THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2003, *56A* (1), 97–115

The use of control groups in artificial grammar learning

Rolf Reber
University of Bern, Bern, Switzerland
Pierre Perruchet
Université de Bourgogne, Dijon, France

Experimenters assume that participants of an experimental group have learned an artificial grammar if they classify test items with significantly higher accuracy than does a control group without training. The validity of such a comparison, however, depends on an additivity assumption: Learning is superimposed on the action of non-specific variables—for example, repetitions of letters, which modulate the performance of the experimental group and the control group to the same extent. In two experiments we were able to show that this additivity assumption does not hold. Grammaticality classifications in control groups without training (Experiments 1 and 2) depended on non-specific features. There were no such biases in the experimental groups. Control groups with training on randomized strings (Experiment 2) showed fewer biases than did control groups without training. Furthermore, we reanalysed published research and demonstrated that earlier experiments using control groups without training had produced similar biases in control group performances, bolstering the finding that using control groups without training is methodologically unsound.

In a now classical procedure described in the literature on implicit learning, participants are exposed to a set of letter strings that are derived from a finite state grammar (see Reber, 1993). They are instructed to memorize these strings. After this exposure phase, new strings are presented, half of them from the same grammar, half of them ungrammatical. The participants have to classify the strings according to their grammaticality: If accuracy of classification exceeds chance performance significantly, the researcher assumes that grammar learning has occurred (see Reber, 1993).

This procedure has been criticized, however, because it is not certain that defining chance level as a baseline is adequate. It is possible, for instance, that characteristics of the materials could result in grammatical items looking more grammatical than ungrammatical items, even

Requests for reprints should be sent to Rolf Reber, University of Bern, Department of Psychology, Muesmattstrasse 45, CH-3000 Bern 9, Switzerland. Email: rolf.reber@psy.unibe.ch

This research was supported by the Swiss National Science Foundation (Grant No. 61-57881.99 to Rolf Reber). We thank Emma Smith for her help in improving language.

http://www.tandf.co.uk/journals/pp/02724987.html DOI:10.1080/02724980244000297

though participants have not been submitted to a prior training session. In order to level this potential drawback, researchers have run control groups in which participants were not exposed to grammatical strings prior to the test. The performance of the control group served as a baseline to assess whether genuine learning occurred in the experimental group (e.g., Altmann, Dienes, & Goode, 1995; Dienes, Altmann, Kwan, & Goode, 1995; Dulany, Carlson, & Dewey, 1984; Meulemans & Van der Linden, 1997).

Redington and Chater (1996) cited five experiments that used control groups. Three of them yielded classification accuracy at chance level, which was at 50% in all studies. Two studies, however, showed above-chance performance of control groups. They also mention an unpublished study by Dienes that showed control group performance as high as .60. Meanwhile, some more studies on artificial grammar learning have been published that used control groups. We found 16 no-training control groups in 14 experiments (Experiments 1 and 4 in Altmann et al., 1995, had two control groups each; our own Experiment 1, presented later, had two control groups). The experiments and the corresponding classification accuracy are shown in Table 1.

It appears that the departure from .50 is often quite substantial, with classification accuracy ranging from .45 to .60. These values suggest that the use of control groups is necessary because the comparison of observed performance to chance performance may result in erroneous inferences about the occurrence of learning. At first glance, assessing whether learning has occurred seems straightforward when a control group has been run. The performance of the experimental group that has been exposed to grammatical strings during the training session is

TABLE 1
Classification accuracy of no-training control groups in 16 artificial grammar learning experiments

Authors and year	*Experiment*	*Accuracy*[a]
Altmann, Dienes, and Goode (1995)	1 (letters)	50
Altmann, Dienes, and Goode (1995)	1 (tones)	49
Altmann, Dienes, and Goode (1995)	2	50
Altmann, Dienes, and Goode (1995)	3	47
Altmann, Dienes, and Goode (1995)	4 (symbols)	51
Altmann, Dienes, and Goode (1995)	4 (syllables)	49
Dienes (in Redington & Chater, 1996)		60
Dienes, Altmann, Kwan, and Goode (1995)	1	46
Dienes, Altmann, Kwan, and Goode (1995)	5	52
Dulany, Carlson, and Dewey (1984)		56
Meulemans and Van der Linden (1997)	2a	49
Meulemans and Van der Linden (1997)	2b	45
Redington and Chater (1996)		57
Present study	1 (low frequency)	45
Present study	1 (high frequency)	51
Present study	2	49

Note: We listed only studies with control groups that (1) did not have prior training in a grammar and (2) used a classification task.
[a]In percentages.

compared to the baseline provided by the control group performance. If the performance of the experimental group is significantly higher than the performance of the control group, grammar learning supposedly has occurred. The amount of learning is thought to be proportional to the size of the difference between the groups.

In this paper, we provide evidence that in some cases, the rationale of this method is flawed. Let us introduce our line of reasoning by exploring why performance in control groups may depart from 50%. A first possibility, raised by Redington and Chater (1996), is that learning in control groups occurs during the test phase. As the participants classify the test strings one by one, they may extract some regularity within the test strings and therefore classify them with increasing better-than-chance accuracy. However, empirical evidence provides no support for this first possibility. Indeed, learning during the test should yield a mean improvement of performance for control groups. In contradiction to this expectation, the mean accuracy of the 16 no-training control groups listed in Table 1 is $M = 50.38\%$, $SD = 4.23$, which is not different from a chance level of 50%, $t(15) = 0.36$. This analysis suggests that learning during the test presumably does not cause departure from chance level of control group performance.

A second possibility, also suggested by Redington and Chater (1996), is that some features of test strings result in judgemental biases. As a consequence, the test strings may be classified with better- (or lower-) than-chance accuracy because some of their features favour (or hamper) the classification of grammatical strings as grammatical and of ungrammatical strings as ungrammatical. It seems plausible that participants use information that indicates some form of regularity within a test string—for example, the simplicity of the string or the recursive loops within a string. We call features that indicate regularity and simplicity non-specific variables because their potential influence on grammaticality judgements exists prior to any specific experimental manipulations. If such a variable is not balanced across grammaticality conditions, there may be a judgemental bias. For example, if recursive loops are more frequent for grammatical test items, they will more likely be classified as grammatical than as ungrammatical items, without any prior training.

If departure from chance level by control participants is due to the action of non-specific variables, then subtracting control group performance from experimental group performance to measure learning presupposes that these variables influence classification in both control and experimental groups to the same degree. In other terms, the tacit postulate on which the validity of a difference score is grounded is that learning of the structurally relevant features by the experimental group is simply *superimposed* on the action of non-specific variables. We refer to this condition as the *additivity assumption*. Is this assumption theoretically warranted? Although the issue is practically never discussed in the literature, we guess that most researchers in the field of implicit learning would presumably assume that training removes or at least attenuates the effects of unspecific variables, in the same time as the genuine structure of the grammar is learned. In other words, it seems theoretically sound to construe learning as the *replacement* of irrelevant biases by some relevant features of the material as a basis for the judgements of the participants. This position is even mandatory whenever one envisions the borderline case in which accuracy of a participant is 100%: In this case, any influence of non-specific variables is necessarily eliminated and replaced by that of grammatically relevant variables.

The incongruity between the way learning is theoretically conceived and the way it is empirically measured has devastating consequences whenever the performance of control groups departs from chance level. Let us suppose, for instance, that in a given experiment in which chance level is at 50%, both experimental and control groups reached an accuracy of 60%. From a theoretical point of view, it is possible that experimental participants have changed their basis of judgement from non-specific variables to the relevant features of the grammar, whereas the performance of control participants is due to their continuing reliance on non-specific variables. However, the conclusion drawn from the application of the subtractive method is that learning failed to occur. In this case, the occurrence of learning remains erroneously undetected. Still more damaging, the reverse error may occur in other circumstances. For instance, if accuracy is at 50% for the experimental group and at 40% for the control group, the usual analysis leads to the conclusion that learning has occurred. However, this difference may simply be due to the fact that experimental participants no longer base their judgements on non-specific variables. These participants may have learnt that non-specific features are irrelevant, without acquiring any genuine features of the grammar, regardless of whether these features are construed in terms of rules, exemplar memory, or fragmentary information.

We present two experiments that address these issues. Experiment 1 had two objectives. First, we wanted to confirm that control groups in artificial grammar learning settings depart from chance because they rely on non-specific information, as we have assumed earlier. Two control conditions did not have any training and therefore were never shown grammatical strings prior to the test. We hypothesized that grammaticality judgements in these control groups would be related to simplicity and regularity. The second objective was to assess whether the additivity assumption is warranted. To this end, we compared the control groups to two additional groups that were exposed to grammatical strings prior to testing. If non-specific features of items influenced classifications of participants in the control and experimental groups to the same degree, the additivity assumption would be supported, and nothing would be wrong with the subtraction method. If, however, non-specific features of items influenced classifications in the control group only, but not in the experimental group, as theoretical considerations suggest, the additivity assumption—and therefore the commonly used subtraction method—would be fatally flawed. To anticipate, we observed that non-specific variables exerted a strong influence on the performance of control groups, whereas this influence was removed or attenuated after grammatical string exposure, hence violating the additivity assumption.

In Experiment 2, we were able to replicate the findings of Experiment 1. Moreover, we were interested in the effect of presenting randomized training items to control participants. Do control groups with training rely on non-specific information, similar to the control groups without training, or do they rely on information from the training session, attenuating the impact of non-specific information? As shown in more detail later, training with randomized items did attenuate the impact of non-specific information. Finally, we present a reanalysis of control groups without training in published research using the regression equations obtained in our experiments. After having shown that the findings presented in this article are quite general, we outline the methodological implications of our study.

EXPERIMENT 1

Method

Participants

A total of 60 undergraduate students of the University of Burgundy in Dijon, France, participated in the experiment. A total of 30 participants of the *experimental group* learned consonant strings derived from an artificial grammar and had to classify 40 test strings (see Table 2) as to whether they were grammatical or not. A total of 30 participants served as a control group (*control group without training*) that classified the grammaticality of test strings without having previously learned the items.

Materials

We constructed 20 training strings and 20 grammatical test strings in accordance with an artificial grammar (see Figure 1). A total of 20 ungrammatical test strings were added that could not be derived from the grammar. Grammatical and ungrammatical test strings did not differ significantly from each other in the following: average associative strength, $M = 7.37$ for grammatical and $M = 7.50$ for ungrammatical items, respectively, $t(38) = 0.23$; chunk strength at anchor positions, $M = 3.40$ and $M = 3.28$, respectively, $t(38) = 0.37$; or similarity, $M = 58.40$ and $M = 61.52$, respectively, $t(38) = 0.48$ (see

TABLE 2
Learning and test items[a], Experiments 1 and 2

LE	*LR*	*LP*	*TG*	*TU*
KZQ	XZK	ZKQ	KXZ	XZQ
XZH	QQQ	XHZ	XZK	XHX
KXZQ	ZXK	KQXZ	KXXZ	KXXX
XHZQ	KKHH	QZXH	XZHK	XHHH
KXZQZ	ZXQQ	XQKZZ	KXXZQ	KZQZQ
KZQZK	HKZZ	KZKQZ	KZQHZ	KZHZQ
XZHHH	HZZQ	ZXHHH	XZHHK	XZHHZ
KXZQZH	XKQKX	ZKQZHX	KZQHZQ	KXZHZQ
XHXXZQ	XQKQXK	XHXXZQ	XHXXXZ	XHXXXX
XHZQHZ	XZXQZZ	XHQZHZ	XHZQZK	XHZXZQ
KXXXXZQ	XZZXQX	QXZXXKX	KXXXZQZ	KXXXHZQ
KXXZQZH	ZXKZXQK	QZXKHZX	KXXZQZK	KXXZQZQ
KXZQZHH	ZKZXXHK	QHHKZZX	KXZQHZQ	KXZQXHZ
XHXXXXZ	HXXQKXZ	XXHXXXZ	XHXZQHZ	XHXHZQZ
XHZQHZQ	KZZZKHH	XHHQQZZ	XHZQHXZ	XHXZQZQ
KXXZQHZQ	KXXZZHH	ZKZQQHXX	KXXZQHXZ	KXXZQHHK
KXZQZHHH	KQQKHKZX	XZZHHQHK	KXZQZHHK	KXZQZHHZ
XHXXXZQZ	ZKHXQZKQ	XZZQXXXH	XHXXZQHZ	XHXXXZQH
XHXZQZHK	QQKZHKHX	HQZXZKHX	XHZQHXZQ	XZHZQXZQ
XZHHHHHK	HXQXZZHK	ZXHHHHHK	XZHHHHHH	XZHHHHHZ

Note: LE = learning items, experimental group, LR = learning items, random group (Exp. 2 only), LP = learning items, pseudo group (Exp. 2 only), TG = grammatical test items, TU = ungrammatical test items.
[a]Low-frequency condition only.

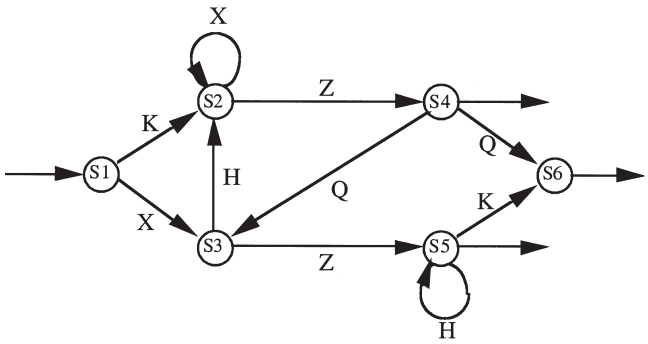

Figure 1. Grammar used in the experiments. Beginning at node S1, this Markovian artificial grammar generates different letter strings. In order to construct such strings, one has to follow the arrows, e.g., from S1 to S2 or to S3, and then to further nodes, until one quits the grammar at one of the endpoints at S4, S5, or S6, respectively. At S2 and S5, arrows return to the same node, giving the option to repeat the same letter. Examples of items generated by the grammar are shown in the columns LE and TG of Table 1.

Meulemans & Van der Linden, 1997, p. 1010, for the definition of these measures). The training items were presented on one page, as a 4 × 5 matrix.

Two sets of letter strings were constructed, one composed of consonants that are frequent in French (high-frequency condition), the other composed of less frequent consonants (low-frequency condition). A total of 17 participants in the control group and 15 participants in the experimental group got the high-frequency letter set, and 13 participants in the control group and 15 participants in the experimental group were given the low-frequency letter set. The set of letter strings for the low-frequency condition is shown in Table 2. The grammar and items with frequent letter strings were constructed by replacing less frequent with frequent letters, according to the following rule:

$H \rightarrow R$; $K \rightarrow S$; $Q \rightarrow T$; $X \rightarrow P$; $Z \rightarrow L$.

By using different letter sets, we were able to assess whether the observed biases replicate with different material conditions. The order of the test strings was reversed for about half of the participants.

Procedure

In the training phase, the experimental group was instructed to memorize the consonant strings for five minutes. It was emphasized that about equal attention should be allocated to each letter string. In the test trial, the participants of the experimental group were told that the strings they had seen before were constructed according to a grammar, and therefore, there was regularity within these strings. The participants were instructed to judge whether the test strings were regular and constructed according to the same rule as the strings seen before. If this was the case, the participants had to answer "yes"; if the string was not regular, they had to answer "no". It was emphasized that the task is not easy. Therefore, the answer should be based on the participant's immediate feeling of the string's regularity. The instruction to the control group was the same, with one exception: As the control group had not seen any training strings, the instruction did not mention any relation to a training trial.

Results

We first examined the grammaticality effect. Means and standard deviations for percentage of accurate classifications are shown in Table 3. A 2 (control group vs. experimental group) × 2

TABLE 3
Percentage of accurate classifications for control and experimental groups of Experiment 1

		% accurate classifications		
Frequency	*Group*	*N*	*Mean*	*SD*
Low	No training	13	44.6	7.2
	Experimental	15	55.6	9.6
High	No training	17	51.3	5.9
	Experimental	15	56.2	8.1

(low vs. high frequency) between-subjects analysis of variance yielded a significant main effect for group. The experimental groups in both conditions were more accurate than the control groups, $F(1, 56) = 15.36$, $p < .001$. The effect of letter frequency and the interaction were marginally significant, $F(1, 56) = 3.20$, $p = .079$, and $F(1, 56) = 2.27$, $p = .137$, respectively.

We used the percentage of endorsement for each item as the dependent variable in multiple regression analyses. This measure was derived by dividing the number of participants who endorsed the item by the total number of participants in the group; this proportion was multiplied by 100. Item grammaticality and different non-specific measures, derived from the test strings, served as independent variables. We used non-specific measures that indicated the simplicity or regularity of items. Johnstone and Shanks (1999) have successfully employed the use of regression analysis, with independent variables derived from the item materials. In contrast to the statistical analysis of these authors, who performed regression analyses for each participant, the dependent variable in our studies was dichotomous, raising problems of estimation with linear regression models (see Gujarati, 1995, p. 540 ff.). Therefore, we performed multiple regression analyses for the scores summed over all participants.[1] There were only moderate correlations, *r*s ranging from –.30 to +.35, between the independent variables. Moreover, the variance inflating factors (*VIF*, see Gujarati, 1995, p. 338 f.) for each variable in these multiple regression analyses were small (all $VIF < 2$), allowing an interpretation of multiple regression findings without problems of multicollinearity.

We entered item grammaticality and five non-specific variables into the regression equation as independent variables:

[1]An important concern is whether findings from a regression analysis over all participants might be due to some extreme data from a minority of participants. Therefore, one would prefer an analysis for each participant, using slopes in an analysis of variance to determine the group differences. Gujarati (1995, p. 542 ff.) listed a number of problems for estimation in linear regression with dichotomous variables. Moreover, he concluded that even if these problems were solved, the linear probability model would not be a very attractive model because the incremental effect of X remains constant throughout. He recommended the use of the logit model, but this model needs a fairly large sample (which is not the case in our study). Cohen and Cohen (1983, p. 240 f.), in contrast, noted that dichotomous variables can be usefully employed in multiple regression models, although this practice is in formal violation to the model. We did an additional analysis, estimating linear regression coefficients for every participant and using the slopes in an analysis of variance. Group differences were virtually the same as those reported for the regression analyses over all participants. Therefore, we report the formally more appropriate analyses over all participants.

1 *Letter number*. This was the number of different letters within a string. The string KXXX contained two different letters, K and X, and therefore got a score of 2. The string KXZQZHHK contained all five possible letters and got a score of 5.

2 *Multiple letter position*. This was the percentage of positions containing letters that occurred twice or more within the same letter string. For example, the string XHX had three positions; two positions contained the letter X that appeared twice whereas the letter H appeared only once. Therefore, the score of 67 was assigned to this string because 67% of the positions contained a letter that appeared twice or more. The string KXZQZHHK got a score of 75, because six positions out of eight positions contained three letters (H, K, Z), each occurring in multiple positions.

3 *Letter repetition* was a dichotomous variable defined by whether a repetition of consecutive letters occurred; a score of 1 was assigned to letter strings containing at least one repetition of consecutive letters (e.g., KXXX), and a score of 0 to strings without repetition of letters (e.g., KZQZQ).

4 *Bigram re-occurrence* was a dichotomous variable defined by whether a bigram reoccurred within the same letter string or not. A string that had the same bigram twice got a score of 1; the other strings a score of 0. For example, the string XHZQHXZQ had one re-occurrence of the bigram ZQ, getting a score of 1. Please note that multiple repetitions of consecutive letters did not count as bigram re-occurrence. For example, the string XZHHHHHK got a score of 0 although the bigram HH re-occurred within the same letter string.

5 *First–last letter identity*. If the first and the last letter of a string were identical, the string got a score of 1, if first and last letter were different, the score was 0. For example, the string KXZQZHHK had identical first and last letters and got a score of 1.

We did not enter item length as a variable because it was controlled experimentally and, moreover, highly correlated with letter number and multiple letter position (rs = .59 and .62, respectively). As letter number and multiple letter position were virtually uncorrelated, r = –.09, much of the variance of the variable length was shared between the two other variables, leading to problems of collinearity.

The results for the multiple regression analyses for control and experimental groups are shown in Table 4. For each condition, we report a multiple regression analysis entering grammaticality and all non-specific variables, with R^2 values at the bottom of each analysis. In the upper half, the analyses for the low-frequency condition are shown; analyses for the high-frequency condition are in the lower half of the table. Let us first examine the determination coefficients for the multiple regression: R^2 values were higher in the two control groups than in the two experimental groups. The high R^2 values for the control groups indicate that control groups did not classify strings randomly.

All non-specific variables showed significant or marginally significant effects in the control groups that were consistent over both the low-frequency and the high-frequency condition. The only inconsistent result between the two control groups were the beta values for multiple letter position, which were marginally positive for the low-frequency control group and significantly negative for the high-frequency control group. We do not know why multiple letter position may have different effects on endorsements in the two frequency conditions.

TABLE 4
Results of the multiple regression analysis for control and experimental groups, for the low- and the high-frequency conditions

Frequency	*Variable*	*Control*	*Experimental*	*Difference*[a]
Low	Lettnumb	–0.35**	0.34*	2.80**
	Mlettpos	0.24^{+}	–0.26	2.49*
	Repetition	–0.66***	0.09	3.26**
	Bigram reocc	0.24^{+}	–0.17	3.06**
	Firstlast	0.19^{+}	–0.18	1.21
	Grammaticality	–0.06	0.18	2.13*
	R^2	0.66	0.35	
High	Lettnumb	–0.24*	0.09	1.19
	Mlettpos	–0.39**	–0.39**	0.20
	Repetition	–0.61***	0.17	3.89***
	Bigram reocc	0.24^{+}	–0.01	1.41
	Firstlast	0.28*	–0.29^{+}	1.61
	Grammaticality	0.04	0.20	1.10
	R^2	0.68	0.36	

Note: Significance levels in the control and in the experimental columns refer to the difference of each slope to zero, as assessed by *t*-tests.
[a] *T* value (df = 78) that resulted from the difference of the two slopes (see text for details).
Lettnumb = letter number; Mlettpos = Multiple letter position; Repetition = letter repetition; Bigram reocc = bigram re-occurrence; Firstlast = first–last letter identity.
*** $p < .001$; ** $p < .01$; * $p < .05$; $^{+}$ $p < .10$.

In the experimental groups of both the low-frequency and the high-frequency conditions, there were no reliable effects of any non-specific variable on endorsements. This means that for all non-specific variables, except multiple letter position in the high-frequency condition, significant or marginally significant effects in the control groups turned into non-significant effects in the experimental groups. In the low-frequency condition, the variable letter number yielded a significantly negative effect in the control group, which turned into a significantly positive effect in the experimental group. Please note that beta values for every non-specific variable, again with the exception of multiple letter position in the high-frequency condition, had reversed signs in the control and experimental groups.

In order to examine these effects further, we calculated separate post hoc regression analyses for grammaticality and each non-specific variable, entering group (control vs. experimental group), non-specific variable, and the interaction term of these two variables as the independent variables, and the percentage of endorsements as the dependent variable. The interaction term indicated the differences of the regression slopes between the two groups. In the rightmost column of Table 4, we report the *T* values of the interaction term. As can be seen from these analyses, all variables, except first–last letter identity, produced significantly different effects in the control group and in the experimental group for the low-frequency condition. For the high-frequency condition, the two groups differed significantly in the effect of repetition on endorsements. Finally, the effect of grammaticality differed significantly in the low-frequency condition, but not in the high-frequency condition.

Discussion

The data for both the low- and the high-letter frequency groups revealed that the endorsements of control participants were highly systematic: The fewer different letters and the fewer repetitions a consonant string had, the more likely it was to be endorsed. These two effects were significant for control groups in both letter frequency conditions. The positive effect of first–last letter identity on endorsements was significant in one control group and marginally significant in the other. The positive effect of bigram re-occurrence on endorsements was marginally significant in both control groups. Interestingly, almost every reliable effect of the control groups was in the direction of higher regularity. As a rule, the control participants were more likely to endorse letter strings that indicated higher simplicity—like lower letter number—or recursive loops—like bigram re-occurrence or first–last letter identity. There is one exception to this rule: Control participants endorsed letter strings *less* if they contained repetitions. Why did they think that repetition could indicate lower grammaticality, whereas other forms of regularity were more readily endorsed? We think that repetition is a very salient form of regularity, probably the most salient form used in this experiment. Participants in experiments follow conversational norms (Schwarz, 1994; Whittlesea & Wright, 1997). This means that participants assume that an experimenter is providing some relevant information. Imagine that an experimenter presents letter strings and tells participants that there is regularity underlying these strings. Moreover, the experimenter tells them that it is not an easy task to find these regularities. From this instruction, control participants may conclude that such an obvious feature like repetition of letters seems not to be the relevant variable that the experimenter is looking for. Hence, the unexpected effect of repetition may be due to the control participants' reasoning based on their assumptions about experiments and experimenters.

The biases were not apparent in the two experimental groups. The differences between control and experimental groups in repetition effects were significant in both frequency conditions. In the low-frequency condition, four out of five differences in effects of non-specific variables were significant. This pattern of findings clearly contradicts the additivity assumption, which posits that performance in a no-training control group can serve as a baseline to assess whether an experimental group is sensitive to grammatical regularities, because both groups would be equally sensitive to non-specific variables.

EXPERIMENT 2

Although control groups without training have been used frequently in artificial grammar learning, some authors have used alternative techniques: Altmann et al. (1995), Perruchet and Pacteau (1990), Reber (1967), and Shanks, Johnstone, and Staggs (1997) added control groups that had to learn random stimuli. Redington and Chater (1996) proposed a crossover design that has been applied by Dienes and Altmann (1997).

However, whatever the experimenter's initial intent, it may be argued that the training of participants with random strings has the side effect of suppressing or at least lowering the effect of unbalanced materials during the test. Indeed, participants may learn from the random

strings that non-specific features are irrelevant. If so, comparing the performance of an experimental group with the performance of a control group trained with random stimuli would be valid. But there are other possibilities. First, it is difficult to assess whether control participants actually learned that non-specific features are irrelevant. Indeed, the need to assess learning in a control group faces us with a problem of infinite regress: In order to make this assessment, we would have to introduce a control group that shows this kind of learning to be less likely. Second, even if one takes for granted that control participants have learned the irrelevance of non-specific features, it must be realized that they are subsequently asked to judge the grammaticality of test items without having benefited from the opportunity of extracting alternative, reliable criteria. Under these conditions, it remains possible that they nevertheless rely on the only features available to them, rather than responding randomly.

Experiment 2 is aimed at addressing two nested questions. The first concerns the question of whether control groups exposed to randomized strings exhibit sensitivity to non-specific features in their grammaticality judgements. If so, the second question pertains to the nature of these biases, and notably to the question of knowing whether these biases are identical to those observed in the control group without training. As an additional manipulation, we also introduced a control group that had to learn strings that were randomized, but respected the letter frequency of the grammatical strings. Indeed, letter frequency is not usually considered as a variable of interest by implicit learning researchers, whether they subscribe to models framed in terms of rule abstraction or memory mechanisms. As a matter of fact, the raw frequency of individual events is now commonly equated between experimental and control groups in other paradigms of implicit learning, such as in the serial reaction times procedures. We anticipated that matching experimental and control groups on this variable in artificial grammar studies could reduce the effects of non-specific variables.

Method

Participants

A total of 91 undergraduate students of the University of Burgundy in Dijon, France, participated in this experiment. Of these, 20 participants learned grammatical letter strings during training (*experimental group*). The other participants were assigned to one of three control groups. A total of 26 participants were in a *control group without training*, similar to the one in Experiment 1; 24 participants were in the *random group*, a control group that learned randomized items; and 21 participants were in the *pseudo group*, a group that learned pseudo-randomized items, constructed in a similar way to the control items of Perruchet and Pacteau (1990).

Materials

The test items were the same as those in Experiment 1, with the exception that only the low-frequency items were used (see Table 2). In addition to the low-frequency learning items used for the experimental groups in Experiment 1, we constructed two sets of randomized learning items. The first set, for the random group, was established by first determining the length (random numbers between 3 and 8) of the consonant strings. Then, one of the five consonants was randomly assigned to each position. There

were no other constraints[2]. The resulting items are shown in the second column of Table 2. For the construction of the learning items of the pseudo group, we started from the learning items of the experimental group. Items for the pseudo group had the same length and the same composition of consonants as the items of the experimental group. We assigned a random position to the first consonant of the experimental learning string, then to the second consonant, and so forth, until all positions were filled. The resulting items are shown in the third column of Table 2. In sum, the random group got only information about the consonants and string length during training. The pseudo group got the same information and additional information about the frequency of each consonant.

Procedure

In the learning session, the random, pseudo, and experimental groups were given the same instructions as those given to the experimental groups in Experiment 1. The control group without training had to complete a symbol–digit test, adapted from the Wechsler Adult Intelligence Scale (Wechsler, 1981) during the learning session. This was a procedural change to Experiment 1, where the groups without training completed the test at the very beginning of the session. In the test session, the participants of the random, pseudo, and experimental groups were given the same instructions as those given to the experimental groups in Experiment 1. The instruction to the control group without training was the same as in Experiment 1.

Results

We first examined the grammaticality effect. Means and standard deviations for the percentage of accurate classifications are shown in Table 5. A one-way analysis of variance yielded a significant main effect for group, $F(3, 87) = 7.88$, $p < .001$. Post hoc tests using Tukey HSD test revealed significant differences between the control group without training and the experimental group ($p = .028$) and the random group ($p < .001$), respectively. All other differences were not significant ($p > .1$).

TABLE 5
Percentage of accurate classifications for control and experimental groups of Experiment 2

	% accurate classifications		
Group	*N*	*Mean*	*SD*
No-training	26	48.8	8.9
Random	24	59.2	5.5
Pseudo	21	53.9	7.5
Experimental	20	55.3	8.2

[2]In accordance with the "truly random control procedure" (Rescorla, 1967), we did not exclude grammatical training items in the control groups with training. As a consequence, one item (XZK) in the random group was also a (grammatical) test item. Although identical learning and test items exist in some studies on artificial grammar learning (e.g., Altmann et al., 1995, Experiments 1 and 2; Dienes et al., 1995, Experiment 5; Dulany et al., 1984; see also Reber, 1993), we now think that randomization should be constrained so that no random training item is identical to any test item. We report analyses with the inclusion of XZK; analyses under exclusion of XZK yielded virtually the same results, which allowed the same conclusions.

As in Experiment 1, the percentage of endorsement for each item served as the dependent variable, and item grammaticality and the same five non-specific measures served as independent variables in a multiple regression analysis. Again, we analysed the differences between groups by examining the interaction effects of each single variable and group on endorsements.

The results for the multiple regression analysis for control and experimental groups are shown in Table 6. For each condition, we report a multiple regression analysis entering grammaticality and all non-specific variables, with R^2 values at the bottom of each analysis. Let us first examine the determination coefficients for the multiple regression: R^2 values were higher in the control groups than in the experimental group. Moreover, the control group without training had a higher R^2 value than the control group with randomized strings, which in turn had a higher R^2 value than the control group with pseudo-randomized strings.

We assessed interactions between groups and non-specific variables. As in Experiment 1, we compared two groups (e.g., experimental group vs. random group) for the difference in slope of one non-specific variable (e.g., letter number). The groups differed markedly as to how they used non-specific variables for their classifications. Replicating the findings of Experiment 1, letter repetition influenced endorsements of participants of the control group without training, but not of the experimental group. The difference between the two groups was significant, $t(78) = 4.01$, $p < .001$. There was a continuous pattern from the no-training control group to the pseudo group: If there was no information in the learning phase, people relied heavily on repetition, $\beta = -.82$. If there was some information, as in the random group, people still relied on repetition, $\beta = -.47$, and the slope was not different from the slope of the control group without training, $t(78) = 0.29$. If frequency information was available, however, which may have added some information about repetition patterns, there was no longer an effect of repetition on endorsements. The pseudo group differed significantly from the control group without training, $t(78) = 4.65$, $p < .001$, and from the random group, $t(78) = 3.06$, $p = .001$, but not from the experimental group, $t(78) = 0.39$.

TABLE 6
Results of the multiple regression analysis for control groups and the experimental group

Variable	*Without*	*Random*	*Pseudo*	*Experimental*
Lettnumb	–0.17^{+}	0.30*	0.48**	0.26
Mlettpos	–0.16	0.17	–0.18	–0.33^{+}
Repetition	–0.82***	–0.47**	0.03	0.08
Bigram reocc	0.02	–0.03	–0.37*	0.11
Firstlast	0.14	0.07	–0.13	0.01
Grammaticality	–0.05	0.28*	0.08	0.20
R^2	0.74	0.53	0.44	0.24

Note: We present average slopes of the multiple regression analysis described in the text. Significance levels refer to the difference of each slope to zero, as assessed by *t* tests. Variables are the same as in Table 4. (Without = without training; Random = random strings; Pseudo = pseudorandom strings; details see text).
***$p < .001$; **$p < .01$; * $p < .05$; $^{+}p < .10$.

For letter number, the two control groups with training behaved more like the experimental group than did the control group without training. The slope of the no-training control group differed significantly from the slopes of the random group, $t(78) = 2.46$, $p = .016$, and from those of the pseudo group, $t(78) = 2.33$, $p = .022$, respectively. The effects of repetition and letter number on endorsements showed that different control groups relied on different information. There were two effects that were not expected and did not yield a meaningful pattern. One was the effect of grammaticality for the random group. A second effect that was difficult to explain was the effect of bigram re-occurrence on endorsements in the pseudo group. There was no such effect in Experiment 1 or in the other groups of Experiment 2. These two effects illustrate that only replicated results or results that show a meaningful pattern suggest real rather than accidental effects.

Discussion

Experiment 2 brought three main findings. First, the significant or marginally significant effects in the control group without training were not apparent in the experimental group, replicating the findings of Experiment 1. Second, the non-specific variables in the control groups with training explained less variance than in the control group without training, but still more variance than in those of the experimental group. The biases of the random group were between those of the control group without training and those of the experimental group, whereas the biases of the pseudo group had more in common with the experimental group than with the control group without training. Providing information about letter frequencies resulted in judgements that were not different from the judgements of the experimental group that learned grammatical strings.

Third, the without training group, but not the random and the pseudo group, differed from the experimental group in accuracy of classification. Hence, different kinds of control in artificial grammar learning yielded different results and would imply different conclusions about grammar learning. If we used only a control group without training, we would find a significant difference and conclude that there was an effect of grammar learning. If we used only a control group with fully randomized or pseudo-randomized strings, we would find no significant difference and conclude that there was no effect of grammar learning. Apparently, introducing randomized learning strings constrained the set of hypotheses that a control participant with training could generate about the kind of features relevant to grammaticality, resulting in higher accuracy.

REANALYSIS OF CONTROL GROUPS WITHOUT TRAINING IN PUBLISHED STUDIES

After we had conducted our experiments, we reanalysed the accuracy of classification reported in published data, using our findings on judgemental biases in control groups. The goal is to show (1) whether or not non-specific variables in the materials of control groups in published research were balanced across grammaticality, and (2) if non-specific variables were not balanced, whether or not this non-matching was severe enough that it may have led to erroneous inferences concerning grammar learning. From the published results on control groups listed in Table 1, 11 studies either reported complete item materials (Altmann et al.,

1995, Experiments 3 and 4; Dienes et al., 1995, Experiment 1; Dulany et al., 1984; Meulemans & Van der Linden, 1997) or referred to a study with published materials (Altmann et al., 1995, Experiments 1 and 2; Dienes et al., 1995, Experiment 5). The latter group of experiments used the grammar and item materials used by Dulany et al. (1984). For the 11 control groups employed in these studies, we calculated the values for the non-specific variables and subsequently computed the predicted accuracy, using our regression equations.

We calculated three predictions of accuracy, two using regression equations from Experiment 1, low- and high-frequency condition, and one from Experiment 2. The dependent variable in our experiments was the percentage of endorsements over all items—grammatical and ungrammatical ones—and not the accuracy of classification, which is the average percentage of endorsements for grammatical items and rejections for ungrammatical items if half of the items are grammatical and half are ungrammatical. Using the resulting regression equations, we calculated the predicted percentages of endorsements separately for grammatical and ungrammatical strings, with the five non-specific variables used in our experiments as independent variables. We then calculated the accuracy of classification by taking the average from the percentage of endorsements of grammatical items and the percentage of rejections of ungrammatical strings; subtracting the percentage of endorsements from one hundred yielded the percentage of rejections. There was one exception: Dienes et al. (1995; Experiment 1) derived their performance by dividing the number of endorsements of grammatical strings by the sum of the number of endorsements of grammatical strings plus the number of endorsements of ungrammatical strings. Furthermore, they derived their reported performance of 46% by averaging the control group's performance on two grammars. We calculated the predictions for this experiment accordingly.

We correlated the predictions from each control group and the average of these three predictions on the one side with the observed accuracy in the rightmost column of Table 7 on the other side. If the materials for the control groups in the published studies were unbalanced in the same way as the materials of the control groups without training in our experiments, we would expect positive correlations between the predictions and the observed data in the published articles. This was indeed the case: The correlations shown in the bottom row of Table 7 were positive and significant. The correlation of the averaged predictions with the observed data was $r(10) = .742$, $p < .01$. This means that unbalanced non-specific features in the materials used in the 11 published experiments resulted in biases in the control groups without training that were highly similar to the biases observed in our own experiments.

After having demonstrated that materials were unbalanced across grammaticality conditions, the question arose as to whether this non-matching was severe enough that it may have led to erroneous inferences concerning grammar learning. Some cases were not problematic because performance of the control group was not lower than performance of the experimental group (Dienes et al., 1995; Meulemans & Van der Linden, 1997, Experiment 2A). In another case, experimental groups showed a performance between 63% and 70% (Dulany et al., 1984). Judgemental biases in control groups are not a problem if accuracy is outside the range of the control accuracy normally observed. From the Altmann et al. (1995) studies, Experiments 1, 2, and 4 seemingly do not pose many problems because control group performances were at chance level, and differences were highly significant. In Experiment 3, however, the below-chance performance of the control group (47%) seems to have been due to the non-matching of grammatical and ungrammatical strings, overestimating the difference between the control

TABLE 7
Predicted and observed accuracies of classification in 11 artificial grammar learning experiments

		Predicted				
		Exp. 1				
Authors and year	*Exp.*	*Low*	*High*	*Exp. 2*	*Average*	*Obs.*
Altmann et al. (1995)	1	53.94	54.38	53.54	53.95	50.0
Altmann et al. (1995)	1	53.94	54.38	53.54	53.95	49.0
Altmann et al. (1995)	2	53.94	54.38	53.54	53.95	50.0
Altmann et al. (1995)	3	46.02	50.00	49.96	48.66	47.0
Altmann et al. (1995)	4	50.06	49.36	49.72	49.71	51.0
Altmann et al. (1995)	4	50.06	49.36	49.72	49.71	49.0
Dienes et al. (1995)	1	44.34	42.76	45.42	44.17	46.0
Dienes et al. (1995)	5	53.94	54.38	53.54	53.95	52.0
Dulany et al. (1984)		53.94	54.38	53.54	53.95	55.5
Meulemans and Van der Linden (1997)	2a	48.44	51.25	50.86	50.18	49.0
Meulemans and Van der Linden (1997)	2b	44.83	47.38	48.96	47.05	45.0
Correlation with Obs.		0.787**	0.692*	0.684*	0.742**	

Obs. = observed; Exp. = experiment number of published study.
**$p < .01$; *$p < .05$.

and the experimental group to some extent. Finally, Meulemans and Van der Linden (1997, Experiment 2B) found a significant difference between the control group without training and the experimental group. The experimental group performed at 53.55%, the control group at 45.10%. As our analysis demonstrated, the control group performance was below chance because non-specific features were not balanced across grammaticality. The authors themselves were quite cautious in interpreting this difference in terms of an effect of grammar learning, but, after showing that grammatical strings were more likely to be endorsed than ungrammatical strings, they used it as additional support of their conclusion that grammar learning occurred. This example shows that a difference between an experimental and a control group may lead to erroneous inferences concerning grammar learning if the performance of the control group lies below 50%.

Our reanalysis demonstrated that our experiments yielded results that hold over a whole range of experiments performed in different laboratories, with different finite state grammars, with different materials (letter strings, syllable strings, symbols, tones), and with participants speaking different native languages. The fact that the control group performances in the published experiments were caused by unbalanced non-specific variables lends strong support to the findings of our experiments and therefore to the conclusion that the use of control groups without training in artificial grammar learning experiments is highly problematic.

GENERAL DISCUSSION

Performance of control groups in artificial grammar learning studies often departs from chance level. We predicted that this phenomenon was due to the fact that participants who

have not been exposed to grammatical strings are sensitive to non-specific variables that are not perfectly balanced between grammatical and ungrammatical test items. This hypothesis was unambiguously confirmed. Our experiments and the reanalysis of previously published studies yielded compelling evidence that grammaticality judgements of control participants without training depended on non-specific features of test items.

Using control performance as a baseline to assess learning in the experimental group was seen as a valid measure among researchers in the field of artificial grammar learning. This method rests on the implicit assumption that the two groups are similarly biased by non-specific variables, with the genuine effect of learning being superimposed on these effects in the experimental group. However, we have shown that this additivity assumption does not hold true, because the performances of participants in the experimental groups was virtually insensitive to the non-specific variables that influenced control participants. Reanalysis of the currently available data suggests that these factors may have led to an underestimate or, more damaging in most research contexts, to an overestimate of the amount of changes due to the factors of interest to most learning theorists. In sum, our survey of published studies using control groups (see Table 1) indicates that comparing the performance of experimental groups to an a priori random value is unsound, and our experiments suggest that taking the performance of a control group as a baseline may be invalid, too. If so, what is the proper way for assessing learning?

We present two methodological conclusions from our results: First, one or several control groups have to be run, but not primarily as a baseline for comparison with the performance of experimental participants. The main function of a control group is to check whether the material is well balanced with regard to non-specific variables. If the performance of the controls is not significantly different from chance, the score of experimental participants can be compared to chance level, or to the actual performance of controls, or both. However, if it turns out that the performance of control participants departs from chance level, there is no proper way to assess the amount of learning in experimental participants. In any case, a difference score would provide a flawed estimation. This principle leaves open which procedure is the best to check that the material is well balanced with regard to non-specific variables. Experiment 2 showed that the influence of non-specific features on grammaticality judgements differed according to the conditions to which control participants were assigned. Indeed, providing minimal information on the materials' regularities during training of control groups limited the use of non-specific information.

This observation leads straightforwardly to our second methodological conclusion: The conditions of training of the control participants have to be as close as possible to the conditions of training of the experimental groups. This principle excludes the running of a no training control group as a valid procedure. Experiment 2 indicated that a control group exposed to random strings was less sensitive to non-specific features, and a control group exposed to pseudo-random strings showed effects of non-specific variables that were comparable to the same effects in the experimental group. Thus, the obvious recommendation is to provide as much information as possible, excluding only the features relevant to the objective of the researcher. Unfortunately, this recommendation does not help to determine unambiguously what aspects of the grammatical material need to be preserved. Here, we reach the limits inherent to any methodological recommendations. Designing control material is a theoretically motivated task. An experimenter interested in rule abstraction, for instance, has to

control all variables that do not have to do with rule abstraction, an endeavour that needs the operationalization of what is meant by abstraction in this experimental context.

To conclude, our studies suggest that the current practice of assessing learning by comparing the grammaticality judgement of experimental groups either to chance level or to judgements of control groups without training is methodologically unsound. Indeed, the latter practice relies on an assumption of additivity between structurally relevant influences and influences of non-specific features, an assumption that turns out to be unfounded. This entails that biases due to test materials do not allow a safe interpretation of the performance of experimental groups. Running a control group is necessary, but with the objective to demonstrate that the test materials do not produce biases that are detrimental to the specific objective of the researcher. To this end, the training of control participants has to be as similar as possible to the training of the experimental group, the only differences pertaining to the aspects relevant to the ultimate objective of the study.

This paper was primarily concerned with artificial grammar learning because this paradigm is currently of widespread use in the literature. However, pending appropriate terminological adjustments, our conclusions are obviously relevant to any learning paradigm. For example, Anastasopoulou and Harvey (1999) demonstrated similar problems in serial reaction time tasks. The problem we raised here is inherent to the very nature of learning. Most theorists presumably construe learning in terms of changes from one set of determining factors to another set of determining factors. By contrast, the effects of training are usually assessed by a difference score, the validity of which is grounded on the implicit assumption that learning works in a purely incremental way, as if structure was simply added on other, non-specific determinants of behaviour. Further methodological refinements are needed in each domain, in order to get a better fit between theory and measurement.

REFERENCES

Altmann, G.T.M., Dienes, Z., & Goode, A. (1995). On the modality independence of implicitly learned grammatical knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 899–912.

Anastasopoulou, T., & Harvey, N. (1999). Assessing sequential knowledge through performance measures: The influence of short-term sequential effects. *The Quarterly Journal of Experimental Psychology*, *52A*, 423–448.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Dienes, Z., & Altmann, G.T.M. (1997). Transfer of implicit knowledge across domains: How implicit and how abstract? In D.C. Berry (Ed.), *How implicit is implicit learning? Debates in psychology* (pp. 107–123). New York: Oxford University Press.

Dienes, Z., Altmann, G.T.M., Kwan, L., & Goode, A. (1995). Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 1322–1338.

Dulany, D.E., Carlson, R.A., & Dewey, G.I. (1984). A case of syntactical learning and judgement: How conscious and how abstract? *Journal of Experimental Psychology: General*, *113*, 541–555.

Gujarati, D.N. (1995). *Basic econometrics* (3rd ed.). New York: McGraw-Hill.

Johnstone, T., & Shanks, D.R. (1999). Two mechanisms in implicit artificial grammar learning? Comment on Meulemans and Van der Linden (1997). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 524–531.

Meulemans, T., & Van der Linden, M. (1997). Chunk strength and artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 1007–1028.

Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of Experimental Psychology: General*, *119*, 264–275.

Reber, A.S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*, 855–863.

Reber, A.S. (1993). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious.* New York: Oxford University Press.

Redington, M., & Chater, N. (1996). Transfer in artificial grammar learning: A reevaluation. *Journal of Experimental Psychology: General*, *125*, 123–138.

Rescorla, R.A. (1967). Pavlovian conditioning and its proper control procedures. *Psychological Review*, *74*, 71–80.

Schwarz, N. (1994). Judgement in a social context: Biases, shortcomings, and the logic of conversation. In M. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 26, pp. 123–162). San Diego, CA: Academic Press.

Shanks, D.R., Johnstone, T., & Staggs, L. (1997). Abstraction processes in artificial grammar learning. *Quarterly Journal of Experimental Psychology*, *50A*, 216–252.

Wechsler, D. (1981). *Manual for the Wechsler Adult Intelligence Scale* (Rev. ed.). San Antonio, TX: The Psychological Corporation.

Whittlesea, B.W.A., & Wright, R.L. (1997). Implicit (and explicit) learning: Acting adaptively without knowing the consequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 181–200.

Original manuscript received 2 January 2001
Accepted revision received 11 January 2002