the Advanced Filter command is typically sparse), even with this assistance there was much scope for error-making, as the Results will confirm. One of the reasons for this choice of task is that the 'Copy To' option in the Advanced Filter dialogue box is not selected by default. Thus after repeated filtering from the same dataset (which, after the first operation, requires only one dialogue box change per attempt), it is likely that it will be omitted (a Skill-level slip). In such an event, the filter mode reverts to 'in-place', hiding all but the filtered cells (the 'lost' data can in fact be recovered by use of the Show All or Unhide commands). Worse, the 'undo' option is not available following execution. This is a disconcerting result, representing the most serious of the many likely difficulties with this menu command.

As mentioned above, the tactic of providing step-by-step instructions was intended to make what would be a protracted series of stabs-in-the-dark, leading to probable failure, into an exercise in confined error-making. While it was therefore unrepresentative of real problem-solving, both the task scenario and information provided were the same for all subjects. This issue will be taken up in the Discussion.

4. The Evaluation Materials

The author's full principles set and Nielsen's heuristics are described in Section 3.3 and 3.5 of Chapter 1, respectively. Only Heuristic and Principle subjects were provided with evaluation materials. Test subjects were given no guidance beyond the on-screen task instructions.

As stated above, part of the remit for this experiment was to restrict the size and scope of both tasks and evaluation materials. To this end, Heuristic and Principle subjects were provided with only those parts of the materials which were most relevant to the (Skill-Rule-Knowledge) task manipulation. These were considered to concern mainly error-handling

and prevention, plus consistency. They therefore comprised three of the ten Molich & Nielsen (1990) heuristics plus two of the 30 attributes[1] from the principles set. Thus 30% of Nielsen's heuristics were to be compared with just 7% of the full principles. (While it is possible and indeed likely that other parts of both sets were relevant to the tasks involved, neither Heuristic nor Principle subjects had access to them).

Heuristics: the three heuristics used were taken in full from Molich & Nielsen (1990), thus being identical in all but numbering to the equivalent materials from Experiment 1 (Appendix A). They were as follows.

Be Consistent (Heuristic 4)
Provide Good Error Messages (Heuristic 8)
Error Prevention (Heuristic 9)

Principles: The two principles (attributes) used were as follows. Their content was identical (in all but numbering) to the equivalents in Experiment 2 (Appendix D).

Error Management (Attribute 8)
Consistency (Attribute 21)

5. Method

5.1 Design

Experiment 3 was a 3 (task level) x 2 (prediction condition) design. The task levels were Skill, Rule and Knowledge. The prediction conditions were Heuristic (N = 8) and Principle (N = 8). Predictions were assessed against a third, Test, condition (N = 7). Only one subject population, novices (different individuals from those in experiments 1 and 2), was involved.

5.2 Subjects

Novice status was assessed by the same questionnaire as used in Experiments 1 and 2 (with additional questions concerning Word and Excel experience), administered at the start of each experimental session. (See Appendix E for a transcript.) Most subjects were University of York Psychology undergraduates. Most were female and aged 18-21. All were paid at hourly rates or received subject credit.

Experiment 3 subjects' mean questionnaire score was 10.22% (median 10, standard deviation 3.64). A one-way between-subjects ANOVA (Table 5-1) showed no significant difference between conditions ($F[2,22]=1.10$, $p=0.35$).

[1] The full principles set comprised 23 principles, some of which had sub-principles or 'attributes', thus 30 attributes in total. **Consistency** and **Error Management** did not have sub-principles.

Heuristic	Principle	Test
10.63	11.25	8.57

Table 5-1. Experiment 3. Mean questionnaire scores per subject, by condition.

A one-way between-subjects ANOVA of the novice questionnaire scores from Experiments 1, 2 and 3 (table 5-2) showed no significant difference between the three novice populations ($F[2,69]=2.23$, $p=0.12$). Thus the increase in novice experience level between Experiments 1 and 2 (unrelated t-test, $p<0.05$ one-tailed) did not continue into this third novice group (Experiments 2 and 3, unrelated t-test, $p=0.14$ one-tailed).

Experiment 1	Experiment 2	Experiment 3
8.98	11.78	10.22

Table 5-2. Experiments 1, 2 and 3. Mean questionnaire scores per novice subject.

In this experiment one post-pilot subject's score (40%) fell outside the novice experience range (0 to 20%); this subject's results were not included in this or the succeeding analysis.

5.3 Procedure

Subjects were assigned to conditions as they volunteered, in order Heuristic, Principle, Test. Heuristic and Principle subjects were presented with the appropriate evaluation materials at the start of the session and were instructed to refer to the materials throughout the session.

For Heuristic and Principle subjects, the object of the experiment was for subjects to predict the usability problems which "people such as yourself" would experience when attempting the identical tasks. To this end, these subjects were asked to imagine how others would fare, given the same on-screen instructions but no other assistance (for example, after failing on the first attempt). Subjects were asked to 'think-aloud', that is, to verbalise their predictions and comments, and were told that they might be reminded to do so.

For Test subjects, the object was to attempt the tasks 'on their own', that is, with no assistance beyond that provided on screen. To this end, these subjects were forewarned that no help would be given unless they completely failed to achieve the goal, and that they should make more than one attempt if necessary; only after repeated failure would the experimenter intervene. Subjects were also asked to verbalise their reactions, and were told that they might be reminded to do so.

In all three conditions the experimenter (the author) sat with the subject and recorded (on paper) his or her error-making behaviour. In the Heuristic and Principle conditions subjects' predictions and other comments were also recorded; in the Test condition close attention was paid to the errors made, while also recording associated comments and reactions. Great care was taken not to direct subjects, but unclear or incomplete comments were more fully elicited. With Test subjects, no discussions or other distractions from the tasks in hand were

entered into; Heuristic and Principle subjects, however, were encouraged to consider the likely causes and consequences of each problem which they envisaged. At the end of each task Heuristic and Principle subjects were asked to look again at the evaluation materials and to add any comments or further predictions which they might have missed.

Each subject worked through the tasks in numerical order; none failed to attempt all seven. After the end of the session, Test subjects were finally offered some feedback on the difficulties experienced (the experiment had been billed as a 'learning software skills' experience). Most subjects (Heuristic and Principle as well as Test) said that they had learnt valuable skills; many reported that they would not have been able to complete task 7 (in particular) without the on-screen instructions.

As stated above, the initial dependent measure was to be the number of problems (errors) either predicted or observed. In order to extract this measure, the experimenter had first to assess any duplication in each subject protocol (the 'within subjects' stage of the problem reduction process), in the same fashion as in Experiments 1 and 2. Since each task could be treated independently for problem-reduction purposes, the second, 'between subjects' stage proved to be far less difficult (involving far fewer problems per task) than in the previous experiments. (The simple nature of most tasks also helped in this regard.) The major difficulty came in the matching of problems between the predicted (Heuristic and Principle) and observed (Test) conditions. It is therefore this stage of the reduction process for which some inter-rater reliability was later sought. During problem identification and matching it became apparent that a number of implicit assumptions had been made concerning both subjects' experience level and what qualified as a problem prediction. These were set down on paper for future reference (see Appendix F for a transcript). The reduction process was also aided by the separation of problem occurrences (reported or observed) from attributed causes (and resulting recommendations). This reporting format has since been seen to be close to that recommended by Cockton & Lavery (1999) in their proposed structured framework for problem extraction (SUPEX).

Once the matching process was complete, severity ratings could be assigned to the observed problems in the Test condition. This was done by the experimenter, taking into account the consequences for the test users of each error.

5.4 Pilot Study

A pilot study had been carried out in order to try out the procedures, tasks and software. This involved six volunteers, two per condition, whose responses were not included in the analysis. As a result of the pilot it was decided to include only seven tasks (a further pair, involving Word style formatting, had been originally designed), and to simplify the Heuristic and Principle recording forms (but leaving space for causes and consequences).

6. Results

6.1 Experimental Measures

In contrast to experiments 1 and 2, this experiment was designed to enable comparison between predicted and observed problems. That is, rather than measuring any one subject's performance out of the total UPTs for that subject population, in this experiment the predictive ability of Heuristic and Principle subjects was assessed against problems observed on the same tasks in the Test condition. Thus it was possible to replace the detection rate as measured in Experiments 1 and 2 with a 'hit rate' based on total *observed* UPTs per task. This relied on assessment of the matches (hits) between predicted and observed problems. It was also possible to measure both the misses (observed problems which were not predicted) and the false positives (predicted problems which were not observed) (adapted from Gray & Salzman 1998)[2]. Table 5-3 sums up the relationship between hits, misses and false positives (FPs).

	Observed	Not observed	
Predicted	HIT	FP	Total predicted
Not predicted	MISS		
	Total observed		

Table 5-3. The relationship between hits, misses and false positives (FPs) for predicted versus observed problems. Adapted from Gray & Salzman (1998).

Out of any set of observed problems, there will be a proportion of hits and a remainder of misses. Out of any set of predicted problems, there will be a proportion of hits and a remainder of FPs. Thus Misses are the difference between total observed and hits, while FPs are the difference between total predicted and hits:

Hits (matches) = UPTs both predicted and observed (4)

Misses = UPTs observed but not predicted (5)

FPs (False Positives) = UPTs predicted but not observed (6)

Misses = Total observed UPTs - Hits (7)

FPs = Total predicted UPTs - Hits (8)

The above absolute measures permit the use of a range of ratio (percentage) measures in addition to hit rate. These include the ratios of hits to misses, FPs to observed UPTs, plus others. Three measures including hit rate were chosen as being best representative of predictive performance of any one subject population:

Hit rate = hits / total observed UPTs (9)

Accuracy = hits / total predicted UPTs (10)

Redundancy = FPs / total predicted UPTs (11)

[2] Gray & Salzman (1998) also proposed that use could be made of the remaining cell in the grid, that is, the problems which are not observed but which could be predicted to be absent (for example, following fixes to previous problems).

Hit rate refers to the proportion of hits (matches between predicted and observed) made against any target number of UPTs (those observed in the Test condition, here). **Accuracy** refers to the proportion of hits compared with the total predicted UPTs (i.e. how well subjects do in relation to their overall predictions). **Redundancy** is a measure of by how much subjects over-predict compared to their overall predictions[3]. While subjects might be expected to attempt to be as accurate (make as few misses) as they can, the reasons for over-predictions (false positives) are more difficult to identify. In particular, it is possible that some FPs are an artefact of the experimental situation and would not otherwise arise. The issue of the potential persistence of false positives will be taken up in the Discussion and later in Chapter 8.

Using the above ratio measures, it was possible to compare the mean performance of Heuristic and Principle subjects both overall and at each task level (within-level comparisons). It was also possible to make comparisons between the three levels for each of the two predictive conditions (between-level comparisons). Similar analyses were performed for the most frequently observed problems and for the Knowledge level problems deemed to be most serious. As a contrast with Experiments 1 and 2, however, we shall begin with the absolute measure of predicted problems.

6.2 Problem Counts Within Levels

A two-way split plot ANOVA of the overall predicted problem counts (Table 5-4) showed neither a significant main effect of condition (Heuristic vs. Principle, $F[1,14]=0.53$, $p=0.48$) nor a condition x task level interaction ($F[2,28]=1.36$, $p=0.27$). Separate pair-wise (Newman-Keuls[4]) tests confirmed that there were no significant between-condition differences at any of the three levels (Skill: $W_2=2.58$; Rule: $W_2=1.95$; $W_2=1.65$). Thus comparison of predicted problems alone (as in Experiments 1 and 2) would once more have failed to reveal significant differences between the two sets of evaluation materials, either overall or at any of the three levels.

Skill		Rule		Knowledge	
Heuristic	Principle	Heuristic	Principle	Heuristic	Principle
5.50 [2.93]	6.88 [1.36]	4.13 [2.30]	3.88 [0.83]	2.63 [1.51]	3.13 [1.55]

Table 5-4. Experiment 3. Mean predicted problems per subject, all problems. Figures in [square brackets] are standard deviations.

However, use of the three ratio measures described in Section 6.1 allows for comparison of different performance measures. Table 5-5 shows the mean hit rate, accuracy and redundancy of Heuristic and Principle subjects at each level.

---

3 Sears (1997) used similar ratio measures, respectively 'thoroughness' and 'validity' for hit rate and accuracy, plus a third measure, 'reliability' (stdev hits / mean hits).

4 Using the Games and Howell procedure for heterogeneous variances (Howell 1997).

	Skill {21}		Rule {5}		Knowledge {14}	
	Heuristic	Principle	Heuristic	Principle	Heuristic	Principle
Ht rate (%)	18.54 [8.09]	19.84 [3.00]	63.54 [22.24]	62.50 [16.67]	8.04 [7.08]	14.29 [5.40]
Accuracy (%)	76.08 [21.55]	61.07 [10.10]	84.93 [22.34]	81.46 [18.07]	39.52 [19.95]	71.88 [26.33]
Redundancy (%)	23.92 [21.55]	38.93 [10.10]	15.07 [22.34]	18.54 [18.07]	60.48 [19.95]	28.13 [26.33]

Table 5-5. Experiment 3. Mean hit rate, accuracy and redundancy per subject, by condition within level, all problems. Figures in {curly brackets} are the numbers of observed UPTs (in the Test condition) at each level. Figures in [square brackets] are standard deviations.

Two-way split plot ANOVAs again showed no significant main effect of condition (Heuristic vs. Principle) for any of the three measures (hit rate: $F[1,14]=0.28$, $p=0.60$; accuracy: $F[1,13]=0.88$, $p=0.37$; redundancy: $F[1,13]=0.88$, $p=0.37$). There was no condition x task level interaction for hit rate ($F[2,28]=0.43$, $p=0.66$), but there were significant interactions for both accuracy and redundancy (both $F[2,26]=5.65$, $p<0.01$). Thus overall neither Heuristic nor Principle subjects managed to predict a greater proportion of the total (of 40) UPTs observed in the Test condition. Neither group of subjects showed a greater tendency to over-predict, and neither were more accurate in their predictions.

However, further pair-wise (Newman-Keuls[5]) tests at each level revealed significant between-condition differences at the Knowledge level, but not at either the Skill or Rule levels, for both accuracy and redundancy (Knowledge: accuracy and redundancy both $q_{0.05}[2,12]=3.08$, $W_2=26.09$); Skill: both $q_{0.05}[2,9]=3.20$, $W_2=19.04$; Rule: both $q_{0.05}[2,13]=3.06$, $W_2=21.98$). Tests for hit rate failed to show such differences at the Knowledge level or (as expected) at the other two levels (Knowledge: $q_{0.05}[2,13]=3.06$, $W_2=6.81$; Skill: $q_{0.05}[2,8]=3.26$, $W_2=7.03$; Rule: $q_{0.05}[2,12]=3.08$, $W_2=21.40$). Inspection of Table 5-5 shows that the direction of the Knowledge level differences on both accuracy and redundancy was opposite to those at the Skill and Rule levels (i.e. in favour of the Principle condition).

Thus on the Knowledge task but not the Skill or Rule tasks, the two principles did enable these novices to be more accurate (and less redundant) in their predictions of the 14 Knowledge-level problems observed in the Test condition than did the three heuristics. However, the proportion of Knowledge level problems correctly predicted was no higher using the principles than the heuristics.

6.3 Problem Counts Between Levels

Table 5-6 shows the three same ratio measures as in Table 5-5, now pivoted to reflect by-level comparisons.

5 Using the Games and Howell procedure for heterogeneous variances (Howell 1997).

	Heuristic			Principle		
	Skill	Rule	Knowledge	Skill	Rule	Knowledge
Hit rate (%)	18.54 [8.09]	63.54 [22.24]	8.04 [7.08]	19.84 [3.00]	62.50 [16.67]	14.29 [5.40]
Accuracy (%)	76.08 [21.55]	84.93 [22.34]	39.52 [19.95]	61.07 [10.10]	81.46 [18.07]	71.88 [26.33]
Redundancy (%)	23.92 [21.55]	15.07 [22.34]	60.48 [19.95]	38.93 [10.10]	18.54 [18.07]	28.13 [26.33]

Table 5-6. Experiment 3. Mean hit rate, accuracy and redundancy per subject by level within condition, all problems. Figures in [square brackets] are standard deviations.

The same two-way split-plot ANOVAs as reported in Section 6.2 showed significant effects of task level (Skill vs. Rule vs. Knowledge) for all three measures (hit rate: $F[2,28]=96.26$, $p<0.001$; accuracy: ($F[2,26]=7.53$, $p<0.01$); redundancy: ($F[2,26]=7.53$, $p<0.01$). As previously reported, there was no condition x task level interaction for hit rate ($F[2,28]=0.43$, $p=0.66$), but there were significant interactions for both accuracy and redundancy (both $F[2,26]=5.65$, $p<0.01$).

Separate one-way within-subjects ANOVAs revealed significant effects of task level on all three measures in the Heuristic condition (hit rate: $F[2,14]=43.69$, $p<0.001$; accuracy: $F[2,12]=9.25$, $p<0.01$; redundancy: $F[2,12]=9.25$, $p<0.01$), but for only hit rate in the Principle condition (hit rate: $F[2,14]=51.03$, $p<0.001$; accuracy: $F[2,14]=2.63$, $p=0.11$; redundancy: $F[2,14]=2.63$, $p=0.11$). Further pair-wise (Newman-Keuls[6]) tests revealed that in both conditions hit rate was significantly higher at the Rule level than at either of the other two levels (Heuristic: Skill/Rule $W_2=19.35$, Rule/Knowledge $W_2=19.08$, Skill/Knowledge $W_2=10.51$; Principle: Skill/Rule $W_2=14.18$, Rule/Knowledge $W_2=14.32$, Skill/Knowledge $W_2=4.86$). However, post-hoc Tukey HSD[7] tests showed that for Heuristic subjects alone accuracy and redundancy were, respectively, lower and higher at the Knowledge level than at either of the other two levels (both accuracy and redundancy, $W_3=28.12$).

Thus whether using heuristics or principles, these novice subjects were able to predict a much higher proportion of observed problems on the Rule tasks than they did for the Skill or Knowledge tasks. Only the heuristics allowed the emergence of relative improvements in accuracy and redundancy on the Skill and Rule levels over the Knowledge level, there being an apparent trade-off between these and the between-condition differences found in Section 6.3.

Taken together, the results from this and the previous Section imply that the same Skill-Rule-Knowledge manipulation which resulted in a Knowledge-level effect for principles over heuristics was, at the same time, having a different Rule-level effect. The Knowledge level difference enabled Principle subjects to be more accurate and less redundant in their predictions of Test problems than were Heuristic subjects, while the Rule level difference

6 Using the Games and Howell procedure for heterogeneous variances (Howell 1997).
7 Since variances for accuracy and redundancy are more similar in this condition (Table 5-6).

enabled subjects in both conditions to predict correctly a higher proportion of those problems than they did at the other two levels. However, Heuristic subjects were less accurate and more redundant on the Knowledge task than they were at other levels. The implications of these findings will be explored in the Discussion.

6.4 Problem Distributions

The previous between-level results are reflected in the distributions of predicted and observed problems at the three levels (See Figure 5.2). Table 5-7 shows that correlations[8] between predicted and observed frequencies were better at the Skill (r=0.51) and Rule (r=0.73) levels than at the Knowledge level (r=0.20). Thus, in general, a frequently occurring Test condition problem was more likely to be predicted by Heuristic and Principle subjects if it was a Skill or Rule problem than if it was a Knowledge problem.

Skill	Rule	Knowledge
0.51	0.73	0.20

Table 5-7. Experiment 3. Correlations between predicted and observed frequencies, all problems.

In contrast to Experiments 1 and 2, where large proportions of problems were predicted once only (and small proportions more than 3 or 4 times), in this experiment the overall distributions of both predicted and observed problems (correlation r=0.67) were less skewed (See Figure 5.2 (a)). However, this was not true of the Knowledge level (Figure 5.2 (d)), where both predicted and observed distributions were less evenly weighted (in this case, in favour of half of the six predicted problems and one of the 14 observed problems).

8 Pearson's r: ranked predicted vs. mean observed problems (both sets of frequencies sorted on predicted scores).

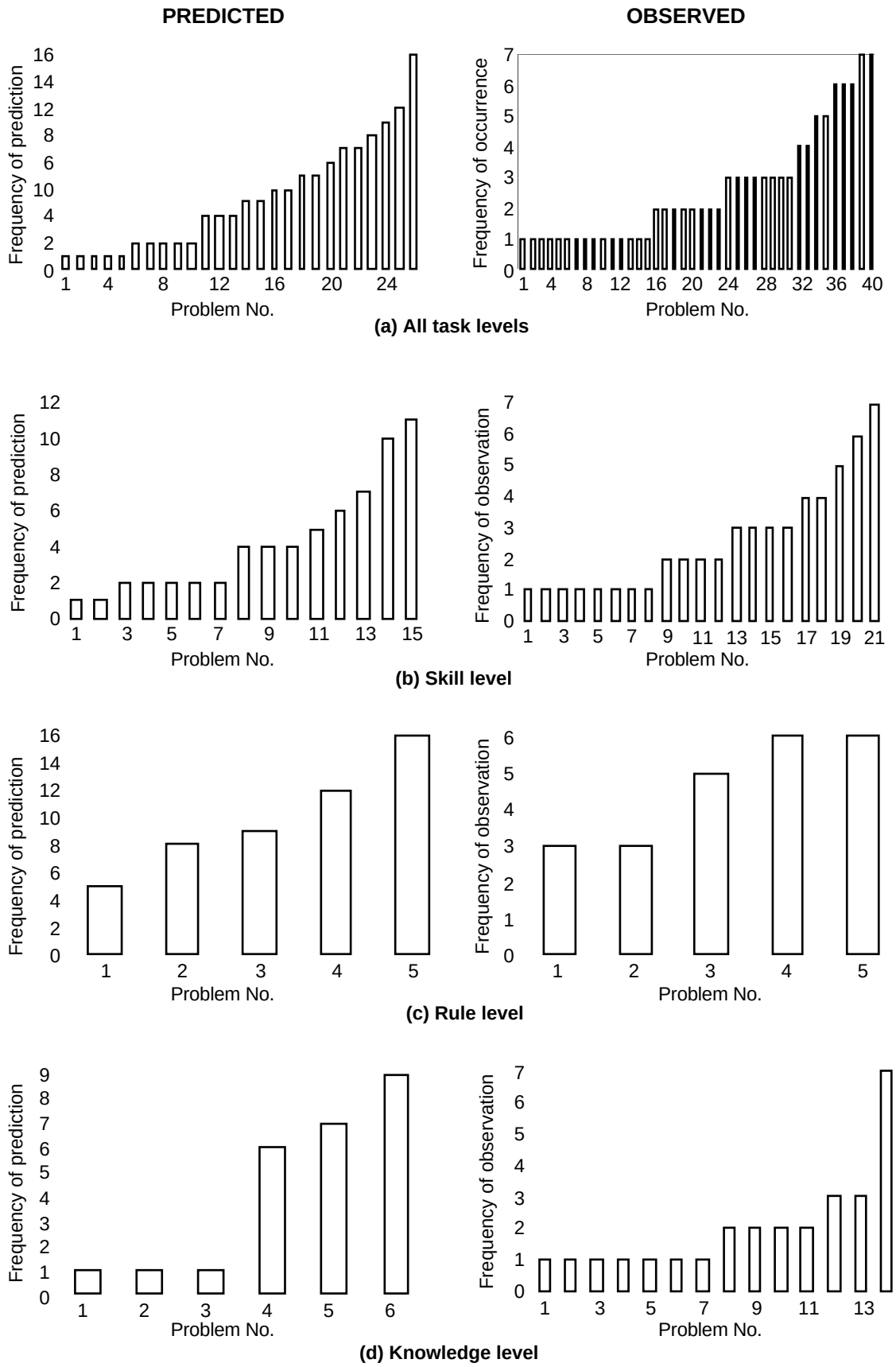

Figure 5.2. Experiment 3. Frequency distributions of predicted (Heuristic and Principle) and observed (Test) problems. Predicted problems exclude False Positives (FPs).

As far as predicted problems are concerned, this pattern is reflected in the overlap (problem sharing) between Heuristic and Principle subjects. Figure 5.3 shows that at the Knowledge level only 20% of predicted problems were shared by subjects in both conditions, compared with 31.6% and 41.2% at the Skill and Rule levels respectively. It also shows that at the Knowledge level a high proportion (45%) of problems were predicted under the Heuristic condition alone.

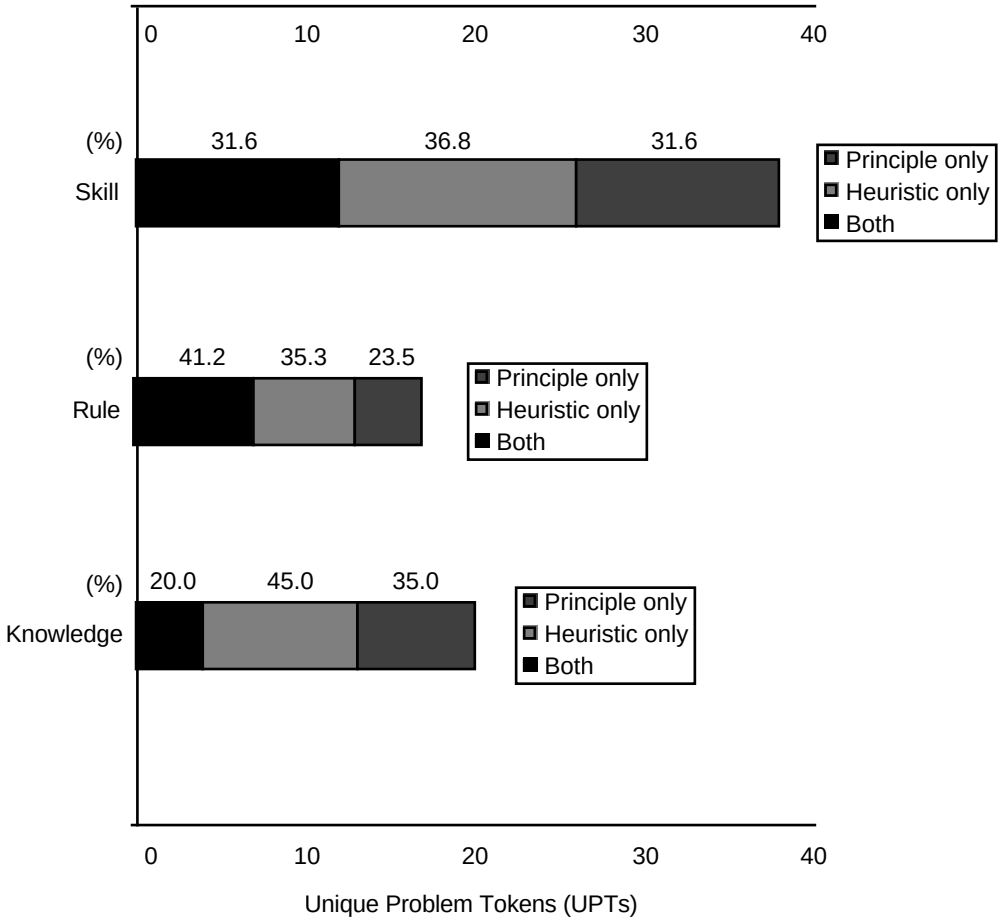

Figure 5.3. Experiment 3. Predicted problem distributions, all problems. Problems include False Positives (FPs).

Taken together, these results imply that the Knowledge task was different from those at the other two levels in exhibiting distribution patterns more in line with Experiments 1 and 2. Higher proportions of both predicted and observed problems featured only once, and a lower proportion of predicted problems were shared, than at the Skill and Rule levels. The likely reasons for this will also be explored in the Discussion.

6.5 Cumulative Curves, Detection Rates and Hit Rates

Figure 5.4 shows the plots of (absolute) cumulative counts for both predicted and matched problems (hits), at each task level and under both conditions. Each graph includes the target (T) number of observed (Test) UPTs at that level. From the Hits curves we can see whether or not the target was reached (i.e. how many of the observed problems at that level were correctly predicted by the cumulative actions of Heuristic and Principle subjects). In contrast, the Predicted curve shows if, and by how much, the predicted problems over-shot the target.

It is clear that only at the Rule level (Figure 5.4(b)) was the target (of 5 problems) reached, and that at the Skill level (Figure 5.4(a)) subjects' cumulative predictions over-shot that target (of 21) while not actually hitting it. At the Knowledge level (Figure 5.4(c)), neither hits nor predictions managed to reach the target (of 14). In every case there was a wide discrepancy between predictions and hits, illustrating the tendency of these subjects to over-predict (the maximum size of each discrepancy being the cumulative FPs for that sample[9]). This separation of predictions from hits implies strongly that hit rate is a more reliable indicator of predictive performance than the detection rate used in Experiments 1 and 2. Before comparing the two, however, we should first introduce a new measure, **prediction rate**, which relates total predictions to target UPTs. Re-working eq (2) (Chapter 4, Section 3) and eq (9) (this Chapter, Section 6.1) gives

Detection rate = mean predictions per subject / total no. of predicted UPTs (12)

Hit rate = mean hits (matches) per subject / total no. of observed UPTs (13)

Prediction rate = mean predictions per subject / total no. of observed UPTs (14)

Table 5-8 shows all three measures for Heuristic and Principle subjects at each task level (hit rates are repeated from Table 5-6).

	Skill		Rule		Knowledge	
	Heuristic	Principle	Heuristic	Principle	Heuristic	Principle
Detection rate (%)	23.75	29.54	30.92	34.91	21.38	27.91
Hit rate (%)	18.54	19.84	63.54	62.50	8.04	14.29
Prediction rate (%)	27.71	33.39	86.46	78.13	18.75	22.32

Table 5-8. Experiment 3. Mean detection rate, hit rate and prediction rate per subject, by condition within level, all problems.

9 There was no under-prediction, and most subjects over-predicted: 15 out of 24 (62.5%) Heuristic predictions and 18 out of 24 (75%) Principle predictions showed non-zero FPs.

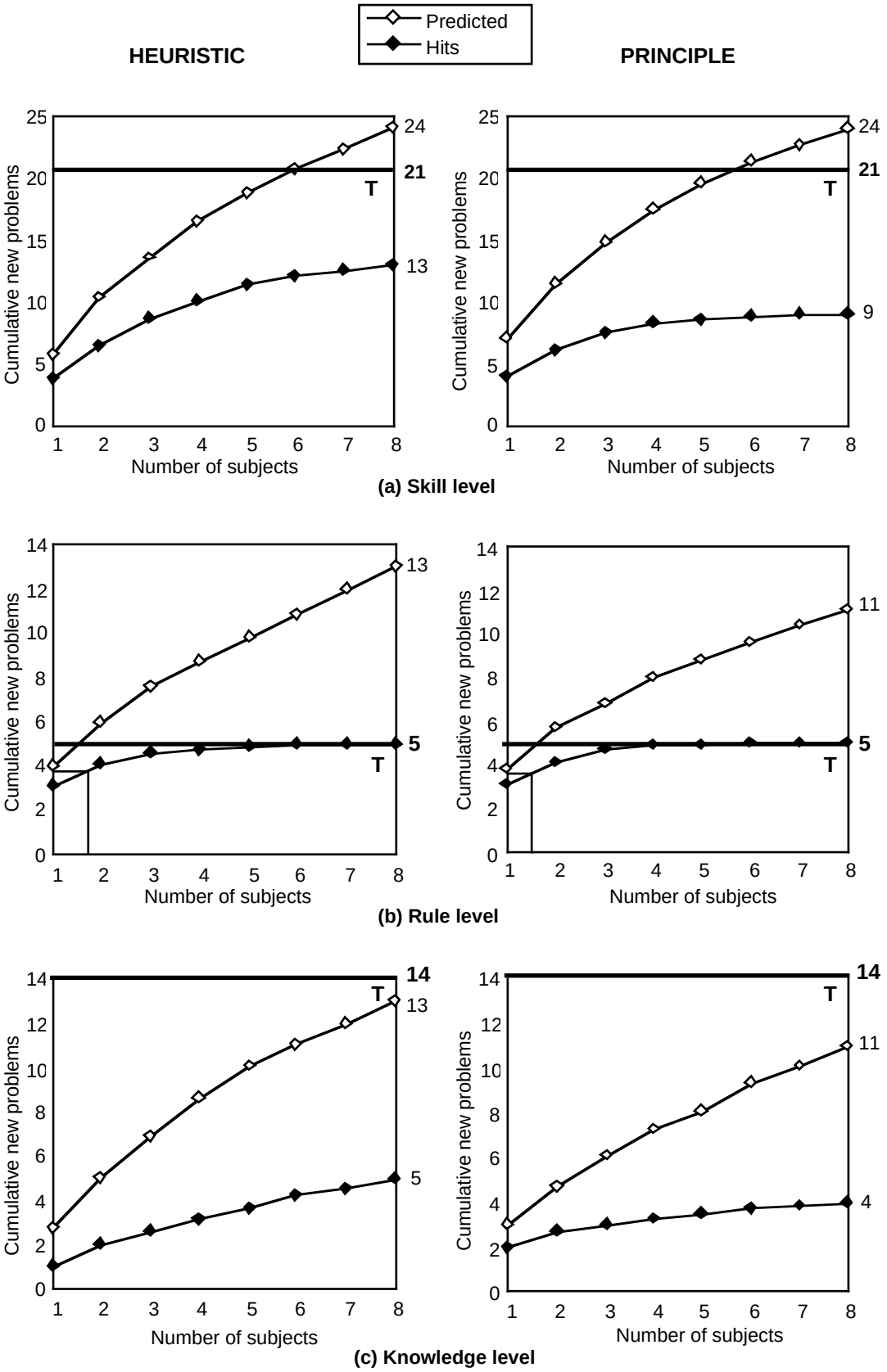

Figure 5.4. Experiment 3. Cumulative problem counts (absolute, not proportional, values), all problems. T is the target (total Test UPTs) at each task level.

We saw in Chapter 4 that detection rates of at least 25% (33% in the Nielsen prediction) are required to support the '3 to 5' claim (75% of total cumulative problems within 3 to 5

evaluators). On this basis, the performance of Principle (and Rule Heuristic) subjects appears to be above the '3 to 5' threshold. However, hit rates at all but the Rule level are lower than 25%, and only at the Knowledge level was there a significant between-condition difference on this measure. As for prediction rates, the fact that they are uniformly higher than hit rates is a result of the over-prediction mentioned above. (The differences between prediction rate and detection rate will be discussed below.)

Figure 5.4(b) showed that only at the Rule level did subjects manage to reach their target of observed problems. In both conditions convergence occurred so early that 75% of target problems were reached within 1 or 2 subjects. (At this level even detection rates are high enough for a '3 to 5'). At the Skill and Knowledge levels, however, we will need to extrapolate the Hits curves in order to see when (if ever) the target would be reached, and how many subjects this would require. Doing this for hit rate and prediction rate will enable us to compare the numbers of subjects required to needed 75% of target problems. We will also compare the resulting curves with their ideal (probability) form, in order to see by how much they diverge from that expected by the theoretical model described in Chapter 4.

Figure 5.5 shows cumulative prediction rates and hit rates at each level and under both conditions. If necessary, curves are extrapolated (using logarithmic regression) until one or both of them reaches 1. Alongside each is the corresponding probability curve, plotted using eq (1) (Chapter 4), that is $P = 1 - (1 - p)^n$ where n is the number of subjects and p is the hit rate for that sample (equal to the height of the y-intercept).

We can see that in each case prediction rate and hit rate continue to diverge. Like the permuted curves from Experiments 1 and 2 (Chapter 4, Figures 4.4), all prediction rate curves (except that for Knowledge Principle) run outside their ideal (probability) equivalents. In contrast, hit rate curves run inside their probability equivalents, but with considerable variation in their points of convergence. (Not all hit rates converge within reasonable limits to afford display of the other curves: where possible, they have been extrapolated to reach 0.75 of the target.)

Again with the exception of the Rule level, the divergence between actual and ideal curves testifies to the skewed nature of the predicted problem distributions (see Figure 5.2). Thus while both Rule-level curves show agreement between actual and ideal, at the Skill and Knowledge levels the actual hit rate curves show such little convergence that to reach 75% of observed problems would require large or very large subject numbers[10]. While the smallest of these, 13 for Skill Heuristic, replicates Nielsen's (1992) estimate for novices (Chapter 4, Section 2.1), the others are outside practical limits for measurement.

[10] R^2 estimates for the extrapolated curves (logarithmic trends) show reliability ranging from 0.96 to 0.99.

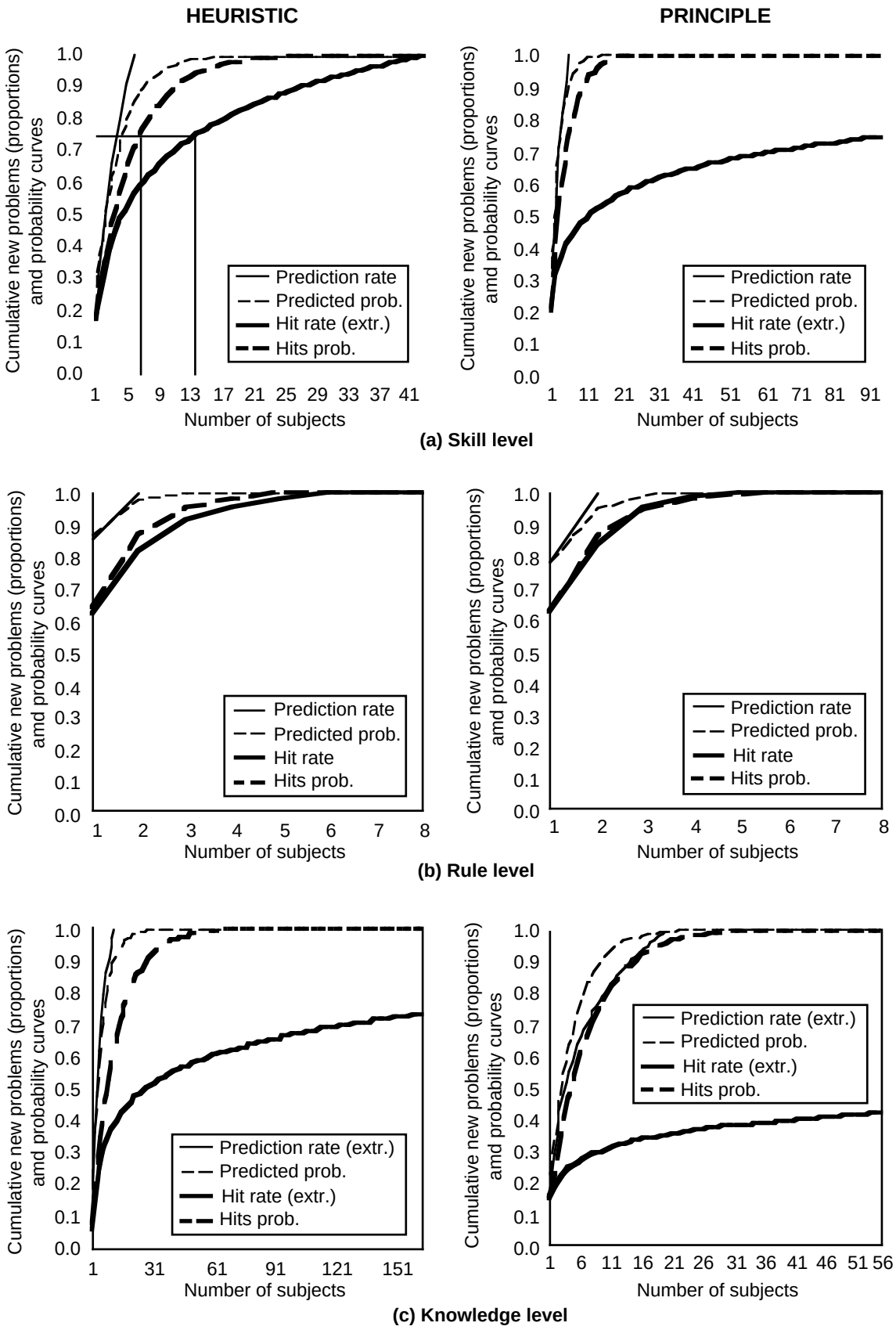

Figure 5.5. Experiment 3. Prediction rate and hit rate: cumulative and probability (prob.) curves, all problems. Some curves are extrapolated (extr.) to show (if feasible) the number of subjects required to reach 75% of observed problems.

Thus with the exception of the Rule level, these novices' predictive performance was, like that of their fellows in Experiments 1 and 2, widely divergent from that required by the model on which the '3 to 5' is based. Only at the Skill level did Heuristic subjects manage to replicate Nielsen's estimate for novices. At the Knowledge level, neither Heuristic nor Principle subjects would have reached the target (nor did Principle subjects at the Skill level). This is so even given the severely restricted nature of both tasks and system functionality compared with the previous experiments. In particular, Principle subjects' combined performance at the Knowledge level was worse than that of Heuristic subjects, even given the former's significantly better accuracy and redundancy. The reasons for these findings will be explored in the Discussion.

6.6 Problem Listings

Table 5-9 shows the complete set of observed problems from this experiment, ranked by frequency of occurrence within task level. Against each problem is shown its frequency of prediction and the severity level (assigned by the experimenter).

At the Skill level, frequently observed problems (representing all four Skill tasks) were nos. 11, 9, 2, 3 and 19. Problems which appear[11] to have been frequently observed but not frequently predicted were nos. 2, 19, 12 and 13. Observed problems which appear to have been frequently predicted but not frequently observed were nos. 1, 6 and 16. Apart from keyboard slips, the under-predicted errors were to do with over-selection (of text or cells), while the over-predictions mainly concerned task 1 (these problems were not predicted to such an extent on later Skill tasks).

At the Rule level, the most frequently observed problems (representing both tasks) were nos. 22, 25 and 24. Problems which appear to have been frequently observed but not frequently predicted were nos. 24 and 26. These were both from task 6.

At the Knowledge level, the single most frequently observed problem was no. 30 (failure to accept filter cell input before attempting to open the Advanced Filter dialogue box). Problems which appear to have been frequently observed but not frequently predicted were nos. 29, 39, 33, 34 and 36. Observed problems which appear to have been frequently predicted but not observed were nos. 37 and 31. The reason why no. 37 (failure to select the 'Copy to Another Location' option) was observed only twice is probably that task 7 required only three filter attempts (this error being more likely to occur after repeated attempts). The high ratio of predicted to observed on no. 31 (incorrect selection of List range) again reflects the tendency to over-predict among these novices.

The next two Sections will summarise some of the above results for the most frequent and most serious of the observed problems.

---

11 Comparing frequency ratios.

Obs. No.	Task No.	Problem description	Freq. Occ.	Freq. Pred.	Sev-erity
SKILL					
11	**3**	**Selected text before or after required word**	**7**	**10**	**2**
9	**2**	**Difficulty finding the move (arrow) cursor**	**6**	**11**	**4**
2	**1**	**Selected (and moved) text before or after required sentence (as well as sentence itself)**	**5**	**2**	**3**
3	1	Did not select spaces before (or after) sentence	4	5	2
19	5	Switched Caps Lock on (while selecting characters to correct)	4	0	1
5	1	Cancelled (correct) sentence selection	3	2	1
12	3	Selected space before (or after) required word	3	1	2
13	3	Selected empty paragraph (marker) before (or after) required words	3	1	2
18	5	Selected too many or too few characters in word (adjacent to, or in, the characters to correct)	3	6	2
4	1	Under-selected sentence (first click in text)	2	2	2
8	2	Cancelled cell selection (clicked outside cells)	2	2	1
10	2	Dropped cell(s) in wrong place	2	4	3
17	5	Selected text before or after word to be corrected	2	0	2
1	1	Difficulty distinguishing required sentences from rest of text	1	4	1
6	1	Dropped (correct) sentence in wrong place (end/middle of sentence)	1	7	3
7	2	Over- or under-selected cells to move	1	2	2
14	3	Under-selected word	1	0	1
15	3	Cancelled word selection	1	0	1
16	3	Held down B key	1	4	2
20	5	Cancelled (characters within word) selection	1	0	1
21	5	Pressed UP/DOWN arrow keys	1	0	1
RULE					
22	**4**	**Tried to use Command+B (on single or multiple cells, select or insertion mode)**	**6**	**12**	**4**
25	**6**	**Tried to use (shift+) arrow keys**	**6**	**16**	**4**
24	**6**	**Difficulty getting into insertion mode (in cell)**	**5**	**8**	**3**
23	4	Did not find correct shortcut (Control+B or Command+Shift+B) without being prompted	3	9	4
26	6	Selected too many or too few characters in cell (adjacent to, or in, the characters to correct)	3	5	2
KNOWLEDGE					
30	***7***	***Failed to accept Filter input (click out of cell; press <return> or <enter>; use tick box in formula bar)***	***7***	***6***	***6***
29	7	Entered [and un-necessarily corrected] lower case filter criteria ('a', 'b')	3	1	3
39	7	Tried to move both Group and Score cells to middle table (after successful attempt)	3	0	3
33	7	[Un-necessarily] deleted previous result (target range) before subsequent attempt	2	0	3
34	*7*	*Put second filter entry ('B') below 'A', not in place of 'A'*	*2*	*0*	*6*
36	*7*	*Appended re-selected (incorrect) range (source or target) to dialogue box field contents*	*2*	*0*	*6*
37	*7*	*Did not select 'Copy to Another Location' option [prior selection of this radio button is not preserved on subsequent attempts]*	*2*	*7*	*7*
27	7	Could not find (Data --> Filter --> Advanced Filter) menu	1	1	2
28	7	Selected menu item below Filter ('Form...')	1	0	1
31	*7*	*Incorrectly selected List range (omitted headings/column/rows)*	*1*	*9*	*5*
32	*7*	*Selected (or specified) one dialogue box range for another (Copy range for Criteria)*	*1*	*1*	*5*
35	7	Clicked dialogue box OK before finished making selections (on attempt subsequent to the first)	1	0	4
38	7	[Un-necessarily] re-selected source ranges (List, Criteria, Copy to) on attempts subsequent to the first	1	0	2
40	7	Under- or over- selected successfully filtered scores (for move to middle table)	1	0	1

Table 5-9. Experiment 3. Observed problem listings at each task level, sorted by frequency of occurrence within level. Obs. No = observed problem no. Freq. Occ. = frequency of occurrence. Freq. Pred. = frequency of prediction. Problems in **bold** were the most frequently observed (ranked 5 to 7). Problems in *italics* were assigned high severity ratings (5 to 7).

6.7 High-Frequency Problems

Since the highest problem incidence at any level was 7, we can define a high-frequency problem as being any problem in Table 5-9 whose frequency of occurrence is between 5 and 7. This yields just seven problems, three from each of the Skill and Rule levels (respectively nos. 11, 9, 2 and 22, 25, 24) plus one from the Knowledge level (no. 30). Table 5-10 shows the mean hit rate, accuracy and redundancy figures relating to these observed problems alone.

	Skill {3}		Rule {3}		Knowledge {1}	
	Heuristic	Principle	Heuristic	Principle	Heuristic	Principle
Ht rate (%)	41.67 [23.57]	54.17 [30.54]	78.13 [28.15]	75.00 [18.90]	12.50 [35.36]	62.50 [51.75]
Accuracy (%)	21.91 [12.19]	22.79 [12.38]	67.50 [25.23]	59.38 [11.68]	2.86 [7.56]	22.92 [21.25]
Redundancy (%)	78.09 [12.19]	77.21 [12.38]	32.50 [25.23]	40.63 [11.68]	97.14 [7.56]	77.08 [21.25]

Table 5-10. Experiment 3. Mean hit rate, accuracy and redundancy per subject, by condition within level, high-frequency problems only. Figures in {curly brackets} are the numbers of observed UPTs (in the Test condition) at each level. Figures in [square brackets] are standard deviations.

Two-way split-plot ANOVAs showed significant effects of task level for all three measures (hit rate: $F[2,28]=6.19$, $p<0.01$; both accuracy and redundancy: $F[2,26]=37.99$, $p<0.001$). There were no overall differences between conditions (hit rate: $F[1,14]=4.02$, $p=0.07$; both accuracy and redundancy: $F[1,13]=2.03$, $p=0.18$). There were no task level x condition interactions (hit rate: $F[2,28]=2.82$, $p=0.08$; both accuracy and redundancy: $F[2,26]=2.15$, $p=0.14$).

However, further pair-wise (Newman-Keuls[12]) tests revealed significant between-condition differences on all three measures at the Knowledge level (hit rate: $W_2=48.26$, both accuracy and redundancy: $W_2=18.53$) but not the Skill level (hit rate: $W_2=29.51$, both accuracy and redundancy: $W_2=13.20$) or the Rule level (hit rate: $W_2=26.11$, both accuracy and redundancy: $W_2=22.24$).

One-way within-subjects ANOVAs revealed significant between-level differences on all three measures in the Heuristic condition (hit rate: $F[2,14]=16.21$, $p<0.001$; both accuracy and redundancy: $F[2,12]=33.42$, $p<0.001$), and on accuracy and redundancy (both $F[2,14]=11.42$, $p<0.01$) but not hit rate ($F[2,14]=0.56$, $p=0.59$) in the Principle condition.

Thus whether using heuristics or principles, subjects were more accurate and less redundant in their predictions of Rule-level high-frequency problems than of those at the other levels (but achieved higher hit rates using only the heuristics). A higher proportion of Principle than Heuristic subjects were able to predict the single high-frequency Knowledge level problem (but not those at the other two levels). As expected, hit rates for high-

12 Using the Games and Howell procedure for heterogeneous variances (Howell 1997).

frequency problems were now sufficiently high to satisfy a '3 to 5' prediction (compare with table 5-5). However, this was at the expense of lower accuracy and higher redundancy.

6.8 High-Severity Problems

In contrast to Experiments 1 and 2, in this experiment subjects were not asked to assign severity levels to predicted (or observed) problems. Severity attributions were made by the experimenter (the author), taking into account the impact on test users of each problem (error). The nature of the tasks at the Skill and Rule levels meant that high severity (5 to 7) could reasonably be assigned to only the six Knowledge level tasks shown in table 5-9. Table 5-11 shows the mean hit rate, accuracy and redundancy in relation to these six problems alone.

	Knowledge {6}	
	Heuristic	Principle
Ht rate (%)	20.83 [14.77]	31.25 [10.68]
Accuracy (%)	47.86 [14.42	69.79 [28.85]
Redundancy (%)	52.14 [14.42]	30.21 [28.85]

Table 5-11. Experiment 3. Mean hit rate, accuracy and redundancy per subject for high-severity Knowledge level problems. The figure in {curly brackets} is the numbers of observed UPTs (in the Test condition) at this level. Figures in [square brackets] are standard deviations.

Separate unrelated t-tests show significant differences between conditions for accuracy and redundancy but not hit rate (hit rate: $p=0.07$; accuracy: $p<0.05$ one-tailed, $p=0.08$ two-tailed; redundancy: $p<0.05$ one-tailed, $p=0.08$ two-tailed). Thus on these problems subjects were more accurate and less redundant, but did not predict more correct problems, using the principles than the heuristics.

In both conditions hit rates for these six problems show improvements over those for all-severity problems (compare with table 5-5). However, the (extrapolated) equivalent cumulative curves (Figure 5.6) show that in neither case was there sufficient agreement among subjects for a '3 to 5'. While that for (a) Heuristic subjects (75% of observed problems found by 12 subjects) was in line with Nielsen's (1992) prediction for novices, curve (b) shows[13] that using the principles it would still have required around 90 subjects to find the same proportion of high-severity problems.

[13] R^2 estimate for the Principle curve was 0.85

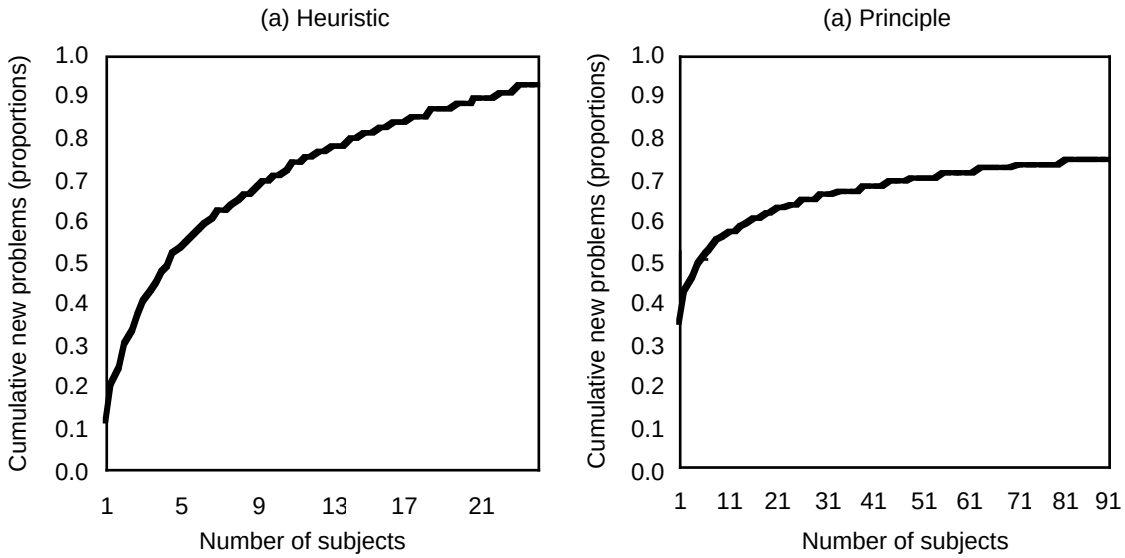

Figure 5.6. Experiment 3. Knowledge level. Hit rates under (a) Heuristic and (b) Principle conditions for high-severity observed problems. Curves are extrapolated to show the number of subjects required to reach 75% of observed problems.

6.9 Inter-rater Reliability

We saw in Chapter 4 that the most difficult part of the problem reduction process is the 'between subjects' stage, that is, where problem accounts from different subjects are combined into a single set of unique problem tokens (UPTs). Hence the inter-rater reliability test reported in Section 6.1 of that chapter concentrated on the 'between subjects' stage of Experiment 1. In Experiment 3, however, each task could be treated independently for the purposes of problem reduction. In addition, the simpler nature of the tasks meant that there were far fewer problems to reduce. Thus the inter-rater reliability assessment performed for this experiment focused on the matching of predicted to observed problems rather than the reductions performed on each set. For comparison with Experiments 1 and 2, the percentage reductions on predicted problems at each task level are shown in Table 5-12.

	Problem Totals (all subjects)		
	Original	Reduced (UPTs)	% Reduction
Experiment 1	444	230	48.20
Experiment 2	222	114	48.65
Experiment 3			
Skill	99	35	64.75
Rule	64	17	73.44
Knowledge	46	20	56.52

Table 5-12. Experiments 1 and 2 problem reductions and Experiment 3 predicted problem reductions.

The test was to assess the manner in which predicted (Heuristic and Principle) problems from each of the seven tasks had been matched against the problems observed on the same task in the Test condition. Unlike the test for Experiment 1 (which used random samples from large problem sets), the complete sets from each task might be used. However, the Experiment 1 test showed that raters could reasonably be expected to

complete only a few of such sets, so only three of the seven tasks, one from each level, were selected. These were tasks 3 (Skill), 4 (Rule) and 7 (Knowledge).

Randomly sorted sets of predicted problem descriptions from each of the three tasks were separately presented on screen to two independent raters (both experienced HCI researchers). The raters were asked to judge if one or more of each problem set were the same as (and *not* "different examples of the same type of thing as") the observed (Test) problems for that task. Those problems judged to match - the rater's hits - were clustered on screen under their respective Test problem description. The remainder - the rater's false positives - were left behind. As many matching attempts as necessary could be made. Raters had available the running software for the three tasks, installed on a separate computer, plus paper copies of the problem sets.

The resulting matchings were compared with those originally generated by the experimenter (the author) for each task. This was done by counting the number of pairs of agreements (and disagreements) between the rater's and the experimenter's matches (and non-matches). Separate tests were performed on each match - no match pairing. These show good overall correlation (inter-rater reliability) between each of the two raters and the experimenter, the mean value for Cohen's κ being 0.69 (Table 5-13).

Task	Rater 1	Rater 2	Mean
3	0.20	0.55	0.37
4	0.92	1.00	0.96
7	0.80	0.67	0.73
Mean	0.64	0.74	**0.69**

Table 5-13. Experiment 3. Inter-rater reliability data (Cohen's κ) for tasks 3 (Skill), 4 (Rule) and 7 (Knowledge). Each cell compares predicted-observed matches for the experimenter and one rater.

This final analysis has shown that the original predicted-observed matchings performed by the experimenter were comparable to those made by two independent raters with less exposure to the tasks. The ratio of matches (hits) to non-matches (false positives) which these figures represent were shown to correlate well on two tasks out of three (very well, on task 4). Thus it is reasonable to conclude that the author's assessment of hits, FPs and misses was not much different from that which might be achieved by other experimenters using the same tasks.

7. Summary of Results

1. The differentiation between Skill, Rule and Knowledge task levels enabled significantly different results to be obtained on the three ratio measures (hit rate, accuracy and redundancy), within and between levels. Within-level differences favoured the Principle condition at the Knowledge level, while between-level differences favoured the Rule level.

2. The within-level differences showed that at the Knowledge level but not the Skill or Rule levels, subjects exposed to the two principles used in this experiment exhibited greater accuracy (more hits) and less redundancy (fewer false positives) than did Heuristic subjects at that level. However, the same subjects did not show higher hit rates (by correctly predicting more observed problems) than those using the three heuristics, at this or any other level.

3. The between-level differences showed that subjects in both conditions achieved much higher hit rates at the Rule level than at either of the other two levels. Heuristic subjects' accuracy and redundancy were worse at the Knowledge level than at the other levels. There appeared to be a trade-off between these differences and those in item 2.

4. The Knowledge level task appeared to be different from those at the other two levels in exhibiting problem distribution patterns more in line with those in Experiments 1 and 2. Higher proportions of both predicted and observed problems featured only once, and a lower proportion of predicted problems were shared between conditions, than at the other levels.

5. Other than at the Rule level, the cumulative performance of both Heuristic and Principle subjects was little better than that of their fellows in Experiments 1 and 2. At both the Skill and Knowledge levels, subjects in both conditions failed to reach the target for observed problems. At both levels Principle subjects performed worse than Heuristic subjects, only Skill Heuristic subjects managing to reach 75% of the target within the limits claimed for novices.

6. Subjects at all levels and both conditions showed a similar tendency to over-predict (thus over-shooting the Skill and Rule level targets). Again with the exception of the Rule level, cumulative curves of prediction rate exhibited a pattern similar to that for detection rate in Experiments 1 and 2.

7. Consideration of the most frequently observed problems showed expected improvements in hit rate at all task levels. On the single Knowledge-level problem that this concerned, Principle subjects' performance was superior to that of Heuristic subjects on all three ratio measures, including hit rate. While subjects in both conditions were more accurate and less redundant at the Rule level, only Heuristic subjects showed better hit rates at this level.

8. Only six observed problems at the Knowledge level were deemed to be of high severity. Principle subjects were more accurate and less redundant than Heuristic subjects in their predictions of these problems; however, there were no significant differences in hit rate. Extrapolating the hit rate curves showed that only Heuristic subjects would have reached 75% of these six high-severity problems within the limits claimed for novices.

9. The results of an inter-rater reliability assessment showed good correlation between the experimenter (the author) and each of two independent raters asked to match predicted and observed problems from one task at each level.

8. Discussion

The underlying rationale for this thesis is that with suitably judged principle-based material, it should be possible for even novice evaluators to identify more usability problems than they would using heuristics alone. Both Experiments 1 and 2 failed to demonstrate this for novices. However, this experiment has shown some limited success in enabling further novices to be more accurate in their predictions of actual (observed) problems using two of the full principles set than were similar subjects given the corresponding three of the ten Molich & Nielsen (1990) heuristics. Consideration of high-frequency problems enabled this principles-heuristic effect to include hit rate (correct predictions to observed problems) as well as accuracy (hits to total predictions) and redundancy (false positives to predictions). This effect was achieved by restricting the scope of both the evaluation materials and the tasks performed. In contrast to the previous experiments, where both systems and tasks were relatively open-ended, in this experiment the type, length and aims of each task were closely constrained. The manipulation of task levels on the Skill-Rule-Knowledge (S-R-K) model enabled different results to be achieved at each level. The results achieved represent a combination of two main factors, namely evaluation materials and task levels.

8.1 The Evaluation Materials

We saw in Experiment 2 that even with prior training those novices were unable to make better use of the full principles set than the ten heuristics. In this experiment it was hoped that by restricting the evaluation materials to just those considered relevant for a given task, success might be achieved for that task alone. The original intention was thus to make use of a single task or task type, in this case involving error recovery. However, the availability of Reason's (1987b, 1990) S-R-K model meant that it proved possible to manipulate task levels. It was thought that the 'added value' offered by the principles over the heuristics might operate most clearly at the more complex, Knowledge, level, with some benefit at the Rule level but little at the Skill level.

In the event, the results only partly support the above hypothesis. The significant differences in accuracy and redundancy between Heuristic and Principle subjects operated entirely at the Knowledge level. This remained so (but now including hit rate) when predictions were considered against the small number of most frequently occurring observed problems, though involving a single problem at the Knowledge level. Similar results (but without a significant difference in hit rate) were also obtained for the six Knowledge-level problems deemed to be of high severity.

The fact that no such differences were found at the other two levels implies something about the differences between the tasks at each level. Though there were consistent differences between the Rule level and the other levels, this did not manifest in any improvements for principles over heuristics at the Rule level. As we shall see in the next Section, this is likely to be due to the simplicity of the Rule-level tasks. The 'added value' of the principles materials is thought to be represented by attribute 8 (Error Management), which focuses on the ways in which errors might be foreseen and thus prevented (or provision made for retraction). The difference between the Knowledge level task and those at the other two levels - the former often insoluble without assistance, the latter easily reversible - is borne out in the lack of principles-heuristics differences at the first two levels. (See Table 5-9 for problem listings and Appendix D for the principles set.)

As far as the Knowledge level is concerned, this result represents an initial improvement over the previous experiments, and some vindication of the author's belief in principles rather than heuristics as evaluation material. It remains to be seen, however, how far this contraction - from full materials down to subsets and from full systems to single tasks - might work in the other direction, and how far different subsets might be applied to different (or even the same) tasks. As mentioned in Chapter 4, it is the skill of the usability engineer to be able to select sufficiently representative tasks that, when combined, make up the whole (or as much of a whole as desired) of the system under test. However, it is the author's belief that whole systems do not flow automatically from collections of tasks.

The lack of a between-condition effect at the Skill and Rule levels implies that material such as the principles set is best applied to complete or more complex tasks. The corollary (supported by the improved accuracy and lower redundancy of Heuristic subjects at these levels) is that the application of short heuristics such as Molich's & Nielsen's may be limited to simple tasks or sub-systems. However, Experiments 1 and 2 showed that letting even experienced subjects loose on full systems (even simulations and half-modules) with full principles materials (even three-page versions) may be to go too far in the opposite direction. This and the issue of task scope will be taken up in Chapter 8.

One of the apparently conflicting results from this experiment was that the combined (cumulative) performance of Principle subjects at the Skill and Rule levels was worse than

that of Heuristic subjects. Only Skill Heuristic subjects managed to reach 75% of their target within the limit claimed for novices (Nielsen 1992), and the largest proportion of single-incidence problems (45%) favoured Knowledge Heuristic subjects (Figure 5.3). Thus it appears that the Principle subjects who were responsible for the Knowledge-level differences were acting mainly alone and in opposition to a greater number of sole Heuristic subjects at the same level.

8.2 The Task Levels

The manipulation of task types achieved a clear differentiation between the Rule level tasks and those at the other two levels. Not only were the numbers of hits considerably higher at this level than at the Skill or Knowledge levels, but the cumulative performance of subjects in both conditions were such that 75% of target problems were found by just one or two individuals. These novices agreed on their predictions of Rule-level problems to such a degree that after just four or five predictions there were no new problems to be found (Figure 5.5 (b)).

Hit rates of 63% and 64% (Table 5-5) would represent exceptionally good performance for even experts (comparable to that for Nielsen's (1992) 'double specialists'). Even the equivalent detection rates of 35% and 31% respectively (Table 5-8) are high enough for a '3 to 5'. However, the way in which these results were achieved was extremely contrived. They were produced by manipulating Skill-Rule task pairs in order to set up an expectation in the subject that a particular keyboard shortcut which worked in one Office application (Word) would work in the same way in another (Excel). It is questionable, therefore, if the Rule tasks represent true 'tasks' at all. (And even without the Skill-Rule manipulation, it is debatable if text or cell emboldening itself qualifies as a 'task'.) Thus the high performance of subjects at this level is easily explicable.

Relative performance at the other two levels was less clear-cut. Heuristic subjects' accuracy and redundancy were worse at the Knowledge level than at the Skill or Rule levels. For high-frequency problems, only Heuristic subjects had better hit rates at the Rule level. And (as mentioned above) Principle subjects' combined performance was worse than that of Heuristic subjects at all but the Rule level. Thus the same manipulation that favoured principles over heuristics at the Knowledge level appeared to work in the opposite direction on the other two levels.

It appears, then, that the relative size and complexity of even just these two principles was having an opposite effect to that desired, such that the heuristics were still able to produce an effect at the less complex task levels. In that sense, these novices were like their Experiment 2 predecessors in failing to make use of the additional material provided. However, it appears that the lack of an effect for the heuristics in the previous experiments may also have been a factor of task scope. This issue will be taken up in Chapter 8.

As to the tasks themselves, it is acknowledged that the provision of step-by-step instructions is unrepresentative of real user-system interaction. This is particularly so in task 7, where real success would come only after trial and error (even with the 'assistance' of on-line help). However, both tasks and instructions were held the same for all subjects, leaving differences in Word and Excel familiarity to account for individual variations in performance. Attempt was made to minimise this effect by the establishment of a set of assumptions concerning subject experience levels (see Appendix F for a transcript); these were also used as the 'base line' assumptions for the inter-rater reliability test in Section 6.9.

More problematical is the order in which tasks were performed. It is likely that subjects' performance on the later tasks was, in part, an effect of their experience of the earlier tasks. For example, task 7 was explicitly introduced as the 'automated' version of task 2, after which the acquisition of Excel drag and drop could be assumed. However, the task type manipulation included the control of the Skill-Rule pairs (tasks 3 and 4, 5 and 6) and the progression from simple to complex tasks; and once again, any effect of task order would have been the same for all subjects. It was decided, therefore, to maintain the same sequence throughout.

The reader may have noticed that for the purposes of the S-R-K manipulation all tasks at the first two levels were treated as if they were of the same type and complexity. The results of separate within-subject analyses for hit rate bear this out at the Rule level but not the Skill level (Rule level (tasks 4 and 6): Newman-Keuls, Heuristic W_2=32.22 ($\overline{X}_1$-$\overline{X}_2$=10.42), Principle W_2=24.10 ($\overline{X}_1$-$\overline{X}_2$=0.0); Skill level (tasks 1, 2, 3 and 5): one-way ANOVAs, Heuristic $F[3, 21]=3.65$, $p<0.05$, Principle $F[3, 21]=7.57$, $p<0.01$). Thus at the Skill level, there was some additional effect of task difference operating in both conditions. Further pair-wise (Newman-Keuls) tests within condition showed this to be chiefly due to fewer correct predictions on task 5 (Word spelling correction). However, pair-wise (Newman-Keuls) tests between conditions showed no significant differences between Heuristic and Principle subjects on any of the three ratio measures, for any of the Skill tasks. Thus whatever differences there were did not affect the lack of a principle-heuristic result at this level. (The differences in hit rate between the Rule and Skill levels will have been greater for task 5 than the other Skill tasks.)

It is also acknowledged that the treatment of task types masks an inconsistency in the approach to problem reduction, compared with the previous experiments. In this experiment each subjects' predictions were compared with the observed problems for that task alone, problem reduction being strictly within-task rather than between-task. Thus problem distributions and cumulative curves represent the summation, rather than combination, of the tasks at each level. By contrast, in Experiments 1 and 2 the 'between subjects' stage of the reduction process attempted to reduce problem totals to their minimum. Thus the percentage reductions listed in Table 5-12 are not strictly comparable.

However, it has just been established that only one task (task 5) was significantly different (in terms of predictions) from the other tasks at the same level.

8.3 Predicted versus Observed Problems

This experiment has shown that the predictive performance of novices on all but simple tasks was still insufficient to achieve Nielsen's (1992) projection for novices. Only on the very contrived Rule level tasks were they inside the '3 to 5' limit, and only at this level did they manage to reach the target of observed problems. On the more representative Knowledge task, even with the help of on-screen instructions they would have failed to reach 75% of target problems within practical limits.

Table 5-14 shows the detection rates and hit rates from all three experiments. It is clear that by comparison with the previous experiments, these novices performed well. On both detection rate and hit rate, performance at all levels outstripped even the experienced subjects in Experiment 1.

	Experiment 1			Experiment 2	Experiment 3 (Novice)		
	Novice	Experienced	All	Novice	Skill	Rule	Knowledge
Detection rate (%)	7.66	10.07	5.56	8.43	26.65	32.92	24.65
Hit rate (%)	-	-	-	-	19.19	63.02	11.17

Table 5-14. Detection rates and hit rates from Experiments 1, 2 and 3, all problems. Experiment 3 figures are means from the Heuristic and Principle conditions.

It is very unlikely that the novices in Experiment 3 were more able than their fellows from the previous experiments (questionnaire scores were not significantly higher than those from Experiment 2). The inter-rater reliability tests from Experiments 1 and 3, plus the similar sharing and distribution patterns from all three experiments, suggest strongly that methodological factors are not the cause of the above differences in performance. The improvement in performance must, then, be to do with the reduction in size and scope of the user tasks. This in turn suggests that the very low detection rates from the first two experiments are a factor of the choice of system and/or task type used.

But Experiments 1 and 2 were like most reported experiments in the UEM literature in relying on predicted rather than observed problem count as the denominator in measures of predictive ability. Had either of these experiments had a datum for observed problems, the 'real' hit rates would have been even lower than those in Table 5-14. (In Experiment 1, the original intention had been to use a deliberately error-laden version of the software used as a basis for comparison.) In Experiment 3, only the expedient of high-frequency filtering enabled hit rates to rise above the '3 to 5' threshold (and those for the high-severity Knowledge problem were at best just comparable with the novice projection). It appears, then, that published reports which use detection rate rather than hit rate are masking a 'true' figure for predictive power which is some 30% to 50% lower than it might be. Even without the speculation that problem totals mask a tendency to categorise at a higher level, this

makes the predictive ability of inspection methods (in particular) appear more unreliable than even Gray & Salzman (1998) would have us believe.

However, this begs one very obvious question, namely, how much better would experienced evaluators do given the same tasks. It is unlikely that a detection rate of 10% is representative of expert performance on reduced-scope tasks (or even on more 'closed' systems than hypermedia browsers and teaching software). The evidence from Experiment 1 suggests that experienced subjects report between two and three times more problems than novices, and that their level of agreement is such that where novices require 14 or 15 subjects to find 75% of a target, they might need fewer than 9 (Chapter 4, Figure 4.4). The above conclusion also assumes that both novices and experts tend to over-predict rather than under-predict compared to observed problems. The former is relatively easy to confirm (but, alas, not in this thesis), by repeating Experiment 3 with suitably experienced subjects. The latter may represent a more worrying feature of comparative UEM experiments, namely that subjects' behaviour under these sorts of experimental conditions is, in part, an artefact of the experimental process. In order to begin to assess this, we would need a much more reliable datum for 'true' usability problems (those observed in a much longer series of user tests or field studies) than that obtained from the seven Test subjects in this experiment. Fortunately, the author has such a set of data, which will be examined in the next Chapter.

Summary of Chapter 5

Experiment 3 showed that predictive performance comparable to that claimed for novices was achieved only on relatively simple tasks, and that given a more realistic task (and a real learning curve), a new set of novices failed to perform within practical bounds. Performance came within that claimed for novices only on high-severity problems on the more complex task. This was so even given the reduced size and scope of the tasks compared to the previous experiments, and allowing for more on-screen guidance than would be typical of such tasks. An inter-rater reliability test implied that the predicted-observed problem matching which was responsible for the author's assessment of hits, misses and false positives was not unrepresentative of other researchers. The finding that subjects at all task levels tended to over-predict lead to a discussion of the reliability of inspection method studies which use detection rate rather than hit rate as their measure of predictive power.

In the next Chapter the results of an earlier study by the author will be used to assess the reliability of observed problem data. This will lead to a re-assessment of the findings from that study in the light of the earlier Chapters.