Abstract

In response to criticism of “subjective threshold” definitions of implicit knowledge (e.g. Holender, 1986; Shanks & St. John, 1994), three experiments investigated possible biases of metacognitive measures of implicit knowledge within the artificial grammar paradigm. The effects of probability training, confidence judgements after blocks of test items, training stimuli rate, instructional manipulation and feedback on Cheesman and Merikle’s (1984) guessing criterion, and Chan’s (1992) zero-correlation criterion were observed. Making frequentist judgements after blocks of trials rather than confidence ratings after each trial led to increased measured amounts of implicit knowledge according to the guessing criterion and reduced overall levels of performance; undergoing probability training led to higher measured amount of implicit knowledge according to the guessing criterion and a lower measured amount of explicit knowledge according to the Chan criterion; warnings of “under-confidence” after each trial caused a measured decrease in implicit learning according to the guessing criterion and an increase in explicit knowledge according to the Chan criterion. Rapid presentation of stimuli in the training phase led to a reduction in overall performance, a reduction in the measured amount of both implicit and explicit knowledge. The significance of the results for the validity of metacognitive measures of implicit knowledge is discussed.

1 Introduction

Central to any debate regarding implicit intelligence, natural or artificial, is the need for a valid measure of whether a cognitive process can be considered implicit. Within the artificial grammar paradigm, subjects are frequently instructed to give confidence ratings on a scale from 50 to 100 during their test phase answers, and those confidence ratings are used to determine the degree to which a subject has metaknowledge (knowledge about their own knowledge) (Dienes, Altmann, Kwan, & Goode, 1995; Dienes & Berry, 1997). A confidence rating of exactly 50 shows that the subject considers their response to a test item to have been a guess. The basis of the paradigm is the view that when subjects display knowledge of the grammar through performance levels significantly greater than fifty percent, and confidence ratings indicate that their responses were regarded as pure guesses, then that knowledge may be considered implicit.

Debate has recently surrounded such “subjective threshold” (Cheesman & Merikle, 1984, 1986; Dienes & Berry, 1997) or metacognitive measures of implicit knowledge. Critiques of various implicit knowledge paradigms (e.g. Holender, 1986; Shanks & St. John, 1994) have drawn attention to the possibility that when subjective report is used to determine the degree of metaknowledge a person has about their knowledge, then the reported degree may or may not reflect the true degree of metaknowledge held. Metacognitive measures of implicit knowledge may be subject to systematic biases, and we cannot assume that a subject has no explicit knowledge simply because they claim to have none. In this way, the subjective threshold has been claimed to be an invalid measure of the implicitness of the knowledge.

Inherent to the subjective threshold model is the need to apply a metacognitive measure with which to gauge the extent of metaknowledge. Various metacognitive measures such as the guessing criterion and zero correlation criterion (Chan, 1992; Dienes et al., 1995) are based upon confidence ratings, or measures of subjective probability, which may be biased under specific conditions (Tversky & Kahneman, 1974; Gigerenzer, Hoffrage, & Kleinbolting, 1991; Reingold & Merikle, 1993; Redington & Chater, 1996).

Rather than being cause to abandon the subjective threshold of implicit knowledge, such criticisms make apparent the need for an empirical investigation of
the subjective threshold as a valid measure of implicit knowledge, and suggest a course of empirical research which might resolve the debate surrounding the validity of metacognitive measures. Through systematic manipulation of possible influences of metacognitive measures, it may be possible to identify conditions under which the subject's metacognitive beliefs have been measured in a biased way or else conditions that influence on the subject's actual metacognitive beliefs. We could thereby begin to identify the circumstances under which metacognitive measures provide a valid measure of implicit knowledge.

The aim of the experiments described in this paper was to explore biases identified in the subjective probability literature and apply them to the case of learning artificial grammars. Overconfidence is the most commonly observed judgemental bias in that literature (Lichtenstein, Fischhoff, & Phillips, 1982; Plous, 1993), this being any situation in which a person claims to be n% sure about each of a whole set of statements being true, and less than n% are in fact true (Lichtenstein & Fischhoff, 1977a/1977b; Yates, 1982; Ayton & Wright, 1994). Metacognitive errors, including especially over-confidence (and the less common underconfidence), have been linked to use of simplifying heuristics that validly apply in only circumscribed conditions (Tversky & Kahneman, 1974; Giggenzer et al, 1991). Because of the heuristic nature of metacognitive judgements, the extent of over- or under-confidence can be influenced by various manipulations, including training and feedback (Peterson, Schneider, & Miller, 1965; Lichtenstein & Fischhoff, 1980; Keren, 1987), and taking metacognitive judgements after blocks rather than individual trials (Gigerenzer et al, 1991). We will consider these major manipulations further in relation to artificial grammar learning. Other manipulations affecting metacognitive judgements not considered further here are context and sample representativeness (e.g. Fischhoff, Slovic, & Lichtenstein, 1978; Teigen, 1983), reaction times (e.g. Gentner & Collins, 1981; Baranski & Petusic, 1994; Boucher, 1996), and the difficulty of items (Lichtenstein & Fischhoff, 1977a/1977b; Griffin & Tversky, 1992; Plous, 1993; Baranski & Petusic, 1994).

Three experiments are described below, in which possible biases of metacognitive measures of implicit knowledge were investigated. The four potential variants of probability judgement investigated were frequentist judgements of confidence after blocks of test items rather than after each item (Gigerenzer, 1991), probability training prior to the artificial grammar test phase, manipulation of training stimuli presentation rates, and the inclusion of additional experimental instructions and feedback.

Gigerenzer et al. (1991) and Giggenzer (1994) suggest that humans have evolved to be well calibrated users of relative frequencies. The primary implication of frequentist models such as that proposed by Giggenzer et al. (1991) is that if people are asked how many times out of a number of items they think they are correct, then they will display less overconfidence than when asked to state their degree of confidence after each item. For example, Keren & Wagenaar (1987) have noted that subjects behave differently in situations involving unique rather than repeated gambles.

Whereas Gigerenzer et al (1991) regard the trial by trial confidence judgements to be biased towards overconfidence, Griffin & Tversky (1992) have suggested that frequentist judgements are biased toward underconfidence. In either case a potential bias may exist in metacognitive measures used in the artificial grammar paradigm. Experiment one aimed to investigate the degree of influence that giving frequentist confidence ratings after blocks of items rather than after individual items would have on metacognitive measures of implicit knowledge within the artificial grammar paradigm.

People may estimate frequencies better after blocks of trials rather than probabilities after single trials simply because they understand frequencies better (Gigerenzer et al, 1991). Can subjects be trained to understand probabilities better? Probability training was also investigated in experiment one, in order to assess untrained subjects’ understanding of the probabilistic significance of a confidence rating set at 50. If subjects were to give lower confidence ratings after probability training, then that would suggest that levels of confidence have previously been exaggerated due to misunderstanding of probabilities, and if subjects were to give higher ratings after probability training then that would suggest that true levels of confidence had previously been concealed by such misunderstanding.

Various training string presentation rates are used within the artificial grammar paradigm, and experiment two was run in order to investigate the possibility that strings presented slowly (once every three seconds in this case) might give rise to qualitatively different confidence/accuracy calibration than strings presented at rapid liminal rates. It is possible that rapid stimuli presentation might cause subjects to give lowered confidence ratings without affecting performance levels, although any such difference would arguably be due to implicit learning rather than any kind of bias. Regardless of the nature of the influence, if subjects in experiment two are found to give lower confidence ratings
when training strings are presented rapidly, then that would provide evidence to support the view that training stimuli presentation speeds should be considered an influence on metacognitive measures of implicit knowledge. The aim of experiment three was to investigate the possible effects of instructional manipulation on metacognitive measures, where additional instructions were shown to experimental subjects which stated that subjects are usually “significantly under-confident”. Experiment three's design included “underconfidence warnings”. The “underconfidence warnings” were simple statements informing the subject that their previous confidence ratings had been underconfident, regardless of actual confidence / accuracy calibration for that subject. The instructions and warnings did not include any explicit instruction that the subject should adjust their confidence ratings, but were intended as an indicator of how additional information might affect subject ratings. If the manipulation of instructions were to give rise to higher confidence ratings, then that would suggest that the apparent degree of subject metaknowledge can be influenced by the presence of information which suggests that their ratings are too low.

The two central metacognitive measures employed in this paper were called the guessing criterion of implicit knowledge, and the zero correlation criterion of explicit knowledge by Dienes et al. (1995). The guessing criterion (Cheesman & Merkle, 1984) defines knowledge as implicit when subjects give a confidence rating of 50, which indicates a “guess”. Implicit knowledge is considered to have been found by the guessing criterion when the number of correct responses associated with confidence ratings of 50 is significantly greater than the number of incorrect responses marked as guesses. The zero correlation criterion (Chan, 1992) defines knowledge as explicit when confidence is related to accuracy. Knowledge can be considered to be implicit by the zero correlation criterion if confidence is not related to accuracy. When the average confidence associated with correct responses is significantly greater than the average confidence associated with incorrect responses, then knowledge is considered explicit by the zero correlation criterion.

1.1 Experiment 1
1.1.1 Method
1.1.1.1 Design
A 2 X 2 between-subjects design was used. All participants were trained on a finite-state grammar. The first variable, probability training, refers to whether participants were run through a probability training session before the artificial grammar test phase. The second variable, rating frequency, refers to whether participants gave confidence ratings after each item in the test phase, or after blocks of twenty-five items. The dependent variables for all participants were classification responses and confidence ratings.

1.1.1.2 Participants
Forty-eight volunteers from the University of Sussex were randomly assigned to one of the four groups, such that each group contained 12 participants.

1.1.1.3 Stimuli
The grammar used was taken from Reber and Allen (1978), and used the letter set; M, T, V, R, and X. An arbitrary limit of 4 recursions was placed on the two recursive nodes in the finite-state grammar, which led to the generation of the 40 string set used, plus 1 additional string arbitrarily discarded by Reber and Allen (1978). The 41 strings generated were between three and six letters in length. Twenty representative sequences were selected from the grammar to form the training set. Reber & Allen constructed non-grammatical items by introducing letter substitutions and string reversals. The training and test strings were exactly those used by Dulaney, Carlson, & Dewey (1984). There were 20 grammatical strings presented during the artificial grammar training phase, and 25 grammatical and 25 non-grammatical strings presented during the test phase. The order of strings within each training set was randomised, and training strings were presented in blocks of six sets.

1.1.1.4 Procedure
All participants received identical artificial grammar training phases. Participants were initially told that they would be shown strings of letters on a computer monitor, and that they were to learn and remember the strings.

Participants were then presented with the training strings for 6 minutes, at a rate of 3 seconds per string. At the beginning of the subsequent test phase, participants were informed that the order of letters in each string was determined by a complex set of rules. Participants were informed that half of the test strings would follow the rules, and that half of the test strings would break the rules. Participants were then presented with 100 consecutive test strings, and were instructed to answer, after each string was presented, whether they thought the strings followed the rules.
Participants in the rating frequency control groups were instructed to enter a confidence rating with each classification response, on a scale ranging from 50 (complete guessing) to 100 (complete certainty). Participants could use any integer in the specified range. Only exactly 50 indicated a literal guess. The confidence range offered to subjects (50-100) was chosen on the basis that when guessing, subjects should expect a resultant mean accuracy of fifty percent, and should therefore give a confidence rating of fifty percent when guessing. It is possible that subjects might actually consider their performance to be less accurate than a guess, but this is not logically possible if the subject understands the meaning of the scale.

Participants in the rating frequency experimental groups were told to enter a confidence rating after blocks of 25 classification responses.

Subjects in the probability training experimental groups were run through a probability training session before the artificial grammar test phase, whereas subjects in the probability training control groups underwent only the artificial grammar training and test phases. The probability training session was comprised of a series of questions presented on the computer monitor, in which the subject was asked the probability of particular outcomes in a variety of hypothetical situations.

The first type of question described a bag containing 100 balls, of which a random proportion were coloured red and the rest coloured blue. The subject was asked to estimate the probability that one ball randomly removed from the bag would be red. The first type of question was repeatedly presented until three consecutive correct answers were given.

The second type of question was of the same format, but the number of red balls was set at 50. The second question type was only presented once, but if it was not answered correctly then the subject was returned to begin again with the first question type.

The third and final question type described a multiple-choice test in which each item had two possible answers. The first question of this type asked how many answers would be expected to be answered correctly, out of 1000, if the person taking the test were to guess on every item. The second question of this type presented the same scenario, but with a 300-item test described. If either question was answered incorrectly, the subject was returned to begin the probability training session again. If both questions were answered correctly, the subject was then instructed to begin the artificial grammar test phase.

Percent correct classification was not significantly different between subjects who were given probability training (m = 61, sd = 7) and those who were not trained (m = 63, sd = 5), t (46) = 1.12, ns. Subjects who gave frequentist confidence ratings after blocks of items (m = 61, sd = 5) had significantly lower percent correct classification than subjects who gave confidence ratings after each item (m = 64, sd = 5), t (46) = 2.04, p = .047. Figure 1, below, shows the mean percent correct classification of the various conditions.

One-tailed t-tests were performed on the mean confidence ratings given after each item, and after blocks of items. Confidence ratings given after blocks of items (m = 59, sd = 8) were significantly lower than those given after each item (m = 68, sd = 9), t (47) = 5.60, p < .0005. Subjects who had answered the probability training questions (m = 61, sd = 9) gave significantly lower confidence ratings than control subjects (m = 66, sd = 9), t (94) = 2.77, p = .008. Figure 2, below, shows the mean confidence levels of the various conditions.

1.1.2 Results
Implicit learning was found in the total subject population by the guessing criterion, \( t(47) = 2.89, p = .003 \). Explicit learning was found in the total subject population by the zero correlation criterion, \( t(47) = 7.30, p < .0005 \).

A significant difference was found between the untrained groups and the trained groups by the guessing criterion, \( F(1,94) = 13.17, p < .0005 \). Implicit learning was not found by the guessing criterion in the untrained groups, \( t(23) = 1.22, \text{ns} \), although implicit learning was found by the guessing criterion in the trained groups, \( t(23) = 2.72, p = .006 \).

A significant difference was found between the rating-by-item groups and the rating-by-block groups by the guessing criterion, \( F(1, 142) = 3.69, p = .057 \). Implicit learning was found by the guessing criterion in both the item groups, \( t(23) = 2.72, p = .006 \), and the block groups, \( t(47) = 2.90, p = .003 \).

A significant difference was found between the untrained groups and the trained groups by the zero correlation criterion, \( F(1, 94) = 7.12, p = .009 \). Explicit learning was found by the zero correlation criterion in both the untrained, \( t(23) = 6.64, p < .0005 \), and trained groups, \( t(23) = 4.03, p = .001 \).

**Discussion**

Implicit learning was not found by the guessing criterion in the untrained groups, although implicit learning was found by the guessing criterion in the trained groups. Although explicit learning was found in both the untrained and trained groups, greater explicit learning was found in the untrained group by the zero correlation criterion. Subjects who had participated in the probability training session gave lower confidence ratings than control subjects. The fact that probability trained subjects gave lower confidence ratings suggests that levels of metaknowledge have previously been exaggerated due to misunderstanding of probabilities, and that untrained subjects do not have metaknowledge which has been concealed by such misunderstanding.

Although implicit learning was found in both item and block groups, greater implicit learning was found in the rating-by-blocks groups, and confidence ratings given after blocks of items were lower than those given after each item. That subjects gave lower confidence ratings after blocks of items is a successful replication of Gigerenzer et al. (1991) within the artificial grammar paradigm, and it suggests that the degree of metaknowledge has been exaggerated in studies where ratings are given after each item. The primary implication for the artificial grammar paradigm is that subjects who give confidence ratings after blocks of items will give lower confidence ratings, and apparently evidence greater implicit learning.

Further analyses, such as confidence/accuracy calibration, response consistency (Reber, 1989), and the Goodman & Kruskal (1954) gamma correlation, are required in order to fully gauge the effect of probability training and confidence judgements after blocks of items on metacognitive measures. After the initial analyses were completed, raw subject data files were lost from experiment one, so a further experiment is required to allow a complete analysis.

### 1.2 Experiment 2

#### 1.2.1 Method

##### 1.2.1.1 Design

A between-subjects design was used. All participants were trained on one of two artificial grammars. The between-subjects variable, stimuli speed, refers to the rate at which subjects in the control and experimental groups were shown the artificial grammar training items. The dependent variables for all participants were classification responses and confidence ratings.

##### 1.2.1.2 Participants

Thirty-eight volunteers from the University of Sussex were randomly assigned to one of the four groups, with sixteen subjects in the control condition and twenty two subjects in the experimental condition.

##### 1.2.1.3 Stimuli

The grammars used were adapted from Reber and Allen (1978) and are shown in Figure 2. The initial string gen-

*fig. 2.* Mean confidence by condition.
oration method was the same as that in experiment one, and the strings were then modified to create two mutually exclusive sets, where grammatical strings generated by one finite-state grammar would be judged non-grammatical if considered to have been generated by the other grammar. In grammar A, the initial legal bigrams were MT, MV, RT, RV, VX, VM, XX, and XM. In grammar B, the initial legal bigrams were XT, XV, VT, VV, RX, RM, MX, and MM. Apart from the initial letter, the grammar was the same as that used by Reber & Allen (1978).

1.2.1.4 Procedure

All participants received artificial grammar training phases. Participants were initially told that they would be shown strings of letters on a computer monitor, and that they were to learn and remember the strings while copying each string on a piece of paper when it appeared. Participants were then presented with the training strings for 10 minutes.

Subjects in the experimental groups (i.e. those in groups 2 and 4) were shown the training strings at a rate of 3 seconds per string. Subjects in the control groups (i.e. those in groups 1 and 3) were shown the training strings at a rate of 0.018 seconds per string. Subjects in groups 1 and 2 were shown a different artificial grammar (grammar A) to subjects in groups 3 and 4 (grammar B). Test phase responses that were correct for grammar A were incorrect for grammar B, and vice versa. Details of the test phase, including instructions, number of test items, and the confidence rating scale used, were identical to those in experiment one.

1.2.2 Results

Percent correct classification (m = 53, sd = 5) was significantly greater than chance, t (37) = 3.10, p = .004, which indicates the presence of learning in the total subject population. Percent correct classification of the rapid stimuli group (m = 49, sd = 4) was significantly lower than the percent correct classification of those subjects shown the slower stimuli (m=56, sd = 5), t (36) = 4.53, p < .0005. Figure 3, below, shows the mean percent correct classification of the two conditions.

Subjects in the experimental group gave significantly higher confidence ratings (m = 71, sd = 7) than those in the control group (m = 62, sd = 11), t (50) = 3.86, p < .0005.

The mean proportional confidence of the total subject population (m = .71, sd = .16) was significantly greater than the mean proportion correct (m = .56, sd = .23), t (167) = 7.99, p < .0005, indicating overconfidence.

The mean proportional confidence of subjects in the control condition (m = .67, sd = .14) was significantly greater than the mean proportion correct (m = .52, sd = .25), t (55) = 4.44, p < .0005, indicating overconfidence in subjects who were shown the rapid training stimuli. The mean proportional confidence of subjects in the experimental condition (m = .73, sd = .17) was significantly greater than the mean proportion correct (m = .58, sd = .22), t (111) = 6.63, p < .0005, indicating overconfidence in subjects who were presented the training stimuli at the rate of 3 seconds per string.

The mean difference between confidence and accuracy in the control condition (m = .16, sd = .26) was not significantly different from the mean confidence/accuracy difference in the experimental condition (m = .16, sd = .25), t (166) = .05, ns, which indicates that mean proportional overconfidence was equal between the two conditions. Figure 4, below, shows the confidence/accuracy calibration of subjects in both conditions.
Implicit learning was not found in the total subject population by the guessing criterion, \( t (37) = .61 \), ns. Explicit learning was found in the total subject population by the zero correlation criterion, \( t (37) = 3.31, p = .001 \).

A significant difference was found between the control group and the experimental group by the guessing criterion, \( F(1,74) = 6.64, p = .012 \). Implicit learning was not found by the guessing criterion in the control group, \( t(15) = 2.02, \) ns, whereas implicit learning was found by the guessing criterion in the experimental group, \( t (21) = 2.27, p = .017 \).

A significant difference was found between the control group and the experimental group by the zero correlation criterion, \( F(1,74) = 16.90, p < .005 \). Explicit learning was not found by the zero correlation criterion in the control group, \( t (15) = 1.68, \) ns, but explicit learning was found by the zero correlation criterion in the experimental group, \( t (21) = 3.05, p = .003 \).

### 1.3 Discussion

Explicit learning was found in the total subject population by the zero-correlation criterion. Implicit and explicit learning were not found in the control group by the guessing and zero-correlation criteria, whereas both implicit and explicit learning were found in the experimental group. Experimental subjects gave higher confidence ratings than control subjects. Subjects who were presented with slow training stimuli showed higher confidence ratings, implicit and explicit learning, whereas those subjects presented with rapid stimuli appear to have diminished confidence and performance levels.

Although the proportional degree of overconfidence was equal between the two conditions, percent correct classification was found to be lower in subjects who had been presented with the rapid stimuli. Such a lowering of confidence and performance levels suggests the possibility that rapid training stimuli presentation causes an impairment of performance, and that subjects are to some degree aware of that impairment, perhaps inferentially. Further experimentation might explore a range of stimuli presentation rates, in order to discover whether there are 'intermediate' rates at which confidence levels are lowered, but performance remains unimpaired.

### 1.3.1 Method

#### 1.3.1.1 Design

A between-subjects design was used. All participants were trained on one of two artificial grammars. The between-subjects variable, instructional manipulation, refers to whether participants were given additional instructions before the artificial grammar test phase. The dependent variables for all participants were classification responses and confidence ratings.

#### 1.3.1.2 Participants

Fifty-one volunteers from the University of Sussex were randomly assigned to one of the four groups, with twenty eight subjects in the control condition and twenty three subjects in the experimental condition.

#### 1.3.1.3 Stimuli

The grammars used were adapted from Reber and Allen (1978), and were the same as those employed in experiment two. The initial string generation method was the same as that in experiment one, and the strings were then modified to create two mutually exclusive sets, where grammatical strings generated by one finite-state grammar would be judged non-grammatical if considered to have been generated by the other grammar.

#### 1.3.1.4 Procedure

The basic procedure employed was the same as in experiment two, except that all subjects were shown the training strings at a rate of 5 seconds per string. Subjects in the experimental groups (i.e. those in groups 2 and 4) were told before the test phase that participants are usually "significantly under-confident", whereas subjects in the control groups (i.e. groups 1 and 3) were given no instructions other than those for the artificial grammar training and test phases. Experimental subjects were also presented with a warning after each test item, appearing on the computer monitor between the classification response and confidence rating, which informed
the subject that their previous ratings had been “under-confident” (regardless of actual confidence/accuracy calibration).

1.3.2 Results

Percent correct classification (m = 57, sd = 7) was significantly greater than chance, t (50) = 7.20, p < .0005, which indicates the presence of learning in the total subject population. Percent correct classification of the control group (m = 57, sd = 7) was not significantly different from the percent correct classification of the experimental group (m=57, sd = 7), t (49) = .22, ns. Figure 5, below, shows the mean percent correct classification of the two conditions.

Confidence ratings given by the experimental subjects (m = 76, sd = 10) were significantly higher than those given by the control subjects (m = 71, sd = 9), t (100) = 2.65, p = .005.

The mean proportional confidence of the total subject population (m = .75, sd = .17) was significantly greater than the mean proportion correct (m = .60, sd = .21), t (193) = 9.23, p < .0005, indicating overconfidence.

The mean proportional confidence of subjects in the control condition (m = .74, sd = .18) was significantly greater than the mean proportion correct (m = .66, sd = .19), t (81) = 3.77, p < .0005, indicating overconfidence. The mean proportional confidence of subjects in the experimental condition (m = .75, sd = .17) was significantly greater than the mean proportion correct (m = .56, sd = .21), t (111) = 9.19, p < .0005, indicating overconfidence in subjects who were shown the underconfidence warnings and feedback.

The mean difference between confidence and accuracy in the control condition (m = .009, sd = .21) was not significantly different from the mean confidence/accuracy difference in the experimental condition (m = .18, sd = .21), t (192) = 3.16, ns, which indicates that mean proportional overconfidence was equal between the two conditions.

Implicit learning was not found in the total subject population by the guessing criterion, t (36) = .91, ns. Explicit learning was found in the total subject population by the zero correlation criterion, t (50) = 4.34, p < .0005.

A significant difference was found between the control group and the experimental group by the guessing criterion, F(1,72) = 6.46, p = .014. Implicit learning was found by the guessing criterion in the control group, t (16) = 2.09, p = .027, but implicit learning was not found by the guessing criterion in the experimental group, t (19) = 1.68, ns.

A significant difference was found between the control group and the experimental group by the zero correlation criterion, F(1,100) = 7.02, p = .009. Explicit learning was found by the zero correlation criterion in the control group, t (27) = 2.94, p = .004, and explicit learning was found by the zero correlation criterion in the experimental group, t (22) = 3.29, p = .002. Figure 6, below, shows the confidence/accuracy calibration of subjects in both conditions.
1.3.3 Discussion

Implicit learning was found in the control group by the guessing criterion, but not in the experimental group, and a difference was found between the control and experimental groups by the zero correlation criterion, although explicit learning was found in both conditions. Confidence ratings given by subjects who had been presented with the underconfidence instructions and feedback were higher than those given by control subjects, even though percent correct classification and mean proportional overconfidence was equal between the two conditions.

The effect of the manipulation on a range of metacognitive measures suggests that the degree of subject metaknowledge may be exaggerated or enhanced when the additional “underconfidence” information is made available to the subject, and that the metacognitive measures appear to be sensitive to underconfidence information in the experimental environment. The result has practical importance for the artificial grammar paradigm, because a large number of subjects run in the experiments are students who may have some broad prior knowledge of confidence effects from introductory psychology courses. Further experimentation might ascertain which elements of the experimental manipulation were primarily responsible for the increase in subject confidence, and how subtle the instructional manipulation can be, while still exerting an influence on metacognitive measures.

2 Conclusions

The criterion of metaknowledge introduces a psychologically interesting distinction between implicit and explicit knowledge in both natural and artificial systems (Clark & Karmiloff-Smith, 1993; Dienes & Perner, 1996). Once such a criterion of metaknowledge has been adopted in order to distinguish between implicit and explicit knowledge, establishment of the validity of metacognitive measures of implicit knowledge is not only of central importance in any discussion of implicit knowledge in humans, but also in discussion of any system which has the capacity for metaknowledge.

The aim of the experiments described in this paper was to investigate manipulations of possible biases which might provide alternative explanations to implicit learning for the differences between subject confidence and accuracy under certain circumstances. The results illustrate situations in which metacognitive measures can be influenced within the artificial grammar paradigm, indicating a constraint on the valid application of such measures.

All of the four possible manipulations investigated had some influence on confidence ratings, Cheesman and Merikle’s (1984) guessing criterion, or Chan’s (1992) zero-correlation criterion of explicit learning.

Making frequentist judgements after blocks of 25 items and undergoing probability training both caused subjects to give lower confidence ratings. Making frequentist judgements after blocks of trials rather than confidence ratings after each trial led to increased measured amounts of implicit knowledge according to the guessing criterion and reduced overall levels of performance. Undergoing probability training led to higher measured amount of implicit knowledge according to the guessing criterion and a lower measured amount of explicit knowledge according to the Chan criterion.

Warnings of “underconfidence” caused confidence ratings to rise. Such warnings, after each trial, caused a measured decrease in implicit learning according to the guessing criterion and an increase in explicit knowledge according to the Chan criterion. Rapid presentation of stimuli in the training phase led to a reduction in overall performance, reduction in confidence, and a reduction in the measured amount of both implicit and explicit knowledge. Further analysis of the data will investigate any effects the manipulations may have had on measures such as response consistency (Reber, 1989) and the gamma statistic (Goodman & Kruskal, 1954).

The measured extent of metaknowledge a subject has depends on two factors: The actual metaknowledge the subject has and how the subject expresses that metaknowledge on a scale, i.e. in the experiments here, converts it into a number. To the extent that a manipulation affects the conversion phase we can talk of biases in the measurement process of metaknowledge (measurement biases). On the other hand, if a manipulation affects the actual degree of metaknowledge, the manipulation does not reveal biases in the measurement of metaknowledge, but conditions under which the metaknowledge is itself biased (metacognitive biases).

These two types of biases are fundamentally different when it comes to using metacognitive measures to interpret the implicit status of knowledge. Measurement biases interfere with the measurement of the extent to which knowledge is implicit. It is important to establish the conditions under which the measurement is not biased.

Presumably, for example, training in probability affects measurement biases - if subjects understand the
scale better, they can use it to reflect their own metaknowledge states more accurately. But it is possible that understanding the nature of probability may also actually influence the true extent of one’s metacognition. That is, metacognitive judgements may be cognitively penetrable by explicit knowledge, making the very implicit or unconscious status of one’s knowledge states dependent on what other explicit beliefs one has:

The process of measurement can change that which is measured in a Heisenbergian fashion. We suggest that this is indeed the case. After probability training the estimates of the amounts of implicit and explicit knowledge are relatively unbiased estimates, albeit estimates of amounts that have possibly changed by the measurement process itself. Similarly, frequency judgements may be less biased estimates according to Gigerenzer, but surely they are less sensitive than trial by trial judgements in the current paradigm because they require the subject to reflect on numerous past mental states. Thus, trial by trial judgements in suitably trained subjects provide the best combination of sensitivity and unbiased estimates of metaknowledge.

A third possibility is that subject mood affects access to knowledge, and that the manipulations investigated in this paper have a demoralising effect on the participants, thereby altering their performance and confidence. Further research is required to assess the impact of the experimental manipulations on subject mood, knowledge, and metaknowledge.

We believe there is a learning system that produces knowledge that is largely unconscious and another learning system that produces knowledge that is largely conscious, the precise conscious status of each piece of knowledge varying with the other explicit beliefs of the subject (Dienes & Perner, 2001).

References

P. Ayton and G. Wright. Subjective probability: what should we believe? In G. Wright & P. Ayton (Eds.), Subjective probability. Chichester: John Wiley & Sons, 1994


G. Gigerenzer. Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In G. Wright & P. Ayton (Eds.), Subjective probability. Chichester: John Wiley & Sons, 1994


S. Lichtenstein and B. Fischhoff. Do those who know more also know more about how much they know? The calibration of probability judgments. Organizational Behavior and Human Performance, 3:552-564, 1977a

S. Lichtenstein and B. Fischhoff. Do those who know more also know more about how much they know? Organizational Behavior and Human Performance, 20:159-183, 1977b


