Confidence Judgements, Performance, and Practice, in Artificial Grammar Learning

Martin Redington, Matt Friend and **Nick Chater**
Department of Experimental Psychology
University of Oxford
South Parks Road
Oxford, U.K.
OX1 3UD
{fmr@sable,matt.friend@magd,nick@psy}.ox.ac.uk

Abstract

Artificial grammar learning is noted for the claim that subjects are unaware of their knowledge. Chan (1992) and Dienes *et al.* (in press) have demonstrated that subjects are unaware in the sense that they lack meta-knowledge. Dissociations between subjects' performance and their confidence in their decisions suggest that the learning mechanism may be in some sense encapsulated from the "confidence system". Here we tested the alternative hypothesis that the confidence system is initially poorly calibrated, or does not know which aspects of the learning mechanism to attend to, by training and testing subjects over four weekly sessions. On all four weeks we found a strong, near-perfect association between confidence and performance for trained subjects, but a dissociation for untrained control subjects. We discuss possible explanations for these results, and previously observed dissociations.

Artificial grammar learning is notable for the controversial claim that subjects acquire knowledge which allows them to distinguish strings which follow the same rules as a set of previously memorised strings, from those which do not, but that they are not consciously aware of this knowledge (Reber, 1967, 1989). However, measuring subjects' conscious knowledge is plagued with the problems of finding suitably sensitive, explicit tests, and of ensuring that the knowledge these explicit tests measure is the same as the knowledge subjects use to perform the classification (see Shanks and St. John, 1995, and commentaries).

Chan (1992) and Dienes, Altmann, Kwan and Goode (in press) set issues of consciousness aside and focus instead on behaviour: Subjects' ability to make confidence judgements about their performance. This tests meta-knowledge, or "what subjects know about what they know". Chan (1992) showed that subjects' confidence in their judgements was unrelated to the likelihood that those judgements were correct. Similarly, Dienes *et al.* (in press) showed that even when subjects thought that they were guessing, their performance was above chance (and untrained control) levels.

Chan (1992) and Dienes *et al.* (in press) propose that the dissociation between confidence and accuracy is evidence that subjects lack meta-knowledge. This suggests that the learning mechanism is to a certain extent encapsulated from the "confidence system", so that its inner workings (*e.g.* the strengths of its rules, or connection weights, or its error signals) are inaccessible.

In this paper, we present a study which aimed to test a possible alternative explanation for this dissociation; that meta-knowledge is in principle available, but simply that extraction of this knowledge requires learning or calibration, and is initially inaccurate.

Dienes and Perner's (in press) discussion of the extent and manner in which neural networks possess meta-knowledge should clarify this point. They suggest that if the "confidence system" is able to observe, and knows the significance of, the output activation of the learning mechanism (which in a network might indicate the extent to which the current test string is consistent with the training strings, *e.g.* Dienes, 1992), then confidence and performance should be correlated. However, if the confidence system can only observe whether the learning mechanisms' output is on or off (where the "better" the test string, according to the learning mechanism, the more likely the output is to be activated), then confidence and accuracy will be unrelated.

In these terms, our alternative hypothesis suggests that the confidence judgement mechanism might be initially ignorant or uncertain of the identity and/or significance of the learning mechanism's output, but that with practice, or the opportunity to observe the learning mechanism more closely, the confidence judgement mechanism will be able to accurately identify the learning mechanism's output.

Dienes and Perner also suggest that if the learning mechanism's output was only accessible to the confidence system on a transformed scale, so that the cut-off for responding "grammatical" was unknown, then accurate confidence judgments could be made, but only by comparing the current transformed values to previous values (*e.g.* if the value is low compared to previous ones, respond 'grammatical' with low confidence, or 'non-grammatical' with high confidence). However, initially, when the sample of previous values is small, confidence may be relatively inaccurate. Extensive experience may allow the construction of a sufficiently

large sample that accurate confidence judgements can be made.

Previous dissociations between confidence judgements and performance were observed during a single test session. If the learning mechanism is genuinely opaque to the confidence system, then these dissociations should be maintained across multiple test sessions. Alternatively, if the learning mechanism is in principle transparent, then with practice, confidence may come to accurately reflect performance.

We tested these conflicting hypotheses by training and testing subjects over a four-week period. The paradigm used was the guessing game (Redington & Chater, 1994). Here, prior to making a grammaticality, and confidence, judgement about each string, subjects are first required to reconstruct the string, guessing each letter until it is correctly identified, and then proceeding to the next letter. This provides a detailed measure of subjects' knowledge of the possible continuations at each point. There were two conditions: one group of subjects memorised the training strings each week, prior to testing, while an untrained control group never saw the training strings.

Method

Design. This was a 2 × 4 × 2 mixed design. The between-subjects factor was training, with subjects randomly assigned to the trained or untrained condition. The within-subjects factors were Week (1–4), and Non-Grammatical String Violation Type (non-grammatical strings contained either non-permissible pairs and non-permissible triples).

Materials. The stimuli were exactly those used by Gomez and Schvaneveldt (1994). There were 18 training strings, and 51 test strings, of which 17 were grammatical, 17 were non-grammatical because they contained illegal pairs of letters (non-permissible pairs; NPP), and 17 were non-grammatical consisting entirely of legal pairs of letters, but in illegal combinations (non-permissible triples; NPT[1]). All the strings were between 3 and 8 letters long. See Gomez and Schvaneveldt (1994) for details of the grammar and exact strings used. The experiment was run on Apple PowerPC's.

Subjects. The 20 subjects were undergraduate or postgraduate students at Oxford University. A small (£20) payment was made for their participation. One subject (assigned to the control condition) did not attend for the final week's session, due to illness.

Procedure. Subjects performed an identical procedure each week for four weeks. The weekly procedure was closely modelled on that used in Redington and Chater (1994). It was stressed before each session that subjects should pay care and attention to the task. Trained subjects then saw the following instructions;

> This is a simple memory experiment. When you press the button labelled 'Start', you will see 18 strings, constructed from 5 different letters. The items will run from three to eight letters in length. Your task is to learn and remember as much as possible about all 18 items. You have 10 minutes. If you have any questions about the task, please ask the Experimenter now. Press "Start" when you are ready to begin.

The learning strings were displayed for 10 minutes, in three left-justified columns of six strings each. The order of the strings was randomised separately for each S. Trained subjects then saw the following instructions:

> The order of letters in the set you saw was determined by a rather complex set of rules. The rules allow only certain letters to follow other letters. Now you will be presented with a set of test strings. Some of these obey the same rules as the the training strings, and some violate these rules in some way. For each test string, you must guess the letters of the string, one at a time, and then indicate whether it obeys the rules or not. You guess letters by pressing the button corresponding to the letter which you think comes next. If your guess is correct, the letter will appear on the screen, and you can proceed to the next letter. If your guess is incorrect, the button will disappear, and you must take another guess. The button labelled 'End' is for guessing that the string is complete. The strings are all between 3 and 8 letters long.
>
> Once you have completed the item, two more buttons will appear, labelled 'Correct' and 'Incorrect'. If you think the item that you have just guessed follows the same rules as the original items, then press 'Correct'. If you think it violates those rules then press 'Incorrect'.

Untrained subjects performed only the test phase. Their instructions closely followed this above, except that they commenced:

> You will be presented with a set of test strings. Some of these obey a certain set of rules, which dictate which letters can follow other letters. For each test string...

These instructions were reiterated verbally before the subjects commenced the test phase. Subjects were also told that after each judgement, they would be asked to rate how confident they were that their decision was correct, on a scale from 50% (guessing) to 100% (absolutely certain).

The test display initially showed a (blank) string display, centred on the screen, and below this, a row of five guessing buttons, labelled from left to right with the appropriate letters (in random order) and 'End'. Subjects guessed by clicking on the appropriate button with the mouse. Following a wrong guess (*i.e.* not matching

[1] Gomez and Schvaneveldt (1994) refer to this kind of error as "non-permissible location". Here we use the more mnemonic term, following Gomez (1996).

the next letter of the current item), the button disappeared. If their response was correct, then the letter was appended to the string display, and all the guessing buttons reappeared for the next letter to be guessed. The 'End' button acted in an identical manner to the other guessing buttons; an 'End' guess was correct if the item was otherwise complete (all its letters had been guessed), and incorrect otherwise. Following a correct 'End' guess, the guessing buttons disappeared, the string display was centred, and two buttons labelled 'Correct' and 'Incorrect' appeared to the right of the string display. After the subject responded by clicking one of these, a dialog box appeared, asking "How certain are you that your judgement is correct?". A drop-down menu allowed the subjects to indicate 50, 60, 70, 80, 90, or 100%. There was no default value; subjects had to indicate a confidence value before proceeding. Once this was done, the display was reset for the next test string.

The 17 grammatical and 34 non-grammatical test items were each presented twice, resulting in 102 trials.

Results

Grammaticality Judgements

Grammaticality judgement scores were assessed in terms of *violation sensitivity*: The proportion of non-grammatical items correctly classified minus the proportion of grammatical items incorrectly classified (correct rejections − misses, see Gomez & Schvaneveldt, 1994). Violation sensitivity was calculated separately for NPP and NPT type violations. Summary statistics for violation sensitivity are shown in Table 1.

	Week			
	One	Two	Three	Four
Trained:				
NPP	.30 (.14)	.77 (.12)	.84 (.13)	.82 (.11)
NPT	.16 (.16)	.24 (.28)	.26 (.23)	.33 (.25)
Control:				
NPP	.23 (.18)	.36 (.19)	.39 (.21)	.38 (.22)
NPT	.10 (.08)	.11 (.10)	.13 (.10)	.18 (.14)

Table 1: Mean Violation sensitivity by Group, Violation Type (NPP and NPT), and Week. Standard deviations are shown in parentheses.

A three-way ANOVA comparing violation sensitivity, with Group (trained or control) as a between-subjects variable, and both Week and Violation Type (NPP or NPT) as within-subjects variables indicated that *all* main effects and interactions were reliable. Trained subjects outperformed controls ($F(1,17) = 19.95, p = 0.0003, MS_e = 0.10$); subjects were more sensitive to NPP than to NPT type violations ($F(1,17) = 63.94, p = 0.0001, MS_e = 0.06$), but this difference was less marked in trained subjects ($F(1,17) = 10.06, p > 0.006, MS_e = 0.06$); subjects in both groups improved over weeks ($F(3,51) = 28.02, p = 0.0001, MS_e = 0.02$); trained subjects improved at a faster rate than controls ($F(3,51) = 7.00, p = 0.0005, MS_e = 0.02$); the increased sensitivity to NPP type violations changed over the weeks ($F(3,51) = 12.46, p = 0.0001, MS_e = 0.01$); there was a reliable Violation Type × Group × Week interaction ($F(3,51) = 3.59, p > 0.02, MS_e = 0.01$).

As in previous guessing game studies (Redington & Chater, 1994) untrained control subjects performed reliably above chance (the lower 95% confidence limit of violation sensitivity was above zero on all four weeks, for both kinds of violation).

For our present purpose, these results serve to confirm that the main experimental manipulations (training, practice, and violation type) have had the expected effect on subjects' grammaticality judgement (performance is improved by training and practice, with more subtle non-grammatical violations being harder to detect). Given this, we can be relatively confident that any effects on meta-knowledge are genuinely due to these manipulations.

The Guessing Game

Subjects' guessing game performance was assessed in terms of $\hat{H}$, the amount of information in their guesses. The less that a subject has learnt during the training phase, the more guesses (and thus feedback) they will take to reconstruct the test items. $\hat{H}$ is a measure of the amount of information in the feedback (via an ingenious argument of Shannon, 1951):

$$\hat{H} = -\sum_{i=1,2,\ldots,n} \hat{p}_i \log_2(\hat{p}_i) \qquad (1)$$

where $\hat{p}_i$, the estimated probability that the subject will require i guesses to identify an element of the sequence, is derived from observed relative frequencies of i guesses being required.

A three-way ANOVA comparing $\hat{H}$, with Group as a between-subjects variable, and both Week and Grammaticality (grammatical, NPP, or NPT) as within-subjects variable revealed effects which predictably paralleled those in the grammaticality judgement data: all subjects needed reliably less feedback to reconstruct grammatical strings ($F(2,34) = 230.37, p = 0.0001, MS_e = 0.01$), but this advantage was greatest for the trained subjects ($F(2,34) = 7.46, p = 0.0021, MS_e = 0.01$); subjects required less feedback in later weeks ($F(3,51) = 102.34, p = 0.0001, MS_e = 0.01$), but the improvements were reliably greater for trained subjects ($F(3,51) = 10.26, p = 0.0001, MS_e = 0.01$), and for grammatical strings ($F(6,102) = 9.92, p = 0.0001, MS_e = 0.004$). There was no reliable Group × Week × Grammaticality interaction for the guessing data ($F(6,102) = 1.55, p = ns, MS_e = 0.004$).

The main effect of Group (trained or control) was reliable on a 1-tailed test ($F(1,17) = 3.09, p < 0.05, MS_e = 0.21$).

These findings closely mirror those for grammaticality judgements (as in previous studies with the guessing game), and serve as further confirmation that the experimental manipulations of training, practise, and violation type did have the predicted effects.

Meta-Knowledge

Subjects' meta-knowledge was assessed according to the guessing criterion (Cheeseman & Merikle, 1984, Dienes *et al.*, in press), and the extent to which confidence was correlated with accuracy (Chan, 1992; Dienes *et al.*, in press).

The Guessing Criterion. According to Cheeseman and Merikle (1984), if subjects score above chance, when they claim to be guessing, then they lack meta-knowledge. Violation sensitivity was calculated for those trials on which subjects rated their confidence at 50%. This constituted only 5% of all judgements with the untrained group making more than the trained group (7% vs. 3%), and both groups making fewer over the four weeks).

There was no indication that the trained subjects performed above chance when they claimed that they were guessing. Violation sensitivity did not differ reliably from zero on any week, for either NPP or NPT type violations (by 1-group t-tests, all p's > 0.05), and 6 of the 8 values were numerically below zero.

There were some indications that control subjects performed above chance when they claimed they were guessing. Violation sensitivity was reliably above zero on Week 4, for NPP type violations ($M = .41, t(8) = 2.17, p < 0.05$), and for NPL type violation, on Weeks 1 and 4 the effect was marginally reliable ($M = .13, t(9) = 1.57, p = 0.075$, and $M = .28, t(8) = 1.77, p = 0.058$). However, in general there was no indication of a consistent, reliable effect.

The Zero-Correlation Criterion. Chan (1992) proposed that if subjects possessed meta-knowledge, then confidence and accuracy should be correlated.

Dienes *et al.* (in press) suggest that instead of testing for a correlation, the difference between subjects' confidence in correct and incorrect judgements can be used as a measure of subjects' meta-knowledge. If this value is reliably greater than zero, then subjects were more confident in correct decisions, and did possess meta-knowledge.

We rejected this approach for the current data, as it doesn't take account of response bias[2]. Instead, we looked at how subjects' violation sensitivity was related to confidence. Table 2 shows the Pearson's correlation coefficient between confidence and mean violation sensitivity by Group, Violation Type, and Week.

	Week			
	One	**Two**	**Three**	**Four**
Trained:				
NPP	.93*	.95*	.96*	.99*
NPT	.75*	.84*	.97*	.56
Control:				
NPP	.21	.29	-.11	.11
NPT	-.87*	-.18	-.17	-.58

Table 2: The correlation coefficient between confidence level and mean violation sensitivity, for trained subjects. The criterion value for a 1-tailed test is 0.73 (see Bruning and Kintz, 1977, p. 174) and values in excess of this are marked *.

These results suggest strongly that as trained subjects' confidence increased, so did their sensitivity to both types of violations. In other words, we found no evidence for a dissociation between confidence and accuracy. There is no indication that the strength of this association increased over the four weeks. The failure to find a reliable correlation on Week 4 for NPT type violations can be reasonably considered as an anomaly, given the strong, highly reliable, positive relationships otherwise observed.

In the results for untrained controls (see Table 2) we see no evidence of an association between confidence and performance (and obviously no indication of an improvement over weeks). Indeed, the only reliable correlation is negative, for NPT type violations on Week 1; the more confident subjects were, the less sensitive they became.

Discussion

By both the guessing and zero-correlation criteria, our results indicate that trained subjects possessed considerable meta-knowledge, on all four weeks of testing. Thus under some conditions, subjects can make accurate confidence judgements about their performance in artificial grammar learning.

How can we reconcile this observation with previously observed dissociations between confidence and performance? One obvious explanation is that the guessing procedure provided a basis for subjects' confidence judgements, which was absent in the standard, "grammaticality judgement only" procedure. Thus subjects might be confident that a string whose letters were relatively easy to guess was grammatical, and confident that

[2]For instance, by Dienes *et al.*'s measure, subjects possessed *negative* meta-knowledge of NPT type violations, as they were less confident when responding non-grammatical to these items. However, at high levels of confidence, subjects were less likely to misclassify grammatical items as non-grammatical, so violation sensitivity to NPT violations nevertheless increased with confidence.

a string which was relatively difficult to reconstruct was non-grammatical. This is a plausible explanation, and some replication of this study without the guessing game paradigm is obviously required.

Dienes *et al.* (in press: Experiment 2) demonstrated a similar case, when subjects had to say which of three test items were grammatical, in a forced choice procedure. Dienes *et al.* point out that the learning mechanism might be strongly encapsulated, with a binary output, but that meta-knowledge could still be *inferred*, simply by being more confident in a choice when it was the only "grammatical" string, according to the learning mechanism.

In the present case, we do not find this a completely convincing explanation. Our choice of the guessing game paradigm was intended to maximise both the amount of attention that subjects paid to the test strings, and to encourage the "confidence system" to focus on relevant aspects of the learning mechanism's output. But the situation here is quite different from that of Dienes *et al.* Even in a standard "grammaticality judgement only" situation, one can imagine that subjects might play an "internal" guessing game, assessing strings on the basis of how unexpected each successive letter is. Indeed, this is how some computational models of artificial grammar learning, simple recurrent networks, function (see Berry & Dienes, 1993), and this prediction information is available to subjects, as guessing game performance demonstrates. The guessing game therefore does not appear to provide additional information, that subjects' might not possess similar amounts of meta-knowledge in its' absence.

A second possible explanation for the positive association is that trained subjects engaged in rule-searching behaviour, rather than implicit learning. Chan (1992) found an association between confidence and performance, in subjects given rule-searching instructions during training. However, whilst this possibility may apply to later weeks, where subjects were aware of the rule-governed nature of the strings, in Week one, when the subjects were naive, a strong association was still observed.

A third possibility is that there is some subtle motivational or procedural factor, which differs between those studies where confidence and performance are associated (Manza & Reber, cited in Dienes *et al.*, in press, and the present study), and those where a dissociation has been observed (Chan, 1992; Dienes *et al.*, in press). For instance, it may be that subjects in the Manza and Reber study were somehow encouraged to play an "internal guessing game" as suggested above, whilst those in Chan's and Dienes *et al.*'s studies made their confidence judgements on some much less reliable basis.

This factor might lead subjects to infer meta-knowledge that is not available from the learning mechanism directly (as Dienes and Perner suggest), or it may cause the confidence system to focus on the appropriate aspects of the learning mechanism (the alternative hypothesis that the present study was intended to test).

Some support for the effect of motivational factors on confidence judgements comes from our control groups results. These subjects did perform reliably above chance, but showed some tendency towards above chance classification when their confidence was 50%, and a clear dissociation between confidence and performance by the zero-correlation criterion. If the guessing game was responsible for the association observed in trained subjects, then we should observe a similar association in controls.

Of course, the range of guessing performance is lower for controls, and so confidence judgements which were cued by guessing performance might be less strongly associated with grammaticality judgement performance. However, this would not account for the reversal of the relationship between confidence and performance that we find for control subjects with NPT type strings; the more confident they were, the less sensitive to NPT type violations they became. The main motivational difference between the two groups was that having seen no training strings, the control subjects had no good external reason to believe that they could accurately classify the test strings.

To conclude, it appears that under some conditions confidence and performance may be highly associated, whilst in others, a clear dissociation may be observed. Gaining a clear understanding of what influences this relationship may allow us to draw strong inferences about the nature of the learning mechanism, and the extent to which its processing and representations are available to, or encapsulated from, conscious awareness. However, for the present, these influences remain far from clear.

Acknowledgements

This research was supported in part by the U.K. Economic and Social Research Council (ESRC). Grant number R000236214.

References

Berry, D. C. & Dienes, Z. (1993). *Implicit Learning: Theoretical and Empirical Issues.* Hove: Lawrence Erlbaum Associates.

Bruning, J. L. & Kintz, B. L. (1977). *Computational Handbook of Statistics.* Glenview, IL: Scott, Foresman and Company.

Chan, C. (1992). *Implicit cognitive processes: Theoretical issues and applications in computer systems design.* Doctoral Thesis. Oxford, UK: University of Oxford, Department of Experimental Psychology.

Cheeseman, J. & Merikle, P. M. (1984). Priming with and without awareness. *Perception and Psychophysics*, 36, 387–395.

Dienes, Z. (1992). Connectionist and memory-array models of artificial grammar learning. *Cognitive Science* , 16, 41–79.

Dienes, Z., Altmann, G. T. M., Kwan, L. & Goode, A. (in press). Unconscious knowledge of artificial grammars is strategically applied. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Dienes, Z. and Perner, J. (in press). Implicit knowledge in people and connectionist networks. In G. Underwood (Ed.), *Implicit Cognition.* Oxford, UK: Oxford University Press.

Gomez, R. L. (1996). *Transfer and Complexity in Artificial Grammar Learning.* Manuscript.

Gomez, R. L. & Schvaneveldt, R. W. (1994). What is learned from artificial grammars? Transfer tests of simple associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 396–410.

Manza, L. & Reber, A. S. (1994). *Representation of tacit knowledge: Transfer across stimulus forms and modalities.* Manuscript submitted for publication.

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

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, 118, 219–235.

Redington, M., & Chater, N. (1994). The guessing game: A paradigm for artificial grammar learning. In *Proceedings of the Sixteenth Annual Meeting of the Cognitive Science Society*, 745–749. Hillsdale, NJ: Lawrence Erlbaum Associates.

Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences*, 17, 367–447.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, 30, 50–64.