Chapter 4: Cumulative Problem Curves

1. Introduction

We have seen that in Experiments 1 and 2 the distributions of unique problems tokens (UPTs) followed a similar pattern, namely relatively little overlap between subjects (sharing of a common core of problems) and a large proportion of problems reported only once. Such a pattern of reporting represents a relatively large set of disparate problems, out of which a smaller number of issues (which may be addressed in a programme of improvements) might be extracted. The approach proposed was to prioritise UPTs on the product of severity rating and frequency of occurrence.

If the degree of problem sharing or overlap in a given set of UPTs is about the same (as is suggested by the findings from Experiments 1 and 2), then it follows that the number of evaluators which will be required to identify a given proportion of the problems with an interface can be determined from the expected sharing profile. This depends on the likelihood of any one problem being found by any one evaluator following some consistent probability. Assessments of the degree to which any one profile follows a predictable pattern can best be made by plotting the combined (additive) problem counts for successive evaluators, taking account of shared problems - cumulative problem curves.

A model based on such assumptions has been used by Nielsen (1992, 1993, 1994d) to justify claims made about the number of experienced evaluators required to find most of the problems with an interface. Nielsen and Landauer (Nielsen & Landauer 1993, Nielsen 1994c) used the same model to predict benefit/cost ratios for typical software projects. (These claims are examined in more detail below.) In this chapter we shall derive cumulative problem curves for Experiments 1 and 2 and compare them with those predicted by this model. The differences found will raise some hypotheses concerning the processes of problem extraction and reduction.

2. Nielsen's 'Three to Five Evaluators' Claim

2.1 Description

The proposition that most of the usability problems in any given interface can be found with as few as 3 to 5 experienced evaluators become something of a given of the usability literature. Before assessing the claim, we should first look at its origins.

Figure 4.1 (reproduced from Nielsen 1993 p156 and 1994d p33) shows the mean cumulative proportion of usability problems per evaluator from the six case studies of heuristic evaluation discussed in Nielsen (1992). It shows that on average 75% of the total UPTs from each of those six studies were found by only 5 evaluators.

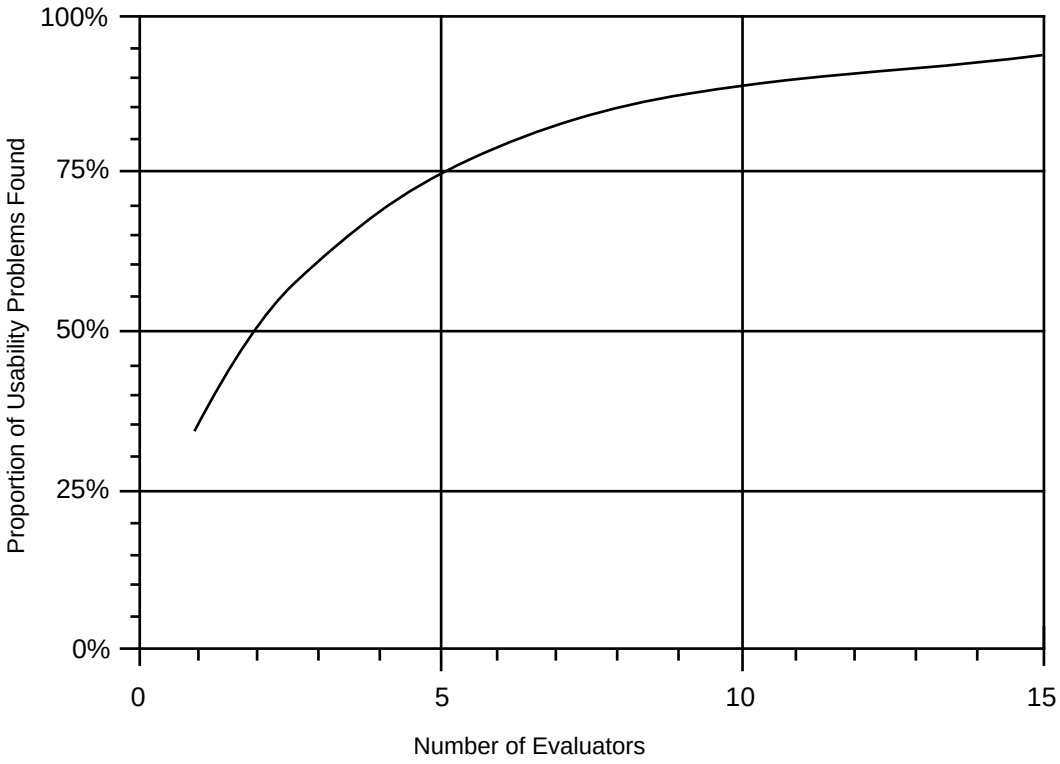

Figure 4.1. Curve showing the proportion of usability problems in an interface found by heuristic evaluation using various numbers of evaluators. The curve represents the average of the six case studies of heuristic evaluation discussed in Nielsen (1992).

With the y-axis at x = 1, the line intercepts at around 33% of total problems.

Source: Nielsen (1993) p156 and Nielsen (1994d) p33.

The place at which such curves as Figure 4.1 begin to converge (reach their possible maximum) represents a "point of diminishing returns" for heuristic evaluation (Nielsen & Molich 1990). Beyond this point the cost of bringing in additional evaluators will not be justified by the decreasing number of new problems found. According to Figure 4.1 and Nielsen & Molich (1990) (aggregating four of the same six studies), that point is reached at about 10 evaluators. In Figure 4.1 this many evaluators uncovered around 90% of total problems.

However, Figure 4.2 (reproduced from Nielsen 1992) shows that this number can vary with evaluator experience. In that paper, Nielsen claimed that convergence occurred rather earlier for "regular specialists", and much earlier for "double specialists" (those with both general and domain-specific experience). Of the six studies there discussed, 75% of total problems were found on average by between 1 and 2 "double specialists", while for "regular specialists" it was by 3 and for novices as many as 14 evaluators.

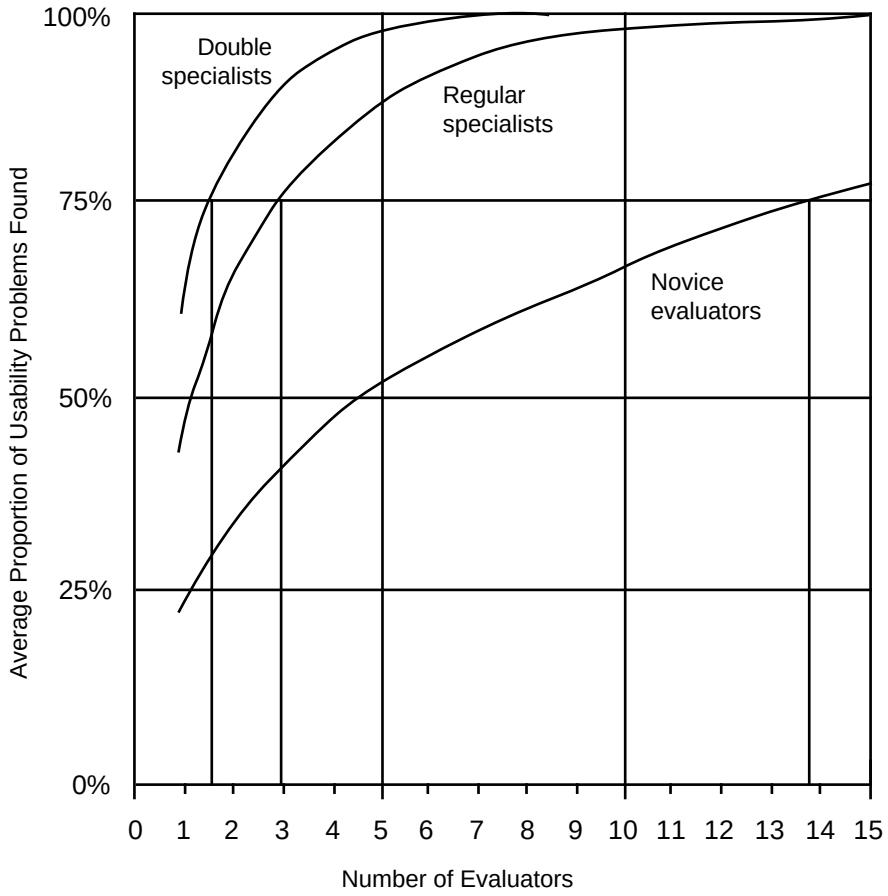

Figure 4.2. Average proportion of usability problems [from six previous studies] found as a function of number of evaluators in a group performing the heuristic evaluation.

The intercepts with the x-axis show the number of evaluators in each group required to find 75% of the total problems reported by that group.

Source: Nielsen (1992).

The claim, then, is that when different evaluators' problems are aggregated, between 3 and 5 experienced evaluators will be required to identify most of the problems found in a heuristic evaluation. Though the precise figure would vary according to experience level, system type and the stage at which evaluation was performed (Nielsen & Landauer 1993), this number is in principle predictable:

> "... there is a nice payoff from using more than one evaluator [in heuristic evaluation], and it would seem reasonable to recommend the use of about five evaluators, and certainly at least three. The exact number of evaluators to use would depend on a cost-benefit analysis, and more evaluators should obviously be used in cases where usability is critical or when large payoffs [due to extensive or mission-critical use] can be expected ...". (Nielsen 1993 p156, Nielsen 1994d p33).

The kind of cost/benefit analysis to use would depend on similar assumptions. Nielsen & Landauer (1993) and Nielsen (1994c) offered projections of the numbers of both heuristic evaluators and think-aloud test users which would be necessary to achieve optimal cost/benefit ratios for typical (medium-large) software projects. For heuristic evaluation (Nielsen & Landauer 1993) and 'discount user testing' this ratio occurred at between 4 and 5 evaluators, while for 'full' user testing it was around 3 evaluators (Nielsen 1994c).

The claim is that heuristic evaluation and "simplified thinking aloud" (think-aloud tests without use of specialist laboratory equipment), can, when confined to specified tasks and restricted usage scenarios, produce benefits comparable to those of the "best" or more perfect methods at greatly reduced cost (Nielsen 1994c). This approach forms the basis for what Nielsen (Nielsen 1989, 1993) dubbed "discount usability engineering". (See Chapter 1, Section 5.3 for a discussion of the role of heuristic evaluation in discount usability engineering.)

2.2 Validation

The only explicit attempt at replication of Nielsen's claim known to the author is described in Virzi's two (1990, 1992) papers. A third paper on which Virzi collaborated, Dumas et al. (1995), provides further support, while Sears's comparison of heuristic evaluation and cognitive walkthrough, Sears (1997), can be considered as a partial validation. Lewis's (1994) replication of Virzi's study will be discussed later.

In the first of these studies, Virzi (1992) found that in three separate think-aloud experiments (the first reported in Virzi 1990), 80% of the usability problems identified were reported within the first five test subjects (5, 5 and 4 subjects respectively), additional subjects introducing decreasingly fewer new problems. Further, where severity ratings were taken, the most severe problems were uncovered "first", that is, within 2 or 3 subjects. Dumas et al. (1995) reported that (in a single think-aloud test) a similar proportion of problems were found by 5 evaluators, but in the same time (the first hour) the same proportion had been detected by less than 3 evaluators. Sears (1997) found that it took only 2 or 3 evaluators to identify the same proportion of serious and intermediate problems as did 5 or more evaluators of minor problems (for each of the three methods there compared).

What all these studies have in common is that they used cumulative counts to investigate the effect of successively adding new problems to a running subject total. Cumulative problem counts are additive totals of unique problems tokens (UPTs) found by successive evaluators, discounting repetitions. Since the order of independent evaluators is not important, this order can be varied or permuted a large number of times, so as to take a 'line of best fit' through the evaluator totals[1]. This results in cumulative curves like those in Figures 4.1 and 4.2. From such curves, assessment can be made of the number of evaluators required to find any given proportion of the problems found by all evaluators taken together. (Expressing cumulative counts as percentages of total UPTs also enables direct comparisons to be made between curves which represent different evaluator populations.)

[1] Virzi (1992) and Lewis (1994) likened this technique to a Monte Carlo procedure.

All three of the above studies can be seen to support Nielsen's claim, in respect of not only the '3 to 5 evaluators' but also the conditions where even fewer evaluators are required. This is so despite differing in the methods and techniques used (think-aloud tests, heuristic evaluation, cognitive walkthrough, plus a new technique called heuristic walkthough), subjects (novices, interface design experts, computer science students) and source of severity assessment (independent experts, experimenter). The combined effect is to provide strong, if not extensive, support for Nielsen's claim.

In this Chapter we shall compare the cumulative curves derived from Experiments 1 and 2 with those used by Nielsen in support of the '3 to 5' claim. The hypotheses raised will be assessed here and taken forward into later Chapters, in particular Chapter 5 (Experiment 3). Before doing so, however, we should first examine the underlying model on which the claim is based.

3. Cumulative Problem Curves and Probability Theory

We shall use a hypothetical (thought) experiment to manipulate the underlying model.

Imagine a lottery in which there are 20 balls to choose from and only one winning ball. Each ball has a unique label, from 'a' to 't'. Balls reside in a large tub, from which our punters (experimental subjects) must pick. Subjects may only pick one ball at a time, and after each subject's selections are recorded the balls are returned to the tub for the next subject. In this thought experiment there are three separate groups of ten subjects (three conditions), namely Low, Medium and High. **Low** condition subjects are each constrained to pick between 2 and 4 balls (mean = 3); **Medium** subjects may each pick between 6 and 8 balls (mean = 7), while each of the **High** subjects can pick from 13 to 15 balls (mean = 14). Thus Low condition subjects have a low chance (= 3/20, or 0.15) of picking the winning ball (letter), Medium subjects an intermediate chance (= 7/20, or 0.35), and High subjects the best chance (= 14/20, or 0.70) of the prize. (There are actually three prizes, one per condition, but subjects are not told this !)

Since choices (picks) are random, each member of each condition has the same chance of picking the winning letter as the 9 others in that group. Thus within each condition the probability (p) of any one subject picking any one letter will be the same. According to binomial probability theory (Lewis 1993, 1994), the cumulative probability (P) of a letter of probability p being picked at least once is given by

$$P = 1 - (1 - p)^n \qquad (1)$$

where n is the number of subjects in each group[2].

[2] This depends on three preconditions: (a) selections are random, (b) subjects are independent of one another, (c) the total number of available choices is the same for each subject (Lewis 1993, 1994).

Since we know the mean value of p for each condition, we can plot the cumulative probability of the winning letter being found by each group. Doing this for

n = [1..10] with

$p_{Low} = 0.15$, $p_{Med} = 0.35$, $p_{High} = 0.70$

produces the three curves 'Low', 'Med' and 'High' shown in Figure 4.3(a). It is clear that because each member of the High group is allowed more selections (from the same total of 20) than the other groups, this group's pool of available new letters will run out very early (producing a winner within the first four subjects). In contrast, the Low group may not find a winner at all, new letters still being available after all ten subjects have picked. Someone in the Med group will probably win, their curve converging at just over ten subjects. (Note that since the number of available choices is always the same, in this case we can compare absolute totals rather than proportions.)

Figure 4.3(b) shows the actual cumulative counts of the numbers of new letters picked by each group. They were produced by randomly selecting one of 20 tokens an (evenly weighted) number of times according to membership of the three imaginary groups. Each curve (Low, Med, High) was plotted by varying the subject order 100 times and plotting the cumulative totals of new tokens (discounting repetitions) selected by the 10 subjects in that group. (With UPTs for lottery balls, this will be the method of production of cumulative curves - adapted from Virzi (1990, 1992) - used throughout this thesis[3].)

It is clear that the 'High', 'Med' and 'Low' curves in (a) and (b) are almost the same. In particular, the points at which the actual curves in (b) converge (after which further subjects found no new tokens) is in close agreement with that predicted by the theoretical model in (a). Further, the points at which the corresponding curves intercept the y-axis are the same relative to the axis. Thus for cases where picking from the available choices *is* random (that is, not favouring any one choice over another), the theoretical model can be said to be a good predictor of usability problem (token) selection behaviour. The corollary is that when choices are *not* random, comparisons with the ideal curves will show by how much real choices have deviated from the ideal.

However, there are limits to this comparison. For the purposes of the above demonstration we assumed both a fixed total number of tokens (20, here) and equal numbers of evaluators (10) between the predicted and actual curves. In reality neither total UPTs nor subject (evaluator) numbers may be equivalent, so that the correspondence between the shapes of the theoretical (ideal) and actual (observed) curves will begin to break down. This is likely to be the case when subject numbers are insufficient to demonstrate convergence within the total UPTs predicted by the ideal model. Such lack of correspondence will be more marked for lower rather than higher values of p.

---

[3] Sears (1997) and Bastien & Scapin (1995) used a different method, taking all available permutations of (1, 2, ...n) members of each subject group.

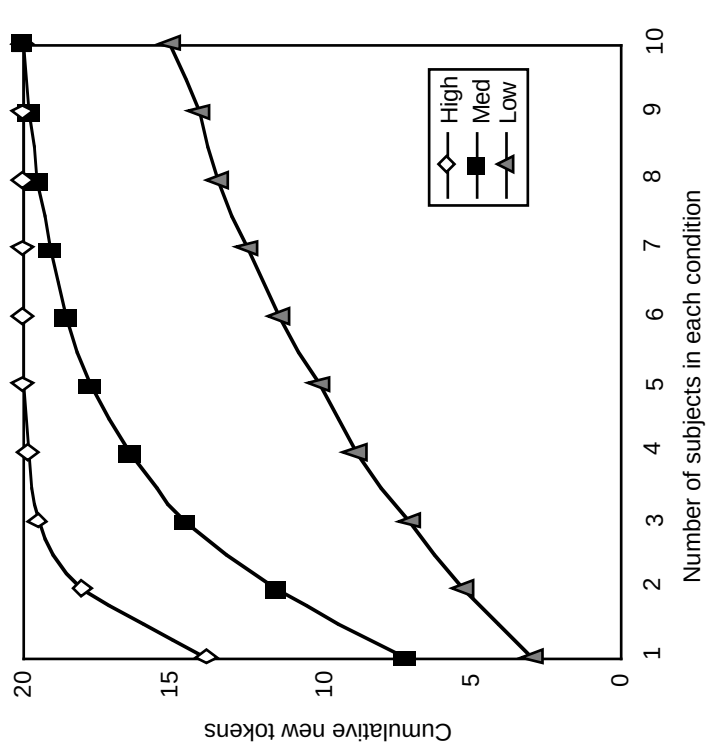

Figure 4.3(b). Cumulative counts of new tokens randomly picked by members of the same three conditions (High, Med and Low) from the imaginary pool of 20 tokens. The number of tokens that subjects may select is constrained by condition membership, as follows:

High: between 13 and 15 tokens (mean 14)
Med: between 6 and 8 tokens (mean 7)
Low: between 2 and 4 tokens (mean 3).

Note that the start point of each line (the y-intercept) is the same as the mean number of tokens per subject in that condition.

See text for explanation of how the curves were produced.

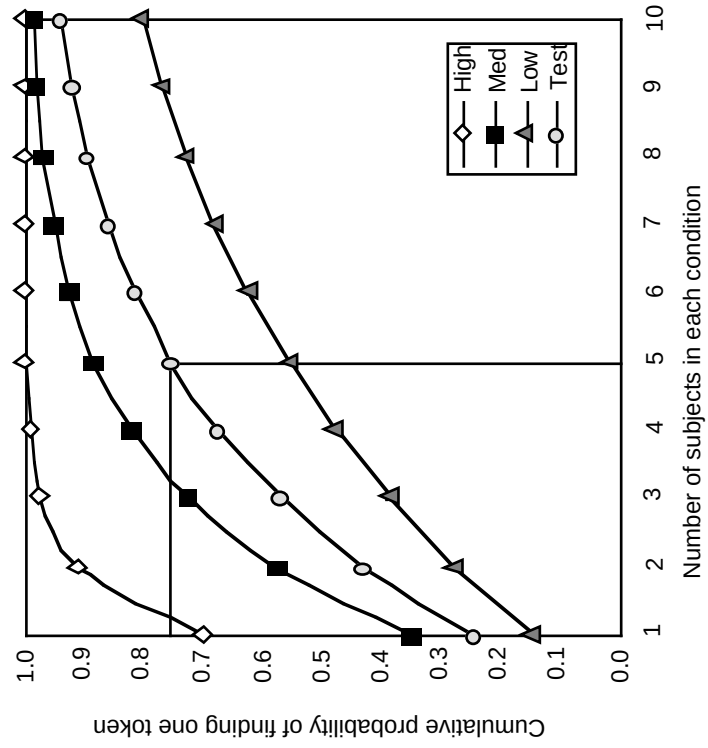

Figure 4.3(a). Cumulative probabilities of members of three conditions High, Med and Low finding one of 20 tokens.

Lines are produced by the function
$y = 1 - (1 - p)^x$
where p is the mean probability of one subject from each condition independently finding any one token, and x is the number of subjects in that condition.

The height of each y-intercept is equal to the value of p for that condition, namely High: 0.70 (=14/20), Med: 0.35 (=7/20), Low: 0.15 (= 3/20).

Plotting the above function with p = 0.245 produces the line labelled 'Test' This represents the lowest value of p which will enable 75% of a sample of randomly selected UPTs to be found by 5 subjects.

Figures 4.3(a) and 4.3(b).

Since in eq (1) when n = 1, P = p, the value of p is also the size of the y-intercept. Thus the shape of any theoretical curve can be easily predicted from the value of p alone (i.e. from the

start point for the curve). This means that *when* we have reason to believe that a particular cumulative curve will follow the ideal model (as do the three curves in Figure 4.3(b)), we can predict its form (and thus the number of evaluators required to find any given proportion of problems) from the expected value of p for that sample (in Figure 4.3(b), 0.15, 0.35 and 0.70). In such *limiting* cases there will be agreement between actual and ideal rates of problem detection, so that p will correspond to the **detection rate** for a given sample:

detection rate = mean problems per subject / total UPTs found by those subjects (2)

The circumstances under which the correspondence between actual and theoretical detection rates will break down are complex, but include situations where both subject numbers and expected p values are low. For the moment, we shall defer further discussion of this issue (and that of the shape of the curves) till later in the thesis and Chapter 8 in particular.

Assuming this correspondence, if we want to know the ideal detection rate which will enable any number of evaluators to find a certain percentage of problems, we can calculate it from the required values of P and n. With P = 0.75 and n = 5, this tells us that a detection rate of 0.245 (≈0.25) will be required to enable 5 evaluators to detect 75% of the total UPTs found by any evaluator sample. This is confirmed by the table of sample sizes in Lewis (1993), reproduced below (Table 4-1), from which we can see that 5 evaluators are required to detect 75% (.75) of problems with a probability of occurrence of .25. The table also shows that 13 evaluators are required to find the same proportion of problems of probability 0.10, and that in order to find 95% of such problems we would need 28 evaluators. (We shall see that many of the cumulative curves generated for Experiments 1 to 3 exhibit similarly low detection rates, thus requiring unrealistically large subject numbers to exhibit convergence within ideal limits.)

Problem probability	Cumulative likelihood of detecting the problem at least once					
	.50	**.75**	**.85**	**.90**	**.95**	**.99**
.01	68	136	186	225	289	418
.05	14	27	37	44	57	82
.10	7	**13**	18	22	**28**	40
.15	5	9	12	14	18	26
.25	3	**5**	7	8	11	15
.50	1	2	3	4	5	7
.90	1	1	1	1	2	2

Table 4-1. Sample size requirements as a function of problem probability and the cumulative likelihood of detecting the problem at least once. (Source: Lewis 1993 p669).

Plotting eq (1) with p = 0.245 produces the curve labelled 'Test' in Figure 4.3(a). This represents a theoretical baseline against which such projections as Figure 4.1 may be assessed. By shifting the y-axis of the latter curve to x = 1, we can now see that this projection requires a value of p which is higher than the 25% predicted by the model, namely around 33%.

For curves that *do* follow the model, we can also determine the effect on p of increasing the number of problems found. By interpolation from eq (1), it can be shown that an increase of x over y problems represents a decrease in the value of p of

$$1 / (1 + (x / y)) \quad (3)$$

of the original value. Plotting a new probability curve with the decreased value of p will then tell us by how much the number of subjects required to find a given percentage of problems would rise as a result of the increased number of problems. (We will later do this for one of Nielsen's own problem sets.)

As stated above, the theoretical (ideal) model just described was used by Nielsen & Landauer (1993) to underpin their predictions of the optimum cost/benefit ratios for typical software product evaluations. In that paper, a version of eq (1) was used to compare actual and fitted curves for 11 heuristic evaluation studies. In this and subsequent Chapters we will compare the cumulative curves derived from the experiments reported in this thesis with the requirements of the same model. We shall see in this Chapter that the cumulative curves from Experiments 1 and 2 exhibit lower detection rates than those expected by this ideal.

4. Experiments 1 and 2: Cumulative Curves by Problem

In this Section we shall compare the cumulative problem curves for Experiments 1 and 2 with those used by Nielsen to support the '3 to 5' claim. We shall also compare the curves with those predicted by the theoretical model. The result of doing so will be to generate some hypotheses concerning the problem reduction process.

Cumulative problem counts were generated for each of Experiments 1 and 2. The source was the single sets of UPTs from each experiment. Each set of UPTs represents the total number of non-duplicate problems against which any one subject's problem count may be compared. (The production of UPT sets is very much a non-trivial process, which will be explored in detail below.)

Mean cumulative counts of new (non-repeating) problems were computed by generating 100 permutations of the subject order (adapting the method used by Virzi (1990, 1992)). This produced the cumulative (permuted) curves shown in Figures 4.4 (a) and 4.4 (b). These show the mean cumulative proportions of the total numbers of UPTs found by subjects in (a) Experiment 1 (by group) and (b) Experiment 2, session 2 (all subjects), respectively. It can be seen that the curves do not converge in the manner predicted by Nielsen and others. In particular, in both cases the number of subjects required to uncover 75% of the total UPTs found by the group concerned was larger than 5, namely 10 experienced and 16 novice (Experiment 1) and 14 or 15 novice (Experiment 2).

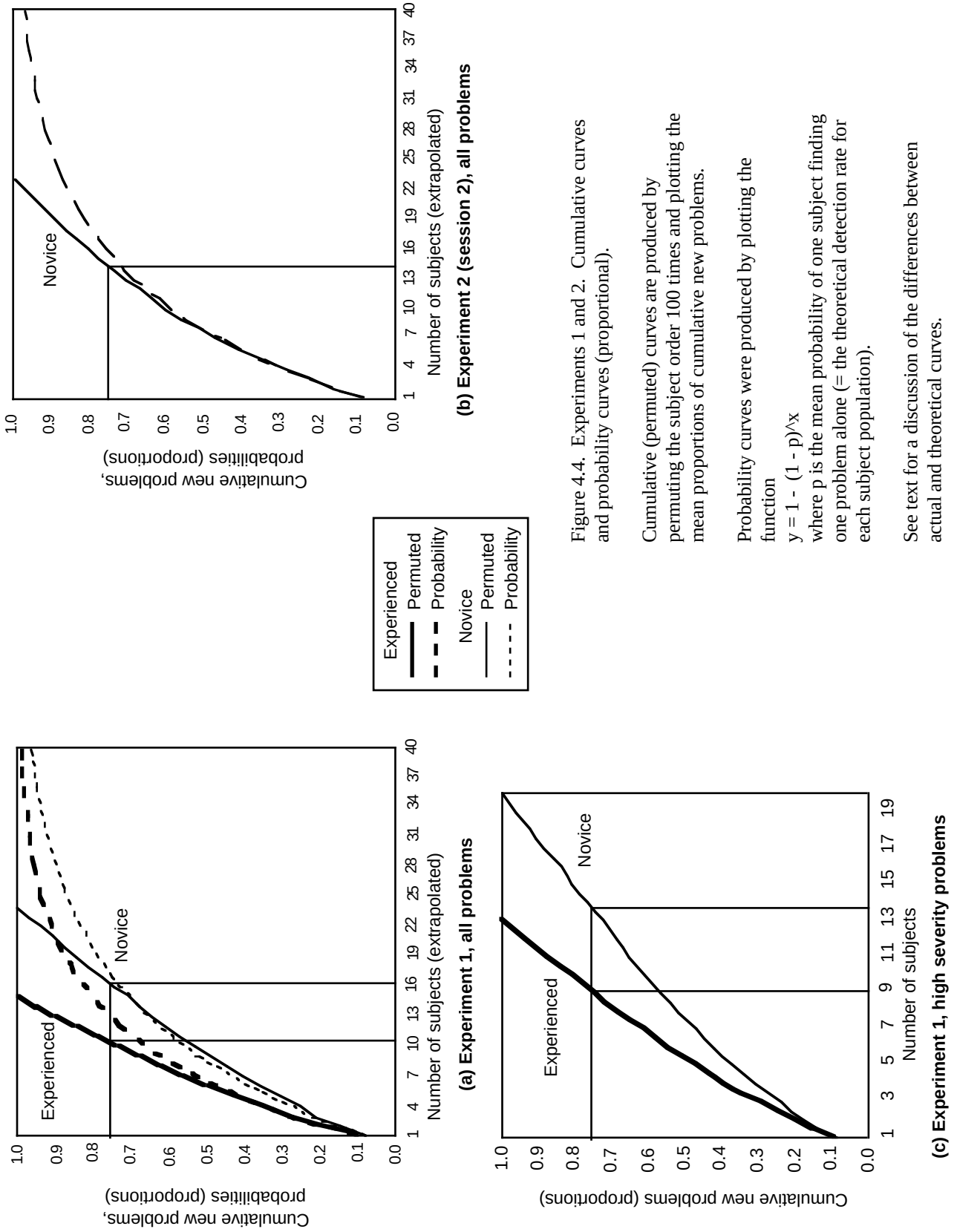

Figure 4.4. Experiments 1 and 2. Cumulative curves and probability curves (proportional).

Cumulative (permuted) curves are produced by permuting the subject order 100 times and plotting the mean proportions of cumulative new problems.

Probability curves were produced by plotting the function
$y = 1 - (1 - p)^x$
where p is the mean probability of one subject finding one problem alone (= the theoretical detection rate for each subject population).

See text for a discussion of the differences between actual and theoretical curves.

Figure 4.4 (a), 4.4(b) and 4.4(c)

The first reason for these discrepancies becomes clear when we consider the detection rates for each experiment (Table 4-2). Rather than the 33% required by the Nielsen projection, these two experiments exhibit detection rates of just 5.6% and 8.4%

respectively, with the experienced subjects in Experiment 1 performing only slightly better than novices in either experiment (at just over 10%). These figures are reflected in the low starting points for the curves in Figure 4.4.

Experiment 1			Experiment 2
Novice	Experienced	All	Novice
7.66	10.07	5.56	8.43

Table 4-2. Detection rates (%) for subject groups in Experiments 1 and 2, all problems.

The second reason for the differences in the number of evaluators can be seen by comparing these cumulative curves with those predicted by the probability theory described in Section 3. Plotting eq (1) with mean p equal to the each of three detection rates in Table 4-2, and extrapolating n until P tends to 1, produces the probability curves in Figures 4.4 (a) and (b). It is clear that the actual and theoretical curves do not agree, but that in this case the former run outside the latter (thus slightly reducing, rather than increasing, the numbers of subjects required to uncover 75% of problem totals, compared with the model). Further discussion of this complex issue will be deferred until Chapter 8.

We saw in the previous Chapters that in Experiment 1 there was a significant effect of problem count for high-severity problems found by experienced subjects (and that no such effect was found for Experiment 2). Therefore it is reasonable to suppose that the cumulative curves for severe problems in Experiment 1 might demonstrate some convergence for experienced subjects. Figure 4.4 (c) shows the curves for high-severity problems in Experiment 1. Comparison with Figure 4.4 (a) shows that the numbers of subjects required to find 75% of high-severity problems was only slightly lower than for all problems (respectively 9 and 14 against 10 and 16, experienced and novice), and that once again there was little convergence. This continued lack of agreement is reflected in the low detection rates for high-severity problems (Table 4-3) and the start points for the curves in Figure 4.4 (c).

Experiment 1		
Novice	Experienced	All
7.18	8.97	4.85

Table 4-3. Detection rates (%) for subject groups in Experiment 1, high-severity problems.

However, these results do confirm the finding of Virzi (1990, 1992) that problems judged to be more serious are found “first”, that is, by a higher proportion of evaluators within a given sample. And while the curves for experienced subjects demonstrate little agreement with either the overall averages in Nielsen (1993) and Nielsen (1994d) (Figure 4.1) or the regular specialists in Nielsen (1992) (see Figure 4.2), those for novices are comparable with Nielsen’s (1992) projection of 14 novices to find 75% of UPTs.

At this point a caution should be introduced regarding detection rate comparisons. Though it is possible to generate reasonable detection rates for the subjects in each cell of the 2 x 3

design of Experiment 1 (and the 1 x 3 cells of Experiment 2, session 2), it must be pointed out that such figures would be derived by dividing relatively large subject population means into relatively small UPT totals. For example, the detection rate for Experienced Principle subjects in Experiment 1 is 24.6%, the ratio of the mean problems per subject (12.80) and total UPTs (52) found by those subjects *alone*. This is a consequence of not having a datum for the number of UPTs that there actually 'were' to be found, against which to compare each cell mean: in the absence of such a figure, we must use the number of UPTs found by each subject population. Thus in Experiments 1 and 2 we do not know whether subjects were missing 'real' problems, inventing or hypothesising 'unreal' ones ('false positives') or accurately predicting 'real' problems ('hits') which would be observed in user testing. This important issue will be addressed in the next Chapter (Experiment 3).

5. The Problem Reduction Process

In order to investigate the possible reasons for the discrepancy between the cumulative curves derived for Experiments 1 and 2 and those used by Nielsen and others, we will examine the stages of the process by which individual subject problems are reduced to a single set of UPTs.

5.1 A Model

Figure 4.5 shows the author's model of the problem reduction process which is believed to be undertaken for any one experiment or user test employing more than 1 subject or test user. The process involves at least three stages, dubbed 'within subjects', 'between subjects' and 'between types'. Subjects (or test users) are numbered S(i), S(i+1) ... S(n).

Within subjects. In the first stage of the process each subject's account is reduced to a set of single usability problems, not yet matched with those of other subjects. This involves the removal of any duplication in individual accounts, so that each subject's problem set incorporates only one instance of each problem. The inputs to the process are the records of the subject sessions, from which individual subject protocols have been extracted. (The records can take many forms, including user-system logs, video and audio tracks, observer notes). The outputs are n sets of non-duplicate problem instances, each of which represents the minimum number of separate problems identified by that subject. Since individual subjects tend to repeat themselves, the identification of duplicate problems is relatively easy, and can be agreed with the subject subsequent to the session (as was done in Experiment 1).

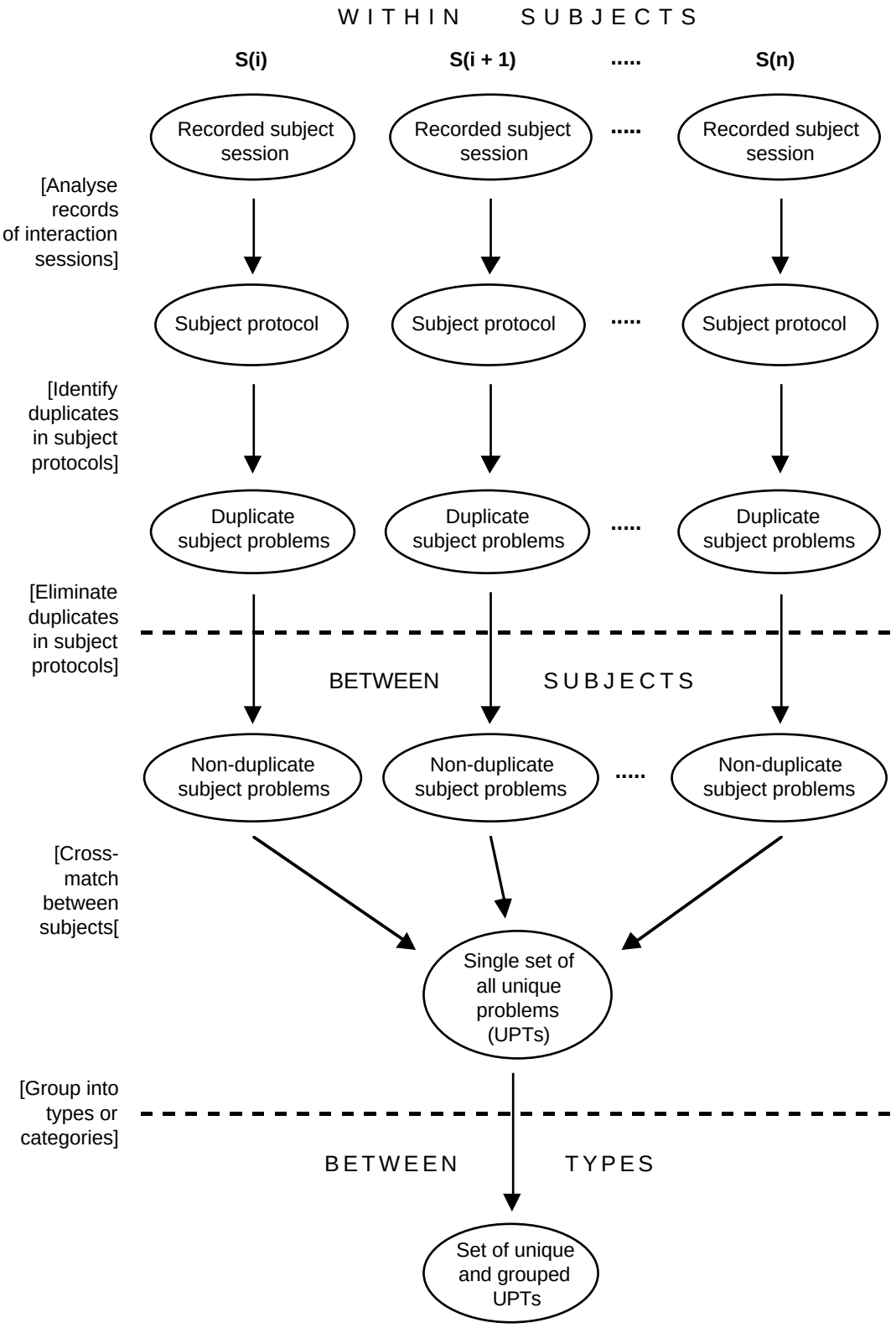

Figure 4.5. The three stages of the problem reduction process. Problems separately identified by subjects S(i), S(i + 1), ... S(n) are reduced to a single set of unique, grouped problems (UPTs).

Between subjects. The second stage investigates the manner in which individual subjects' problem sets overlap with those of other subjects. This involves identification of issues which have been uncovered by more than one subject (i.e. repeating instances of the same problem from the emerging set of UPTs). The inputs to the process are the n sets

of non-duplicate problem instances. The output is the single set of all unique problem tokens (UPTs) found collectively by all subjects. This is a lengthy business, involving the cross-matching of a large number of problem instances. Since different subjects tend to make similar but not identical points, the process of deciding whether particular problem accounts are the same, different or overlapping (and, crucially, to what degree they overlap) is extremely difficult. Worse, some problem instances will need to be split apart while others are re-combined, requiring a return to the 'within subjects' stage. Thus the first two stages can be seen as an iterative process in which individual problem accounts are matched and re-matched into (*one* version of) a final set of UPTs.

Between types. The third stage involves the organisation of the resulting set of UPTs into groups or categories. This is necessary in order to give structure to what may be a large set of relatively disparate usability issues. If guiding heuristics or principles have been used, the organisation will naturally reflect these groupings, but in principle the UPT set can be organised on any pragmatic basis. The output from this final stage will assist the identification of possible causes and remedies for each cluster of usability issues, such that recommendations can be made for (re)design or other changes.

5.2 An Illustration: Experiment 2

As an illustration of the full reduction process, we shall follow the three stages of the model from an individual subject's problem account to the final set of UPTs for Experiment 2. The problem account is contained in the report form that appears in Appendix J, from a (novice) subject ('Heuristic 3') in the Heuristic condition. (In this experiment each subject's protocol was entered directly in to a prepared spreadsheet by the subject her/himself.) The wording of the problem descriptions (with only minor editing for spelling and punctuation) is reproduced in Table 4-4. For the purposes of this illustration only, the author has added clarifications in [square brackets]. Problems 4, 5, 6, 13, 17 and 18 relate to the 'covariation principle' exercise of which the sample screen illustrated in Figure 3.1 of Chapter 3 forms a part.

Problem No.	Problem description
1	On title page - Social Psychology the modules - there is no indication that program is loading automatically; it seems as though you should have to click something to start
2	Video clips are difficult to hear and interpret
3	Explanations of the possible forms of attributions in videoclip examples are not very comprehensive
4	Page "what influences attributions" should come before examples to provide a better picture of the relevance of the clips
5	Use of term A and S very confusing in the Covariation Information page
6	No explanation about how low consistency is known as an unstable attribution [In the Covariation Information page]
7	Contents page does not make it clear that the first four text boxes form a continuum and that all information can be accessed from one box ["Text boxes" represent the Contents page listing; only the first three are linked together]
8	References on the "Further readings" page[s] should be accessible from the pages to which they are relevant
9	There is no facility to search by a word in order to find information about a specific topic
10	It is not possible to go from a word in the glossary directly to the page which relates to that word
11	Should be links between the "further reading" and the references section, to allow explanation of relevant references to be accessed from their reference information
12	The "GO contents" [pull-down menu] option returns the user to the Social Psychology contents page, not the Attribution Theory contents page [Requiring the user to enter the Attribution Theory half module again]
13	On the "Covariation Principle" page, only consistency is highlighted in bold type, whist all three of the terms do appear on the glossary
14	Words which appear in the glossary are not highlighted every time they are mentioned
15	It is not clear that the "Further readings" are listed under different headings [There are four sets of readings, on separate pages]
16	Further readings on later topics can only be accessed by scanning through the further readings for original topics [There is no Contents page for the four sets of readings]
17	The situations created by drawing up parameters in the Covariation Information page are non-sensical and confusing
18	It is possible to highlight both high and low conditions in the attribution exercise

Table 4-4. Experiment 2. Problem descriptions of a single subject (Heuristic condition, subject 3). Text in [ ] are the author's clarifications. The only reductions that were finally applied by the author were to combine pairs **13** & **14**, **15** & **16**, thus reducing this subject's total to 16. See Appendix J for a transcript of the report form.

On the first pass of the 'within subjects' stage, the 18 problems reported by this subject were reduced to a tentative 11. This involved combining together problem pairs 3 and 17, 5 and 6, 7 and 15, plus the four problems 8, 10, 11 and 16. However, in the 'between subjects' stage, it became clear that unambiguous matches could not be made with other subjects' problem reports without an unacceptable level of grouping (this particular subject made several related points). Hence on the second pass of the first stage, the 18 problems were reduced to just 16 (by combining pairs 13 and 14, 15 and 16), a decrease of 11%. In the final pass of the second stage, most of these 16 were matched with one or more problems from other subjects. (See Appendix H for a complete listing.) In the 'between types' stage, every matched set was later assigned to one attribute from the principles set (one of many possible categorisations which could be applied).

UPT No.	Attri-bute	Problem description	Problem No.*	Reported by subject (condition, no.)	
5	FO	"What influences attributions" should come before examples	4	Heuristic 3	
29	FP	Covariation information: too much repetition with different selections, situations with some parameter combinations make no sense	17	Heuristic 3	Control 5
34	TLS	Covariation explanations arising from video clip(s) unclear, unrelated to rest of material, boring, inadequate, repetitive	3	Heuristic 3	Control 6
					Control 3
					Heuristic 6
					Principle 1
					Principle 5
35	TLS	Covariation example terminology (A, S) inadequately explained	5	Heuristic 3	Control 3
					Control 7
					Heuristic 1
					Heuristic 2
					Heuristic 5
36	TLS	Causal attribution: terminology (e.g. consistency) explanations missing	6	Heuristic 3	Heuristic 6
60	LN	Provide (selectable) Contents for Further Reading	15/16	Heuristic 3	Control 7
					Heuristic 6
63	LN	Not clear (in Contents) that first three sections run into each other : use discrete topics (separate links from one section to the next)	7	Heuristic 3	Heuristic 6
69	MS	"Go contents" menu item returns to top level, not this module opener	12	Heuristic 3	
73	MS	Provide links from text to Further Reading and further information	8	Heuristic 3	Heuristic 8
74	MS	Provide two-way links between References and Further Reading	11	Heuristic 3	
75	MS	Provide links from Glossary terms to main text	10	Heuristic 3	
77	MS	Provide a finder, search by keywords, topic index	9	Heuristic 3	Control 5
					Heuristic 7
87	PCL	Video sound quality poor (background noise): provide transcript	2	Heuristic 3	Control 4
					Heuristic 6
					Heuristic 8
					Principle 5
					Principle 8
89	FEE	No indication when module is loading (from top level page)	1	Heuristic 3	Principle 5
92	OT	Can highlight both high and low conditions (attributions exercise)	18	Heuristic 3	
104	AC	Missing emphasis for glossary terms (e.g. consensus, distinctiveness)	13/14	Heuristic 3	Control 3

Table 4-5. Experiment 2. The problem descriptions from the subject ('Heuristic 3') in Table 4-4, re-numbered and matched with those of other subjects (* column 'Problem No.' shows the original numbering). Nos. 104 and 60 correspond to pairs 13 & 14, 15 & 16 in Table 4-4. See Chapter 3, Section 2.2 for an attribute listing, and Appendix H for the full problem set for this experiment.

Table 4-5 shows how the now matched and re-numbered problem reports from this subject contributed to the final set of UPTs for this experiment. Against each UPT no. appears the principle set attribute to which it was assigned, the problem description (now edited to describe the essential features), and all subject(s) reporting that problem (for clarity, subject 'Heuristic 3' is listed separately). (Matches with Table 4-4 are shown in the fourth column).

For example, comments on the poor quality of the video sound (UPT 87, problem 2) were offered by six subjects including 'Heuristic 3', but only this subject noticed that in the covariation exercise both high and low conditions could (sometimes) be highlighted together (UPT 92, problem 18).

The reader might object that there is far more correspondence between these problem accounts than the author allows. For example, UPTs 73, 74 and 75 might be clustered into a single UPT to do with linking main text and reference material, while UPT 36 might be included in the group numbered 35. However, this subject's problem set was chosen for illustration since it represents an extreme example of the tendency to make *similar* but not *identical* points. In that sense, 'Heuristic 3' was untypical of the majority, whose problem accounts were generally much less ambiguous. (The reduction for all subjects was from an original total of 222 problems to a final set of 114 UPTs; this represents a decrease of 48% compared with this subject's 11%). The degree to which the author's problem reductions may be untypical of other researchers will be addressed in the next Section.

6. Experiments 1 and 2: Internal Validity

6.1 Inter-rater Reliability

It may be that the above unwillingness to 'lump together' related problem accounts is unrepresentative of that performed by other researchers. Given the identical subject protocols, other experimenters might arrive at entirely different UPTs with radically different totals. That is, it is possible that the results in Experiments 1 and 2 are an artefact of the author's particular perspective on what qualifies as a separate usability problem. In order to test this proposition, an inter-rater reliability assessment was performed on the problem reduction data from Experiment 1.

We have seen that the problem-sharing profiles of the two previous experiments were very similar, even though based on different totals and recorded in different fashions (by the experimenter in Experiment 1, by subjects themselves in Experiment 2). Table 4.6 shows that the reductions performed on the two problem totals were almost identical. Hence an inter-rater test might be performed on the data from either experiment. Experiment 1 was chosen because its detection rates were the lower (5.6% compared to 8.4%: see Table 4-2). We have also seen that the 'between subjects' stage is the most difficult part of the reduction process. Hence the reliability test was performed on that stage of the reduction process from this experiment.

	Problem Totals (all subjects)		
	Original	Reduced (UPTs)	% reduction
Experiment 1	444	230	48.20
Experiment 2	222	114	48.65

Table 4-6. Experiments 1 and 2. Overall problem reductions, all subjects.

The reliability test was to assess the manner in which the original problem total for Experiment 1 (444) had been reduced down to the single set of (230) UPTs in Table 4-6. Since the first total was so large, random sets of about 30 problems[4] from all subjects in this experiment were separately presented to two independent raters from the York HCI research group (both postgraduate researchers). Rater 1 had been a subject in this experiment, rater 2 had not. Both raters were asked to go through the same process for each set of 30 problems as the experimenter had done for all 444, by separating them into matched clusters on screen. (Raters had available the running software for this experiment, installed on a separate computer, plus paper copies of the problem sets.) The rater's task was to judge whether one or more problems in each set were the "same as" (and *not* "different examples of the same type of thing as") other problems in the same set. As many matching attempts as necessary could be made. Each rater could complete as many sets of 30 as she/he wished, managing 3 and 4 sets respectively.

The resulting matchings for each set of 30 problems were compared with those originally generated by the experimenter (the author). This was done by counting the number of pairs of agreements (and disagreements) between each of the raters' and the experimenter's matches (and non-matches). The experimenter's matches were those which would have been made for only those same 30 problems (that is, disregarding between-set matchings and the remaining 12 problem sets). 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 0.66 (Table 4-7(a)).

Problem Set	Rater 1	Rater 2	Mean
1	0.64	0.64	0.64
2	0.53	0.34	0.43
3	0.85	0.85	0.85
4	-	0.78	0.78
Mean	0.67	0.65	**0.66**

Table 4-7(a). Experiment 1. Inter-rater reliability data (Cohen's κ) for the 'between subjects' stage of the problem reduction process. Each cell compares data from the experimenter and one rater, for the same random sample of 30 problems from the full set for this experiment.

Table 4-7(b) compares the equivalent percentage reductions for each set of 30 problems. These show that the mean reduction performed by the experimenter (22.6%) was comparable to that achieved by each of the two raters (mean 24.0%).

[4] As many as would fit legibly onto a double-A4 monitor screen.

Problem Set	Experimenter	Rater 1	Rater 2	Mean
1	22.86	25.71	25.71	25.71
2	11.11	19.44	30.56	25.00
3	37.84	27.03	27.03	27.03
4	18.75	-	12.50	12.50
Mean	**22.64**	24.06	23.95	**24.01**

Table 4-7(b). Experiment 1. Problem reductions (%) for the experimenter and the same two raters as in Table 4-7(a).

Based on these few samples, the level of problem reduction produced by the experimenter has been shown to be comparable to that achieved by two independent raters with relatively little exposure to the software. The problem matchings which these figures represent were shown to correlate well (Cohen's κ is a stringent test of agreement and non-agreement). Thus it is reasonable to conclude that the problem reduction approach used by the author was not so much different from that of other experimenters as to render the results in this Chapter unreliable. This assertion will be taken up in the Discussion.

6.2 Cumulative Curves by Type

We have seen that the cumulative curves for Experiment 1 did not follow the pattern predicted by Nielsen and others, in that in Experiment 1 more than 5 experienced subjects (10 for all problems, 9 for high-severity problems) were required to uncover 75% of their total UPTs (Figures 4.4 (a) and (b)). This was attributed to a low probability of each problem being found by any one subject alone, compared with the 33% required for a '3 to 5'. It follows, then, that a problem count which resulted in a lower number of UPTs would be more likely to produce curves of the form claimed. Put more simply, lowering the number of sole problems that there are to be identified increases the chance that any one evaluator will uncover a problem already found by someone else.

To test this hypothesis, new cumulative curves were generated for Experiments 1 and 2, this time using only the respective problem categories derived from the two versions of the principles set used in these experiments. Counting was now done not by problem *instances* (out of 87 for novices and 188 for experienced in Experiment 1; out of 114 for novices in Experiment 2), but by problem *types* (out of 19 for novices and 23 for Experienced in Experiment 1; out of 14 for novices in Experiment 2). The numbers of types reflect the categories - principle attributes - into which problems had been grouped (but not every attribute was involved in each subject population). The resulting curves are shown in Figures 4.6 (a) and (b).

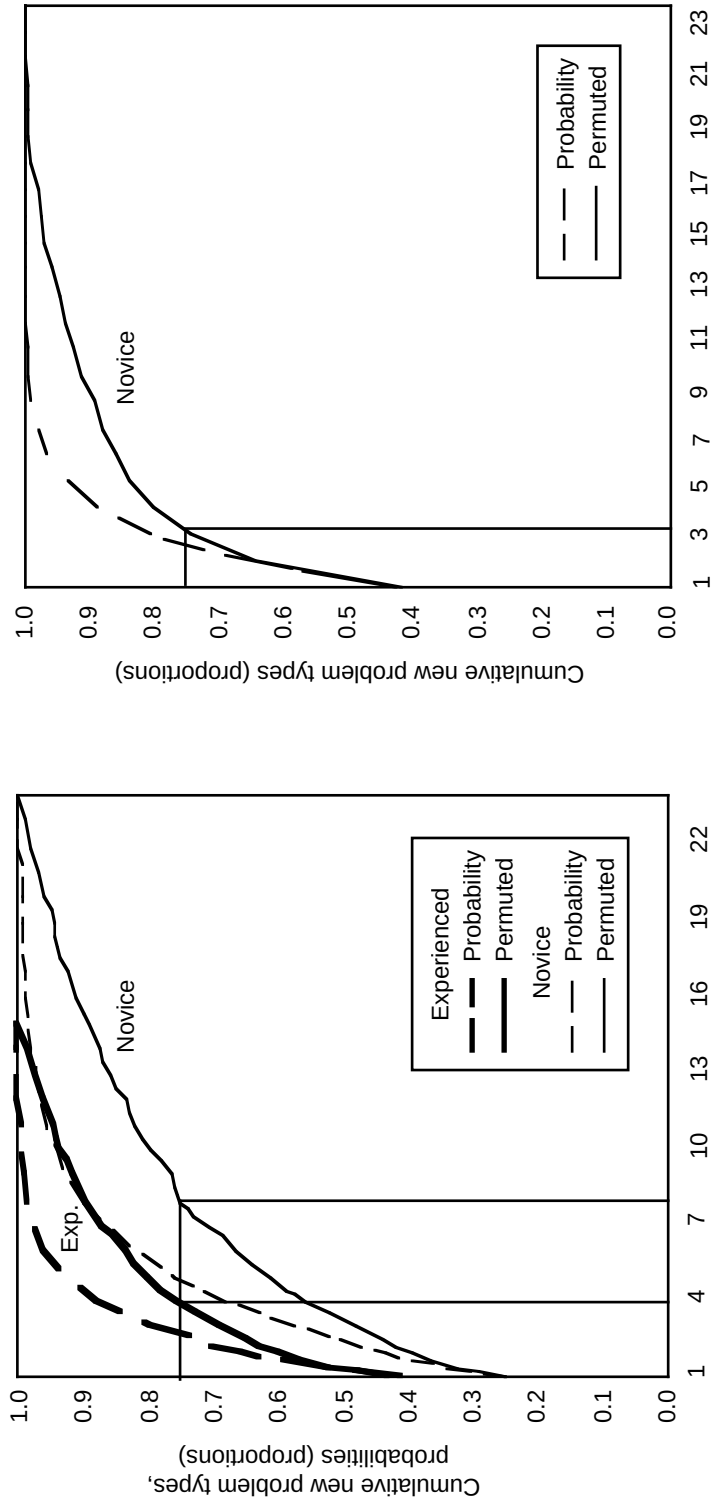

Figure 4.6. Experiments 1 and 2. Cumulative (permuted) curves and probability curves, by problem type (principles set attributes).

The total number of categories (types) used in each case varies according to the subject population:
Experiment 1: Experienced 23, Novice 19
Experiment 2: Novice 14.

See text for a discussion of the differnces between permuted and probability curves.

Figure 4.6(a) and 4.6(b).

It is clear that counting by problem types has enabled the numbers of both experienced and novice subjects required to uncover 75% of their respective (all-severity) totals to be considerably reduced from those in Figure 4.4 (a) and (b). The figure for experienced subjects in Experiment 1 has been reduced from 10 to 4, while that for novices has been halved in Experiment 1 (from 16 to 8) and reduced five-fold in Experiment 2 (from 14 or 15 to just 3). In particular, the figure for experienced subjects is now in line with Nielsen's

prediction. Comparing the actual (permuted) and probability curves shows that the former now run inside rather than outside of their theoretical equivalents (unlike those in Figure 4.4). This is the pattern expected by the model (predicted by Nielsen & Landauer 1993 and confirmed by Virzi 1990, 1992; to be further discussed in Chapter 8). Note that the start points for both the Experiment 1 Experienced curve (detection rate = 41%) and the Experiment 2 Novice curve (detection rate = 42%) are now well above the prediction of 33% for a '3 to 5'.

We have seen, then, that counting by *type* rather than *instance* has enabled the proportion of problems found by any one subject to be raised sufficiently to emulate the Nielsen projection for experienced evaluators. By the simple expedient of higher-level categorisation, in this case on lines dictated by the principles set, we have been able to reproduce the predicted effect for experienced subjects in Experiment 1. The effect for novices in both experiments was further reduced from that predicted, particularly so in Experiment 2. It can be shown that any categorisation at the same level, that is, into similar numbers of groups or frequencies, would produce comparable results.

It is the author's belief that the iterative nature of the problem reduction process means that implicit problem categorisations may come to dominate the whole process. This is believed to be an unwitting result of a natural tendency to combine together (schematise) related items at a higher level. In fact it may be impossible *not* to do this, since the essential 'between subjects' stage of the process involves cross-matching of problems into 'same as' groups (Figure 4.5). The author agrees with John & Mashyna (1997, p1019) that there may be a tendency to conflate problem instances with problem types. The result of this will be to drive down final problem totals, making the contributions of individual evaluators appear larger, and cumulative curves converge earlier, than would otherwise be the case.

This raises the following speculation: that the effect reported by Nielsen and others might also be a result of problem grouping or categorisation, occurring at an early stage and persisting into final problem totals. If this is so, then the detection rates (and cumulative curves) used to support the '3 to 5' claim will be higher than they would be if the separation of problem instances took place at a lower level. This in turn requires that the problem lists on which these and other results are based would admit of different interpretations, different matchings, out of which a higher number of final UPTs might emerge. This final proposition will be addressed in the next Section.

7. External Validity: Two UEM Studies

In this Section we shall inspect the problem lists which appear in two selected papers, in order to see if an alternative interpretation of the usability issues which they represent could produce different (and possibly higher) problem totals. The goal will be to increase some of

the UPT totals on which the claims of Nielsen and others may be based, such that the corresponding detection rates might fall to a level sufficient to refute the '3 to 5' claim.

Unfortunately there are very few published results which contain problem listings in sufficient detail for us to do this. Typically only illustrative or representative accounts are included. However, there are exceptions, one being Nielsen's own (1994b) paper on the number of subjects required for a think-aloud test. A second paper, Wright & Monk (1991), included selected problem scenarios, but these authors have kindly allowed access to the original data on which the scenarios were based. Further, both of these papers made use of the same underlying model as described above. We shall look at each of the two papers in turn.

7.1 Nielsen (1994b): Estimating the Number of Subjects Needed for a Thinking Aloud Test

In two separate studies by Nielsen of the use of the 'think-aloud' method by non-specialists, the combined mean detection rate for single subjects was 29%, rising to 86% for 6 subjects (81% for 5 subjects). The cumulative curves for each study fitted closely with the theoretical model used in Nielsen & Landauer (1993)[5] and described in Section 3, generating an estimated probability value of p=0.282. Based on these findings, Nielsen speculated that early rather than late user testing would find more usability problems per test subject, and for iterative design and testing he recommended a relatively small number of users per test. Once again, 75% of problems in these two studies were identified by running 4 and 5 subjects respectively.

The single-subject detection rate of 29% is well within the projected figure of 33% for heuristic evaluation, and supports the view that even non-specialists (here computer science students) can profitably conduct think-aloud studies. (The test subjects are not specified.) The p-value of 0.282 is also well above the minimum of 0.25 required by the model (see the curve labelled 'Test' in Figure 4.3(a)). However, these results mask two issues: first, the total set of UPTs for each study (from which the detection rates were computed) was the union of all problems in the students' test reports, rather than a 'known' base for that application; second, the arbiter of the degree of problem reduction applied to each report (both 'within' and 'between' reports) was not the student experimenters themselves (and was, presumably, the paper author).

As to the first issue, the paper shares with most of the literature the questionable assumption that the result of running even a large number of objective user tests (let alone subjective inspections such as heuristic evaluation) represents the sum total of the problems 'with' an interface. (This crucial question will be addressed in Chapter 5 and further debated in Chapter 8.) The second issue concerns the additional level of problem reduction

[5] Described as a Poisson model in Nielsen & Landauer (1993). In that version, p is written as λ.

which is introduced by employing more than one evaluator (of test users' accounts and/or data) in think-aloud tests. Not only will different test *users* produce different problem accounts, but different *evaluators* will disagree on what those accounts represent. Thus the final set of UPTs will be the product of some combination of both users and evaluators (depicted in Jacobsen et al. 1998a, 1998b as an interaction). Thus it is tempting to believe that the usability problems presented in the Nielsen (1994b) paper may reflect Nielsen's perspectives as much as those of its student evaluators.

With these comments in mind, we shall take a close look at the problem descriptions in the paper. The goal, once again, is to see if these accounts might allow for a higher number of separate usability problems or "single design aspect[s]" (Nielsen 1994b p627) than are contained in these descriptions.

The Appendix to the paper (Nielsen 1994b pp394-397) lists 23 problem descriptions, 9 for the first study (of a "commercial word processor") and 14 for the second (a "shareware outliner"). We shall take the first five problems in order of presentation.

Study 1, problem 1
The arrow keys had no effect in this word processor even though they were present on the keyboard. This problem was corrected in version 5 of the application.

This may be acceptable as a single problem, but with several possible consequences.

Study 1, problem 2.
Users had problems learning the standard cut/copy/paste commands. First, the commands only worked if something had been selected. Second, the copy command gave no feedback, leading some users to wonder if it had any effect.

Two separate problems may be detected, one to do with the need for prior selection (a global issue with several likely ramifications, hence a good candidate for *later* categorisation), the other concerned with lack of feedback (perhaps also a global problem here).

Study 1, problem 3.
Because this version of the program only supported one open document at a time, it was not possible to open a new document before the previous document had been closed. Closing a document is a very indirect action if one has the goal of opening a document, so many users tried the open command first. Upon seeing it was grayed out in the menu, they often realized that it was disabled for some reason ... [Goes on to suggest a remedy involving a dialogue box].

Given possible conflation of effect(s) with inferred cause(s), and a need for more precise quantification ("many", "often") this may be acceptable as a single problem.

Study 1, problem 4
Due to yet another underlying functionality limitation, the word processor could not support undo for the global replacement operation. The interface correctly informed users about the limitation but did so using such scary language that several users did not dare use this feature at all.

Two problems may be detected, the first concerning the lack of global replacement undo, the second the style of error message (also likely to occur elsewhere).

Study 1, problem 5
The "two-cursor problem" [citation] of having one cursor indicating the text insertion point and another indicating where the mouse is pointing. Some users assumed that they would be typing at the mouse pointer and were confused when the text appeared elsewhere on the screen. One subject tried to move the insertion point to a new location by direct manipulation (clicking on it with the mouse pointer and trying to drag it to a new location).

Two problems may be detected. First, some users confused the cursor position and insertion points. Second, one user tried to use 'drag and drop' to shift the insertion point. In the former, cause (two moveable indicators) and effect (typing at the cursor) are closely connected; in the latter, a much more general cause (drag and drop use is hard to infer) may be at work.

The remaining 4 problems in this first list may be acceptable as single problems, making an increase of 3 on the first study. From the second list of 14 problems, between 4 and 8 additional problems might be added in similar fashion.

The most conservative of these alternative interpretations shows that it might be possible to add an additional 7 (3+4) problems to the original total of 23, a combined increase of 30.4%. Using eq (3), this can be shown to represent a decrease in p to 0.77 of its original value. Using either choice of original values (estimated=0.282, actual=0.29), the new value of p approximates to 0.22. Plotting eq (1) with this value increases the number of subjects required to find 75% of the combined problems from the two studies from 4 (p=0.29) to between 5 and 6 (p=0.22). Thus the result of increasing the problem count by only this 30% would be to increase the number of required evaluators by up to 50%. (This disproportionately large increase is a consequence of the relatively steep gradient of the cumulative curves at the lower p values.)

7.2 Wright & Monk (1991): A Cost-effective Evaluation Method for Use by Designers

In two studies of the use of a (then) new method dubbed "cooperative evaluation", Wright & Monk showed that software designers could perform effective evaluations of early prototypes. The method (described in full in Monk et al. 1993) involves having users work through set tasks, explaining to the designer what they are doing and asking questions. The designer allows the user to make mistakes, and makes use of the user's questions to elicit further information about a potential usability problem. Unexpected behaviour and user comments are viewed as symptoms of usability problems (Monk et al. 1993 p133).

In each of the two studies reported, trainees (postgraduate software engineering students) without previous human factors experience performed cooperative evaluations of single subjects' use of a real or prototype system. The evaluations took the form of a "behavioural record used in conjunction with think-aloud protocols" (Wright & Monk 1991 p891). As mentioned above, the first author of the paper provided access to some of the original data from the first study, allowing amplification of the sample user scenarios contained in the paper. This Section will therefore concentrate on the first study.

The first study involved 27 trainee evaluators in 13 teams (12 of two and one of three). Each team evaluated a single (different) user's 2-hour interactions with the same system, in this case a bibliographic database called REF (first described in Wright & Monk 1989). Teams were randomly assigned to perform either a cooperative evaluation or a conventional think-aloud test. However, this manipulation was not successful (some members of the think-aloud teams were "freely interacting" with their user), so the distinction was not used in the results. Each trainee wrote a report independently of other trainees, and all individually reported problems were listed in the form of scenarios. For the purposes of analysis, the final UPT set used was the union of the sets of problems identified by individual trainees. This yielded a total of 29 different problems, ten of which were later classified as serious by the paper authors. On average, each team identified 9.6 (33%) of the 29 total problems and 4.5 (45%) of the 10 serious problems.

Wright & Monk also compared the teams' performance with those previously identified by the same authors in a series of five 2-hour studies of another single user's interactions with the same system. These 5 sessions (reported in Wright & Monk 1989) had involved a variety of recording methods and task prescriptions, including cooperative evaluation (session 1) and re-enactment or retrospective protocol of the critical incidents from a previous session (session 5). This had yielded 40 problems, 29 of which were found by the trainees in the 1991 study. Of the remaining 11 problems, only 4 were considered to have been missed by the trainees (the other 7 related to aspects of the system not evaluated). Thus the trainees had between them found around 90% of the problems previously identified by the two specialists.

In support of their claim (Wright & Monk 1991 p910) that cooperative evaluation is a cost-effective method on a par with Nielsen's "discount" approach, the authors employed yet another version of the theoretical model bound up in eq (1). As they say, a problem with a probability of p=0.3 of being detected in one attempt has a very good chance (p=0.75) of being detected in four attempts (Wright & Monk 1991 p903). Thus detection rates of 33% for all problems and 45% for serious problems appear to be good candidates for a '3 to 5'. However, two issues may be identified. First, like Nielsen's (1994b) study, the above detection rates refer to the total number of problems reported by all trainees, rather than the 33 (29+4) which were available for identification. Second, also like the Nielsen study, the arbiters of the final set of UPTs found by all evaluators were the paper authors rather than the evaluators themselves.

On the first issue, an average of 9.6 problems per team represents 29.1% of the 33 available problems, still well within the 25% to 33% but not a per evaluator figure (this would be 14%). (The equivalent figure for serious problems would be around 22%.) As to the second, the evaluator effect (Jacobsen et al. 1998a, 1998b) was probably reduced by having joint teams do the problem identifications, but without the trainee reports this is impossible to verify.

Though this study is unusual in listing problems detected and missed (by eight or more groups in each case), the sample scenarios are insufficient to enable the sort of detailed inspection performed on the Nielsen (1994b) paper reported in the previous Section. However, the additional data enabled further analysis. The data took two major forms: a transcript of the protocol from the original cooperative evaluation session (session 1) from Wright & Monk (1989), plus the 29 problem descriptions (including summary listings) extracted from the Wright & Monk (1991) trainee reports. Unfortunately, the full 40 problems against which the 29 scenarios were compared could not be found (nor could the trainee reports).

The goal of the thesis author's analysis was to see if different or additional problems could be extracted and/or reduced from either the original protocol or the problem lists. In order to avoid biasing the latter possibility, the author was careful not to look at the 29 extracted problems before inspecting the protocol. The result of the 'first cut' of the protocol analysis was a list of 41 problems, produced without having seen the system (but with the aid of sample screens from Monk et al. 1993). The author then reduced this list down to a set of 35 grouped and prioritised problems, this time with the aid of the trainee set of 29 (it proved impossible to proceed otherwise).

Since 35 and 4 is 39, it is tempting to believe that the sum of the thesis author's problems and the 'missing' 4 not reported by the trainees represents a near-match against the original 40. If that is so, the author's reduction did not exceed the total identified by Wright & Monk. However, comparing the author's 35 with the trainee 29 revealed only 20 matches, 15 to be accounted for, yet there were a maximum of 11 (40-29) problems remaining. *If* this analysis is correct, an additional 4 problems had been found by the author which were not present in the original 40. Unfortunately, at such a distance from the original experiments it is not possible to confirm or refute this speculation.

The effect of this incomplete analysis has been to suggest that the detection rates used to support the claim for the cost-effectiveness of the cooperative evaluation method *appear* to be lower by some 4% than when calculated from the available totals rather than those of the evaluator populations. An alternative problem reduction, from a single subject protocol, *may* also have identified an additional 10% over and above the total identified by these authors. (The previous Section showed that such apparently small differences may have a disproportionate effect on estimates of cumulative performance.)

8. Summary of Results

1. An underlying model was used by Nielsen (several sources) and Nielsen & Landauer (1993) to support the claim that between 3 and 5 experienced evaluators could identify 75% of the problems in a heuristic evaluation. The same model was also used to support a series of cost/benefit projections based on typical software products.

2. This model was shown to depend on certain limiting assumptions based on probability theory. In such cases the behaviour of cumulative problem curves could be predicted from the probability of any one evaluator finding one problem alone. This was then the same as the theoretical detection rate for a given sample.

3. The cumulative problem curves from both Experiments 1 and 2 exhibited patterns different from those on which the Nielsen claim was based. In Experiment 1 the numbers of experienced and novice subjects required to uncover 75% of problems were greater than the respective 5 or 14 claimed. This was also so for high-severity problems. In Experiment 2 the number of novices required to find the same proportion of problems was nearer 14.

4. The start points of these cumulative curves were shown to correspond to lower detection rates than the 33% predicted by Nielsen (or the lower figure of 25% required by the model). The shapes of the curves also deviated from this ideal (theoretical) form in ways to be discussed later in the thesis.

5. A model of the problem reduction process, necessary to combine separate problem accounts into a single set UPTs, was introduced and illustrated with reference to the problem report of a subject from Experiment 2.

6. The results of an inter-rater reliability assessment showed good correlation between the experimenter (the author) and each of two independent raters (HCI researchers) asked to perform a problem reduction on random samples of problems from Experiment 1.

7. It was shown that in Experiments 1 and 2 cumulative curves of the form predicted by Nielsen and others could indeed be generated by assigning problems to categories derived from the principles set. This led to the speculation that other researchers may have allowed implicit categorisations to influence the early stages of the problem reduction process.

8. Attempt was made to show that the problem descriptions in two published papers might admit of different interpretations. In the first case, Nielsen (1994b), the effect was to raise problem counts sufficiently to increase the number of think-aloud test subjects required to find 75% of problems by up to 50%. The second (Wright & Monk 1991) might have admitted of a speculative and unconfirmed increase of 10% on the problem total.

9. Discussion

9.1 Model and Theory

Many of the conclusions in this Chapter have rested on the differences between actual detection rates and the 33% required by the '3 to 5' claim (that only 5 evaluators can between them identify 75% of problems). It has been shown that lowering detection rates beyond this point will start to push up combined evaluator numbers to a level sufficient to refute the claim. The requirement for the claim is, then, that analyses of problem totals should reveal detection rates of 33% or higher (at least 25% for the ideal model).

However, Nielsen himself has not always kept to this requirement. In Nielsen & Landauer (1993), the same analysis used to produce the cost/benefit ratios uses in this paper also generated estimates of the optimal numbers of heuristic evaluators and testers for a typical project which, at 16 and 15 respectively, were "[...] much larger than obtained for our earlier 'discount usability engineering' recommendation of using about five heuristic evaluators or test users" (Nielsen & Landauer 1993 p212). Even given the avowed variation in p (or λ) values, this increase in expected evaluator numbers is well beyond Nielsen's other projections.

Another shift away from the claim occurs in the attempted replication by Lewis (1994) of Virzi (1990, 1992). While supporting the general finding that additional evaluators find fewer new problems, Lewis (1994) found no correlation between severity, or "problem impact" (as rated by observers of user tests), and rate of discovery. This refutes Virzi's finding (and that of Experiments 1 and 2) that severe problems are discovered by a higher proportion of evaluators. Lewis's view of Virzi's studies was that the latter may not have fully distinguished between problem frequency and severity, thus proposing that the two be treated as independent until proven otherwise. (In such a case the strategy of prioritising on frequency x severity would seem to be a good tactic.) The issue of severity assignment will be taken up in Chapter 8.

While the theoretical model holds for cases where problem identification is random (i.e. each problem has the same chance of being detected), it will begin to break down when there is a preponderance of evaluators agreeing on small clusters of problems. (This is just *one* of the limiting conditions on which the model depends.) We saw that such a deviation from the

ideal occurred in Experiments 1 and 2. But the tendency to disagree about what constitutes a usability issue (let alone any differences on severity) is the very phenomenon which is addressed by the declared need for multiple evaluators; once again, the question is not that evaluators will identify different problems, but to what degree. Evaluator agreement (or disagreement) has only recently been isolated as a measurable phenomenon by Jacobsen et al. (1998a, 1998b). This issue will be returned to in Chapter 8.

9.2 Problem Reduction

The three-stage problem reduction model introduced in Section 5.1 represents a beginning of the process of understanding what even a single evaluator must go through in order to incorporate multiple user reports into a single problem total. Elaborations of the model will be required when multiple evaluators (of more than one user) are introduced, and when (as in Nielsen 1994b and Wright & Monk 1991) 'second level' evaluations (of multiple evaluator reports) are added. In the ideal case, inter-evaluator reliability would be tested at each and every stage of the process, including initial protocol analysis and even the final problem grouping. Clearly, this would be both practically and financially unfeasible, even for experimental purposes, so it will be important to focus resources on the most important stage(s) of the process. The inter-rater reliability assessment in Section 6.1 represents one attempt to do this.

The proposition that problem grouping may come to dominate the whole process rests on the view that categorisation is a natural consequence of the necessary matching which must be undergone in the 'between subjects' stage. This is doubly so because it is the over-riding goal of the usability assessment business to make pragmatic recommendations concerning interface design. Indeed, the skill of the usability engineer is to be able to identify the inter-relationships between different issues and the ways in which recommendations that address one issue (or set of issues) may affect others. However, it is the author's view that such groupings should be left until as late as possible in the process, in order to avoid the conflation of causes and consequences of which the Nielsen (1994b) paper may be an example. As we have seen, it is only too easy to run together types and instances, thus unwittingly pushing up the cumulative curves.

What, then, would be the author's 'stopping point', beyond which it is not worth attempting to separate issues, and to start 'lumping together' ? The definite cases that can be offered are typographical and terminology instances, which are simply too trivial (and easy to fix) to be worth counting separately. The author's view is that categorisation beyond the lowest levels should wait until the latter stages of the process, when grounds for grouping and prioritisation can be established. This does not deny the value of setting out specific goals for an inspection, as much as for user testing, but the value of guideline review or heuristic evaluation is that the scope can be set much wider. The issue of evaluation task scope is addressed in Experiment 3 (Chapter 5).

9.3 Reliability

The inter-rater reliability figures presented in Section 6.1 are an attempt to demonstrate for Experiment 1 that the most important stage of the reduction process was not biased by the experimenter's view of what contribute separable usability issues. While some success was achieved, it should be pointed out that this was the only stage for which reliability assessment was attempted, leaving out of account the start and end of the reduction process. As to the latter, problem types were used in order to demonstrate the effect of categorisation on detection rates, and in the additional analyses performed for Experiments 1 and 2. However, at the other (start) end of the process, no video or audio records of the subjects or user-system interactions, from which to take independent raters' accounts, were made in any of Experiments 1 to 3. We have seen that in Experiment 2 subjects entered their own written accounts into a spreadsheet, and that the problem distributions for Experiments 1 and 2 were remarkably similar (though based on different problem totals). Therefore the author is fairly confident in asserting that the lack of video and/or audio would not make much difference to the results, though in hindsight it would be preferable to have recorded at least a sample of user sessions. (This issue will be returned to in Chapter 8.)

Having obtained a subject record, it still remains to be transcribed, even before any problem extraction is attempted. This too is a possible source of experimenter bias, if far less so than the recording process. One advantage of not using video or audio is that this long and tedious business is avoided (not that this is sufficient excuse).

As to the inter-rater figures obtained, they do show good evaluator-rater correlation in the degree of problem reduction carried out on samples from Experiment 1. However, the object of these matchings were the short (up to 35 words) problem descriptions which had themselves been created by the experimenter. These too are subject to any biases in the way in which single-issue problems are encapsulated. Thus another stage of validation would be to have different evaluators codify agreed problems in different ways.

Summary of Chapter 4

Some evidence has been presented that the problem distribution and sharing profiles exhibited by Experiments 1 and 2 were not a result of a particular perspective by the author on what qualifies as a separate usability problem. The pattern underlying both sets of results may represent a more general feature of problem extraction, identification and reduction than is suggested by the low detection rates and late-converging cumulative curves from these two experiments. It is speculated that the discrepancy between these results and those of other researchers can be attributed to an unwitting tendency to group together separate usability issues at an early stage in the problem reduction process.

> "It appears, then, that the identification of usability problems by inspection may be a more lengthy and expensive process than has been portrayed. The proposition arising from the current studies is that up to twice as many experienced evaluators than has been predicted may be necessary to uncover the same proportion of problems with an interface as has been claimed." (Connell & Hammond 1999 p628).

In the next Chapter we shall continue to test this proposition against the results from a very different evaluation study to those in the first two experiments. Experiment 3 uses a range of closely prescribed user tasks in order to investigate the effect of task scope on detection rate. It uses further cumulative curves to assess predictive performance, while making the important distinction between predicted and observed problems.