THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2003, *56A* (1), 125–130

Why untrained control groups provide invalid baselines: A reply to Dienes and Altmann

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

Dienes and Altmann argue that an untrained control group provides a reliable baseline to measure artificial grammar learning. In this reply, we first provide a fictitious example to demonstrate that this assessment is faulty. We then analyse why this assessment is wrong, and we reiterate the solution proposed in Reber and Perruchet (this issue) for a proper control. Finally, we point out the importance of these methodological principles in the context of implicit learning studies.

In their comment, Dienes and Altmann (this issue) raise two main concerns. First, they argue that any difference in classification between an experimental group and an untrained control group reflects the fact that experimental subjects have acquired content more positively correlated with the experimenter's grammar than when they started learning. Hence, by definition, this difference provides a reliable measure of learning. In our reply, we show that this intuitively appealing argument is misguided. Second, they propose to measure biases by using regression analyses instead of controlling them, as we proposed in our article. We will present some caveats against using the regression analysis and justify the solution presented in our article.

WHAT IS WRONG WITH AN UNTRAINED CONTROL GROUP AS BASELINE?

This section is aimed at demonstrating that a part of the difference between an experimental group and an untrained control group can be unrelated to the experimenter's grammar. Of course, we do not allude here to the unavoidable part of error variance that plagues any behavioural measures, but to a systematic part of variance inherent to the use of an untrained control group.

Requests for reprints should be sent either to Rolf Reber, Department of Psychology, University of Bern, Muesmattstasse 45, CH-3000 Bern 9, Switzerland, Email: rolf.reber@psy.unibe.ch or to Pierre Perruchet, LEAD/CNRS, Faculté des Sciences, Université de Bourgogne, 6 Bd Gabriel, Dijon, F-21000, France.
Email: pierre.perruchet@u-bourgogne.fr

© 2003 The Experimental Psychology Society
http://www.tandf.co.uk/journals/pp/02724987.html DOI:10.1080/02724980244000314

Starting from an example

Imagine a grammar that generates trigrams in which the middle letter is either a T or an L. Subjects are assigned either to an experimental group, which is exposed to a set of grammatical trigrams (e.g., STX, RLL, . . .), or to a no-training control group. Then subjects from both groups are shown one of the three sets of test strings listed in Table 1, with standard classification instructions. At first glance, the three sets are equally well designed: Grammatical and ungrammatical strings are paired, and, within a given pair, the trigrams differ from each other only by their middle letter, which is the key feature of the grammar. A priori, any difference in responding should reveal sensitivity to the nature of the middle letter. Let us assume that subjects are a priori biased to judge as grammatical the test strings comprising no repetition (a bias indeed demonstrated in Reber & Perruchet, this issue). The accuracy of control subjects exposed to Test A is 40%, as shown in Table 1. On the other hand, Reber and Perruchet's experiments suggest that experimental subjects will no longer be sensitive to this kind of bias at the time of testing. If one assumes for the sake of simplicity that experimental subjects have failed to learn anything about the middle letter of the trigrams, their classification accuracy is about 50%.

In keeping with Dienes and Altmann's line of reasoning, one would conclude from the better performance of experimental subjects that they have acquired some content that is more positively correlated with the experimenter's grammar than at the start of training. However, a closer look will reveal that the difference between experimental and untrained control groups brings *no information at all*. Why?

TABLE 1
Simulation of performance of untrained control subjects, as a function of test strings

	Test					
	A		*B*		*C*	
Actual status	*Strings*	*Response*	*Strings*	*Response*	*Strings*	*Response*
Grammatical	VTX	G	VTX	G	VTX	G
	SLX	G	SL*T*	G	SLX	G
	RTT	NG	RT*X*	G	RTT	NG
	VLL	NG	VLL	NG	*T*LL	NG
	TTS	NG	TTS	NG	*V*TS	G
Non-grammatical	VXX	NG	VXX	NG	VXX	NG
	SSX	NG	SS*T*	NG	SSX	NG
	RMT	G	RM*X*	G	RMT	G
	VXL	G	VXL	G	*T*XL	G
	TVS	G	TVS	G	*V*VS	NG
Accuracy[a]		40		50		60

Note: In Test A, there are more repetitions within grammatical items than within non-grammatical items. In Test B, the letters T and X (in italic characters) have been switched, resulting in the same number of repetitions in both categories of items. In Test C, the letters T and V (in italic characters) have been switched, resulting in fewer repetitions within grammatical items than within non-grammatical items.
[a]In percentages.

The reason becomes evident when one considers Tests B and C in Table 1. These tests are very similar to Test A, they differ from Test A only by the inversion of two irrelevant letters (shown in italic characters in Table 1). However, following exactly the same assumptions as before to generate performance, results with Test B exhibit no difference between experimental and control groups. Moreover, with Test C, the control group outperforms the experimental group. In the terms of Dienes and Altmann, one would conclude from Test B that subjects have not learned anything, and from Test C that subjects have acquired content that is more *negatively* correlated with experimenter's grammar. There is a striking contradiction between all of these conclusions, given that the knowledge of our fictitious subjects (both experimental and control) was left unchanged in the three simulations. The only change pertained to irrelevant aspects of the material.

In our simulation, subjects' knowledge was given a priori and served to generate performance. In real experimental settings, one has to infer subjects' knowledge from performance. Our example shows that the very same subjects' knowledge may generate different effects on performance, depending on minor changes of contextual aspects of the test material. Therefore, it logically follows that it is impossible to reason backward and to infer the nature of the subjects' states from the comparison between the performance of an experimental group and that of an untrained control group. There is something wrong in the principles underlying the comparison advocated by Dienes and Altmann, but *what* exactly is wrong?

We do not object to the Dienes and Altmann's definition of learning as "acquiring content more positively correlated with the experimenter's grammar than when the subject started learning". This definition allows us to include into learning contents any by-product of the grammar, which the experimenter may a priori be unaware of. However, Dienes and Altmann added: "By positively correlated, we mean that the content when used to classify *some set of items* will produce classification responses correlated with the responses required by the experimenter's grammar" (italics are ours). What is wrong is to assume that any difference in responding between *some set* of grammatical and ungrammatical test items—that is, between the limited samples of items actually used in experimental studies—relates to the experimenter's grammar.

Indeed, judgements of typical samples of grammatical and ungrammatical items are generally sensitive to some forms of content unrelated to the grammar. This is the case in our example. The grammar we used is mute with regard to repetition of letters. The rate of letter repetition is not even a remote by-product of the grammar. This claim is not a theoretical option, something that depends on a "theory of what type of learning one is interested in", as Dienes and Altmann repeatedly claim, but an objective, pre-theoretic assessment. Indeed, trigrams conforming to the grammar (i.e., with T or L as a middle letter) include the same proportion of letter repetitions as do randomly generated trigrams.[1] As a consequence, the fact that letter repetition is not balanced across grammatical and ungrammatical items is due to a

[1]Because the grammar is very simple, this assertion can be checked mathematically. With 26 letters, there are 26^3, i.e., 17,576 different trigrams, including 26∗26∗2, i.e., 1352 pairs of contiguous identical letters. When the middle letter is limited to a fixed set of 2, there are 26∗2∗26, i.e., 1352 different trigrams, including 26∗2∗2, i.e., 104 pairs of identical letters. The ratio (2/26) is identical in the two cases, indicating that the grammar does not introduce a bias in the number of letter repetitions. With complex grammars, a computation may be difficult to carry out. However, a computer simulation generating all the possible issues may generally meet the same objective.

bias of sampling, which favours either grammatical or ungrammatical items in unpredictable ways, as shown in Table 1. Because this factor does not influence experimental and control subjects to the same extent, it generates a difference between experimental and control groups that is uncorrelated with the experimenter's grammar. Therefore Dienes and Altmann's a priori assignment of this difference to a genuine learning effect is contradictory with their reference to the experimenter's grammar that defines learning.

Reframing the problem and its consequences

The primary source of the invalidity of untrained control group is the existence of non-specific variables. Non-specific variables are variables that affect classification responses before any training. Indeed, the performance of untrained control subjects is not random: Reber and Perruchet's experiments showed that these subjects classify items as grammatical and ungrammatical in a highly systematic way, using, for instance, letter repetition or the number of different letters within a string as information for the classification task. Of course, it may turn out that non-specific variables are related to the grammar. A grammar can generate strings that differ from random strings with regard to letter repetition or number of different letters within a string. For the sake of simplicity, we consider here only those non-specific variables that are unrelated to the grammar (such as letter repetition in our example).

Whenever non-specific variables are balanced across grammatical and ungrammatical test strings, there is no problem, as illustrated with the Test B in Table 1. Judgements are biased equally for grammatical and ungrammatical strings, and the final discrimination score is unaltered. Such a situation would arise if results were averaged over a huge number of experiments run with different materials. However, we have shown that this was often not the case in individual experiments: Grammatical and ungrammatical items are generally not equivalent with regard to non-specific variables, whatever the care given to the selection of the material. Recall that in our example, constructing grammatical and ungrammatical trigrams in such a way that they differed only with regard to the key feature of the grammar (i.e., their middle letter) does not prevent a severe problem of balance across grammaticality. As a consequence, the score of untrained subjects varies *randomly* around the 50% value. As a matter of fact, the conclusion of our survey of earlier studies in Reber and Perruchet is consistent with this scenario: The departure of untrained control subjects from 50% is quite substantial (with classification accuracy ranging from 45% to 60%), although the *mean* accuracy is 50.38% over 16 independent groups.

Given this situation, Reber and Perruchet suggested that a difference score could still be valid, but only if a certain assumption—which we refer to as the additivity assumption—is fulfilled. Dienes and Altmann assert that "Taking a difference score in no way assumes the additivity assumption, contrary to the repeated (though never justified) claims of R. Reber and Perruchet". The justification is crystal clear. A difference score amounts to posit $L = E - C$, where L stands for learning, E for experimental group, and C for control group. As we have shown, unpredictable effects of non-specific variables affect the performance of untrained controls. In order to keep the equation valid, we have to assume that E is affected by these variables to the same extent. This is what we called the *additivity* assumption, because the assumption is that learning of the structurally relevant features by the experimental group is simply *superimposed* on the action of non-specific variables. Reber and Perruchet showed empirically

that the additivity assumption does not hold, because experimental subjects were not affected by non-specific variables to the same extent as were untrained control subjects. It follows that a difference score between experimental and untrained control subjects is methodologically unsound and may lead to erroneous inferences.

HOW TO OBTAIN A RELIABLE BASELINE?

Although this argument demonstrates that the difference between an experimental group and an untrained control group does not provide a reliable measure of learning, it could be argued that some supplementary analysis might help to solve these difficulties. Dienes and Altmann suggest that we have ourselves shown that this is possible, through the use of regression analyses. Of course, regression analyses may bring valuable information in such situations. However, a prerequisite for the use of regression analyses is the a priori identification of the variables that are potentially influential. One may claim that for artificial grammar learning, the non-specific variables are now known reasonably well. However, people react very flexibly to the materials at hand (e.g., Schwarz, 1994; Whittlesea & Wright, 1997), so that it is actually impossible to gain *exhaustive* knowledge about what information subjects are using to classify stimuli. Furthermore, the standard finite state grammars involving consonant letters as primitives are more and more supplemented by studies using varied and often more natural material. We do not yet know whether the number and the nature of non-specific variables differ from one situation to the other. Hence, designing control conditions that prevent the influence of any bias appears far more preferable than using methods that assume that the biases are exhaustively identified.

In Reber and Perruchet, we proposed that a control group trained with materials as close as possible to the grammatical material provides a reliable baseline to assess learning. The reason is that we have shown, in two experiments, that training with randomized materials reduces the effects of non-specific variables, and that their residual effects are all the more similar in the two groups as the conditions of training are closer. This makes the additivity assumption increasingly valid. In addition, we recommended that performance of control subjects is as close as possible to chance level. Dienes and Altmann found this recommendation "curious", and wonder: If "the only valid control group is . . . one in which subjects respond randomly . . . , why not just compare trained subjects' responding to a chance baseline?" The response is straightforward: Chance performance by control subjects indicates that the test material is unbiased, or that the biases no longer influence performance at the time of testing. If trained control subjects do not perform at chance level, this means that some undetected biases are still in operation, hence casting doubt on the actual meaning of a difference between experimental and control subjects, and a fortiori, between experimental subjects and a chance baseline.[2]

We do not claim that this is the only valid method. For instance, the cross-over design proposed by Redington and Chater (1996) addresses the concerns raised here in a more elegant way (see Dienes & Altmann, 1997, for an example). The main advantage of our proposal is that

[2]Note that, arguably, control performance that is above or below chance level could be taken as baseline when experimental and control groups are trained in very similar conditions, as the additivity assumption—that non-specific variables have similar effects on experimental and control performance—is likely to be fulfilled in this case.

it is easier to design at relatively low cost. Our main objective was to argue against the use of untrained control groups, and not to explore the full range of alternative solutions. Moreover, we do not claim that following our recommendation offers an absolute guarantee for methodological soundness. It remains possible, for instance, that random performance in the control groups results from the influence of several non-specific variables acting in opposite directions. Of course, replications of studies with different materials and conditions are necessary.

IS THIS DISCUSSION REALLY IMPORTANT?

It could be tempting to discard these points as methodological refinements deprived of any implications. On the contrary, we believe these methodological issues to be fundamental. First of all, we think of the principles involved here as quite general across the whole learning area. The details of the arguments and illustrations would obviously differ from one situation to another (in fact, even for artificial grammar learning studies, other arguments could be worked out), but we are aware of no situation in which a control group without training should be sound. Note that in conditioning research, in which methodological matters have been extensively explored long ago, the use of untrained control groups has been abandoned for three or four decades (see Prokasy & Kumpfer, 1973). Careful consideration of these principles is especially important in implicit learning research, because the reported difference between experimental and control groups is often small. In these conditions, an even minor bias may be erroneously interpreted as evidence for learning.

The methodological principles recalled here are pre-theoretic, because they apply regardless of the researcher's theory of learning. Dienes and Altmann seemingly inferred from our arguments that theories are not important to us. We do believe that no theory can save an ill-designed experiment. But of course, our focus on pre-theoretic principles does not mean that we think that theories are irrelevant in designing experiments. The need for a control group that is submitted to some form of training is a methodological prerequisite. By contrast, as we claimed in our paper, "designing control material is a theoretically motivated task".

REFERENCES

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

Prokasy, W.F., & Kumpfer, K.L. (1973). Classical conditioning. In W.F. Prokasy & D.C. Raskin (Eds.), *Electrodermal activity in psychological research*. New York: Academic Press.

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

Schwarz, N. (1994). Judgment 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.

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 26 February 2002
Accepted revision received 22 April 2002