The Quarterly Journal of Experimental Psychology

ISSN: 0272-4987 (Print) 1464-0740 (Online) Journal homepage: https://www.tandfonline.com/loi/pqja20

On the relationship between task performance and associated verbalizable knowledge

Dianne C. Berry & Donald E. Broadbent

To cite this article: Dianne C. Berry & Donald E. Broadbent (1984) On the relationship between task performance and associated verbalizable knowledge, The Quarterly Journal of Experimental Psychology, 36:2, 209-231, DOI: 10.1080/14640748408402156

To link to this article: https://doi.org/10.1080/14640748408402156

The Quarterly Journal of Experimental Psychology (1984) 36A, 209–231

On the Relationship between Task Performance and Associated Verbalizable Knowledge

Dianne C. Berry and Donald E. Broadbent
Department of Experimental Psychology, University of Oxford, U.K.

Three experiments explore the relationship between performance on a cognitive task and the explicit or reportable knowledge associated with that performance (assessed here by written post-task questionnaire). They examine how this relationship is affected by task experience, verbal instruction and concurrent verbalization. It is shown that practice significantly improves ability to control semi-complex computer-implemented systems but has no effect on the ability to answer related questions. In contrast, verbal instruction significantly improves ability to answer questions but has no effect on control performance. Verbal instruction combined with concurrent verbalization does lead to a significant improvement in control scores. Verbalization alone, however, has no effect on task performance or question answering.

GENERAL INTRODUCTION

In the case of manual skill it is generally accepted that certain crucial aspects of human performance are unavailable for introspective report. We cannot adequately describe all that we can do, and, conversely, we cannot do all that we can describe. Recent evidence, however, suggests that this dissociation is not limited to the field of manual skills. The suggestion of a distinction between different types of knowledge or thought processes is becoming increasingly popular in cognitive psychology. Dichotomies such as implicit/explicit, conscious/unconscious and verbal/non-verbal are being applied at all levels, from perception (Turvey, 1974; Marcel, 1983) and memory (Jacoby and Dallas, 1981) to concept learning (Reber, Kassin, Lewis and Cantor, 1980; Kellogg, 1982) and deductive reasoning (Wason and Evans, 1975; Evans, 1980). Most of these authors posit some form of dissociation between an individual's performance on a given task (whether

Requests for reprints should be sent to D. C. Berry, Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford, U.K. OX1 3UD.

This work was supported by the Social Science Research Council. The sugar production and personal interaction systems were originally programmed by Dr Peter Naish.

this involves recognizing a word or solving a problem) and the explicit or reportable knowledge associated with that performance.

The present series of experiments stems from research reported by Broadbent (1977) and Broadbent, Fitzgerald and Broadbent (1982) on the control of complex systems. Broadbent (1977) found that individuals practising on a simple model of a city transportation system improved in ability to make the right decisions but not in ability to answer questions about the relationships within the system. The task involved subjects having to control the number of people using bus transport and the number of empty car-parking spaces by varying the time interval between buses and the price of the car parks. Using the same task, Broadbent, Fitzgerald and Broadbent also found that subjects improved in ability to control the system with practice (that is, they took fewer attempts to reach the specified targets) but did not improve in ability to answer questions. In addition, it was reported that overall control performance was not correlated with questionnaire score and that verbal explanation had no effect on task performance.

These results at first suggest a conclusion which parallels that reported by the authors mentioned above—that of a dissociation between an individual's performance on a given task and the explicit verbalizable knowledge associated with that performance. Unfortunately, however, there are several methodological and interpretational difficulties with the earlier Broadbent experiments, which make it impossible to draw such a strong conclusion. Firstly, the sample size in the 1977 study was very small, the total number of subjects tested being only 12. Secondly, the verbal explanation given to subjects in the second study consisted of a written questionnaire with the correct answers filled in, together with a justification for each answer. No attempt was made to match this explanation to the control task directly, nor to ensure that it had been adequately understood by the subjects. Thirdly, subjects were given written questionnaires before attempting the control task. Recent unpublished experiments by the present authors, however, show that completion of a written questionnaire can affect subsequent task performance. Finally, performance was assessed on the basis of the number of attempts to reach a pair of specified targets. Individuals who were worse at controlling the system therefore had had far more experience with the task at the time of answering the post-task questionnaire than those who were better.

The present study has been designed with these criticisms in mind. Two different tasks are employed here; they are described in detail in the next section of the paper. Both are computer-implemented and require subjects to reach and maintain specified target values of an output variable by varying a single input variable. In one task, subjects take on the role of manager of a simple sugar production factory and are required to reach and maintain a specified level of sugar output by varying the number of workers employed. The other task requires subjects to interact with a

"computer person". The communication is based on a fixed set of adjectives describing various degrees of intimacy of personal interaction. Subjects are told to shift the behaviour of the person to the *Very Friendly* level and to attempt to maintain it at that level.

These tasks, like manual skills, require sustained performance rather than the making of a single response to a single stimulus. Unlike manual skills, however, they require cognitive or intellectual decision making. Although the two tasks appear very dissimilar, the equation relating production level to work force is the same as that relating the responses of the computer person to those of the human subject. This equation (which is described further in the next section) comes from a class of systems described in engineering science as requiring adaptive control. There is therefore no single stimulus–response mapping. The same action can result in different responses, depending on the existing state of the system. On this basis the tasks should bring out aspects of performance that would not normally be tapped by single response measures. The sugar production and personal interaction guises provide two superficially very different frameworks. We should therefore be able to detect whether subjects are responding to the underlying algorithm or merely to some superficial characteristic.

The major purpose of the present study is to see whether there are grounds for applying the "performance/verbalizable knowledge" distinction to control tasks such as these. Performance here refers to the ability to control the sugar or person task. All subjects interact with the computer for a fixed number of trials, and the performance measure is thus the number of these trials on which the target value is reached. Verbalizable knowledge refers to task-related information that is (or has been) conscious and is potentially expressable in language. It is assessed by written post-task questionnaire. In order to increase the chance of tapping specific task-related knowledge, the questionnaires are made up of three different types of questions. These are described later in the paper.

A basic problem with research in this area is that it is very difficult to prove that task performance and verbalizable knowledge are unrelated in a given situation. Simply comparing performance on one task under one set of conditions with an individual's related verbalizable knowledge is not sufficient. Failure to find a positive relationship can always be attributed to methodological inadequacies. A stronger approach is to introduce factors which might have differential effects on task performance and question answering. Three experiments therefore examine the effects of task experience, verbal instruction and concurrent verbalization. By looking at how each of these affects control performance, question answering and their relationship to each other, we can hope to discover more about the way in which these tasks are performed and particularly about the role of explicit verbalizable knowledge.

The Tasks

Sugar Production

Subjects were instructed to imagine that they were in charge of a sugar production factory in an underdeveloped country. They were told that they could control the rate of production simply by changing the size of the work force, ignoring all other factors. The initial level of production was 6,000 tons, and the task was to reach and maintain a target output of 9,000 tons. The size of the work force could be varied in 12 discrete steps, so that on any one trial it was possible to employ between 100 and 1,200 workers. The starting level was 600 workers. Subjects typed in digits between 1 and 12 to represent the number of hundreds of workers they wished to employ. Only integers were accepted. The level of production was related to work force by the equation, $P = 2 \times W - P1$, where W = the number between 1 and 12 typed in to represent the size of the work force, $P1$ = the number, again between 1 and 12, representing the previous sugar output (that is, the previous sugar output divided by 1,000) and P = the number representing the current sugar output. The computer added 1, 0 or -1 on a random basis to P on each trial. The final P value was then converted to thousands of tons of sugar output. (The computer simply added three zeros to P.) There were 12 possible outputs, ranging between 1,000 and 12,000 tons. The lower and upper bounds were fixed at 1,000 and 12,000 tons, respectively. If the equation therefore resulted in a sugar output of less than 1,000 tons, an output of 1,000 tons was displayed; similarly, if the equation resulted in an output of more than 12,000 tons, an output of 12,000 tons was displayed. Subjects were aware of these upper and lower bounds. On each trial, the current work force and production level were displayed numerically on the VDU. In addition, a graphic display was shown, with the target value represented by a line stretching horizontally across the screen. Each successive period's output was shown by a dot plotted with output on the vertical axis and time on the horizontal.

The nature of the equation was such that there was not a unique output associated with any one input. The resulting output depended on the previous output as well as the new work force figure. A work force of 600, for example, would produce more, or less, sugar depending on whether the existing production level was 4,000 or 8,000 tons, respectively. An illustrative series of input and output values is shown in Table I. The above equation was chosen, and the random element included, to ensure that subjects would exercise continuous control. Subjects might, through chance, hit the target value early in a series of trials. They would be unlikely, however, to remain on this target if they repeatedly entered the same input value. Subjects interacted with the system for a fixed number of trials. Each entry of an input value into the computer counted as one trial. The performance measure was therefore the number of these trials

on which the target output value was reached. The scoring criterion was set so that a sugar output of 8,000, 9,000 or 10,000 tons counted as being on target. This meant that despite the random element it was still possible to control the system 100% of the time. Subjects were not aware of this criterion, however.

Personal Interaction

Subjects were told that they would be meeting a computer person named Clegg. They would communicate with Clegg through a typewriter keyboard and VDU. There were 12 possible grades of behaviour, these being *Very Rude*, *Rude*, *Very Cool*, *Cool*, *Indifferent*, *Polite*, *Very Polite*, *Friendly*, *Very Friendly*, *Affectionate*, *Very Affectionate* and *Loving*. These adjectives had numerical equivalents, which corresponded to the 12 levels of work force and sugar output. Responses took the form of typing the initial letter or letters of the words describing the styles of behaviour the subject wished to show. Clegg initiated the interaction by displaying one of the three adjectives centred on *Polite*. Subjects were instructed to shift Clegg's behaviour to the *Very Friendly* level and to maintain it at that level. On each trial Clegg's and the subject's responses were displayed on the VDU. These scrolled up the screen so that on any one trial it was possible to see the responses made on the previous six trials. The equation relating Clegg's responses to those of the subject was the same as that relating sugar production to work force. (The tasks were mathematically equivalent.) Clegg therefore amplified any change in human decision in a retaliatory

Table I

An Illustrative Series of Inputs and Outputs for Both the Sugar Production and Personal Interaction Tasks

Sugar Production		Personal Interaction	
Work force	Sugar output in tons	Your behaviour	Clegg's behaviour
	6,000		polite
700		very polite	
	8,000		friendly
900		very friendly	
	10,000		affectionate
800		polite	
	7,000		very cool
1,000		friendly	
	12,000		loving
900		polite	
	6,000		very rude
1,000		indifferent	
	12,000		very friendly
1,000		very friendly	
	9,000		friendly

fashion. Clegg would make a very different response to an input of *Polite*, for example, depending on whether Clegg was previously *Affectionate* or *Cool*. An illustrative series of inputs and outputs is shown in Table I. A computer response of *Friendly*, *Very Friendly* or *Affectionate* was scored as being on target, again to allow for the random element.

Attempts were made to make the person and sugar tasks equivalent. Because of the way the tasks were presented on the monitor, however, this was not possible in all respects. Although for both tasks the number of possible inputs and outputs was restricted to 12 each, the information available to subjects at any one time was different. In the sugar task, on any one trial, subjects could see the present work force and sugar output values displayed numerically, as well as a graphical display of all the previous output values. In the person task, they could see both previous input and output values, but only for the last 6 trials.

The Questionnaires

The structure of the questionnaire was the same for both tasks. Each questionnaire was divided into three sections. The first section consisted of two multiple-choice questions. The questions described a starting state—for example, "the initial output is 6,000 tons and the initial work force is 600"—and then gave a change of state—for example, "you increase the work force to 900". Subjects were then asked to predict the resulting output from a given set of six alternatives, for example, "(1) 6,000 tons, (2) 9,000 tons, (3) between 6,000 and 9,000 tons, (4) less than 6,000 tons, (5) more than 9,000 tons, (6) equally likely to be any of these". One question involved predicting the result of an increase in the size of the input variable, the other a decrease. In both cases the correct answer was a clear member of one of the six alternatives. (The answer to the above question, for example, is 12,000 tons, with (5) being the correct alternative.) The questions for the person task were very similar, with the equivalent adjectives substituted for the given work force and sugar output levels. The only difference was that for the sugar task subjects were given the initial work force and sugar output levels (as would appear on the computer screen), whereas for the person task they were initially given only Clegg's behaviour (for example, Clegg's initial level of interaction is *Polite*.) (Although the initial work force value was not taken into account in the computer's calculation of the new output value, it was not thought that providing this "redundant" piece of information would adversely affect subjects' questionnaire performance. The initial work force value was also present at the start of the computer task. Moreover, the equivalent information was not provided on the Clegg questionnaire, and, as will be seen, scores on the latter questionnaire were no better than those on the former.)

The three questions in the second section were even more closely matched to the computer task, in that in each case subjects were given a sequence of three pairs of inputs and outputs and were asked to make a numerical prediction of the next output, given a new input. In the case of sugar production, subjects were given a graph showing each sequence, as well as the numerical values. In the case of personal interaction they were given the sequences of adjectives as they would appear on the screen (starting with Clegg's behaviour). These three questions were scored as being correct if the chosen values either corresponded to, or were one value higher or lower on the scale, than the predicted value. The final section consisted of a general question, which asked subjects to describe how they went about attempting to reach and maintain their target values. They were encouraged to write whatever came into their heads and not to worry about wording or grammaticality.

The above questions are more closely matched to the control task than those used in the earlier Broadbent experiments. In the latter case, subjects were merely asked to make qualitative predictions of the direction of change of an outcome variable, given a change in one of the input variables. It is hoped that asking these three different types of question will increase the likelihood of tapping whatever task-related information subjects have available to them. It seems likely that if an individual does have any verbalizable knowledge concerning his or her performance, or the way the system reacts, the above questions should show some evidence of this.

All of the following experiments were run on a 380Z microcomputer and VDU.

EXPERIMENT 1

This experiment examines how practice with either the sugar production or personal interaction task affects ability to control the system as well as ability to answer questions related to the system. More specifically, it looks at whether individuals who receive two sets of 30 trials perform better on the second set than individuals given only one set of trials and whether this difference in experience is reflected in their questionnaire scores.

SUGAR PRODUCTION

Method

Subjects

The 24 volunteer subjects were Oxford University graduate and undergraduate students, aged between 19 and 35.

Design

Subjects were randomly allocated to one of two groups. Group 1 received one set of 30 trials, followed by a post-task questionnaire. Group 2 received two sets of trials, followed by the questionnaire. It was not possible to use a within-subjects design for this experiment. Unpublished experiments by the present authors have shown a significant decrement in control performance following completion of a written post-task questionnaire. Questionnaires could therefore not be given to subjects before or between sets of trials.

Procedure

Subjects were given written instructions explaining the nature of the system and the task. Starting levels were kept constant for all subjects—a sugar output of 6,000 tons and a work force of 600. Following the thirtieth trial, subjects in Group 1 were asked to complete a written post-task questionnaire. It was stressed that this referred to the task and that subjects should base their answers on their experience. Subjects in Group 2 were told that they would be given another set of 30 trials on the same system, and the starting levels would again be 6,000 tons and 600 workers. Following the final trial, they were asked to complete the post-task questionnaire. They were given the same instructions as subjects in Group 1. Neither group was aware at the time of controlling the system that they would later be asked to answer questions about it.

Results

A sugar output of 8,000, 9,000 or 10,000 tons was scored as being on target. This criterion was chosen to allow for the slight random element. Using this criterion, it was possible for subjects to be on target 100% of the time. The number of trials on target was totalled for each subject for each set of 30 trials. For Group 1, the scores ranged between 3 and 14, with a mean of 8.75. For Group 2, scores on set 1 ranged between 4 and 14, with a mean of 8.58. Scores on set 2 ranged between 8 and 22, with a mean of 16.17. (To aid the interpretation of these results, chance performance has been determined at 3.4 trials on target out of 30. This number was obtained by running 500 simulated sessions, each of 30 trials, in which the subjects chose any one of the 12 responses with equal probability.) Performance was also scored separately for the last 5 trials of each set. Mean scores were 1.5 for Group 1 and 1.42 (set 1) and 3.92 (set 2) for Group 2.

At the outset, most subjects adopted the strategy of increasing the work force if production was below target, decreasing if above target and maintaining the same number if an adequate state arose. For many of these subjects this initially led to a cyclic pattern of large overproductions followed by large underproductions. With practice, however, the size of these deviations around the target line decreased.

Post-task questionnaires were scored out of a maximum of five. Mean scores were 1.75 and 1.67 for Groups 1 and 2, respectively. Looking at the questions individually, there were no consistent patterns within or between

groups. None of the five scored questions was particularly well answered. Answers to the final general question were not very informative, and no attempt was made to analyse these objectively. Many subjects did not answer the question at all, protesting that they were "unable to put it into words". Others wrote general statements such as "increasing and decreasing the work force to change the sugar level".

Statistical comparisons of particular interest were those between the questionnaire scores of subjects in the two groups and between their control scores for the five trials immediately preceding completion of the questionnaire. In both cases non-parametric statistics were used, as subjects could only score one of six possible values (0–5). A Mann Whitney U test showed no significant difference between the questionnaire scores of subjects in the two groups, $U = 65.5$, $n1 = n2 = 12$, $p > 0.1$. Questionnaire performance was equally low in both groups. In contrast, there was a highly significant difference between groups in control scores on the five trials immediately preceding the questionnaire, $U = 6$, $n1 = n2 = 12$, $p < 0.001$. Subjects in Group 2 were on target on an average of 3.92 (78%) of these trials. In comparison, those in Group 1 were on target on only 1.5 (30%) of these trials. This differences in control scores was not confined to the last five trials. An independent t test showed a highly significant difference between the overall control scores for subjects in Group 1 and the overall control scores for set 2 for subjects in Group 2, $t(22) = 4.40$, $p < 0.001$.

A possible explanation for the comparability of the questionnaire scores in the two groups is that some critical amount of learning occurred before the end of the first set of trials. This would then be reflected in the questionnaire scores of subjects in Group 1. These scores would not be expected to differ from those of subjects in Group 2. Two features of the results argue against this. The first is the highly significant difference between groups in scores for the five trials immediately preceding completion of the questionnaire. The second is the overall low level of questionnaire performance. To provide a more complete picture, however, a further group of 12 subjects was asked to complete questionnaires. These subjects had had no previous experience with the system. They were simply told that the questionnaire referred to a simple sugar production factory in an underdeveloped country and that the rate of production in this factory was controlled by changes in the size of the work force. They were asked to base their answers to the questions on what they might intuitively expect to happen in such a situation. The final general question was not included on this version of the questionnaire. The mean questionnaire score for subjects in this group was 1.42. A Kruskal–Wallis analysis showed that these scores did not differ significantly from those of subjects in the other two groups, $H = 1.06$, $p > 0.1$. Practice at controlling the system therefore had no effect on ability to answer questions.

To investigate the relationship between task performance and "ver-

balizable knowledge" further, questionnaire scores were correlated with overall control performance. (Set 2 scores were used for Group 2.) Spearman rank correlation coefficients were -0.30 and -0.25 for Groups 1 and 2, respectively. These negative coefficients did not reach significance at the 0.05 level (probably because of small sample size.) Rank correlations between performance on the last five trials and questionnaire scores were also negative and again failed to reach significance. Finally, for subjects in Group 2, questionnaire scores were also correlated with improvement in performance between sets 1 and 2. In this case a highly significant negative correlation was found, $r_s(10) = -0.75$, $p < 0.001$.

PERSONAL INTERACTION

Method

The method here was the same as that used for the sugar production task, with the exception that the appropriate personal interaction instructions were substituted for those concerning sugar production. Clegg initiated each set of trials by displaying one of the three adjectives centred on *Polite*. The subjects were 24 Oxford University graduate and undergraduate students. None had participated in the previous experiment.

Results

A computer response of *Friendly*, *Very Friendly* or *Affectionate* was scored as being on target (again to allow for the random element). As with the sugar task, the majority of subjects initially adopted the strategy of increasing the intimacy of their behaviour if Clegg was below target, decreasing it if above target and maintaining the same behaviour if Clegg was on target. This frequently led to Clegg jumping from one end of the intimacy scale to the other. Again, the size of these deviations around the target decreased with practice. Successful subjects learned to choose input values that were not too far on the scale from Clegg's given behaviour. In addition, a few subjects seemed to employ specific substrategies in that once successful, they would consistently respond with a specific behaviour given a certain level of behaviour from Clegg. For example, they might respond *Affectionate* whenever Clegg was *Loving* or *Very Polite* whenever Clegg was *Polite*. This only occurred on a minority of trials, however.

The number of trials on target was totalled for each subject for each set of 30 trials. For Group 1, scores ranged between 4 and 18, with a mean of 9.92. For Group 2, scores on set 1 ranged between 5 and 18, with a mean of 10.5. Scores on set 2 ranged between 10 and 26, with a mean of 17.58. An independent t test showed the difference in scores between subjects in Groups 1 and 2 (set 2) to be highly significant, $t(22) = 4.01$, $p < 0.001$. Performance was scored separately for the last five trials of each set. Mean

scores were 1.58 for Group 1 and 1.83 (set 1) and 4.0 (set 2) for Group 2. The difference between groups in scores on the five trials immediately preceding the questionnaire was again highly significant, $U = 12$, $n1 = n2 = 12$, $p < 0.01$.

Post-task questionnaires were scored out of a maximum of five. Mean scores were 1.58 and 1.50 for Groups 1 and 2, respectively. A Mann Whitney U test showed no significant difference between the scores of subjects in the two groups, $U = 72$, $n1 = n2 = 12$, $p > 0.1$. Again, looking at the questions individually, there were no consistent patterns within or between groups. None of the five scored questions was particularly well answered. Unlike in the case of the sugar task, nearly all subjects answered the final general question. The majority of answers were non-informative, however, consisting of statements such as "responding in the desired way" or "being more or less friendly as necessary". Only two subjects mentioned Clegg's "overreactive nature". Again, no attempt was made to relate these general statements objectively to task performance. Although the final general question was included on questionnaires used in the two subsequent experiments, responses to these are not discussed further in this paper. This does not appear to be a particularly efficient method of tapping an individual's specific task-related knowledge.

Questionnaire scores were correlated with control performance (set 2 scores were used for Group 2). Spearman rank correlation coefficients were -0.32 and -0.51 for Groups 1 and 2, respectively. Correlations between performance on the last five trials and questionnaire score were -0.42 and -0.54. None of these was significant at the 0.05 level (again probably because of the small sample size). The correlation between questionnaire score and improvement in performance between sets 1 and 2 (for subjects in Group 2) also failed to reach significance, $r_s = -0.26$, $p > 0.1$.

Discussion

Despite the dissimilar appearance of the two tasks, the results were remarkably consistent. In both cases practice improved ability to control the system yet had no effect on ability to answer questions. There was no evidence for a positive association between control ability and questionnaire score. All correlations were negative, although not significantly so.

The low level of questionnaire performance seems a little surprising at first. The task involved cognitive or intellectual decision making, and in both cases the inputs and outputs took a verbal form. It is unlikely, however, that the poor performance can be easily attributed to the quantity or quality of the particular questions asked. The questions were simply worded and matched to the control task. Moreover, two additional questionnaires have been devised for the sugar production and personal interaction tasks, and results from experiments using these have shown

similar levels of performance to those found with the original version. In one of these questionnaires the number of questions asked was doubled but the original form retained. (None of the questions was the same as any of those on the original version.) In addition, the number of multiple-choice alternatives for questions in the first section was reduced from six to four. A group of 12 subjects (who had not participated in the above experiment) was given two sets of trials with either the sugar or person task, followed by both the new questionnaire and the original questionnaire (the order of presentation being counterbalanced). The results showed that doubling the number of questions did not affect the proportion of questions answered correctly. Mean scores across the two tasks were 1.58 (out of 5) for the original version and 3.0 (out of 10) for the new version.

Changing the form of the questions asked also failed to improve question-answering performance. The original questionnaire always involved prediction of an outcome value, given a change in input state. In contrast, questions on the second revised questionnaire required subjects to state (given a certain situation) which input value would be needed to bring the outcome value to target. Twelve naive subjects were given two sets of trials with either the sugar production or personal interaction task, followed by the questionnaire. Performance on the new questions, however, was even lower than on the original version, the mean score being 0.67 out of a possible 5.0.

It seems that subjects are not able to access specific task-related information in a form that will allow them to answer these post-task verbal questions. Many claimed during their debriefing interviews that they were, in fact, operating on the basis of "some sort of intuition", making responses because they "felt right". It is possible, therefore, that whatever is learned during task performance is not verbalizable. This view is consistent with a series of experiments by Reber (Reber, 1967; Reber, 1976; Reber and Allen, 1978; Reber et al., 1980) on the implicit learning of artificial grammars. Reber has suggested that, under certain conditions, "complex structures such as those underlying language, socialisation, perception and sophisticated games are acquired implicitly and unconsciously". It is difficult, however, to know exactly what subjects in the present study learned with practice and how this diverges from explicit verbal knowledge. There may have been some critical verbal component to performance which could not be tapped by any of the forms of questioning described above.

Given these two alternatives, a relevant question concerns the efficacy of pre-task verbal instruction. If an individual is given detailed verbal instruction on how to reach and maintain a specified target value, will this result in enhanced control performance? Can verbal information be successfully applied to the task even though verbal access to task-related information appears to be limited? Experiment 2 examines whether control performance and questionnaire score are affected by verbal instruction.

Again, both the personal interaction and sugar production tasks are used. Broadbent, Fitzgerald and Broadbent (1982) reported no effect of verbal explanation on control scores using the simulated transport system. Their explanation consisted of a written questionnaire, complete with correct answers. In addition, a written justification was given for each answer. No attempt was made, however, to match this explanation to the control task, nor to ensure that it had been adequately understood.

EXPERIMENT 2

PERSONAL INTERACTION

Method

Subjects

The 24 subjects were paid volunteers from the Oxford University subject panel. They were aged between 18 and 45. None had participated in Experiment 1.

Design

Subjects were randomly allocated to one of two conditions, these being "training" (T group) and "no training" (NT group). Both groups of subjects received two sets of 20 trials, followed by a post-task questionnaire. The training group was given verbal instruction following the first set of trials.

Procedure

The procedure for subjects in the NT group was similar to that for subjects in Group 2 in Experiment 1. The only difference was that in the present experiment each set consisted of 20 rather than 30 trials. The initial procedure for subjects in the T group was the same as for those in the NT group. Following the first set of trials, those in the T group were given detailed verbal instruction on how to get Clegg to reach and maintain the *Very Friendly* level. The instructions were standardized for all subjects. They were slowly read by the experimenter from a typewritten sheet. They began with a brief statement explaining how Clegg reacts in an overreactive manner, exaggerating any response made by the subjects. This was followed by illustrative examples demonstrated on the VDU. It was pointed out that Clegg responds according to the difference between the subject's position on the intimacy scale and Clegg's previous position. The examples included instances where the subject's response was several scale positions away from Clegg's position, and where the subject's response was near to Clegg's. The instructions not only explained how Clegg would react to certain moves made by the subject, but also included advice on the best moves to make given certain levels of behaviour from Clegg. The nature of the random element was also explained. Subjects were allowed to ask questions and every attempt was made to ensure that they had understood the explanation. Subjects then received the second set of 20 trials followed by the unexpected questionnaire.

Results

Mean scores were calculated for each set of trials for each of the two groups. These are shown in Table II. For the training (T) group, the

Table II

Mean Performance (Number of Trials on Target) and Questionnaire Scores for the Person and Sugar Tasks for Experiments 2 and 3

	Person Task			Sugar Task		
Group	Set 1 (max = 20)	Set 2 (max = 20)	Questionnaire (max = 5)	Set 1 (max = 20)	Set 2 (max = 20)	Questionnaire (max = 5)
Experiment 2						
no training no verbalization	6.0	7.83	0.92	5.58	6.83	1.58
training no verbalization	5.75	9.17	3.42	4.67	7.0	3.58
Experiment 3						
no training verbalization	7.92	8.0	1.17	4.5	6.67	1.58
training verbalization	5.58	12.0	4.67	5.17	13.33	3.42

means were 5.75 (set 1) and 9.17 (set 2). For the no training (NT) group, the equivalent means were 6.0 and 7.83. A two-factor analysis of variance showed a significant practice effect, $F(1, 22) = 10.49$, $p < 0.01$, but no significant effect of training, $F(1, 22) = 0.27$, $p > 0.1$. The interaction also failed to reach significance, $F(1, 22) = 0.95$, $p > 0.1$. Mean scores for the last five trials were 1.25 (set 1) and 2.58 (set 2) for the T group and 1.16 (set 1) and 2.25 (set 2) for the NT group. Although the analysis showed an overall significant practice effect, the amount of improvement shown by subjects in the NT group was not as great as that shown by subjects in Group 2 in Experiment 1. The difference was probably due to the smaller number of trials per set in the present study and the different subject population.

In contrast to the insignificant effect of training on control performance, subjects in the T group scored significantly higher on the post-task questionnaire. Mean scores were 3.42 and 0.92 for the T and NT groups, respectively. A Mann Whitney U test showed this difference to be highly significant, $U = 7.5$, $n1 = n2 = 12$, $p < 0.001$. Questionnaire scores were again correlated with control performance. Spearman rank correlation coefficients were -0.19 and -0.31 for the T and NT groups, respectively. Neither coefficient was significant at the 0.05 level. Questionnaire scores were also correlated with improvement in performance between sets 1 and 2. The resulting rank coefficients were -0.53 (T group) and 0.03 (NT group). The coefficient for the T group just failed to reach significance at the 0.05 level. Subjects in the T group were divided into the six who showed the least improvement in performance between the two sets of trials and the six who showed the most improvement. The mean questionnaire score for those showing the least improvement in performance was 4.17, compared with 2.67 for those showing the most improvement. A Mann Whitney U test showed the difference in questionnaire scores to be statistically reliable, $U = 4$, $n1 = n2 = 6$, $p < 0.05$.

SUGAR PRODUCTION

Method

The method for this task was the same as for the person task, with the exception that all instructions referred to sugar production rather than personal interaction. The 24 naive subjects were again paid volunteers from the Oxford University subject panel. The instructions given to the training group were standardized and were read to the subjects by the experimenter. The wording of the instructions and the examples given were kept as close as possible to those given for the person task. The instruction therefore again included examples demonstrating how the system would react following certain changes of input, as well as advice on the best moves to make given certain levels of work force and production. Examples were again given on the VDU, with the experimenter referring to the graph and the numerical values.

Results

Mean scores were calculated for each set of trials for each of the two groups. These are shown in Table II. For the T group, the means were 4.67 (set 1) and 7.0 (set 2). For the NT group the equivalent means were 5.58 and 6.83. A two-factor analysis of variance showed a significant practice effect, $F(1, 22) = 6.12$, $p < 0.05$, but no significant effect of training, $F(1, 22) = 0.1$, $p > 0.1$, and no significant interaction, $F(1, 22) = 0.56$, $p > 0.1$. Mean scores for the last five trials were 1.33 (set 1) and 1.83 (set 2) for the T group and 1.83 (set 1) and 2.17 (set 2) for the NT group.

Again, there was a clear effect of training on post-task questionnaire scores. Mean scores were 3.58 and 1.58 for the T and NT groups, respectively. A Mann Whitney U test showed this difference to be highly significant, $U = 9$, $n1 = n2 = 12$, $p < 0.001$. Questionnaire scores were correlated with control performance. Spearman rank correlation coefficients were 0.21 and -0.17 for the T and NT groups, respectively. Questionnaire scores were also correlated with improvement in performance between sets 1 and 2. Rank coefficients were 0.1 (T group) and -0.43 (NT group). None of the above coefficients was significant at the 0.05 level.

Discussion

The major finding from this experiment was that detailed verbal instruction had no significant effect on control performance. It cannot be argued that the instruction was not adequately understood, however, as the trained subjects scored significantly higher on the post-task questionnaire. These individuals appeared to have the knowledge but would not, or could not, apply it to the control task. This result is in accordance with a study by Cooke (1965). Cooke found that even though students controlling a process could report some knowledge about the process, their non-verbal behaviour showed that they did not use this knowledge in controlling the process. He suggested that reportable knowledge is not necessarily used in action.

There is no obvious methodological reason for the fact that subjects in the present study failed to apply their newly acquired knowledge to the control task successfully. Attempts were made to match the wording of the instruction to the task, and examples were demonstrated on the VDU. The instruction not only explained how the system would react following certain changes of input, but also included advice on the best responses to make given certain levels of production or intimacy. It is unlikely that the lack of effect of training on control performance can be attributed to any problem that is specific to the particular instructions given. Rather, it

appears that these tasks might not be performed in a way that allows relevant verbal information to be easily implemented.

Given that verbal information cannot be easily incorporated into control performance and that whatever is learned with practice might not be verbalizable, an interesting question concerns the effect of concurrent verbalization. Several studies have shown that overt verbalization can influence the way a task is performed (see Ericsson and Simon, 1980, for a good review). Verbalization has, for example, been shown to result in more efficient solutions to the Tower of Hanoi problem (Gagne and Smith, 1962), to improve concept learning (Bower and King, 1967) and to facilitate transfer from a "concrete" to an "abstract" version of the Wason selection task (Berry, 1983). It is possible, therefore, to make two predictions concerning the effects of concurrent verbalization on performance in the present study. Firstly, for subjects who have received verbal instruction, overt verbalization throughout the second set of trials might induce them to implement the newly acquired verbal knowledge. If the verbal instruction directs attention towards certain critical features of the task, the verbalization requirement might keep attention on these salient features and hence result in improved control performance. Secondly, for subjects who have not received verbal training, overt verbalization might change the way in which they carry out the task, hence affecting their control performance and its relationship to verbal knowledge. Experiment 3 explores these possibilities.

EXPERIMENT 3

PERSONAL INTERACTION

Method

The method was very similar to that used in the previous experiment. The only difference was that before beginning the second set of trials, subjects in both the T and NT groups were told, "This time I would like you to think aloud while you are performing the task. Try to give a reason for each response which you make before you type it into the computer." All verbalizations were tape-recorded. Subjects were not asked to verbalize concurrently while completing the post-task questionnaire. The 24 subjects were again paid volunteers from the Oxford University subject panel. None had participated in the previous experiments.

Results

Mean scores were calculated for each set of trials for each of the two groups. These are shown in Table II. For the T group, the means were 5.83 (set 1) and 12.0 (set 2). The equivalent means for the NT group were 7.92 and 8.0. A two-factor analysis of variance showed no overall effect of

training, $F(1, 22) = 0.59$, $p > 0.1$, a highly significant practice effect, $F(1, 22) = 10.06$, $p < 0.01$, and a highly significant interaction, $F(1, 22) = 9.53$, $p < 0.01$. The same pattern of results was observed for performance on the last five trials. Mean questionnaire scores were 4.67 and 1.17 for the T and NT groups, respectively. A Mann Whitney U test showed the difference in questionnaire scores between subjects in the two groups to be highly significant, $U = 0, n1 = n2 = 12, p < 0.0001$. Subjects in the T group in the present experiment also scored higher on the questionnaire than did those in the T group on the person task in Experiment 2, $U = 24$, $n1 = n2 = 12$, $p < 0.01$. Because of insufficient variation in questionnaire scores within each group, it was not possible to analyse these as a function of control performance.

There was a qualitative difference between the concurrent verbalizations of subjects in the two groups. Subjects who had not received verbal instruction tended to talk at a very general level, either simply stating the inputs and outputs or giving reasons such as "He's affectionate, that's too friendly for my liking. I don't like people to be affectionate when you haven't known them for long." In contrast, those in the training group usually gave reasons more in line with the verbal explanation, for example, "VA, so if I come to A, in theory he should come down to VF." This subjective impression was confirmed by three independent judges. These judges were familiar with the task and with the nature of the verbal training given. They were given the unmarked transcribed protocols and were asked to decide whether each protocol came from a trained or an untrained subject. In addition, they were asked to rate their confidence in their decisions and to provide any reasons for their judgements. The judges did not confer. A protocol was classed as coming from either a trained or an untrained subject where at least two of the judges made this decision. Using this criterion, protocols from all 12 trained and from 9 of the untrained subjects were correctly placed by the judges. Mean confidence ratings were higher for correct than for incorrect judgements and higher for trained than for untrained subjects. The two reasons most frequently given for a protocol coming from a trained subject were, firstly, that subjects would predict the outcome of a response before entering it into the computer, and, secondly, that subjects would refer to their and Clegg's respective positions on a single intimacy scale. In contrast, the reason most often cited for a protocol coming from an untrained subject was that of talking in "real-world terms".

SUGAR PRODUCTION

Method

The method here was the same as for the person task, with the exception that all instructions referred to sugar production rather than to personal interaction. The

24 subjects were paid volunteers from the Oxford University subject panel. None had participated in the previous experiments.

Results

Mean scores were calculated for each set of trials for each of the two groups. These are shown in Table II. For the T group, the means were 5.17 (set 1) and 13.33 (set 2). The equivalent means for the NT group were 4.5 and 6.67. A two-factor analysis of variance showed a significant effect of training, $F(1, 22) = 20.84$, $p < 0.01$, a significant practice effect, $F(1, 22) = 34.59$, $p < 0.01$, and a significant interaction, $F(1, 22) = 11.66$, $p < 0.01$. The same pattern of results was observed for performance on the last five trials. Mean questionnaire scores were 3.42 and 1.58 for the T and NT groups, respectively. A Mann Whitney U test showed this difference in questionnaire scores to be highly significant, $U = 14$, $n1 = n2 = 12$, $p < 0.001$. Unlike the person task, questionnaire scores of subjects in the T group in this experiment were not significantly higher than those of the T group in Experiment 2, $U = 64.5$, $n1 = n2 = 12$, $p > 0.1$. Questionnaire scores were correlated with control performance. Spearman rank correlation coefficients were -0.80 ($p < 0.01$) and -0.07 ($p > 0.1$) for the T and NT groups, respectively. Questionnaire scores were also correlated with improvement in performance between sets 1 and 2. Ranks coefficients were -0.53 and -0.22 for the T and NT groups, respectively. Neither was significant at the 0.05 level.

Again, it was possible to differentiate qualitatively between the protocols of subjects who had received the verbal instruction and those who had not. As with the person task, those in the latter group tended to talk at a more general level, either merely describing the inputs and outputs, or making comments such as "The same work force is now producing less. I think a few of them must have gone on strike." Three independent judges (the same as those used for the sugar task) correctly placed protocols from 10 of the trained and 10 of the untrained subjects. Mean confidence ratings were again higher for correct than for incorrect judgements and higher for trained than for untrained subjects. The reason most frequently given for a protocol coming from a trained subject was again that of prediction, namely that subjects would predict the outcome of a response before entering it into the computer. The two reasons most frequently cited for a protocol coming from an untrained subject were, firstly, testing the system (that is, making a response to see what would happen) and, secondly, talking in real-world terms.

Throughout the above experiments, Spearman correlations have been calculated using the Pearson product moment formula with ranked values. Ferguson (1976) suggested that this is a convenient method to use when tied scores are involved. It was therefore possible to average the cor-

relations across experiments using the method suggested by Fisher (1946). The particular correlations used for this were those between questionnaire scores and scores on the immediately preceding full set of trials. Coefficients were available for all conditions except for those in the person task in Experiment 3. Correlations were not calculated in this experiment, as there was insufficient variation in questionnaire scores. A total of 10 coefficients was therefore available, 9 of them negative in value. Averaging the 10 coefficients together produced a small but significant negative correlation between overall control performance and questionnaire score, $r(87) = -0.31, p < 0.001$. There is some evidence, therefore, that individuals who were better at controlling the systems were actually worse at answering the questions.

Discussion

Taken together with the results from Experiment 2, it appears that verbal instruction and concurrent verbalization combined to produce a significant improvement in control performance. Neither factor alone was sufficient to increase performance levels. In contrast, verbal instruction was sufficient to increase post-task questionnaire scores. For the person task, this effect was strengthened when the instruction was combined with concurrent verbalization. For the sugar task, however, concurrent verbalization had no additional effect on questionnaire score over that produced by the instruction. In both cases verbalization alone had no effect on question answering.

Concurrent verbalization was therefore not uniform in its effects. Verbalization alone failed to improve control performance or question answering. The null effect on questionnaire score, together with the zero correlation between these scores and control performance, suggests that the verbalization did not result in access to the type of information necessary to answer the post-task questionnaire. Verbalization combined with prior instruction, however, did lead to a significant improvement in control scores. The critical factor seemed to be the availability of relevant verbal information at the time of verbalization.

One possible explanation for the beneficial effect of the combined instruction and concurrent verbalization is that the verbalization requirement induced subjects to rehearse specific elements of the verbal instruction overtly. Inspection of the protocols, however, shows that although there was a qualitative difference between the verbalizations of the trained and untrained groups, only 7 of the 24 trained subjects explicitly mentioned critical aspects of the instruction such as the importance of the size of the distance between the existing machine output and the new input or the need to choose a response half-way between the present output and the desired output. Moreover, the relatively poor control scores of the

trained subjects in Experiment 2 cannot be attributed to their having forgotten the explanation. They scored significantly higher on the post-task questionnaire than did subjects who had not received the prior training.

An alternative explanation for the combined instruction and verbalization effect is that verbal instruction directed attention towards certain critical features of the task. The subsequent verbalization requirement kept attention on these salient features, and irrelevant aspects were ignored. This account would not necessarily entail overt verbalization of these critical features. This view is consistent with an experiment by Berry (1983) in which subjects attempted to solve a "concrete" version of the Wason selection task. Following an initial trial, they were provided with a minimal explanation of the correct solution. This improved performance on subsequent concrete trials, but there was no evidence of transfer to a logically equivalent abstract task. Concurrent verbalization during the trials following the explanation, however, led to significantly higher scores on the subsequent abstract task. Bainbridge (1979) has also suggested that certain types of verbalization might change the way in which a task is done by forcing concentration on critical task components.

GENERAL DISCUSSION

The experiments reported in this paper have three major implications. Firstly, given that verbal access to task-related information is limited to some degree, assessing post-task knowledge solely by means of written questionnaire will not give a true picture of an individual's competence. We need to develop alternative methods of assessing different forms of post-task knowledge. Experiments currently in progress are addressing this problem. One approach has been to look at transfer to other systems varying in conceptual similarity to the original task.

Secondly, providing an individual with detailed verbal instruction that is understood and later remembered is not necessarily sufficient to improve task performance. A person must be induced to implement this verbal information and incorporate it into control performance. Experiment 3 demonstrates that concurrent verbalization is one successful technique for this. In this situation, however, it is not sufficient merely to ask a person to "think aloud". It is necessary to guide the content of the verbalization, for example by providing preceding verbal instruction.

The final point concerns what these experiments tell us about the hypothesized "performance/verbalizable knowledge" dissociation. Experiments 1 and 2 show that practice significantly improves ability to control the sugar production and personal interaction tasks, but it has no effect on ability to answer related questions. In contrast, verbal instruction significantly improves ability to answer questions, yet it has no effect on control performance. These differential effects, together with the overall

significant negative correlation between control performance and questionnaire score, are theoretically important. They argue against the assumption that performance and verbalizable knowledge might, in fact, be positively related, but that these experiments are somehow not sensitive enough to show this. One interpretation of these findings, therefore, is that these tasks might, under certain conditions, be performed in some implicit manner with individuals not being verbally aware of the basis on which they are responding. Although this interpretation is appealing, we cannot be totally confident that the written questionnaire was an adequate test of subjects' "verbal knowledge". It may be that there are two different types of performance, one corresponding to the control task, the other to the questionnaire. Both could be considered as having verbal and non-verbal components, although the relative weighting would probably differ in the two cases. Whichever interpretation is accepted, it does not seem to be particularly productive to talk in terms of there being a general dissociation between task performance and verbal knowledge. Such a general description cannot convey the full complexity of the experimental results, nor of the likely underlying relationship.

If we accept that these tasks might, under certain conditions, be performed in some implicit manner, the major problem for future research is to discover more about the actual operations involved. By definition, we cannot expect the subjects themselves to have any unique insight. Moreover, we cannot ask them to monitor their future performance, as such an instruction is likely to disrupt the learning process (Reber, 1976; Reber et al., 1980; Berry and Broadbent, in preparation). It is necesary to discover not only what form of representation this implicit knowledge might take, but also the conditions under which we might expect this form of learning to occur. Experiments in progress are addressing these and other issues.

REFERENCES

Bainbridge, L. (1979). Verbal reports as evidence of the process operator's knowledge. *International Journal of Man–Machine Studies*, **11**, 411–436.

Berry, D. C. (1983). Metacognitive experience and transfer of logical reasoning. *Quarterly Journal of Experimental Psychology*, **35A**, 39–49.

Bower, A. C. and King, W. L. (1967). The effect of number of irrelevant stimulus dimensions, verbalisation and sex on learning biconditional classification rules. *Psychonomic Science*, **8**, 453–454.

Broadbent, D. E. (1977). Levels, hierarchies and the locus of control. *Quarterly Journal of Experimental Psychology*, **29**, 181–201.

Broadbent, D. E., Fitzgerald, P. and Broadbent, M. H. P. (1982). Conscious and unconscious judgement in the control of complex systems. Available in electronic form in the British Library experimental electronic journal, *Computer Human Factors*.

Cooke, J. E. (1965). Human decision in the control of a slow response system. Unpublished D.Phil. Thesis, University of Oxford.

Ericsson, K. A. and Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, 87, 215–251.

Evans, J. St. B. T. (1980). Current issues in the psychology of reasoning. *British Journal of Psychology*, 71, 227–239.

Ferguson, G. A. (1976). *Statistical analysis in psychology and education*. New York: McGraw-Hill.

Fisher, R. A. (1946). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Gagne, R. and Smith, E. (1962). A study of the effects of verbalization on problem solving. *Journal of Experimental Psychology*, 63 (1), 12–18.

Jacoby, L. L. and Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General*, 110 (3), 306–340.

Kellogg, R. T. (1982). When can we introspect accurately about mental processes. *Memory and Cognition*, 10 (2), 141–144.

Marcel, A. J. (1983). Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology*, 15, 197–237.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behaviour*, 5, 855–863.

Reber, A. S. (1976). Implicit learning of synthetic languages; The role of instructional set. *Journal of Experimental Psychology: Human Learning and Memory*, 2, 88–94.

Reber, A. S. and Allen, R. (1978). Analogy and abstraction strategies in synthetic grammar learning: A functional interpretation. *Cognition*, 6, 189–221.

Reber, A. S., Kassin, S. M., Lewis, S. and Cantor, G. (1980). On the relationship between implicit and explicit modes of learning a complex rule structure. *Journal of Experimental Psychology: Human Learning and Memory*, 6 (5), 492–502.

Turvey, M. T. (1974). Constructive theory, perceptual systems and tacit knowledge. In W. B. Weimer and D. S. Palermo (Eds.), *Cognition and the symbolic processes*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Wason, P. C. and Evans, J. St. B. T. (1975). Dual processes in reasoning? *Cognition*, 3, 141–154.

Revised manuscript received 20 May 1983