

Does Guessing Incorrectly Impair Fact Learning?

Harold Pashler¹, Michael C. Mozer²,

Nicholas J. Cepeda^{1,3}, Doug Rohrer⁴, and Shana K. Carpenter¹

¹Department of Psychology, University of California, San Diego, La Jolla, CA

²Department of Computer Science, University of Colorado, Boulder, CO

³Department of Psychology, York University, Toronto, Ontario

⁴Department of Psychology, University of South Florida, Tampa, FL

ABSTRACT

In learning new information through testing with feedback, is it best for the learner to guess when unsure, or does that cause errors to become engrained? On day 1, subjects learned 80 facts, and were then tested on each. When a subject reported being unable to give an informed response, on a randomly chosen half of all trials the computer insisted upon a guess and provided corrective feedback (either immediately, in Experiment 1, or at a delay, in Experiment 2). Subjects returned for a test on day 2. Forced guessing neither impaired nor facilitated day 2 recall. However, items that spontaneously elicited erroneous guesses on day 1 were more likely to be correctly recalled on day 2 than were items on which the subject refused to guess. Why would spontaneous guessing bode well, while forced guessing had no effect? An explanation in terms of error-correction learning algorithms is suggested, and simulations are presented to show the potential viability of this account. Practical implications are also discussed.

More than a century of research on learning and memory has provided students, instructors, and designers of computer-aided instruction systems with rather little concrete guidance on how to promote learning and retard forgetting. This conclusion is true even for the acquisition of facts, foreign language vocabulary, and other learning tasks with simple and clear criteria of success. One reason for the lack of practical translation from memory research may be that contemporary researchers have not focused on the sorts of concrete procedural choices that arise in real learning situations (Pashler, Rohrer, Cepeda, & Carpenter, 2007). Instead, investigators have tended to focus more on the nature of the materials to be learned, or the type of test used to assess learning—both of which are likely, in real-world contexts, to be dictated by the task at hand.

Does Guessing Inhibit Learning?

The present article examines one particularly obvious and concrete question that confronts learners frequently, but which has been the focus of little study: when individuals have studied some information but have not yet fully mastered it, should they guess when presented with a quiz item on which they cannot answer with any confidence? Or should individuals avoid producing low-confidence guesses? The importance of the question is amplified by the fact that quizzing is normally far more useful in potentiating learning than re-studying or re-reading (for a review, see Roediger & Karpicke, 2006). The question also arises commonly in daily life; for example, when one is not sure how to spell a word, one has the choice of whether to guess, or to avoid doing so until reliable information is available.

In addition to its practical import, the question also bears on important theoretical issues about the mechanisms of learning. The learning theorist Edwin Guthrie famously claimed that people “learn what they do” and consequently, according to Guthrie, making an error stamps in undesirable stimulus-response associations (e.g., Guthrie, 1952). Guthrie assumed that this would happen even when the learner recognized that he or she had made an error. The idea that producing errors has undesirable consequences was also advocated by researchers in the Skinnerian tradition (e.g., Taber, Glaser & Schaefer, 1965; cf. Vargas, 1986). Further, understanding the conditions giving rise to voluntary guessing and the conditions under which guessing is beneficial is likely to serve as an important source of constraint on computational models of learning and memory. We consider this issue in the General Discussion.

Present Experiments

In each of the two studies described below, subjects learned a set of 80 facts they were not likely to have encountered before training. Then, within the same session, they were quizzed on each of these items. When the subject indicated that “I have no idea”, on a randomly chosen half of the trials, he or she was required by the computer to type in a guess. Regardless of whether a forced guess was required or not, the subject was always shown the correct response to each item. Subjects returned the next day and were tested on all items without feedback.

This design allows us to pose a number of unanswered questions about the effect of guessing. First, we can compare performance on (a) trials where the subject spontaneously answered – *and was wrong*; (b) trials where the subject responded correctly; and (c) trials

where the subject did not spontaneously answer. Intuition would seem to suggest that guessing wrong (an *error of commission*) would be associated with worse performance on that item the following day, because the erroneous answer would need to be unlearned. Second, and perhaps most pertinent to the questions posed above, we can ask whether being forced to guess affects the degree of learning seen the following day. The Guthrie account—according to which the production of an erroneous guess strengthens learning that will compete with correct learning triggered by the corrective feedback—obviously implies that it will be harmful, assuming that it often elicits an error (as the data will show).

Experiment 1

Method

Subjects. 65 students from the University of California, San Diego, participated in a two-session experiment.

Materials and Stimuli. A list of 80 not-well-known facts was assembled. For each fact, there was a corresponding question and correct answer or answers. For example, one fact was “Richard Nixon set up the only hamburger stand in the South Pacific during World War II.” and the corresponding question was “Who set up the only hamburger stand in the South Pacific during World War II?”, to which “Richard Nixon” was the correct answer.

Design and Procedure. The overall method is summarized in Figure 1. The subject participated in two sessions separated by one day. The first session was a training session. This consisted of one sequential presentation of the facts, followed by a quiz on all the items. In the presentation, the subjects saw each of the 80 facts in a random order, for four seconds per fact. Before the quiz, subjects were told that on some trials the computer would insist

that they respond, and were cautioned that when this happened, they were not to give a “joke answer” (pilot work indicated subjects sometimes did this in the forced guess condition). For every fact, the subject was first asked the question corresponding to that item. If the subject wished, he or she could click “I have no idea”. If the subject did respond, he or she was required to indicate one of five confidence values ranging from “Very Low” to “Very High”. When the subject clicked on “I have no idea”, the computer randomly determined (with probability .5) whether the item would be assigned to the Forced Guess condition (in which case the subject was required to guess), or the Continue condition (in which case the subject was not required to guess). In either case, the next event to occur was that the computer provided feedback on the item, showing both the question and the correct response for 1 second.

On the second day, the subject was tested on all the items in a random order. In this phase, an answer was required on all trials, and no feedback was given.

Results and Discussion

Each of 65 original participants was tested on 80 items, except for one subject who was tested on only 79 items due to computer malfunction, resulting in a total of 5199 items tested.

Day 1 performance

On Day 1, subjects spontaneously provided an answer on approximately 69% of all trials (3590 of 5199, 69.1%). On the remaining approximately 31% of trials (30.9%) where

subjects reported having “no idea” of the answer, the computer required Forced Guessing on 804 trials (15.5%), and did not require a guess on 805 trials (15.5%).

Of the 3590 responses produced without any compulsion, 2151 (59.9%) were correct. On the other hand, forced guesses were very rarely correct. Of 805 forced guesses, only 11 (1.4%) were correct. In short, when participants said “I have no idea”, they were generally correct in this assessment.

Confidence and Day 1 Performance

When subjects responded without compulsion, they indicated their confidence, and the distribution of these ratings is seen in Table 1. Pooled and per-participant analyses of Day 1 responding indicated a very similar relationship between confidence and success, as seen in Figure 2. Responses made with highest confidence were correct 86.2% of the time, whereas responses made with lowest confidence were correct 8.5% of the time. Confidence was highly correlated with performance, by both pooled ($r^2 = .997$) and per participant scoring ($r^2 = .988$, $p < 0.001$).

Day 2 Overall Performance

When they returned for a final test on Day 2, subjects were required to respond on all trials, and they were correct on 53.1% of these. This is about 10 percent below the overall Day 1 performance (59.9%). Several factors likely contributed to this difference, including (a) new learning taking place when feedback was provided, as happened on all trials, (b) forgetting, and (c) the fact that all items, including the difficult items for which subjects did not volunteer a response in Day 1, were included in the Day 2 test.

Day 2 Performance as related to Day 1 Performance

Table 2 shows performance on Day 2 as a function of the subject's response to an item on Day 1. For items answered correctly on Day 1, Day 2 accuracy averaged 91.8%. For items answered incorrectly on Day 1, Day 2 accuracy averaged 32.7% . For items eliciting "I don't know" on Day 1, Day 2 accuracy averaged 21%. The first of these three values (Day 2 accuracy given spontaneous Day 1 accuracy) was significantly greater than both the second ($F(1,64) = 995.0, p < 0.0001$) and the third ($F(1,64) = 987.3, p < 0.0001$) .

With regard to the second and third values, Day 2 performance was better for items that elicited spontaneous errors on Day 1 than for items that elicited "I have no idea" on Day 1 ($F(1,64) = 26.1, p < 0.0001$) . This effect, which is similar to the "high confidence hypercorrection effect" previously observed by Butterfield and Metcalfe (2001; see also Pashler, Zarow, and Triplett, 2003), represents roughly half a standard deviation ($d = 0.74$), a medium-sized effect according to the criteria of Cohen (1988).

Effects of Forced Guessing

When subjects indicated "I don't know", on a randomly chosen half of the trials, the computer insisted that they produce a guess (as we have seen, these guesses were scarcely ever correct). The Continue condition exhibited marginally better performance (21%) than the Forced Guess Condition (18.3%) on day 2, but this difference was not statistically significant ($F(1,64) = 2.44, p = 0.12$).

Day 2 performance for Day 1 spontaneous guesses with the lowest possible confidence (29.7%) was superior to either Forced Guess or Continue "I have no idea"

conditions. This advantage of ~9% was statistically significant $F(1,64) = 12.8, P < 0.001$, ($d = 0.63$).

Confidence and Day 2 Performance

Not surprisingly, confidence of correct Day 1 responses was also positively related to correctness on Day 2 (pooled analyses $R^2 = .95$; per-participant $R^2 = .81, p = .037$). Less obviously, for items eliciting a spontaneously produced error on Day 1, there was a trend for confidence on Day 1 to show a positive relationship to correctness on Day 2 (pooled $R^2 = .84$, ; per-participant $R^2 = .66, p = .093$). That is, greater confidence in spontaneously produced Day 1 errors was associated with greater accuracy on Day 2. One might have expected instead that when subjects spontaneously produced the wrong response, the more confidence they had in this response, the more difficult it would be to unlearn it and replace it with the correct response (Butterfield & Metcalfe, 2001).

Summary of Results

This study examined the consequences of being forced to guess in a fact learning task on trials in which the subject has said that he or she has “no idea” of the answer. Two basic findings emerged. First, being forced to guess has no significant effect on the accuracy of recollecting the correct response the following day. The data offer little support for the Guthrie hypothesis that producing an error automatically “stamps it in.” Guessing erroneously does not seem to have any notable cost for learning, or if it does, the costs are approximately canceled out by some comparable benefit. Of course, this conclusion may be limited to the case where immediate corrective feedback is provided, as in the present study.

Second, willingly-offered incorrect responses are a positive prognostic sign, in the sense that they indicate a greater likelihood of ultimately achieving correct learning, as compared to a refusal to guess (cf., Butterfield & Metcalfe, 2001). This, too, would seem surprising if one assumes that incorrect guesses reflect associations that will need to be unlearned before the correct linkage can be developed.

Experiment 2

In the first experiment, feedback was provided immediately after subjects responded within the test in Session 1. One might suppose that the harmlessness of forced guessing could be limited to this case in which corrective information is provided without delay. This question is addressed in Experiment 2.

Method. Experiment 2 was just like Experiment 1, except that feedback was provided only after the completion of the quizzing on all items in Session 1. After subjects had responded to all 80 questions, the list of 80 facts was shuffled and each of the facts (which connoted the question and correct answer) was displayed for 4 seconds per fact.

Results.

42 participants were tested on 80 items resulting in a total of 3360 items.

Day 1 performance

On Day 1, subjects spontaneously provided a response on approximately 72% of the trials (2422 of 3360). On the remaining 27.9% of trials where subjects did not spontaneously respond, the computer required Forced Guessing on 483 trials (14.4%), and assigned 455 trials (13.5%) to the Continue Condition (going straight to provide feedback). Of the 2422 responses produced without any compulsion, 1418 (58.5%) were correct.

As in Experiment 1, forced guesses were rarely correct. Of 483 forced guesses, just 14 (2.9%) were correct.

Confidence and Day 1 Performance

The distribution of confidence responses is shown in Table 3. As in Experiment 1, pooled and per-participant analyses of Day 1 responding both yielded a very similar and strongly positive relationship between confidence and success. Responses made at highest confidence were correct 82.5% and responses made at lowest confidence were correct 8.6%. As in Experiment 1, confidence was highly correlated with performance for both pooled ($r^2 = .989$) and per participant scoring ($r^2 = .983$, $p = 9.2e-4$).

Day 2 Performance

On Day 2, subject's accuracy averaged 56% (1882/3360). Table 4 shows performance on Day 2 as a function of responses to a given item on Day 1. For items answered *correctly* (and spontaneously) on Day 1, Day 2 accuracy averaged 93.8%. For items answered *incorrectly* (and spontaneously) on Day 1, Day 2 accuracy averaged just 32.3%. For items to which subjects responded "I have no idea" on Day 1, Day 2 accuracy averaged 22%. The first of these numbers was significantly different than both the second ($F(1,41) = 709.7$, $p < 0.001$) and the third ($F(1,41) = 438.3$, $p < 0.001$).

As in Experiment 1, Day 2 performance was better for items on which subjects volunteered an incorrect response on Day 1 as compared to items that elicited "I have no idea" responses ($F(1,41) = 11.5$, $p = 0.002$), $d = 0.41$.

Confidence and Day 2 Performance

Not surprisingly, for correct Day 1 responses, confidence was related to correctness on Day 2: the higher the confidence, the better the Day 2 performance (pooled analyses $R^2 = .87$; per-participant $R^2 = .96$, $p = .004$). Less expectedly, and as in Experiment 1, there was a trend for confidence on Day 1 to have a positive relationship to accuracy on Day 2 for items that elicited voluntary but incorrect responses on Day 1 (pooled $R^2 = .62$; per-participant $R^2 = .50$, $p = .18$), echoing Butterfield and Metcalfe (2001).

Effects of Forced Guessing

In this experiment, subjects did marginally *worse* on Day 2 for items on which they were not forced to guess (22.4%) as compared to items on which they were forced to guess (26.1%), but this difference was not statistically significant $F(1,41) = 0.77$, $p = 0.386$.

Day 2 performance for Day 1 spontaneous guesses with the lowest possible confidence (32.0%) was superior to either Forced Guess or Continue “I have no idea” conditions. This ~6% advantage was statistically significant, $F(1,41) = 7.7$, $p = 0.001$, ($d = 0.39$).

Discussion

As in Experiment 1, there was no significant effect of being forced to guess when subjects thought they lacked the correct answer. While it was our suspicion that in the absence of immediate feedback, forced guessing might well have deleterious effects, the results actually showed a trend in the opposite direction, with a slight benefit for forced

guessing (in Experiment 1, the trend ran in the opposite direction). Putting the results of the two experiments together, it seems reasonable to conclude that forced guessing does not impair fact learning to any notable degree, when feedback is provided fairly soon¹ (but not necessarily immediately). As in the previous experiment, we also find that spontaneous errors of commission are more likely to be corrected, as manifested on a later test, than are errors of omission (cf. Butterfield & Metcalfe, 2001).

General Discussion

The conclusions of the present study have both practical and theoretical interest. We begin with the practical implications for instruction.

Practical Implications

The results provide some reassurance that requiring a learner to produce factual information even when the learner is unsure does not seem to produce a detrimental effect on the ability to profit from feedback. This is important, because there is a growing body of evidence indicating that requiring retrieval promotes learning and retention, as compared to rereading or other modes of practice that do not require retrieval (Roediger & Karpicke, 2006), and this “retrieval practice” will, of course, occasionally produce incorrect answers. It should be pointed out, however, that the lack of any harmful consequences of erroneous guessing may well have important boundary conditions remaining to be determined. It seems

¹ We also ran another experiment like the two presented here, but in which no feedback was provided. The results showed day-2 performance very near zero for both the Continue and Forced Guess conditions, which is not surprising in light of prior findings on the necessity of feedback after errors (Pashler, Cepeda, Wixted, & Rohrer, 2005).

easy to imagine several ways in which the present findings might not generalize to other important situations, of which the following are a few examples.

First, it may be that there are forms of non-declarative memory for which producing an error does indeed “stamp in” responses. For example, it seems intuitively plausible that poorly chosen actions might have lasting negative consequences in motor learning, e.g., producing bad golf swings or tennis serves.

Second, even for verbal materials, there may be certain types of learning tasks in which familiarity may play an especially critical role, and for which incorrect guessing might therefore be harmful. Examples are pronouncing and spelling words. In these cases, performance may be limited by different factors than in fact memory, as studied in the present report. When a fact has not been well learned, learners appear to be unable even to retrieve any candidate answers. By contrast, in spelling and pronunciation, it may often be easy to think of plausible candidates but hard to choose amongst them, and familiarity may play a key role in this choice. For this reason, a self-produced error may render the wrong spelling or pronunciation more familiar, and have a long-lasting negative effect (see Jacoby & Hollingshead, 1990, for evidence that being exposed to other people’s spelling errors impairs subsequent spelling performance).

Third, it should be noted that our subjects were clearly aware that the responses they were producing were almost certainly errors (indeed, the forced guessing manipulation was imposed only on trials in which the subject said they had no idea of the answer). Making errors in situations where errors are easily recognized as such by the subject might not be as harmful as guessing in more ambiguous situations. Thus, while the current results find little

harm in guessing in the situation where the error rate is highest, it does not automatically follow that guessing in cases with lower error rates will necessarily also prove harmless.

This discussion points out the fact that further empirical efforts will be needed before cognitive psychology can really offer learners and teachers reliable advice about the pros and cons of guessing in a wide range of situations.

Theoretical Implications

The results also bear on basic questions about the underlying mechanisms of fact learning. In previous work on testing effects, we have found it helpful to model human associative learning within the neurocomputational framework of error correction learning (Mozer, Howe, & Pashler, 2004; see also Howard & Kahana, 2002). In the remainder of this article, we suggest that the same framework may provide an illuminating perspective on the present results.

There are four basic phenomena that need to be accounted for in the current pattern of findings (for simplicity, here we refer to the Day-1 test as “Test 1” and the final delayed test as “Test 2”):

1. Overall accuracy improves from Test 1 to Test 2.
2. Accuracy on Test 2 is higher if the item was correctly retrieved on Test 1 as compared to the case in which the item elicited an error.
3. Accuracy on Test 2 is higher if the subject ventured a guess on Test 1 and was wrong, as compared to when the subject was unwilling to guess.
4. Forcing a guess on Test 1, after the subject has declined to guess, does not affect accuracy on Test 2.

Phenomenon 3 is puzzling if one makes the intuitively reasonable assumption that when someone guesses erroneously, this means that the learning they have achieved is in fact erroneous, and therefore must be unlearned before the correct learning can take place. We will see shortly that simulations in the error-correction learning framework allow for a very different interpretation of the phenomenon.

One interpretation of Phenomenon 2 was offered by Butterfield and Metcalfe (2001), who noted that after feedback, people are more accurate at remembering items on which they produced wrong responses with high confidence. (We observed trends in the same direction in the present study, as well.) Butterfield and Metcalfe suggested that this may be caused by what they termed *hypercorrection*, in which the surprise resulting from finding out that a high-confidence response was wrong actually promotes subsequent memory storage once the corrective feedback is available.

We offer an alternative account of these four findings, which does not rely on the assumption that surprise potentiates memory storage. The basis for our account is rather counterintuitive: within a connectionist learning framework, incorrect guesses (as compared to not guessing) may actually be a sign of having achieved a greater degree of partial learning *of the target-relevant material* than would be indicated by a failure to respond.

A Model of Feedback-Based Learning

We begin by sketching a qualitative account of the phenomena observed in our experiments. Our account depends on the assumption that each fact stored in long-term memory can be characterized by the notion of a *memory trace strength*, which reflects the degree to which the fact is stored. The trace strength determines the probability of recall and

also the amount of additional learning required for robust retention. The expected trace strength starts at zero and increases with each training trial (thus explaining phenomenon 1).

Using the notion of trace strength, we can characterize why the type of response an individual makes on Test 1 (correct, error, no guess) predicts performance on Test 2. Simply, Test 1 performance reflects the strength of a memory trace, and the trace strength on day 1 is correlated with the trace strength on day 2 (Test 2). Thus, phenomenon 2 arises because a correct Test 1 trial implies a stronger memory trace than an incorrect Test 1 trial; and a stronger memory trace at time of Test 1 implies a stronger memory trace at the time of Test 2, even allowing for memory decay or interference (assuming that these mechanisms operate independently of the Test 1 response type).

Our account of phenomenon 3 relies on the fact that willingness to guess is correlated with trace strength. With a very weak trace, no guess is ventured. With a strong trace, the correct response will be produced. With an intermediate trace strength, a guess will be ventured but it will often be wrong. Thus, guesses are not an indication that the system is “below zero” in the amount of target-relevant learning that has taken place, as one might have supposed. Instead, guesses actually index a partial state of learning of the relevant material.

Phenomenon 4 will be explained by assuming that the memory trace is affected only by supervised training—i.e., being shown the correct answer—and not by the act of guessing (consistent with the model of Mozer, Howe, & Pashler, 2004).

Error-Correction Learning Model

We implemented our qualitative account in an associative neural network, which takes the question portion of a fact as its input and produces the answer as its output. The trace strength of an

association in a neural network is difficult to assess, because memories are distributed over weights in the network. We thus suppose that the *magnitude of the network output* (e.g., the length of the output activity vector) might serve as a proxy for trace strength. Because networks are usually initialized with small weights, the network response is small in magnitude at first, but gradually grows as the weights grow. This notion of trace strength will be used by the model to determine when to respond, and with what degree of confidence to respond. Although it seems as if we might be able to account for the four key phenomena within this framework, a simulation is necessary, because the simulation might not play out as we hope. For example, the network might produce incorrect responses that are large in magnitude, and that would imply an error trial that is difficult to correct with subsequent training (and hence work against phenomenon 3).

We evaluate an extremely simple neural network model, a linear associator with α input units and α output units, trained on a set of stimulus-response paired associates. The stimulus and response items are represented as distributed vectors in $\{-1,+1\}^\alpha$. The vectors are generated at random, with +1 and -1 values equally likely. The network also includes α biases on the output units. The network is trained via a supervised-learning gradient descent procedure (i.e., the LMS or error-correction learning rule; Widrow & Hoff, 1960) to minimize the mean squared error. At the start of each simulation, all weights in the network are initialized to be zero. Free parameters included the learning rate of the neural net for each item i , η_i , and the number of distinct facts to be learned, μ .

Simulation procedure

The set of μ facts is presented to the network in three blocks, where each block is one pass through the facts to be learned. In the first block, supervised training is performed. In the second block, corresponding to Test 1, each fact is tested and then trained. In the third block, corresponding to Test 2, each fact is tested. Within a block, facts are presented in random order, with a different order for each block. We did not model decay or interference effects

occurring in the time interval spanning Test 1 and Test 2.

Interpreting network output

On each test trial, the network output must be classified as correct, error, or no guess. To perform this classification, we define a set of *well-formed states*, vectors in the α -dimensional output space that serve as possible responses. We calculate the distance of the actual network output to each of the well-formed states, and choose the well-formed state that has the smallest distance. The well-formed states consist of: (1) all possible target response vectors, (2) an additional set of v distractor response vectors randomly placed at corners of the α -dimensional hypercube, and (3) the origin. The origin corresponds to the no-guess state. To control the proportion of no-guess responses, we allow for a scaling factor on the distance to the origin.

To summarize, the network produces a response vector \mathbf{r} , and from this output a discrete response \check{r} is chosen according to $\check{r} = \operatorname{argmin}_{i=0\dots\mu+v} d_i$ where $d_0 = \|\mathbf{r}\|^2 / \sigma$ for the origin state, and $d_i = \|\mathbf{s}_i - \mathbf{r}\|^2$ for each of the corner states defined by vector \mathbf{s}_i , $i = 1 \dots \mu + v$; σ is the scaling factor on the origin.

This response read-out rule is very much like feeding the output of a linear associator into an attractor net to clean up the noisy output vector and interpret it in terms of one of the well-formed states (Sitton, Mozer, & Farah, 2000; Zemel & Mozer, 2001). Thus, the read-out rule has an interpretation in terms of neural net dynamics.

In a linear network with α input units, α arbitrary associations can be learned exactly if the input vectors are linearly independent. However, with the read-out rule we describe, the network can be loaded with an even larger number of associations, because it does not have

to produce exactly the right output; it only has to get closer to the correct well-formed state than any other well-formed state. Thus, although our simulations select $\mu > \alpha$, network accuracy can still asymptote at 100%.

Simulations

Our simulation consisted of a training set of $\mu = 40$ facts. We cut in half the number of facts relative to the human experiment due to the time each simulation required. Each simulation involved 200 replications (“subjects”). The replications differed in the random vectors used for training and the random ordering of examples within a block.

In each simulation, we adjusted model parameters to achieve an approximate match to the distribution of responses on Test 1 and Test 2 observed in the human study, and then examined the distribution of correct responses on Test 2 conditional on Test 1 response type.

Picking learning rates is always difficult, because the optimal learning rate depends on the size of the network, and simulation results can be sensitive to the choice of learning rate. For linear nets with α inputs and a bias weight, training error on the current example will go to exactly zero if a learning rate of $\eta = 1 / (\alpha + 1)$ is used. Thus, any larger of a learning rate will overshoot the target. However, the chosen learning rate should be smaller still because perfect learning on the current example often results in large interference with recent examples. Thus, we used a default learning rate of $\eta = 1 / 2 (\alpha + 1)$ which seemed to work well across a range of network sizes and simulations.

Simulation 1

For Simulation 1, we picked $\sigma = .45$, $\alpha = 30$, $v = 300$. The results are summarized in Table 5. Column 2 is the human data, column 3 is the simulation result. Figure 4 plots Test 1

and conditional Test 2 probabilities (the first and third groups of results in the table).

The simulation shows a good qualitative match to the data. Probability of the three response types is about equal for Test 1, and the probability of a correct response on Test 2 conditional on the response type of Test 1 is largest for Test 1 correct, then Test 1 incorrect, then Test 1 no guess. Although the difference on Test 2 accuracy between Test 1 error and Test 1 no guess—let's call this the *confidence effect*—is only 5%, it is a reliable difference due to the fact that the simulation involves so many replications. Thus, the simulation result is compatible with our hypothesis that the magnitude of the associative output can be used to decide whether to make a guess, and a larger magnitude output is more likely to be correct. We later test this hypothesis explicitly, but first deal with the fact that the Test 2 correct conditional probabilities do not quantitatively match the human data.

Simulation 2

In a neural net, interference between two stimulus vectors increases as the angle between the two vectors decreases. By generating stimulus-response associations for training at random, most stimulus vectors will be nearly orthogonal to one another, but by chance, some stimuli will overlap others, resulting in more interference, and making associations involving these stimuli intrinsically more difficult to learn. We conjecture that the results of Simulation 1 are due to some items being more difficult than others. The more difficult items are the ones for which no guess is given on Test 1, and which are less likely to be correct on Test 2. One piece of evidence in support of this conjecture is that the confidence effect became less pronounced when we increased the dimensionality of the input from $\alpha = 30$ to $\alpha = 40$, which yields stimulus vectors more likely to be orthogonal to one another.

In Simulation 2, we ensured that some items would be more difficult than others by assigning half of the associations a learning rate equal to that used in Simulation 1, and half of the associations a learning rate which was $1/3$ of that used in Simulation 1. Other parameters were the same as Simulation 1, except σ was reduced to 0.30. Figure 5 shows the result, which is an excellent quantitative fit to the human data.

Our explanation of this result consists of two arguments. First, as an association is trained, the output magnitude increases as a function of both the learning rate and the number of training trials. Second, the distribution of responses changes as a function of output magnitude: when the magnitude is small, no-guess responses are produced, when the magnitude is large, correct responses are produced, and in between error responses are produced.

The first argument is supported by Figure 6 which shows the mean network output

magnitude, $\|r\|$, as a function of both number of training trials and learning rate. The Figure indicates that the output magnitude grows with the amount of learning that has transpired. The second argument is supported by Figure 7, which shows the frequency of no-guess, error, and correct trials, as a function of the output magnitude. As conjectured, the smallest magnitude responses are no-guess trials, the largest magnitude responses are correct trials, and error trials are in between.

The results we obtained were robust to a variety of manipulations, which we summarize here.

Weight initialization. In the simulations reported above, weights were initialized to zero. We explored two alternative weight initialization procedures, and neither had a qualitative impact on the results. One scheme was to choose the initial weights from a Gaussian distribution and then renormalize by scaling such that the L1 norm of the fan-in weight vector is 1.0. Another scheme was to set the initial weights for simulation $n+1$ to be the final weights of simulation n .

Response selection. In the simulations reported above, the chosen response was the well-formed state closest to the network output. We also explored a response selection rule in which the probability of choosing response i was inversely related to the distance of the network output from v_i . That is, instead of choosing the closest well-formed state as the response, responses were chosen according to a softmax function or Luce choice rule.

Training set size (μ). We explored training set sizes ranging from $m = 20 \dots 100$, and as long as the network size was scaled to be on the same order, the results were similar. This is because with μ on the same order as α , the amount of interference between vectors is constant.

Variation in learning rates. In Simulation 1, learning rates of all associations were equal. Simulation 2 consisted of easy and hard associations, defined by a large and small learning rate, respectively. We also explored simulations in which the learning rates were chosen randomly from uniform or Gaussian distributions. The qualitative pattern of results was the same as Simulation 2.

Number of distractor responses (v). In Simulations 1 and 2, we included $v=300$ additional well-formed states to serve as distractor responses. Varying v has a very minor effect on performance. As v increases, the error rates increases slightly. However, even with $v=1000$, only a miniscule proportion of the 2^{α} corners of the response hypercube are populated. Thus, landing near a distractor is unlikely.

Discussion of Model

The basic confidence effect, i.e., $p(\text{test 2 correct} \mid \text{test 1 error}) > p(\text{test 2 correct} \mid \text{test 1 no guess})$, occurred robustly in all simulations. The magnitude of the effect was smaller for some parameter settings, but in only a few cases did the pattern disappear or reverse, and these cases were pathological for other reasons (e.g., initial performance at floor, final performance at ceiling).

Variability in association difficulty seems to be the key factor in obtaining the effect. In our simulations, association difficulty was affected by two factors. First, by random choice of stimulus and response vectors, some associations suffered interference due to their similarity to other associations in the pool (more specifically, the interference is caused by similarity in the stimulus space). Second, by varying individual association learning rates, we made some items intrinsically more difficult to learn.

Other than the requirement of variability in association difficulty, our model is a standard neural network with a fairly standard set of assumptions for how responses are read out of the network. These assumptions are based on the notion of trace strength. Trace strength is typically conceived of as a property of connections or weights in the brain's neural network, but the construct we proposed is based on the output activity from the neural network, and is thus a source of information that individuals might feasibly use in metacognitive judgements. We showed how this notion of trace strength produces a plausible interpretation of how individuals decide when to guess. But most importantly, we showed that in this account, willingness to guess is a reflection of degree of learning.

Concluding Summary

The present article presented the results of two studies examining the consequences for retention when subjects were tested on some facts that they had just learned, sometimes to the point of being “forced” to guess on facts that they were completely unsure about--and were then given feedback on all of the facts (either immediately or at a short delay). The results showed that a willingness to guess the wrong answer on Day 1 was associated with better performance on the final test, and that being forced to guess did not have adverse effects on the final test. The results provide some limited reassurance that use of retrieval tests early in instruction may not have negative consequences, even when they elicit a great many errors. We also described a possible theoretical interpretation of the results based upon error-correction learning. By assuming that (a) error-correction learning is not initiated until feedback is present, but also that (b) erroneous responding may emerge as the system achieves partial learning *of the relevant information*, we saw that a reasonable explanation of

the pattern of results was possible. In future research, it will be interesting to see if it is possible to develop additional tests of this analysis, and to contrast it with competing accounts.

AUTHOR NOTE

This work was supported by the Institute of Education Sciences (US Department of Education, Grants R305H020061 and R305H040108 to H. Pashler) and National Science Foundation (Grant BCS-0720375, H. Pashler, PI; and Grant #SBE-0542013, G. W. Cottrell, PI). Phil Starkovich was responsible for programming the experiments, and Shirley Leong carried out data analysis.

Correspondence concerning this article should be addressed to Hal Pashler, Department of Psychology, 0109, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0109. E-mail: hpashler@ucsd.edu

References

- Butterfield, B. & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 1491-1494.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cunningham, D. J., and Anderson, R. C. (1968). Effect of practice time within prompting and confirmation presentation procedures on paired associate learning. *Journal of Verbal Learning and Verbal Behavior*, 7, 613-616.
- Guthrie, E. (1952). *The Psychology of Learning (Rev. Edition)*. New York: Harper.
- Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46, 269–299.
- Jacoby, L. L., & Hollingshead, A. (1990). Reading student essays may be hazardous to your spelling: Effects of reading incorrectly and correctly spelled words. *Canadian Journal of Psychology*, 44, 345-358.
- Marx, M. H., & Witter, D. W. (1972). Repetition of correct responses and errors as a function of performance reward or information. *Journal of Experimental Psychology*, 92, 53-58.
- Mozer, M. C., Howe, M., & Pashler, H. (2004). Using testing to enhance learning: A comparison of two hypotheses. *Proceedings of the Twenty Sixth Annual Conference of the Cognitive Science Society* (pp. 975-980). Hillsdale, NJ: Erlbaum Associates.

- Pashler, H., Cepeda, N., Wixted, J., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 3-8.
- Pashler, H., Rohrer, D., Cepeda, N., & Carpenter, S. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review*, 14, 187-193.
- Pashler, H., Zarow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 29, 1051-1057.
- Roediger, H. L. & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1, 181-210.
- Sitton, M., Mozer, M. C., & Farah, M. J. (2000). Superadditive effects of multiple lesions in a connectionist architecture: Implications for the neuropsychology of optic aphasia. *Psychological Review*, 107, 709-734.
- Taber, J. I., Glaser, R., & Schaefer, H. H. (1965). Learning and programmed instruction. Reading, MA: Addison-Wesley.
- Vargas, J. (1986). Instructional design flaws in computer-assisted instruction. *Phi Delta Kappan*, 738-744.
- Widrow, B. & Hoff, M. E. Jr. (1960). Adaptive switching circuits. *IRE WESCON Convention Record*, 4, 96-104.
- Zemel, R. S., & Mozer, M. C. (2001). Localist attractor networks. *Neural Computation*, 13, 1045-1064.

Table 1

Experiment 1: Day 1 Success conditionalized on Confidence (Spontaneous Answering)

	Confidence				
	0	1	2	3	4
Day 1 Success, pooled	8.5%	26.0%	43.4%	66.3%	86.2%
Day 1 Success, per participant	8.4%	26.2%	42%	59.4%	86.6%
% of Day 2 Trials (5199)	1.2%	1.4%	3.1%	4.2%	31.5%

Table 2

Experiment 1: Day 2 Success conditionalized on Subject's Response to that Item on Day 1

	Day 1 Outcome				
	Spontaneous Response		"I don't know" Response		
	Correct	Wrong	Overall	Forced Guess	Continue
Day 2 Success, pooled	91.8%	32.7%	19.6%	18.3%	21.0%
Day 2 Success, per participant	90.7%	35.8%	21.0%	18.7%	22.7%
Pooled Count	2151	1439	1609	804	805
% of Day 2 Trials (5199)	38.0%	9.1%	30.9%	15.5%	15.5%

Table 3

Experiment 2: Day 1 Success conditionalized on Confidence (Spontaneous Guessing)

	Confidence				
	0	1	2	3	4
Day 1 Success, pooled	8.6%	21.2%	47.9%	67.0%	82.5%
Day 1 Success, per participant	11.4%	19.7%	45.2%	61.6%	82.6%
% of Day 2 Trials (3360)	1.2%	1.2%	3.1%	6.1%	30.6%

Table 4

Experiment 2: Day 2 Success conditionalized on Subject's Response to that Item on Day 1

	Day 1 Outcome				
	Spontaneous Response		"I don't know" Response		
	Correct	Wrong	Overall	Forced Guess	Continue
Day 2 Success, pooled	93.8%	32.3%	24.3%	26.1%	22.4%
Day 2 Success, per participant	92.7%	35.7%	27.3%	28.6%	25.7%
Pooled Count	1418	1004	938	483	455
% of Day 2 Trials (3360)	42.2%	29.9%	27.9%	14.3%	13.5%

Table 5

	Human Data	Simulation 1 Result	Simulation2 Result
Test 1 Correct	0.42	0.36	0.37
Test 1 Error	0.26	0.33	0.30
Test 1 No Guess	0.32	0.32	0.33
Test 2 Correct	0.54	0.66	0.53
Test 2 Error or No Guess	0.46	0.34	0.47
Test 2 Correct Test 1 Correct	0.92	0.78	0.96
Test 2 Correct Test 1 Error	0.35	0.61	0.36
Test 2 Correct Test 1 No Guess	0.20	0.56	0.20
Test 2 Correct Test 1 Forced Guess	0.19	0.56	0.20
Test 2 Correct Test 1 No Forced Guess	0.23	0.56	0.20

Figure 1a. Basic procedure on Day1 and Day2 of Experiments 1 and 2. See Figure 1b for the “Forced Guess Procedure” that occurs on approximately half of all trials on which subject declines to respond.

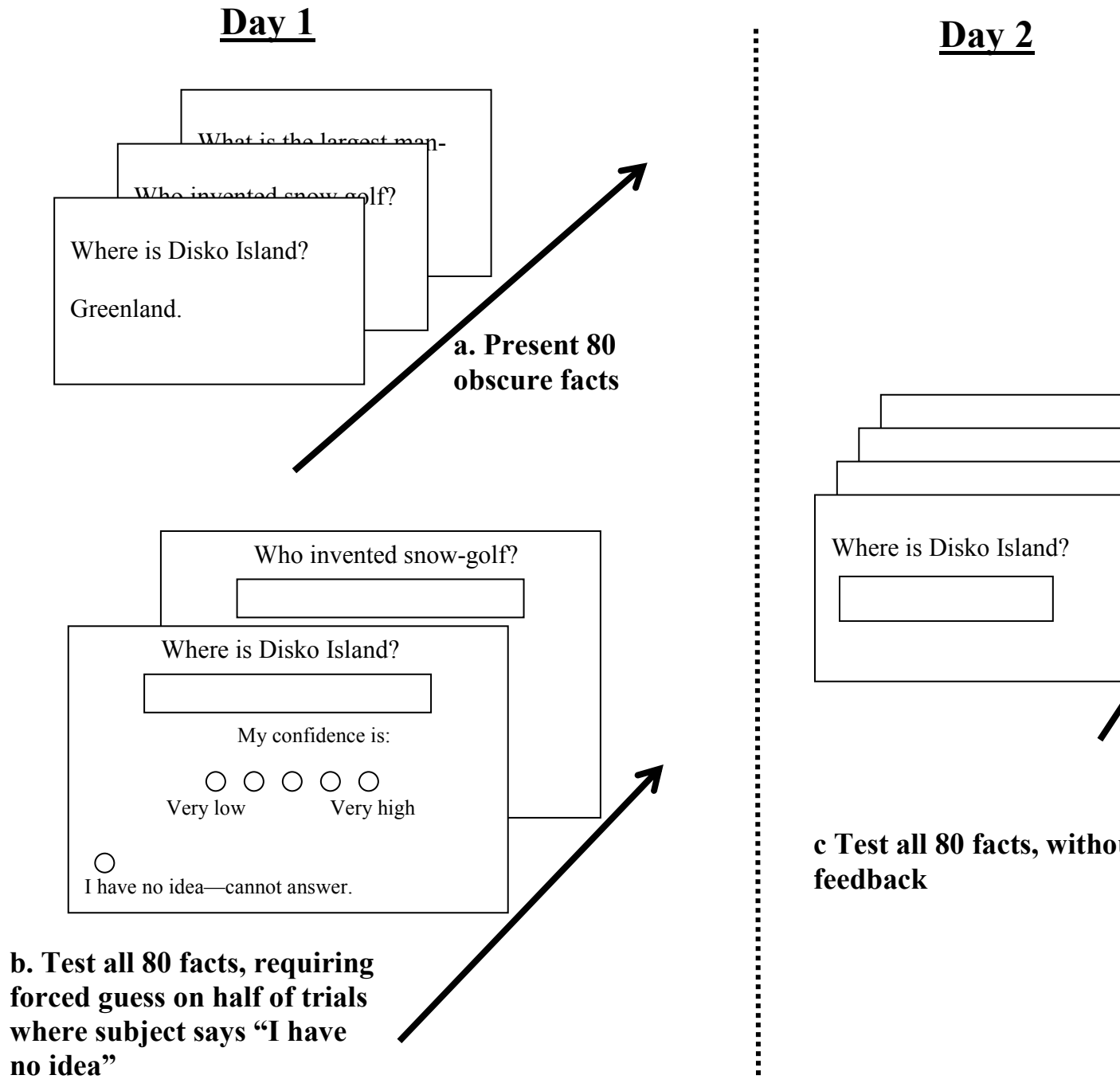


Figure 1b. Procedure on Day 1 trials where subject initially declines to respond, stating that they “have no idea”. On a random half of trials, the subject is “forced” to guess. Feedback is provided on all trials (immediately in Experiment 1; at the conclusion of the test run through all the items in Experiment 2).

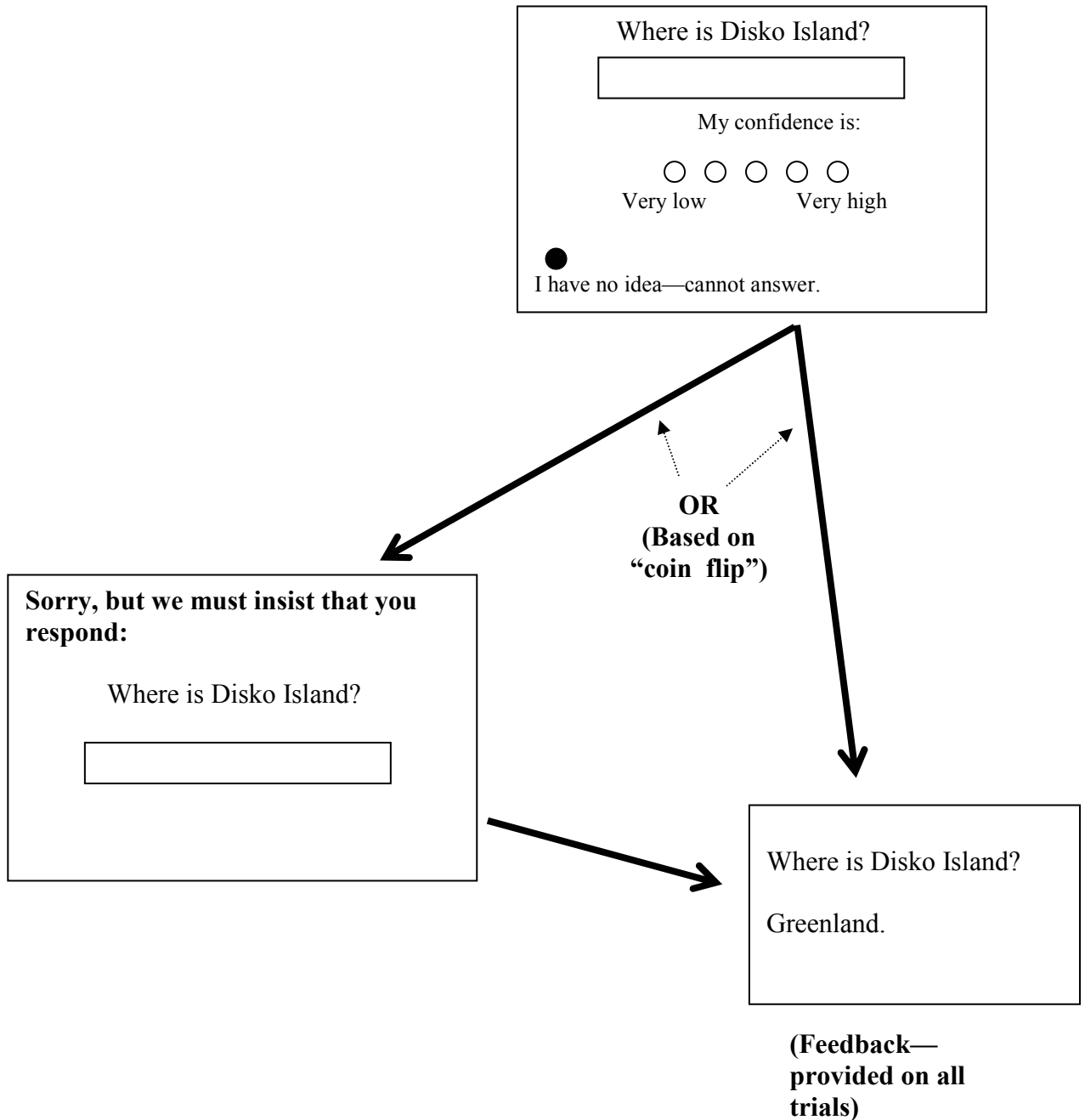


Figure 2

Experiment 1: Day 1 Performance by Confidence

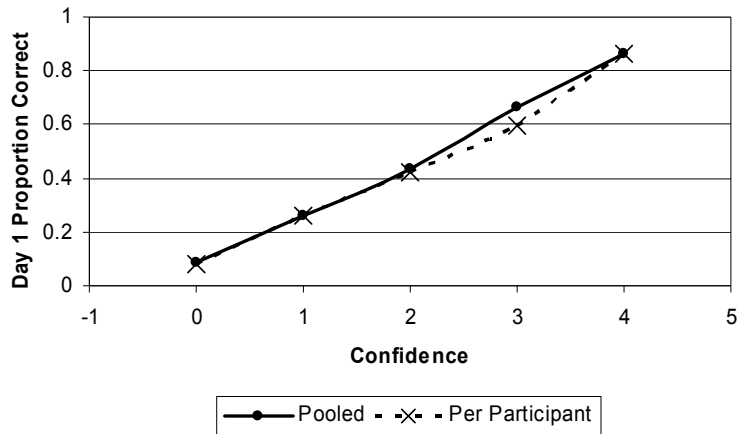


Figure 3

Experiment 2: Day 1 Performance by Confidence

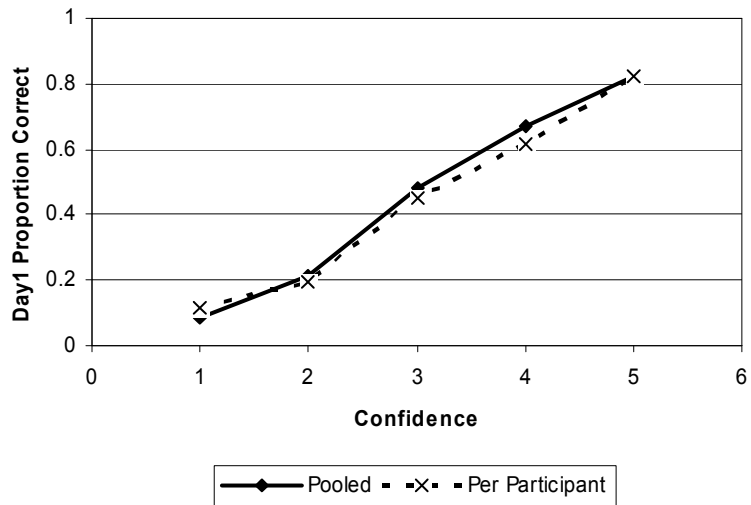


Figure 4

Simulation 1: Probability of Test 1 correct(c_1), Test 1 error(e_1), Test 1 no guess (ng_1), and probability Test 2 correct conditionalized on Test 1 correct ($c_2|c_1$), error ($c_2|e_1$), or no guess ($c_2|ng_1$).

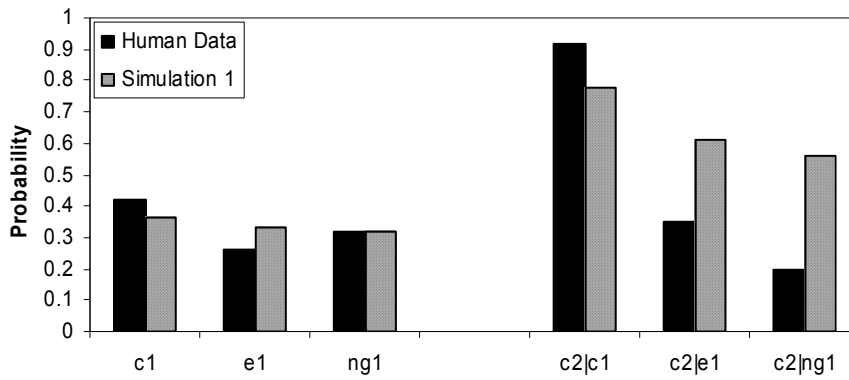


Figure 5

Simulation 2: Probability of Test 1 correct(c_1), Test 1 error(e_1), Test 1 no guess (ng_1), and probability Test 2 correct conditionalized on Test 1 correct ($c_2|c_1$), error ($c_2|e_1$), or no guess ($c_2|ng_1$).

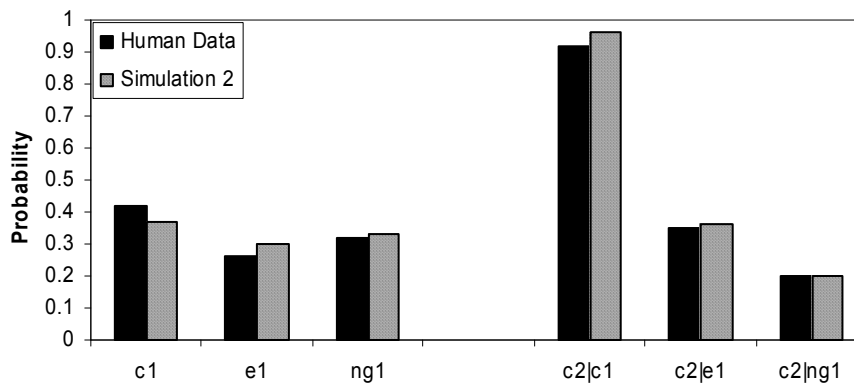


Figure 6

Magnitude of neural net response prior to training as a function of training trial and learning rate. Solid line is for the low learning rate, dashed line is for the high learning rate.

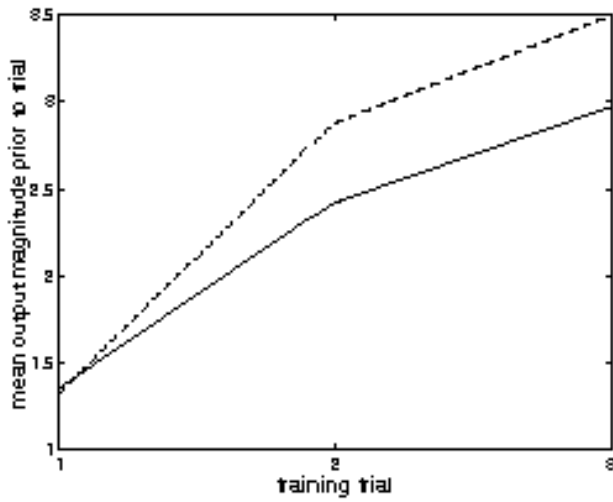


Figure 7

Frequency of no-guess (dotted line), error (dashed), and correct (solid) trials, as a function of response magnitude. These curves are based on the responses following 1 and 2 training trials.

