Int. J. Human–Computer Studies (1998) **49,** 717–742
Article No. hc980224

Modelling self-confidence in users of a computer-based system showing unrepresentative design

P. Briggs, B. Burford and C. Dracup
Division of Psychology, University of Northumbna at Newcastle, Newcastle open Tyne, NE1 8ST, UK

(Received 14 May 1997 and accepted in revised form 1 July 1998)

While a great deal of research has demonstrated that users' self-efficacy beliefs have a major impact upon both their attitudes to technology and their performance, a related construct, self-confidence, has been largely ignored within the domain of human–computer interaction. This is surprising given the vast literature on the calibration of confidence which can be found within the judgment and decision literature. In this study, 60 participants were asked to complete a novel computer-based task, and to provide measures of self-confidence in terms of their anticipated performance at each stage of the task. Users' confidence judgements showed sensitivity to their rate of improvement on the task, but were poorly calibrated with actual performance at each stage. Furthermore, confidence judgements were insensitive to the complexity of the individual task components, even though these different components led to very different levels of performance. The style of computer interface was also found to affect anticipated performance independently of actual performance. The ability of existing models of confidence judgement to deal with these data is discussed.

1. Introduction

The major focus for human-computer research throughout the 1980s was the specification of design factors which lead to improved performance—be it faster learning, fewer errors or increased use of system functionality. In this performance-driven culture, termed "cognitive Taylorism" by Frese (1987), scant regard was paid to the beliefs that people could hold about the computer systems they used, or about themselves as users of those systems.

We now know that such beliefs can have a strong influence upon a user's attitudes and behaviour and more attention has been paid to whether or not the user trusts the system (e.g. Muir, 1987; Lee & Moray, 1992), or whether or not they view themselves as competent users of the system (e.g. Compeau & Higgins, 1995). Of particular importance is the user's belief in his or her ability to successfully complete the task at hand. This belief has been shown to predict the adoption of advanced technology (Hill, Smith & Mann, 1987), and also the extent to which users will engage in active system exploration (e.g. Zuboff, 1988). It has most often been labelled "self-efficacy", but we argue below that a distinct, but closely related construct, self-confidence, is also useful in understanding human–computer interactions.

1071-5819/98/110717 + 26 $30.00/0

Self-efficacy is best construted as "a belief in one's capability of performing a specific task" (Bandura, 1986). It is derived from a variety of sources, including external task-relevant cues (e.g. task complexity), and internal, self-referent cues (e.g. task-relevant knowledge, arousal). Differences in self-efficacy beliefs can thus reflect bona fide differences in skill level, but they can also reflect differences in personality, motivation and differences in the task itself (Gist & Mitchell, 1992). Self-efficacy has been shown to have a major impact upon performance in a variety of domains, with self-efficacy ratings being positively correlated with job satisfaction, commitment and quality, and negatively correlated with absenteeism and poor timekeeping (McDonald & Siegal, 1992). From the outset, self-efficacy was seen as having a motivational component, in that people will be more prepared to take on tasks that they feel they are able to perform, and this has been a major focus for self-efficacy work within the domain of human–computer interaction.

Within HCl, self-efficacy has typically been defined in terms of "people's expectations of being able to use computers", and is often assessed by asking users to rate themselves against statements such as "I will never understand how to use a computer" (Hill, *et al.*, 1987), or this, taken from Compeau and Higgins (1995) computer self-efficacy scale: "I could complete the job using the software package ⋯ if I could call someone for help if I got stuck". Such scales give either a general assessment of perceived computer ability, or a specific assessment of likelihood of success in a particular task: however, they generally encompass a rather broad prediction of overall performance (see Gist & Mitchell, 1992). Nonetheless, such measures have proved useful. For example, Hill *et al.* found that the likelihood of using computers was in part determined by such efficacy beliefs. Furthermore, these beliefs were found to operate independently of previous experience with computers, and also independently of the perceived instrumental value of learning to use computers. Concluding their study, Gist and Mitchell (1992) state that:

> "experience *per se* does not directly affect subsequent behaviour regarding further adoption of computer technology; rather, only through changes in perceived efficacy does experience with computer technology lead to a higher likelihood of technology adoption."

Other studies support and extend such observations. For example, Burkhardt and Brass (1990) found that high self-efficacy (as measured by statements such as "I have the capability to effectively use computers in my job") was related to early adoption of new technology, and further, that such "early adopters" had what was effectively "*a recipe for increased network centrality and power*". Gist, Schwoerer and Rosen (1989), also showed the importance of self-efficacy beliefs in training. In a field experiment involving University Managers, they found that modelling approaches to training in new technology were more successful than tutorial approaches because of the way in which modelling influenced self-efficacy beliefs, which in turn effected performance. They concluded that watching a model perform specific computer software operations could enhance participants' beliefs about their own ability to use the system. These and other studies (e.g. Bar Tal, 1990; Carlson & Grabowski, 1992; Mitchell, Hopper, Daniels, George-Falvey & James, 1994) lend general support for the view that self-efficacy is an important component in human–computer interaction.

One limiting aspect of this research, however, already indicated, is that self-efficacy is most commonly measured using general statements of perceived ability, which allows comparison of one individual with another, but which means that (1) the measure is relatively insensitive to short-term changes in one's beliefs about the extent to which one feels in control of any system, and (2) the self-efficacy measure cannot be properly calibrated with actual performance. In other words, ratings of self-efficacy based upon statements of the "I can use computers effectively" type can be *correlated* with performance—but not calibrated in a way which tells us whether or not people's beliefs about their abilities are accurate. In order to achieve this, we would need some measure of efficacy which is defined in terms of anticipated or expected performance on a particular task, rather than in general reference to one's own ability. However, we cannot justify defining *self-efficacy* in this restricted fashion, since most researchers make it clear that self-efficacy, which encompasses a very broad range of predictors of performance, and which includes a number of self-referent cues (such as personality, motivation, arousal), must be clearly differentiated from any notion of *expectancy* (e.g. Gist & Mitchell, 1992).

Happily, there is a distinct, albeit related literature which deals with expectancy or anticipated performance. This is the literature concerning *confidence*.†

This literature encompases a large number of studies which are based upon the premise that confidence can be defined in terms of anticipated performance, and that this can be measured and calibrated against actual performance. In the main, this literature has shown that people are very often overconfident in their abilities, anticipating performance levels which greatly exceed their actual achievements. This has been demonstrated comprehensively in general knowledge and forecasting tasks (e.g. Fischhoff, Slovic & Lichtenstein, 1977; Fischhoff & MacGregor, 1982), and has been shown to apply to a range of professional judgements, such as those found in engineering (Kidd, 1970) and law (Wagenaar & Keren, 1986). In general, confidence has been shown to be dependent upon feedback from past performance (e.g. Einhorn & Hogarth, 1978; Te'eni, 1990), although overconfidence can be maintained in the presence of outcome feedback (Subbotin, 1996). However, confidence is also determined by the way in which a particular judgement is framed (Gigerenzer, Hoffrage & Kleinbölting, 1991), of which more later.

One issue of great interest here, concerns how confidence judgements change in the process of human–computer interaction, and how these changes affect performance. It is likely, for example, that in the course of learning to use a computer-based system, or even in the course of completing a specific computer-based task, that minor difficulties, or small unexpected events would affect confidence which in turn may affect subsequent performance and motivation. Although few studies have followed changes in confidence over time, an intriguing set of experiments by Lee and Moray (1992, 1994) demonstrated, using a time-series analysis, that fluctuations in users' confidence and trust can jointly

†Note that there are inconsistencies in both the self-efficacy and the confidence literatures. Specificaly, there are examples of self-efficacy defined in terms of anticipated performance, and examples of confidence defined by general statements of efficacy. However, to date the work on *computer efficacy* has made no attempt to calibrate anticipated performance against actual performance, but has rather relied upon rating scales of the "I am comfortable using a computer" type. We believe that this paper presents the first attempt to utilize calibration methods within the HCl domain.

determine the extent to which a systems operator will switch between manual and automatic modes of process control.

In the 1992 study, operators were asked to control a simulated orange juice pasteurization plant, so as to balance the competing goals of productivity and safety. They were free to switch between two modes of control—automatic and manual—for any of the three-component processes in the plant, and the system was set up so that a fault could be introduced into one of the component processes. The introduction of a fault was generally associated with poorer performance—although in many cases recovery was swift. It was also associated with decreased trust, which proved harder to re-cover from. Surprisingly, operators would often respond to this reduced trust, by placing *more* reliance upon the system–switching from manual to automatic control. Lee and Moray suggested that this move away from manual control might be caused by a loss of the operator's confidence in their own control abilities, and set out to investigate this relationship further in their 1994 paper.

In their later study, Lee and Moray made a number of observations concerning the relationship between trust, self-confidence, and the allocation of function. Firstly, during fault-free operation of the orange juice plant, operators tended to rely on manual control. During this phase, trust declined and confidence rose steadily, as performance improved. A fault was then introduced to one of the sub-components, which primarily affected either manual or automatic control. When the fault affected manual control, confidence declined while trust increased, and this led to a switch in the allocation of function with nearly complete automatic control. Conversely, when the fault affected the automatic controller, trust in the system gradually decreased, as confidence increased. Lee and Moray were able to model this process, creating a time-series model in which the use of automatic controllers depended upon the difference between trust and confidence, but also upon past use of the automatic controller—indicating a certain reluctance of the operator to change strategy.

These findings mirror some of the observations found in the decision-making literature. For example, the early reliance on manual control, accompained by high confidence judgments may be related to the overconfidence effects reported earlier, while the reluctance of the operator to change strategy, even though the accumulated evidence indicated change, could be construed as another form of confirmation bias (Wason, 1960).

Lee and Moray have shown that it is possible to model trust, and have also shown that it is possible to use measures of trust and confidence in order to model allocation of function. However, it should be noted that their measures of confidence, despite being defined as "anticipated performance during manual control", utilized a general rating scale which allowed comparison of one user with another, but which prohibited calibration of anticipated and actual performance.

It should certainly be possible to provide a more precise model of how and why confidence judgements change over the time course of any human–computer interaction, and this is the focus of our study. Confidence—as anticipated performance—is a judgement not only of the difficulty of the task at hand, but also of the extent to which one's experience and ability matches up to that task. Such judgments are difficult to model, although a number of useful attempts have been made. One which seems particularly relevant to issues raised in the human-computer domain—because it deals explicitly with the issue of task design, is the ecological approach (e.g. Gigerenzer, *et al*., 1991, Juslin, 1994, 1995, Gigerenzer & Goldstein, 1996).

Ecological approaches assume that an individual's internal model of the world faithfully reflects that individual's real-world experience. In this view, overconfidence is a function, not of any cognitive bias, but of a *sampling* bias, which can be naturally occurring, but which can also occur as a function of *unrepresentative design*—which is an interesting issue for HCl. To understand the ecological viewpoint, consider one of the best known studies—that of Gigerenzer. Hoffrage and Kleinbölting (1991), who based their model of self-confidence upon the processes in operation when answering two-choice general knowledge questions of the form: "Which city has the largest population: Bonn or Heidelberg?".

Gigerenzer *et al.* argue that people initially try to solve this problem by a simple act of recall. Should they be able to retrieve the information, they will be completely confident in their answer, but in most cases people do not "know" the answers to such questions, and so they have to use a process of inference. In Gigerenzer's terms, they construct a probabilistic mental model (PMM) which relies upon a reference class of objects and a network of inferential cues. For example, in the case of Bonn and Heidelberg, respondents may use the reference class 'cities in Germany' and the cue 'does the soccer team play in the Bundesliga?' as an indicator of city population. If this cue is a good cue (i.e. if it has high *cue validity* in the sense proposed by Rosch, 1978), then people will tend to answer correctly. Furthermore, if this cue is perceived to be a good cue (i.e. has high *perceived validity*) then people will be confident in their answers. Of course, it may be that the cue has low actual validity, but high perceived validity—in which case people will show overconfidence in their responding. Supporters of the ecological approach believe that, in most contexts, appropriate reference classes are selected, and that perceived cue validities are well calibrated with actual cue validities. But there are may occasions where a task or problem is designed so as to be unrepresentative of the real world. In the example above, it is usually the case that the best soccer teams come from larger cities ⋯ but it would be easy to design a quiz question for which this cue would be misleading. Gigerenzer *et al*., argue that this type of unrepresentative design is responsible for the typical finding that people are overconfident in their answers to individual questions.

In marked contrast, people are usually accurate in judging their performance over a quiz as a whole. In other words, people show accurate confidence judgements when asked to evaluate their performance over a run of events. Why should this be? According to Gigerenzer *et al*., the probabilistic models which are constructed are likely to involve reference classes of previous quizzes, and to utilize cues concerning frequencies of past successes. In other words, when asked to make a judgment about performance over a range of items, people will base their response on past success rates in similar task. If we assume that this quiz is typical of other quizzes in containing "tricky" questions, then confidence judgments of overall performance are likely to be accurate. Of course, in the circumstance that trick questions are not introduced, and individual questions genuinely are representative of the underlying knowledge domain, then single-event judgements would be appropriate, and frequency judgments would show *underconfidence*. Overall, however, Gigerenzer *et al.* believe that frequency judgements (i.e. judgements based on a run of events) tend to be better calibrated since the cues in the reference class activated by the probabilistic mental model are likely to have higher validity.

If we now turn to the specific issue of how to model confidence judgement in a HCl task, then a number of specific issues present themselves. The main issue concerns the fact that the ecological models described above have been developed to account for confidence judgement in declarative, fact-based tasks—there are as yet, no models of confidence judgment in the acquisition and execution of procedures, and yet procedural learning is a key aspect of human–computer interaction.

Consequently, the emphasis which ecological models place upon the overconfidence which results from single-event judgments may be inappropriate, since these judgments are based upon the kind of declarative knowledge which is only really important in the very early stages of procedural learning (Anderson, 1983). Put simply, most procedural skills are acquired through practice, which involves repetition of events, and so these events are more usefully viewed collectively, rather than individually. It is likely, then, that confidence in the human–computer context is more appropriately modelled in terms of frequency judgements (success over a run of events) rather than single event judgments. If this is the case, then confidence judgements in the human–computer domain should *not* show the same pattern of overconfidence so commonly reported in the decision making literature, since (1) frequency judgements are generally well-calibrated† and (2) feedback, in terms of the success or failure of each initiated procedure, is usually immediate. This is in complete contrast to the experimental quiz question paradigm described earlier, and in contrast to many professional judgments, where feedback is usually delayed, and is not always available for each individual decision.

Thus, for reasonably experienced users, we would predict that initial confidence levels would be based on a probabilistic mental model which contains a reference class of previous computer-based tasks, plus any real-world tasks which are cued by the system itself. The cue validity for this reference group is a function of the extent to which the current or target task is representative of the other tasks encountered. If the target task shows representative design (i.e. is fairly representative of computer tasks in general), then the user's confidence judgments will be well-calibrated (assuming their individual experience is fairly typical), but if the target task is particularly difficult, then overconfidence will result. Of course this will change over time, as the user incorporates experience of the target task into their PMM, which effectively means the introduction of a new, highly valid cue, which in turn will lead to better calibrated confidence judgments.

This is the prognosis tested in the current study. Experienced users are given a novel task to complete, which involves building up a 4 × 4 grid of simple objects which matches a given template. They do thus via a process of object manipulation which incorporates foreground and background selection, object selection, object-colouring and object-rotation. The task was of *unrepresentative design*—i.e. set up differently from most systems in that the interface was particularly inconsistent. This was achieved by making the syntax for each of the four objects within the grid unique. Note that although this degree of inconsistency is unusual, it is by no means rare to find that similar interface objects have different control specifications. To take an example from the Apple Mac desktop, most documents may be opened by clicking on the document icon, but some documents (e.g. those imported from another application package) may only be opened

†Note that a few studies have reported underconfidence in frequency judgments, and these are briefly reported in the discussion section of this paper.

within the word processing application itself—even though the icons for these different documents might look very similar.

By designing the task in such a way that it is unrepresentative of other computer-based tasks, we are creating a situation in which the PMM initially activated by the end user should have very low cue validity, and should lead to overconfidence. In other words, we are engineering a situation in which the initial reference class invoked by the user will be inappropriate, whatever the initial level of experience of the end user. This is useful if we wish to gain some insight into the ways in which users adjust their PMMs and therefore their confidence judgments, throughout a task. As the user gains experience with the target task, so the user's performance predictions should gradually become well-calibrated with actual performance. In the next phase of the study, users are asked to move on to a second version of the same task, involving a new set of four objects, each with a new control syntax. We predict that users base their confidence judgments for this transfer task upon a PMM which includes performance on the first task—thereby avoiding the initial overconfidence predicted for the first task. Note that performance is predicted to drop off on transfer, but this drop in performance should be accompanied by more conservative estimates of success. The specific predictions for both the target and the transfer tasks are represented graphically in Figure 1.

It is expected that these predicted changes over time would take place for any inconsistent system, irrespective of type of interface, and so three different interfaces to the same underlying system are presented in the current study, in order to check on the robustness of the predicted effects. However, as we have seen, the ecological view places considerable emphasis upon the issue of task design, and so interface design should

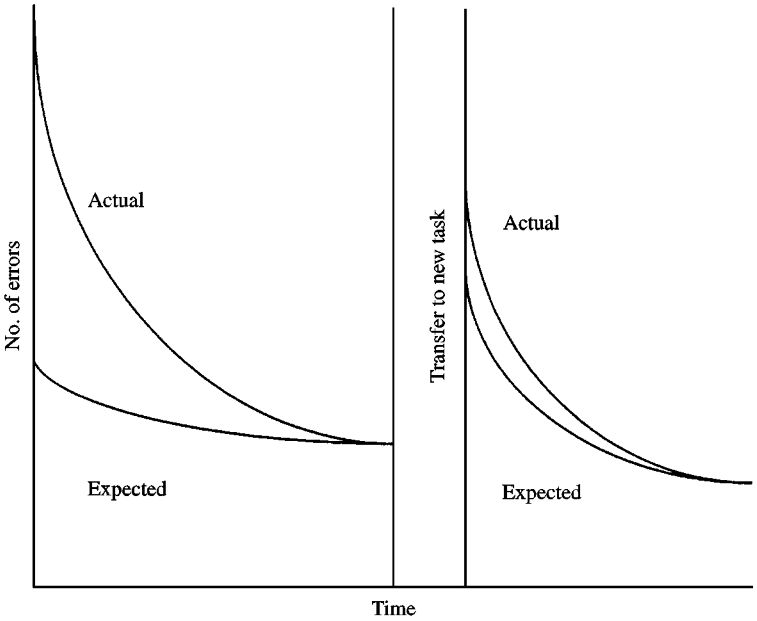

FIGURE 1. Predicted changes in expected and actual performance over time.

actually have some independent effect upon underlying confidence levels. We have argued that for computer-based tasks, confidence will be based largely upon a combination of past experience with other computer systems, and upon additional reference groups deliberately cued by the interface itself. Given that the task we have created is unfamiliar, the associations triggered by the interface will be limited, but nonetheless, there should be differences between the reference groups generated by relatively meaningful forms (menus, icons) and unmeaningful forms (keys) of input. There may also be individual differences in past successes with, say, icon versus keyed input systems, both of which may influence initial confidence judgments [see Te'eni (1990) for an example of the ways in which direct manipulation interfaces can generate increased confidence]. Taken together, these issues imply that while the overall pattern predicted in Figure 1 is expected to be robust across interfaces, there is an important additional hypothesis. This is that the more "meaningful" command-labels present in icon and menu systems will generate higher absolute confidence levels, particularly during the initial stages of the task.

2. Method

2.1. DESIGN

The experimental design was basically a 3 × 2 mixed factorial, with the levels of the between-subjects factor, interface, being icon, menu and keyboard input, and those of the within-subjects factor being the first, and second grids (target task and transfer task). A third factor was created from the two dependent measures of error taken-predicted error and actual error, which allowed for analysis in a 3 × 2 × 2 Anova. In addition, since a major factor of interest was the change in performance over time, the errors made per row of each grid, and per cell of each row were considered. In addition to participants predicting the number of errors they would make for each cell, they were also asked to make posterior estimates of the total number of errors made per grid. Actual performance was measured in terms of errors made per cell.

2.2. PARTICIPANTS

Sixty undergraduate and postgraduate volunteers were paid £3 to take part in the experiment, which lasted between 30 and 40 mins. All were recruited from the University of Northumbria at Newcastle, and all had some basic experience of word-processing and statistical software, although none could be considered "expert" users. Participants were randomly assigned to one of the three interface conditions (20 participants per condition).

2.3. THE TASK

The task was written in Supercard v1.5 from Silicon Beach Software and run on an Apple Macintosh IIci. The start display consisted of two 4 × 4 grids, the right-hand (target grid) empty and the left-hand (object grid) filled with symbols of different colour and orientation from one of two different symbol sets.

- Set 1 (Animals): cat, rabbit, hen, fish
- Set 2 (Transport): car, train bike, plane,

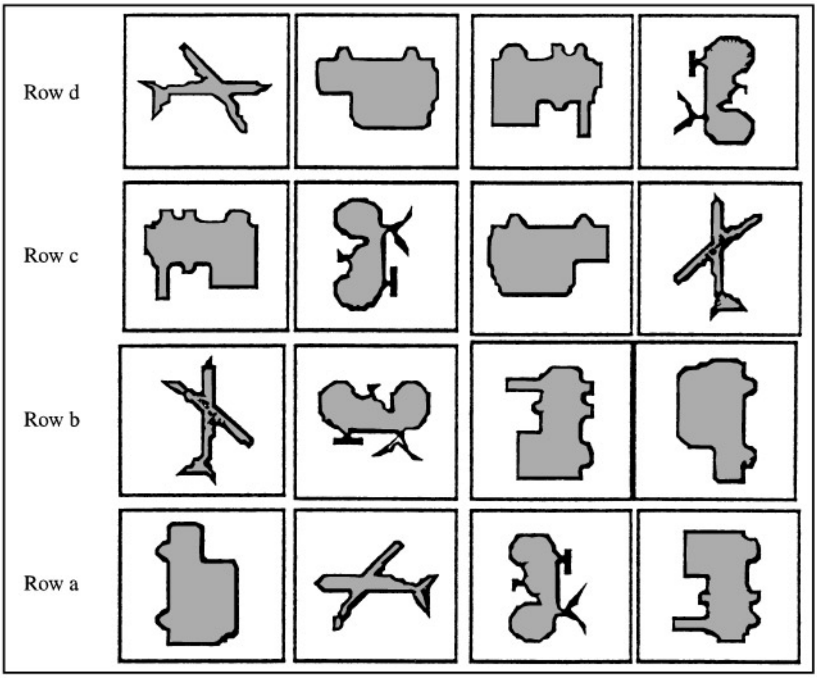

FIGURE 2. Example of a grid containing transport symbols.

An example of the grid with the Transport symbol set is given in Figure 2. Participants were asked to reproduce the completed grid by "filling in" the spaces in the empty grid with the various symbols, and manipulating and colouring them as appropriate. Each candidate was asked to complete two grids, one of each symbol, set. The order in which participants received these symbol sets was counterbalanced, with half of the participants starting with the animal set in grid 1, and transferring to transport in grid 2, and half starting with transport and transferring to animals.

Symbols and backgrounds could be one of four colours (red, blue, green, yellow for Animals, deep red, deep blue, orange, purple for Transport) providing a possible 12 combinations of colours. Eight orientations were also possible (up, down, left and right with mirror reversal either true or false), which meant that there were 96 combinations possible for each symbol. The grid was constructed to ensure that the same combination did not appear twice, and that each symbol appeared only once in each row.

The user constructed the target grid by issuing a series of commands for background and foreground selection, symbol selection, colour selection and orientation (rotate right, rotate left, flip vertically, flip horizontally). Participants were told that items would have to be constructed using a particular sequence of commands, but were not told the precise nature of those sequence. However, they were told that they would receive auditory feedback for both correct (fanfare) and incorrect ("boing") selections. Participants were allocated to one of three different interface conditions, and these are described below.

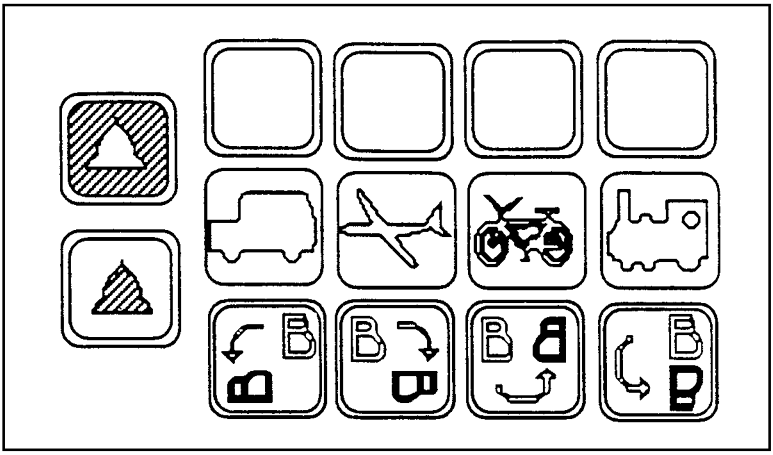

FIGURE 3. Control window as presented in the icon condition.

2.3.1. Icon

In this condition, the standard Apple menu bar was hidden and all functions were presented as buttons in a single window at the bottom of the screen (see Figure 3). The usual protocols were observed for these buttons: selection was by mouse click and was followed by the button being briefly highlighted. In the case of the mode buttons (foreground vs. background) the currently active mode was indicated by a darker border around that button and was therefore always visible to the participant.

2.3.2. Menu

In this condition, participants were presented with four menus at the top of the screen: Area (containing the foreground vs. background mode controls). Symbol, Colour and Orientation. Menu items were selected by a standard click and drag procedure, with the currently active mode indicated by a tick against that item in the Area menu.

2.3.3. Keyboard

In this condition the menu bar was hidden, and participants were given no on-screen guidance at all. However, a separate keyboard map was provided for reference (see Figure 4). A list of controls and their associated keys was also provided. Controls were given the same names as in the menu condition, and were spatially grouped. No continuous feedback was available as to the currently active mode.

2.3.4. Task constraints

As mentioned earlier, this task was particularly challenging in being highly inconsistent. Thus, the order of commands accepted as correct varied for each symbol (although it was consistent for that symbol). Any attempt to follow the wrong symbol syntax was

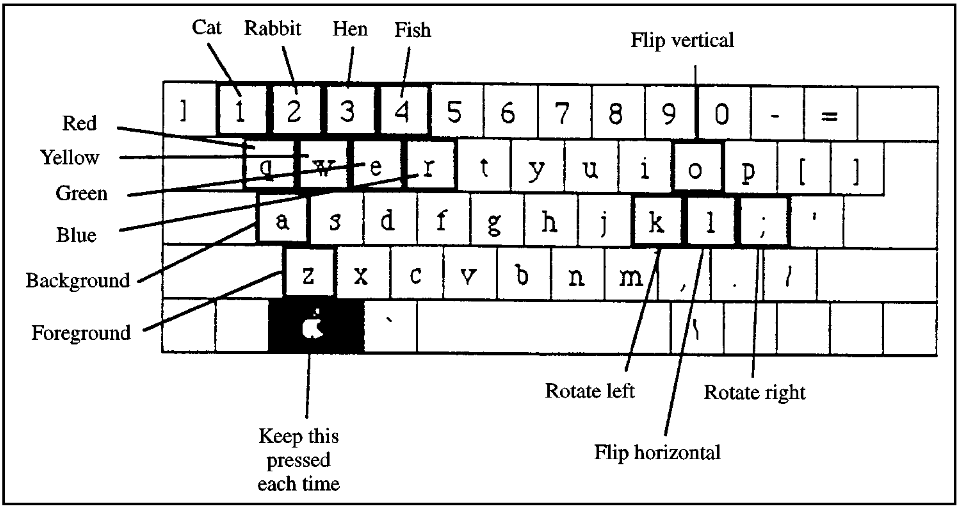

FIGURE 4. Keyboard map given to participants in the key condition.

signalled as an error. The eight legal syntaxes were as follows (note some variation in complexity, which will be discussed latter.

- cat: foreground, symbol, colour, background, colour, foreground, orientation.
- rabbit: foreground, symbol, orientation, background, colour, foreground, colour.
- hen: background, colour, foreground, symbol, colour, orientation.
- fish: background, colour, foreground, symbol, orientation, colour.
- car: foreground, symbol, background, colour, foreground, orientation, colour.
- plane: foreground, symbol, background, colour, foreground, colour, orientation.
- bike: foreground, symbol, colour, orientation, background, colour.
- train: foreground, symbol, orientation, colour, background, colour.

A couple of pragmatic rules applied for all symbols: (1) the correct mode (foreground or background) had to be selected before any other function, and (2) the correct symbol had to be selected before any symbol manipulation. Participants were not made aware of these rules beforehand, and failure to follow these was also signalled as an error.

2.3.5. Data recording

All data were recorded automatically into user-specific logs by the computer: Time-stamps and errors were recorded for each action performed by the user. User confidence in terms of predicted performance was also recorded on-line. A pop-up window asked participants to select the number of errors they believed they would make on the next cell from a choice of 0, 1, 2, 3, 4, 5 or "more than 5" errors. Participants in all conditions used the mouse for this operation.

Posterior performance estimates were also recorded at the end of each grid via a similar pop-up window. In this case, though, participants were asked to use the keyboard to enter the number of *cells* on which they had made each number of errors.

These frequencies for 0, 1, 2, 3, 4, 5 and "more than 5" errors were also recorded in the participant's log.

2.4. PROCEDURE

On arrival participants were seated in front of the system, which displayed a launcher screen of three text buttons reading "practice", "First Grid" and "Second Grid". Participants were told that they would set two grids on the screen, the right one empty and the left one containing pictures, and that their task was to complete the empty grid to look identical to the one on the left. They then used the mouse to click on "Practice" which involved a brief, four item session to allow users to become familiar with the controls, but not with the actual symbols, or the restricted syntax of the main task. Examples of the response scales were included in the practice session, with the prediction scales appearing before the second, third and fourth items, and *a posterior* estimate window appearing at the end of the row. Participants were not asked to make realistic error estimates at this stage, but were simply asked to familiarize themselves with this mode of input.

On completion of the practice grid the participant was returned to the launcher screen. At this point the principle of the control syntax was explained to participants, who were told that an error would be committed each time a specific sequence was violated. Participants then clicked on "First Grid". The first prediction scale was presented immediately the grids were on the screen, and at this point the definition of an error was reiterated. Participants then worked through the grid from bottom left to top right, by completing the four cells in each row before moving on to the next row. At the end of the grid the posterior estimates window was presented, and it was ensured that participants understood exactly what was required. Participants were not, however, advised that their response should sum to 16 (a logical constraint).

Following the first grid, the participant was again returned to the launcher. It was explained that for the next grid the symbols and the sequences would change, but that the controls and all other aspects of the task would remain the same. The participant was asked to click on "Second Grid", and complete the task in the same way as the first.

3. Results

The following analyses are largely focussed upon the error predictions, which are taken as indices of user confidence, and the corresponding performance errors for each item. These data are examinable at a number of levels, by cell, by row and by symbol. However, retrospective error estimates are also described. The results are reported as follows: firstly, ignoring possible effects of interface, reported changes in confidence and actual changes in performance over time are tested against the predictions made earlier; secondly, the effects of type of interface are considered; and thirdly, an analysis by symbol is reported, which provides useful information concerning the way in which frequency data is used to modify confidence judgments.

3.1. CHANGES IN SELF-CONFIDENCE OVER TIME

The predicted pattern was shown earlier in Figure 1. Essentially, users were expected to show high overconfidence when taking the early portion of grid 1, and then improved calibration, with confidence levels approaching actual performance levels. Upon transfer to grid 2, the same early pattern of overconfidence was not expected to re-appear, since users should base their judgements upon a more appropriate reference group—i.e. both confidence and performance estimates should indicate poor initial performance on this grid. The data relevant to this prediction are shown in Figures 5 and 6.

Figure 5 shows the estimated and actual errors made for each row of grid 1—with the insert showing the early changes which occurred within the first row. A number of points stand out. Firstly, the overall pattern shows underconfidence, rather than overconfidence—participants generally think they will make more errors than they actually do, with a predicted mean of 4.0 errors per cell, and an actual mean of 3.6 errors per cell. Secondly, the early pattern is as predicted, with users showing overconfidence for the very first item (depicted in the inset). But it is surprising to note how quickly this changes to a pattern of underconfidence. This observation is statistically supported by a 3 (interface) $\times 2$ (type of error) $\times 4$ (cells) analysis of variance for these first four items (see the appendix) which reveals that the switch from overconfidence to underconfidence is meaningful, as indicated by the interaction between type of error (predicted vs. actual) and item [$F(3,171) = 29.4$, $p < 0.0001$], and is further supported by t-tests which show that actual performance is significantly worse than estimated performance for the first item in the row ($p < 0.0001$), but that actual performance is significantly better than estimated performance for the last item in that first row ($p < 0.0001$), as shown in the inset to Figure 5.

As predicted, this pattern does not reappear for grid 2. Figure 6 shows that users anticipated the drop in actual performance with a similar drop in predicted performance. The pattern for the whole grid is still underconfidence, with a predicted mean error per item of 3.8 and an actual mean error of 2.6. There is no sign whatsover of a final convergence between the performance estimates and the actual performance measures.

The overall pattern of underconfidence is statistically supported. Looking at performance across the two grids, a 3 (interface) $\times$ 2 (grid) $\times$ 2 (type of error) analysis of variance (see the appendix) revealed a significant difference between estimated and actual errors [$F(1,57) = 47.0$, $p < 0.0001$]. There was also a significant difference between the grids, with predicted and actual errors lower for grid 2, as expected ($F1,57) = 44.9$, $p < 0.0001$); plus an interaction between grid and type of error [$F(1,57) = 8.6$, $p < 0.01$] which was caused by the effect reported earlier of initial overconfidence in grid 1 (i.e. estimated errors initially lower than actual errors), which attenuated the overall pattern of underconfidence for that grid. These data are shown in Figure 7.

The pattern of overconfidence which emerged very swiftly during the first few items of grid 1, was also reflected in the retrospective performance estimates collected at the end of each grid. Median values were calculated for each subject (a measure of central tendency which allows the inclusion of the $5 +$ category), and the mean of these medians was found to be 2.94, as compared to a comparable mean of 2.89 for predicted estimates, and 1.93 for actual performance errors. Thus, users' retrospective beliefs about their performance reflect their prospective expectations of failure rather than their actual successes.

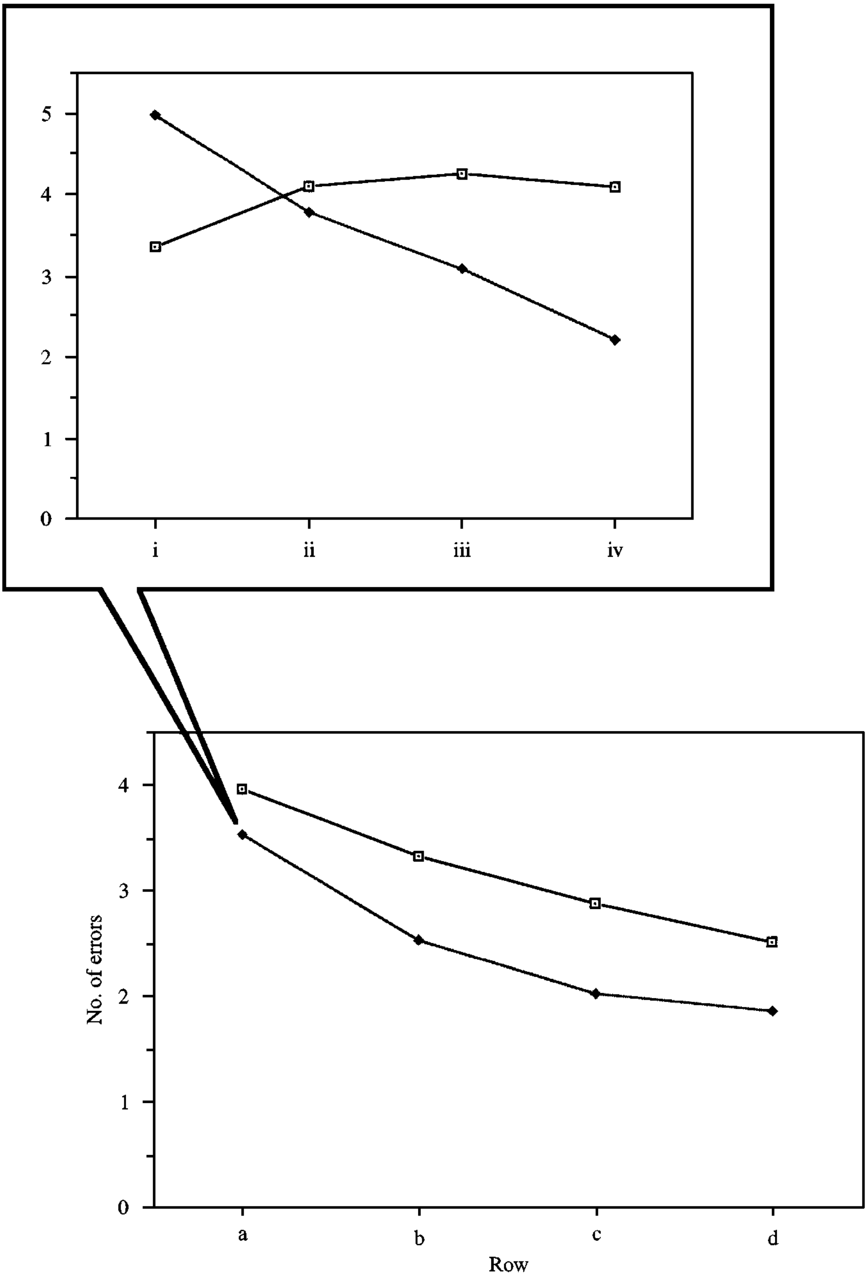

FIGURE 5. Changes in estimated and actual error over time for the four rows of grid 1. The inset details changes over the first four items of that grid (i.e. row *a*). —□— Estimate; —◆— Actual.

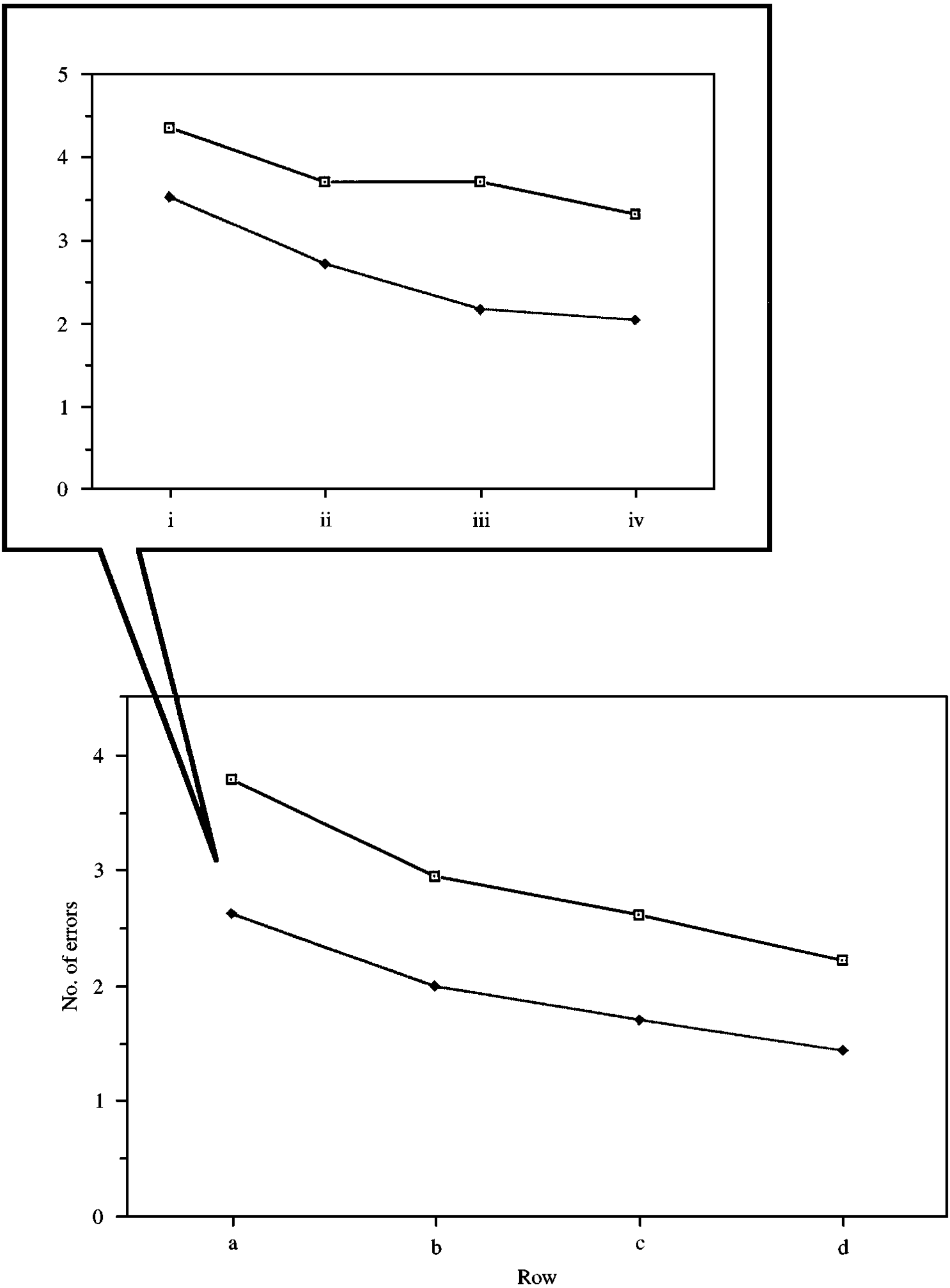

FIGURE 6. Changes in estimated and actual error over time for the four rows of grid 2. The inset details changes over the first four items of that grid (i.e. row *a*). ——□—— Estimate; ——◆—— Actual.

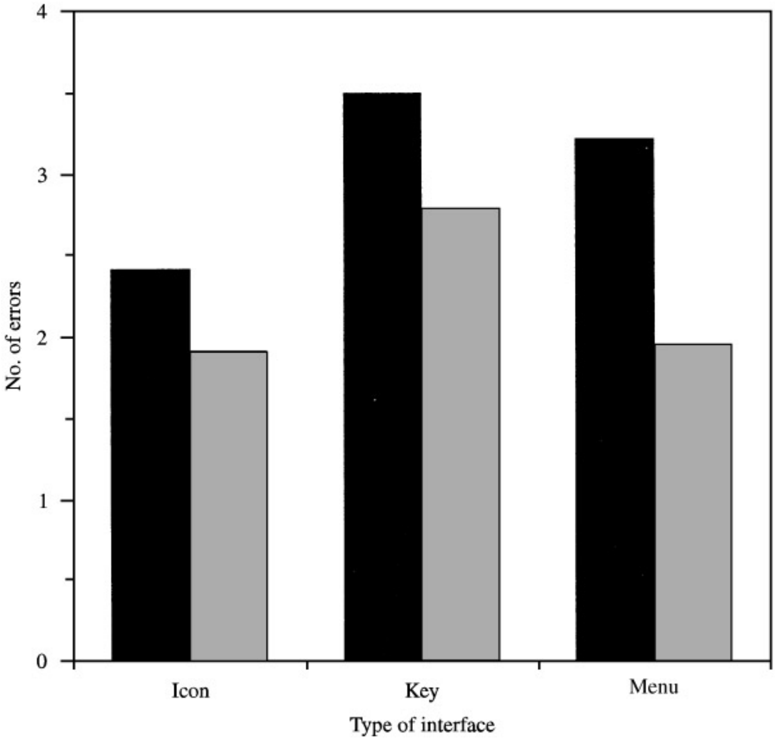

FIGURE 7. The effects of interface on estimated and actual error. ■ Estimate; ▨ Actual.

3.2. THE EFFECTS OF INTERFACE

The *pattern* of confidence changes across time was essentially the same irrespective of interface, as predicted. In other words, these effects are relatively robust, as predicted. However, the degree of underconfidence show by the users of the three systems did vary, as indicated by a significant interaction between interface and anticipated vs. actual error [$F(2, 57) = 3.6$, $p < 0.05$]. The pattern of results is shown in Figure 7. Looking firstly at users' actual performance, described in terms of the mean number of errors made, then it is clear that interaction via keyed input (shortcut commands) leads to a higher error rate than either menu based ($p < 0.01$) or icon based ($p < 0.01$) interaction. The latter two are associated with almost identical low error rates which is in keeping with the general literature on the superiority of GUIs over command input systems. The more interesting finding is that this pattern is not reproduced in the confidence data—i.e. there is an independent effect of interface on confidence, which is not mediated by task performance. This is demonstrated by the users of the menu-based system, who show nothing like the confidence that their performance merits—i.e. for some reason this system is wrongly perceived as being particularly difficult. The statistical analyses show that users clearly have far more confidence in the icon based system than in either the menu or key-based systems ($p < 0.05$, $p < 0.01$, respectively) while these last two could not be statistically separated.

Means taken from the median posterior estimates of errors made throughout the task showed once again that key users showed least confidence (mean error of 3.35), followed by icon users (mean of 3.09), and then menu users (mean of 2.35).

3.3. ANALYSIS BY SYMBOL

Given that each symbol had a different control syntax, and that these varied in complexity, it was anticipated that some symbols would generate more errors than others. An interesting issue, therefore, is whether or not confidence judgments reflect these differing levels of difficulty. This issue is pertinent to the discussion made earlier concerning the importance of frequency-type judgments in acquiring a procedural skill. If the user is basing confidence judgments on cues concerning frequency of successes across the task as a whole, rather than on the cues activated by individual items, then we would predict that confidence judgments would be insensitive to individual item difficulty. Alternatively, if confidence judgments are based upon the cues activated by single items, then the PMMs activated should contain partial knowledge of the syntax, and also frequentistic knowledge of previous success *on that item*, with the result that confidence judgments should come to reflect item difficulty.

Figure 8 shows the overall picture for grids 1 and 2. Note that the actual errors indicate a wide fluctuation in item difficulty—with symbols such as car and rabbit

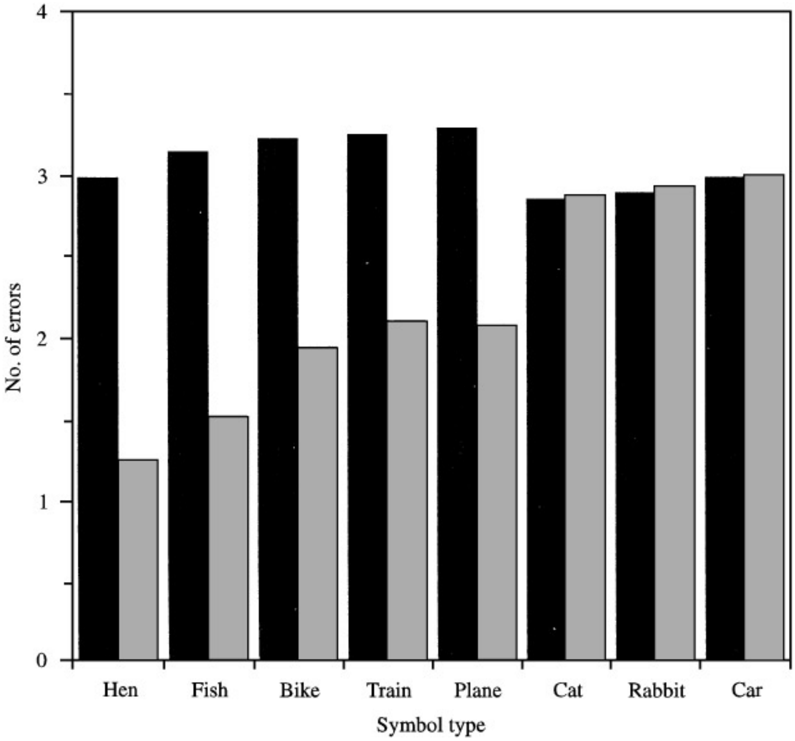

FIGURE 8. Estimated and actual error for individual items. ■ Estimate; ▨ Actual.

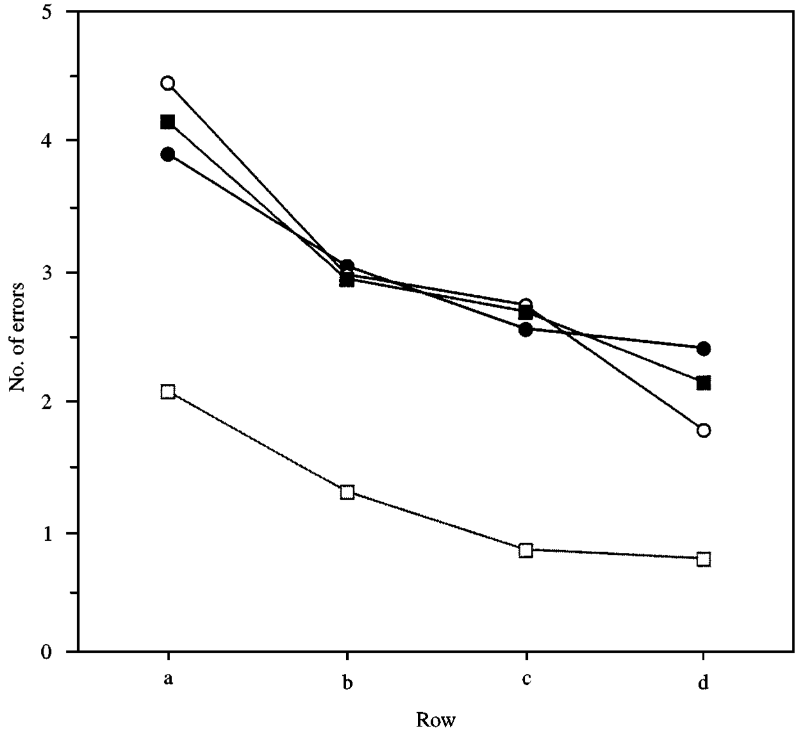

FIGURE 9. Changes in estimated and actual error as a function of practice, for a particularly difficult and a particularly easy item (grids 1 and 2 combined). —●— Estimate, car; —○— Actual, car; —■— Estimate, hen; —□— Actual, hen.

generating many errors, in marked contrast to symbols such as fish and hen which generate few. It is interesting to note that the syntax for the difficult items involved two mode switches: from foreground to background and then back to foreground, whereas the easy items involved one switch only.

Despite major differences in difficulty between symbols, there are no significant differences in the predicted error. Indeed, these confidence judgments are remarkably stable across the range of symbols. Of course, Figure 8 shows mean predicted errors for individual symbols across the task as a whole, which may potentially mask a growing sensitivity to individual item difficulty. So as a further demonstration that confidence judgments are not based upon actual item difficulty. Figure 9 plots the actual and estimated error over time for a particularly difficult and a particularly easy item. What is striking about this plot is the close fit between the estimated errors of both items, indicating no sensitivity whatsoever to individual item difficulty. This is clear evidence, then, that users are not basing their confidence judgments upon cues specific to individual items or events.

4. Discussion

The main purpose of this study was to see whether ecological models of self-confidence, present in the decision-making literature could be adapted to provide a suitable framework for understanding and predicting changes in the confidence of the computer user. With this is mind, there are four main findings which merit discussion: firstly, the predicted finding that users' judgments initially showed overconfidence, but were swiftly modified by feedback from the current task. Secondly, the unexpected observation that users' judgments showed underconfidence overall. Thirdly, the finding that users' judgments lacked any sensitivity to the difficulty of specific items; and fourthly, the predicted finding that there were differences in performance expectations between interfaces.

The predictions made earlier, based on the Gigerenzer *et al.* model, are correct in one important regard: users are basing their initial performance estimates upon an unrepresentative reference class, and then basing their later predictions upon a reference class which includes information concerning successes and failures in the current task. As predicted, users were initially highly confident that they could conduct the task with relatively few errors. In other words they were unprepared for the difficulty of the task before them, having presumably based their initial estimates upon a reference class of more "usable" (i.e. more consistent) computer-based operations. This overconfidence was swiftly modified, as users took account of their early failures, and re-assessed the difficulty of the task. Later judgments, including those about a second, transfer task, were made on the basis of updated data concerning successes and failure in the current task. This is indicated not by a convergence in estimated and predicted error, as anticipated, but by the similarity in gradient between the learning curves of estimated and predicted errors.

However, the predictions made earlier are also wrong in one important regard. It was expected that users' estimates of error would gradually converge with their actual errors, i.e. that users' confidence judgments would become better calibrated over time. However, there was no sign of convergence in the data: the dominant finding was that users quickly moved to a pattern of underconfidence (i.e. they expected to make more errors than they actually did make), and sustained this throughout.

Although this was unexpected, we had previously predicted that the patterns of *overconfidence* commonly reported in the decision-making literature would not be observed with a computer-based task, since the latter is predominantly procedural in nature which should allow confidence judgments to be based upon a frequency count of successes and failures within the task itself. In other words, we argued that, in the human–computer context, users' confidence judgments would be well-calibrated with real performance. Instead, we find that users' judgments show underconfidence, but great sensitivity to the rate of improvement.

Underconfidence in judgment is occasionally observed with very easy items (Lichtenstein, Fischhoff & Phillips, 1982), and is also reported in the revision of opinion literature (see Edwards, 1968, Erev, Wallsten & Budescu, 1994) in which people are asked to estimate the probability of a hypothesis following observation of the relevant data. In addition, as we have already discussed, Gigerenzer *et al.* believe that underconfidence can be observed when participants are asked to make frequency judgments of their performance over the task as a whole. However, Gigerenzer *et al.* account for underconfidence in the latter case, by arguing that people base their evaluations for the current task upon rates of success or failure in *unrepresentative* previous tasks. Such an explanation cannot really account for the data reported here, since users do seem to be updating their judgments on the basis of feedback from the *current* task, which is, of course, perfectly representative of itself. An ecological model may still be appropriate here if we could assume that participants simply did not have enough task experience to generate well-calibrated confidence judgments, in other words convergence between estimated actual performance levels would occur given sufficient time. However, the data don't suggest that this is likely.

As stated earlier, a fundamental tenet of the ecological viewpoint is that people generally use appropriate reference classes when making confidence judgments. However, there are other models of confidence jugdment which assert that such judgments are prone to systematic bias, and these could offer plausible accounts of the underconfidence observed in our task. Two such models are relevant to the data observed here, by Griffin and Tversky (1992) and by Jones, Taylor-Jones and Frisch (1995).

Griffin and Tversky (1992) state that people will show underconfidence when the weight (credibility) of the evidence used to support a judgment in high and the strength (extremeness) of the evidence is low. Although the concepts of weight and strength are rather slippery, one interpretation of their model might be that in circumstances where there are many components to a task, each of which is given reliable feedback (weight–high), but none of them are particularly salient or meaningful (strength–low) then confidence judgments may be overly conservative. This seems quite appropriate to the task described here, since participants are given a good deal of credible information concerning their performance, but none of it may stand out as being particularly salient. In other words, users are swamped by a great deal of accurate feedback concerning the errors they are making–but they are left without a strong impression about where they are going wrong.

Jones *et al.* (1995) would offer an alternative account in which frequency (as opposed to single event) judgments are subject to availability bias, such that instances which are particularly salient, vivid, distinctive or easy to retrieve are more likely to be sampled. In order to account for the data presented here, we would have to argue that particularly difficult items are somehow more distinctive, which seems plausible. Translated into Gigerenzer's terms, their model would assume that a PMM is constructed from the items most easily brought to mind $\cdots$ in this case the most difficult items, and that this would subsequently generate underconfidence in jugdment.

While such cognitive bias explanations seem promising, it is impossible to distinguish them empirically from an alternative account, in which underconfidence appears as an artifact of the type of confidence measurement we are taking. This account addresses the way in which people's internal self-monitoring processes map onto the error judgments requested of them. It is possible that people monitor their performance using criteria which do not map precisely onto the kinds of error judgments which we have asked them to make. For example, they may register some measure or impression of overall success, such as the the number of items completed without any error, but they may fail to log the number of individual errors made in the completion of any one item. In this case, how are they to make their estimates of the likely number of errors on the next item? Their self-monitoring processes may yield an evaluation of performance which correctly reflects rate of improvement, and gives rise to an increasing feeling of confidence, but this could not reasonably be specified in terms of a particular number of errors per item. Users would be unable to give actual error estimates, but would instead have to translate some general impression of confidence into the terms of the error-scale provided for them. In other words, the absolute value of their judgment (in this case indicating underconfidence) would be irrelevant, although the changing pattern of these judgments over time could be meaningfully compared with changes in actual performance over that same period.

The issue of measurement is acknowledged as a real problem in the confidence literature [e.g. by Gigerenzer *et al.* (1991) and by Weaver (1990)]. It has already been noted that estimates of confidence made in terms of percentages (I am 70% confident my answer is correct), yield rather different results from estimates made in terms of frequency (I will get 7 answers correct), and although there are good theoretical reasons why these measures should produce different results, it is always possible that patterns of over- and underconfidence are actually measurement artifacts, as argued by Erev *et al.* (1994). Of course, such problems only arise when trying to calibrate some measure of confidence with real performance—i.e. when trying to decide whether or not people are over or underconfident, and so these problems can be sidestepped if research is focussed upon relative measures—such as the changing patterns of confidence over time, or comparisons of the different degrees of confidence participants experience with different versions of a task, or with different sub-components of a task.

With regard to this last point, we move on to the interesting observation that participants in the current study were unable to distinguish between the items they found easy and those they found difficult in terms of their performance estimates. While such insensitivity would be expected at the start of the session, it was surprising to find that users were still unaware of which items caused the most problems by the time they had completed an entire picture grid, as shown in Figure 12. This, coupled with the observation that users were very sensitive to their rate of improvement on the task lends additional support to the idea that users are monitoring their progress in terms which do not properly distinguish between items, and translating this internal assessment into an error-per-item estimate in order to comply with the constraints of the task. Alternatively, participants are actually monitoring the errors they make on different items, but are then finding the items highly confusable and subsequently make performance predictions which are anchored on the most difficult items (as predicted by the Griffin and Tversky and Jones *et al.* models). Either way it does beg the question of whether or not people always monitor their performance so crudely.

Every day experience tells us that people are often aware of those functions they can carry out easily, and those which cause them difficulty. A secretary may have no problem in formatting text, but may experience difficulty in using style sheets, and it would seem ludicrous to suggest that he wouldn't be aware of this. But there is probably still an issue of *level* of discernment or sensitivity here. Our secretary may seem self-aware when it comes to these major functional distinctions—but would he also be sensitive to his past performance on the different formatting functions, or even on the different commands which make up each formatting sequence? This is not a trivial issue, since his confidence in these sub-tasks may actually determine his future behaviour (e.g. he may try to avoid those functions he perceives as difficult), and so some understanding of the factors which affect discrimination in self-monitoring processes would be useful. But what *does* determine the level at which people will monitor successes and failures? Diversity and discriminability between the elements would seem one likely factor, and length of experience with the system another. It is also possible that *choice* plays a significant role in determining the sensitivity of self-monitoring process, since confidence judgments can only really have a functional utility with choice-driven behaviour. But it does seem likely that people may only make the distinction between function A, which they usually execute successfully, and function *B*, which they usually bungle, if they have a choice about whether or not to actually execute *A* or *B*. In our study users had no such choice, which may in part account for their insensitivity to item difficulty.

These suggestions are simply speculative at present, and merit further investigation but they do show the advantage of trying to monitor confidence judgments more closely. While the many publications on self-efficacy have demonstrated the consequence of different efficacy judgments on motivation and performance, they offer no real account of the development of such beliefs. In contrast, this preliminary study of confidence judgment demonstrates that measures of anticipated performance can reveal systematic distortions in people's ability to monitor performance, and this in turn suggests a wholly different research agenda from that associated with the self-efficacy literature. Of course the precise relationship between self-efficacy judgments and anticipated performance (confidence) judgments remains unclear, and this too would merit further research.

We turn, finally, to the differences between interfaces observed in this study. It was predicted that users who were given a keyed input system would be less confident than users of either the menu or icon-based systems. In fact users of the key-based system did show low confidence—but then the performance of this group was by far the worst of three, and so this was not in itself a particularly interesting finding. The more interesting finding is that confidence judgments for the users of the menu-based system were particularly poor. In other words, users predicted very high error rates, whereas in fact they performed very well, making very few errors. This means that the design of an interface has an independent contribution to make to a user's self-confidence, over and above its capability to influence performance. Indeed, given that in this study, the task itself was totally unfamiliar, it is possible that the independent contribution of interface design to the user's assessment of confidence is rather underplayed in this task. With a more familiar, or even more meaningful task, the commands and syntax of a well-designed graphical user interface should generate a rich reference class upon which to base judgments of confidence—which would mean that the independent effect of the interface upon user confidence should be much greater. This begs a number of questions concerning whether or not designers should promote confidence in the use of a system over and above realistic expectations of performance, or whether designers should aim for well-calibrated users who know their limitations and who may consequently seek to improve their control of a system. Some of the self-efficacy findings cited earlier suggest that users are more likely to explore a system fully if they are initially fooled as to just how easy the system is to master—but once again these are issues which must be resolved in further studies.

As for the current study, we have shown that it is possible to adapt techniques and models from the decision-making literature so as to apply them to the human–computer context, but there are a number of problems. These models deal predominantly with judgments about declarative knowledge; and must be modified to take account of the rather different constraints which operate in a world of skills and procedures, and there is also the issue of measurement, for while definitions of confidence in terms of predicted performance are common in the decision-making literature, it is by no means clear that users are able to offer performance predictions in the precise terms requested by the experimenter—and we have seen that this can lead to real difficulties of interpretation.

The limitations of the current study are obvious. We have constructed an artificial task of highly unrepresentative design in order to test a certain model. As a consequence the specifics of our results are unlikely to generalise to other, real-world tasks. However, the endeavour has been useful, in terms of understanding—or beginning to understand—just how computer users come to judge their own performance. We have seen that self-confidence may be subject to systematic bias: that it may potentially be boosted or lowered, depending upon the reference-classes adopted by each individual user; and we have also seen that the computer interface has an important role to play in this process.

References

ANDERSON, J. R. (1983). *The Architecture of Cognition*. Cambridge, MA. Harvard University Press.

BANDURA, A. (1986). *Social Foundations of Thought and Action: A Social Congnitive Theory*. Englewood Cliffs NJ: Prentice-Hall.

BAR TAL, Y. (1990) The effect of personal use of computers on employees-preception of control over work. *Social Behaviour*, **5,** 103–115.

BURKHARDT, M. E. & BRASS, D. J. (1990). Changing patterns of change: the effects of a change in technology on social network structure and power. *Administrative Science Quarterly*, **35,** 104–127.

CARLSON, R. D. & GRABOWSKI, B. L. (1992). The effects of computer self-efficacy on direction following behavior in computer assisted instruction. *Journal of Computer Based Instruction*. **19,** 6–11.

COMPEAU, D. R. & HIGGINS, C. A. (1995). Computer self-efficacy: development of a measure and initial test. *MIS Quarterly*, **19,** 189–211.

EDWARDS, W. (1968). Conservatism in human information processing. In B. KLEINMUNTZ, Ed. *Formal Representations of Human Judgment*. New York: Wiley.

EINHORN, H. J. & HOGARTH, R. M. (1978). Confidence in judgement: persistence of the illusion of validity. *Psychological Review*, **85,** 395–416.

EREV, I., WALLSTEN, T. S. & BUDESCU, D. V. (1994). Simultaneous over- and underconfidence: the role of error in judgment processes. *Psychological Review*. **101,** 519–527.

FISCHHOFF, B. & MACGREGOR, D. (1982). Subjective confidence in forecasts. *Journal of Forecasting*. **1,** 155–172.

FISCHHOFF, B., SLOVIC, P. & LICHTENSTEIN, S. (1977). Knowing with certainty: the appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, **3,** 552–564.

FRESE, M. (1987). Preface to M. FRESE, E. ULICH, & W. DZIDA, Eds. *Psychological Issues of Human Computer Interaction in the Workplace*. North-Holland: Elsevier.

GIGERENZER, G., HOFFRAGE, U. & KLEINBÖLTING, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychological Review*, **98,** 506–528.

GIGERENZER, G. & GOLDSTEIN, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, **103,** 650–669.

GRIFFIN, D. & TVERSKY, A (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, **24,** 411–435.

GIST, M. E. & MITCHELL, T. R. (1992). Self-efficacy: a theoretical analysis of its determinants and malleability. *Academy of Management Review*, **17,** 183–211.

GIST, M. E., SCHWOERER, C. & ROSEN, B (1989). Effects of alternative training methods on self-efficacy and performance in computer software training. *Journal of Applied Psychology*, **74,** 884–891.

HILL, T., SMITH, N. D. & MANN, M. F. (1987). Role of efficacy expectations in predicting the decision to use advanced technologies: the case of computers. *Journal of Applied Psychology*, **72,** 307–313.

JONES, S. K. TAYLOR-JONES, K. & FRISCH, D. (1995). Biases of probability assessment: a comparison of frequency and single-case judgements. *Organizational Behaviour and Human Decision Processes*, **61,** 109–122.

JUSLIN, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter guided selection of almanac items. *Organisational Behaviour and Human Decision Processes*, **57,** 226–246.

JUSLIN, P. (1995). Can overconfidence be used as an indicator of reconstructive rather than retrieval processes? *Congnition* **54,** 99–130.

KIDD, J. B. (1970). The utilization of subjective probabilities in production planning. *Acta Psychologica*, **34,** 338–347.

LEE, J, & MORAY, N. (1992). Trust, control strategies and allocation of function in human–machine systems. *Ergonomics*, **35,** 1243–1270.

LEE, J. & MORAY, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human–Computer Studies*, **40,** 153–184.

MCDONALD, T. & SIEGALL, M. (1992). The effects of technological self-efficacy and job focus on job-performance, attitudes and withdrawal behaviours. *Journal of Psychology*, **126,** 465–475.

MITCHELL, T. R., HOPPER, H. DANIELS, D., GEORGE FALVEY, J. & JAMES, L. R. (1994). Predicting self-efficacy and performance during skill acquisition. *Journal of Applied Psychology*, **79,** 508–517.

MUIR, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man–Machine Studies*, **27,** 527–539.

ROSCH, E. (1978). Principles of categorization. In E. ROSCH & B. B. LLOYD, Eds. *Cognition and Categorization*. Hilsdale, NJ: Erlbaum.

SUBBOTIN, V. (1996). Outcome feedback effects on under- and overconfident judgment (general knowledge tasks). *Organizational Behaviour and Human Decision Processes*. **66,** 268–276.

TE'ENI, D. (1990). Direct manipulation as a source of cognitive feedback: a human–computer experiment with a judgment task. *International Journal of Man Machine Studies*. **33,** 453–466.

WAGENAAR, W. A. & KEREN, G. (1986). Does the expert know? The reliability of predictions and confidence ratings of experts. In E. HOLLNAGEL, G. MANEINI & D. WOODS, Eds. *Intelligent decision support in process environments*. Berlin: Springer.

WASON, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, **12,** 129–140.

WEAVER, C. A. (1990). Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning Memory and Cognition*, **16,** 214–222.

ZUBOFF, S. (1988). *In the Age of the Smart Machine: The Future of Work and Power*. New York: Basic Books.

Paper accepted for publication by Associate Editor, Dr. S. Wiedebeck

Appendix: Analysis of variance tables

TABLE 1

Analysis of variance on actual and anticipated errors for the first four items of learning grid 1, across the three types of interface (icon, menu, key)

Source	Df	Sum of squares	Mean square	*F*	*P*
Interface	2	52.829	26.415	4.916	0.0107
Subject (group)	57	306.294	5.374		
Item	3	68.573	22.858	15.900	0.0001
Item*Interface	6	15.471	2.578	1.794	0.1031
Item*Subject (group)	171	245.831	1.438		
Error	1	22.969	22.969	6.532	0.0133
Error*Interface	2	24.237	12.119	3.447	0.0386
Error*Subject (group)	57	200.419	3.516		
Item*Error	3	210.073	70.024	29.420	0.0001
Item*Error*Interface	6	22.796	3.799	1.596	0.1510
Item*Error*Subject (gp)	171	407.006	2.380		

TABLE 2

Analysis of variance on actual and anticipated errors for learning grids 1 and 2 across the three types of interface (icon, menu, key)

Source	Df	Sum of squares	Mean square	*F*	*P*
Interface	2	39.013	19.501	7.108	0.0018
Subject (group)	57	156.418	2.744		
Grid	1	10.134	10.134	44.943	0.0001
Grid*Interface	2	0.530	0.265	1.175	0.3160
Grid*Subject (group)	57	12.853	0.225		
Error	1	40.477	40.477	47.020	0.0001
Error*Interface	2	6.184	3.092	3.592	0.0339
Error*Subject (group)	57	49.068	0.861		
Grid*Error	1	1.092	1.092	8.631	0.0048
Grid*Error*Interface	2	0.385	0.192	1.521	0.2272
Grid*Error*Subject (gp)	57	7.211	0.127		

TABLE 3

Analysis of variance on actual and anticipated errors for the first four items of learning grids 2, across the three types of interface (icon, menu, key)

Source	Df	Sum of squares	Mean square	*F*	*P*
Interface	2	50.954	25.477	3.829	0.0275
Subject (group)	57	379.212	6.653		
Item	3	105.617	35.206	29.050	0.0001
Item*Interface	6	5.146	0.858	0.708	0.6438
Item*Subject (group)	171	207.238	1.212		
Error	1	161.008	161.008	63.971	0.0001
Error*Interface	2	18.529	9.265	3.681	0.0314
Error*Subject (group)	57	143.463	2.517		
Item*Error	3	8.242	2.747	2.287	0.0804
Item*Error*Interface	6	8.371	1.395	1.162	0.3292
Item*Error*Subject (group)	171	205.388	1.201		