Psychon Bull Rev (2011) 18:998–1005
DOI 10.3758/s13423-011-0113-x

BRIEF REPORT

Effects of testing on learning of functions

Sean H. K. Kang · Mark A. McDaniel · Harold Pashler

Published online: 1 June 2011

Abstract Is learning of a complex functional relationship enhanced by trying to predict what output will go with a given input, as compared to studying an input–output pair? We examined learning of a bilinear function and transfer to new items outside the trained range. Subjects either saw the input–output pairs (study-only condition) or attempted to guess the output and then saw the pair (test/study condition). The total study times were equated, and motivation was enhanced with a monetary bonus. Performance was markedly better for the test/study condition, both within the trained range and in the transfer test. This benefit of testing during training was observed on a criterial test administered shortly after training. Testing has long been shown to enhance the explicit learning and retention of verbal material; our present findings reveal a novel domain for which testing can also be advantageous—that is, function learning.

Keywords Testing effect · Function learning

Studies of explicit memory have often found that the requirement to retrieve information enhances later memory more than does restudying of the material. This was first observed in studies of recall of word lists (e.g., Abbott, 1909; Allen, Mahler, & Estes, 1969; Lachman & Laughery, 1968). It has also been found in studies of foreign-language vocabulary learning. For example, when subjects were given 5 s to try to retrieve the English equivalent of an Eskimo word, and then saw both words for 5 more seconds, later memory performance was better than when the English and Eskimo words were both available for 10 s (Carrier & Pashler, 1992).

This benefit of testing over restudying—henceforth, the *testing effect*, also sometimes referred to as *retrieval practice*—has been found with learning of facts (McDaniel & Fisher, 1991), of face–name pairs (Carpenter & DeLosh, 2005), and even of visuospatial information (Carpenter & Pashler, 2007; Kang, 2010; see Roediger & Karpicke, 2006, for a review). The present article asks whether the benefits of testing are confined to tasks requiring literal reproduction or, rather, generalize to more complex forms of learning, such as function learning. Function learning is important in a wide range of tasks. Examples include learning perceptual–motor associations (see Koh & Meyer, 1991), predicting job performance from cognitive assessments, forecasting economic states (such as relating interest rates to inflation rates), predicting harvests from rainfall amounts, and many more (Hammond, 1955; Hoffman, 1960). Furthermore, continuous (functional) relations between causes and effects are fundamental to the formation of intuitive theories about how the world works, and these intuitive theories guide subsequent decisions about how to control our environments (Hammond & Stewart, 2001; also see Murphy & Medin, 1985). Here, we examine the learning of numerical functions.

In the present study, subjects were taught a bilinear function that mapped numbers between 1 and 100 onto a numeric output (see Fig. 1). One group of subjects learned the relationship of input to output values by observing pairings of inputs with their corresponding outputs (*study-only*). Another group was presented first with the input value and had to guess or predict the

S. H. K. Kang (✉) · H. Pashler
Department of Psychology, University of California, San Diego,
La Jolla, CA 92093–0109, USA
e-mail: seankang@ucsd.edu

M. A. McDaniel
Department of Psychology, Washington University in St. Louis,
St. Louis, MO, USA

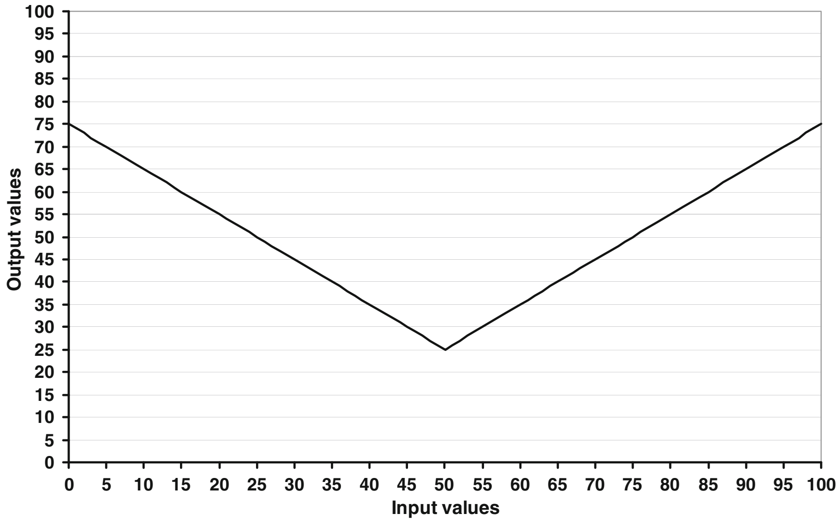


Fig. 1 The relationship between input and output values for the function-learning task

appropriate output value before they were shown the correct answer (*test/study*). The durations of training for both groups were equated, and motivation was enhanced with monetary bonuses. A criterial test administered after training assessed learning of the input–output function for the input values presented during training, and also generalization to input values outside the trained range.

Examining testing effects (retrieval practice) in this paradigm allows for a richer analysis of the benefits of testing on learning and retention than has been possible using the verbal materials mentioned at the outset of this article. We can examine the outcomes of testing on two levels of information: learning of the individual input–output training pairs and learning of more abstract information about the relations among the pairs (i.e., the function rule). With regard to learning of individual training pairs, based on the documented benefits of testing for learning associative information (e.g., Carpenter & DeLosh, 2005; Rohrer, Taylor, & Sholar, 2010), a clear prediction is that the test/study group should display better learning of each input–output pair presented during training than should the study-only group.

Less certain is the degree to which testing might promote learning of the relation (the function rule) among the input–output pairs. The question here is whether testing will stimulate abstraction of the underlying functional relation among the input–output pairs. It is important to note that acquisition of the trained input–output pairs does not necessarily require or stimulate acquisition of the abstract relation among the pairs. For instance, in function-learning paradigms using the test/study procedure used here (previous studies have not included a study-only group), an associative-based exemplar model has been proposed, with this model learning only the input–output pairs and no relational information among the pairs (DeLosh, Busemeyer, & McDaniel, 1997; McDaniel & Busemeyer, 2005). This associative-learning model is not without merit: It accounts for human learning in these paradigms as well as or better than a number of rule-learning models (see McDaniel & Busemeyer, 2005), nicely accommodates interpolation performance, and captures the extrapolation performance of some individuals (see DeLosh et al., 1997, Fig. 6).

For present purposes, these considerations lead to the strong prediction that testing will augment learning of the trained input–output pairs, and that this augmented learning will not be accompanied by acquisition of more abstract relational information. This prediction would be supported if the test/study group showed more accurate performance on trained points but not on transfer points relative to the study-only group, along with near-zero slopes in responses to the transfer points (consistent with an exemplar-learning model; Busemeyer, Byun, DeLosh, & McDaniel, 1997; DeLosh et al., 1997).

Alternatively, in many cases human learners do display some extrapolation (but not perfectly so) to tested transfer points (Carroll, 1963; DeLosh et al., 1997; Wagenaar & Sagaria, 1975), thereby suggesting that learners have acquired some notion of the relations among the trained input–output pairs. If this were the case in the present study, the observed slopes in the subjects' predictions on the transfer points should generally be nonzero. If testing does promote acquisition of the relations among the input–output points, then several possible outcomes would be expected. First, the test/study group should perform better on the transfer points than the study-only group in terms of (1) reducing mean error on their transfer responses and (2) producing a slope that is more extreme (i.e., closer to the given function slope) than that evidenced by the study-only group.

Method

Subjects and design

A total of 53 undergraduates from the University of California, San Diego, participated in return for course credit. The data from 3 subjects were discarded because the subjects did not complete the experiment. The analyses reported here are based on data from the remaining 50 subjects. The subjects were randomly assigned to one of two conditions, a *study-only* or a *test/study* condition (25 subjects in each condition). The final test was the same for all subjects, and included the within-subjects manipulation of whether the tested input value lay within the trained region or outside of it (*trained* vs. *transfer*, respectively).

Stimuli

The function to be learned was a V-shaped piecewise linear function $f(x)$, defined on the interval $0 < x \leq 100$ (see Fig. 1; used by McDaniel, Dimperio, Griego, & Busemeyer, 2009). The input and output values were integers. For values of $x \leq 50$, $f(x) = 75 - x$. For values of $x > 50$, $f(x) = x - 25$. In the training session, 51 specific input–output pairs (with the input lying between 25 and 75 inclusive) were selected individually for each subject; the subject was presented with 50 trials four times, for a total of 200 presentations during the study session.[1] In the testing session, subjects saw 100 input values; 51 of these were the same input values presented during training (all values from 25 to 75), 24 were the lower input values that had not been seen during training (all values from 1 to 24), and 25 were the upper input values that had not been seen during training (all values from 76 to 100). The order of the input values presented during training and test was randomized for each subject. Input values were always presented to subjects graphically, as bar heights (see Fig. 2). There were numeric codes at the bottom and top of the figure, but no numerical information was ever presented to, or elicited from, subjects. As seen in the figure, there were three bars. The leftmost bar was the input bar; its height represented the input value. The middle bar was the response bar. When present, its height was adjustable by the subject using the mouse. The rightmost bar was the output bar, and it was used to display the correct output value.

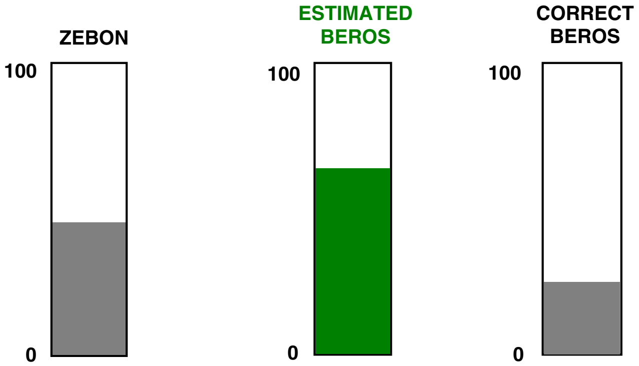


Fig. 2 A sample screen display during the training phase

Procedure

Subjects were tested individually in soundproof rooms. To ensure motivation in all conditions, subjects were told that they were going to play a game in which the top three scorers would each receive a cash bonus of $25. The cover story was that NASA scientists had discovered two new elements and an alien organism on Mars, and that the organism absorbs one of the elements (Zebon) while releasing the other (Beros). The goal for the subject was to figure out the relationship between the amount of Zebon absorbed and the amount of Beros released. In the training phase, subjects received four blocks of training on the same input–output pairs. The order of the pairs was randomized for each block. In the study-only condition, the input and output bars were displayed on the screen simultaneously for 6 s (the response bar was not shown), and the subject was told to observe the input–output pairings to learn the relationship between the two elements. In the test/study condition, the subject saw the input bar, along with the response bar (labeled "Estimated Beros"). Subjects responded by using the mouse to manipulate the height of the response bar and then releasing the mouse when they were satisfied with their answer. If the subject did not release the bar within 4 s from the onset of the trial, the computer displayed the message "too slow!" After subjects responded, the output bar was displayed for the remainder of the 6-s period, and subjects were told to pay attention to this feedback. After the training phase, subjects watched a 20-min video as a filler task, followed by a test of their function learning.

In the test phase, the 100 input values were presented in a random order for each subject. The display consisted of the input bar and the response bar. Subjects were told to estimate the amount of Beros (output) given a particular amount of Zebon (input) and to indicate their response by adjusting the height of the response bar using the mouse. There was no time deadline, and no feedback was provided (the output bar never appeared). After the subject made each response, the program proceeded to the next trial.

[1] A programming oversight led to 51 (instead of 50) input–output pairs being assigned to the trained range. For the 200 training trials, the computer randomly selected 50 out of the 51 pairs for presentation during each block of 50 training trials.

Results

Test performance was broken down into two nonoverlapping ranges of input values: The *trained* range refers to the input values that were presented during training (values between 25 and 75); the *transfer* range refers to input values below 25 and above 75. The α level for all analyses was set at .05.

As an overall measure of a subject's performance on each of these test ranges, the mean squared error (*MSE*) was computed. This is the average of the squared deviations between the true output values and the subject's estimated output values. Figure 3 shows the *MSE*s for subjects in the study-only and test/study conditions. A mixed ANOVA indicated that as expected, the *MSE* was smaller for the trained range (M = 129.25) than for the transfer range (M = 700.17), $F(1, 48) = 106.59$, $\eta_p^2 = .69$. In addition, test/study training yielded a smaller *MSE* (M = 307.70) than did study-only training (M = 521.72), $F(1, 48) = 11.52$, $\eta_p^2 = .19$. The interaction between the two factors was not statistically reliable.

Figure 4 shows the means and standard deviations of subjects' responses to each of the input values, for the study-only group (top panel) and the test/study group (bottom panel). The graphs reveal several things. First, the standard deviations are all larger for the study-only group, as expected from the *MSE* results described above. Second, it is apparent that generalization to the transfer items above the trained range was better than transfer to the items below the trained range. This observation was confirmed by a mixed ANOVA with Input Range as a two-level factor (lower and upper transfer), which revealed a main effect of input range, $F(1, 48) = 5.22$, $\eta_p^2 = .10$. This likely occurred because the upper segment, but not the lower segment, reflects a positive linear slope, and people have a positive linear bias in function learning (Busemeyer et al., 1997). Supporting this interpretation, better generalization along the lower segment is found when the piecewise linear function is concave (i.e., the lower segment has a positive linear slope) than when the piecewise linear function is convex (as used herein, where the lower segment has a negative linear slope; McDaniel et al., 2009, Exp. 1).

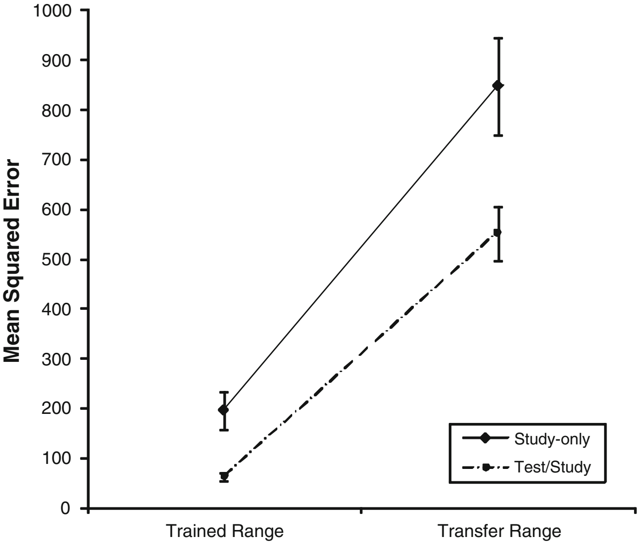


Fig. 3 Mean squared errors of the study-only and test/study conditions in our experiment. Error bars show the standard errors of the means

To gain insight into the degree to which subjects learned something about the topography of the overall function (i.e., the V shape), we next computed the slopes of the subjects' responses for the lower and upper trained ranges and for the lower and upper transfer ranges (see Table 1 for the means). Perfect learning of the function would be indicated by slopes of +1 and −1 for the upper and lower segments, respectively. We conducted a mixed ANOVA with two within-subjects factors (Range, trained vs. transfer; Segment, lower[2] vs. upper) and one between-subjects factor (Training Condition, study-only vs. test/study). Overall, the slopes for the test/study group were more extreme than those for the study-only group, $F(1, 48) = 12.27$, $\eta_p^2 = .20$, and more extreme for the trained than for the transfer range, $F(1, 48) = 14.89$, $\eta_p^2 = .24$, indicating better learning of the underlying function by the test/study group and for the trained range. In addition, a marginally significant interaction between range and segment, $F(1, 48) = 3.92$, $p = .053$, $\eta_p^2 = .08$, indicated that while learning of the function for the upper and lower segments was equivalent for the trained range, for the transfer range, learning was worse in the lower segment than in the upper segment.

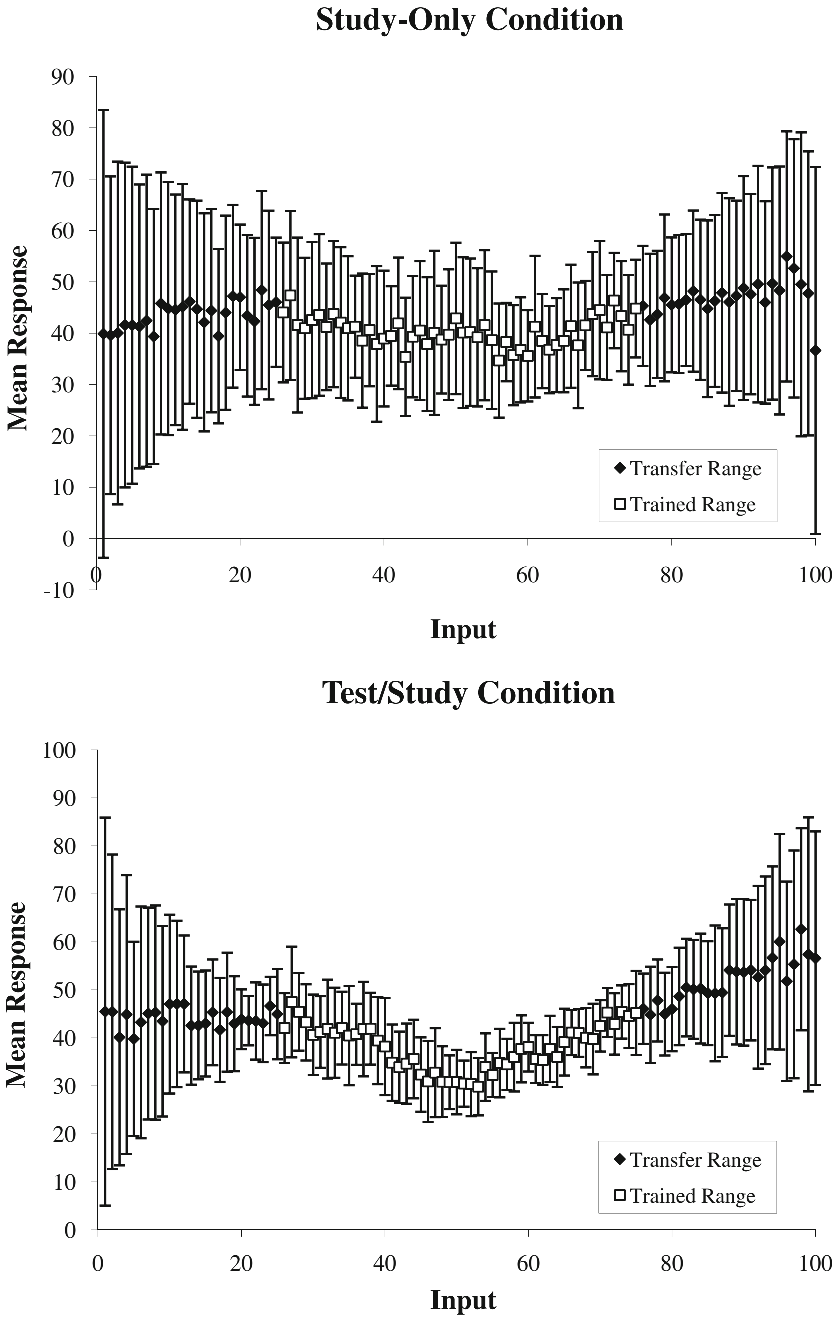
The slopes were also submitted to a one-sample *t*test to determine whether they were significantly different from zero. For the study-only group, the slopes across the range and segment conditions were not reliably different from zero, although the lower trained segment came close [$t(24) = 1.96$, $p = .062$]. For the test/study group, on the other hand, except for the lower transfer segment ($t < 1$), their slopes were reliably nonzero (*t*s > 3.32).

Our final analyses focused on the individuals in each group who showed nonzero slopes in their test performance (which we defined as slopes exceeding a value of 0.15 in the appropriate direction[3]). In the trained

[2] Response slopes for the lower segments (i.e., input values ≤ 50) were reversed in sign for this analysis.

[3] The cutoff was set at 0.15 because we wanted to be conservative and not count fairly flat nonzero slopes as being indicative of learning of the underlying function. Subjects with slopes in the inappropriate direction were counted as having zero slopes, because in both cases there was a failure to learn the relational information among the points.

Fig. 4 Means and standard deviations of the subjects' responses in our experiment



range, significantly more subjects with test/study training than with study-only training gave test responses that reflected nonzero slopes, for both the lower [$\chi^2(1, N = 50) = 5.09$] and upper [$\chi^2(1, N = 50) = 17.57$] segments of the trained range (see Table 2 for the distributions). A remaining question is, given individuals in either training group who showed some learning of the trained range that mirrored the topography of the function (i.e., negative

Table 1 Mean slopes of output responses as a function of training condition and input range

Training Condition	Trained Range		Transfer Range	
	Lower Segment ($25 \le x \le 50$)	Upper Segment ($50 < x \le 75$)	Lower Segment ($x < 25$)	Upper Segment ($x > 75$)
Study only	–0.20 (0.10)	0.16 (0.12)	0.24 (0.23)	0.15 (0.19)
Test/study	–0.63 (0.08)	0.62 (0.05)	0.01 (0.27)	0.56 (0.17)

Standard errors of the means in parentheses

Table 2 Distributions of subjects as a function of training condition and slope of test responses for the trained range

Trained Range	Slope	Training Condition	
		Study-Only	Test/Study
Lower	Zero	10	3
	Nonzero	15	22
Upper	Zero	13	0
	Nonzero	12	25

slopes for the lower segment and positive for the upper segment), did the likelihood that those individuals would extrapolate the topography to the transfer range vary across training conditions? Chi-square tests indicated that the proportions of learners who continued to display nonzero slopes from training to transfer points did not differ significantly across the training conditions (see Table 3 for the distributions), for both the lower [$\chi^2(1, N = 37) = .003$] and the upper segments [$\chi^2(1, N = 37) = 1.30$].

Discussion

The results clearly establish that testing enhances function learning. Consider first learning of the individual input–output points that were presented during study. Several indices of criterial test performance converged on the conclusion that training that incorporated testing (followed by study or feedback) promoted more accurate learning of training points than did a more standard training procedure in which learners just studied the input–output points. Learners in the test/study condition were able to produce output values in response to the given input values that better approximated the actual output values (i.e., with less error) than were learners in the study-only condition. This more accurate reproduction of the individual output values corresponded to the test/study group also displaying a more accurate reproduction of the overall shape of the functional relationship. As indicated by visual inspection of Figure 4 and by statistical comparisons of the slopes of learners' responses across the trained range, the topography of the responses (along the trained range) for the test/study condition (bottom panel) better reflected the V-shaped function than did the response topography for the study-only condition (top panel).

Table 3 Distributions of subjects who exhibited nonzero slopes for the trained range as a function of training condition and slope of test responses for the transfer range

Transfer Range	Slope	Training Condition	
		Study-Only	Test/Study
Lower	Zero	9	13
	Nonzero	6	9
Upper	Zero	1	6
	Nonzero	11	19

Only the subset of subjects who exhibited nonzero slopes for the trained range was included in this table

The slopes of the outputs for the study-only condition were not only more shallow (than for the test/study condition), these slopes were not significantly different from zero. Thus, as a group, study-only learners were not even able to gain an overall impression of the V-shaped pattern reflected by the function. Individual learner performance indicated, however, that the group analysis masked the fact that about half of the learners in the study-only condition did produce outputs for trained points that had nonzero slopes and reflected a V shape (negative for the lower training region and positive for the upper training region). Still, test/study training produced a much higher proportion of learners whose outputs reflected the V shape (nonzero slopes).

The results just discussed are not decisive, however, regarding the extent to which test/study training promoted better learning of more abstract relational information (the function rule) underlying the input–output training points. More accurate learning of the individual input–output points alone could of course produce a topography of responses that well approximates the V-shaped function. The performance for the test items outside the trained range (transfer) helps to inform us on this issue. Formal exemplar models indicate that if the subject learns only the input–output points during training and then applies only the individually learned exemplars to support responses on transfer points (outside the trained range), performance along the transfer region is fairly flat; that is, there is no extrapolation of the slope inscribed by the training points (see DeLosh et al., 1997, for modeling and examples of individuals displaying this pattern of extrapolation). Accordingly, if test/study training promotes only learning of individual training points and these points are used individually to generate transfer responses, then the advantage observed for that training procedure for trained points would be eliminated (or dramatically reduced) on the transfer points (outside the trained range). Furthermore, the slopes of the outputs in the transfer region would be flat (see DeLosh et al., 1997).

Generally, neither of the above possibilities materialized. Transfer responses better approximated the output values generated by the trained function after test/study than after study-only training. Especially telling are the results that slopes were significantly more steep after test/study training and that only test/study training produced transfer responses with slopes significantly greater than zero (in the upper transfer region only). These patterns suggest that test/study training fostered more learning of the abstract relational information—the function rule—than did study-only training. Still, the learning of such information was modest at best, even for the test/study training. At the group level, neither training procedure produced extrapolation in the lower training range that mirrored the negative slope of the function. Only for the upper extrapolation range, where the slope was positive, did training produce transfer that was qualitatively in line with the function. Both of these extrapolation patterns likely reflect in part learners' preexperimental positive linear bias (Busemeyer et al., 1997). This positive bias cannot completely explain the more accurate extrapolation supported by test/study training, because the study-only condition tended to display a more positive slope in the lower transfer region (where the function was negatively sloped) and a more shallow positive slope in the upper transfer region (where the function was positively sloped). Thus, it appears that information about the shape of the function acquired by the test/study learners was also influential in guiding transfer.

The above conclusion does not necessarily imply that test/study significantly altered the degree to which subjects who displayed some learning during training (i.e., showed nonzero slopes for the trained points) were likely to have gained an appreciation of the functional relations among the points. Indeed, the results suggest that though test/study increased the number of individuals who displayed enough learning of the training points to approximate the V-shaped topography of the function, the proportions of those learners (who had learned something about the training points) who also displayed nonzero slopes for the corresponding transfer region did not differ across the test/study and study-only conditions. This result is based on only one of several possible training regimens. Here, learners experienced 51 different training points repeated an average of four times each. It is uncertain whether test/study might foster more (or less) relational learning, were training to include fewer input–output points, but with more repetitions of each training point.

While our results demonstrate that function learning benefits from testing, it is worth noting that the causal mechanism underlying this testing effect may be qualitatively different from those described in the extant literature. Previous studies on the testing effect have typically examined how retrieving information from memory affects later retention of that information, and dominant theoretical explanations of the effect have focused on the strengthening or elaboration of the memory trace due to retrieval (e.g., Bjork, 1975; Carpenter, 2009; McDaniel & Masson, 1985). It is not obvious how such explanations would apply to the present situation of function learning, because subjects in the test/study condition might not necessarily have retrieved previous exemplars from memory during training (and certainly, early on in training there would be nothing to retrieve).

An alternative approach would be to conceptualize the present effects in terms of learning models, rather than in terms of memory/retrieval processes. For instance, from the perspective of contemporary neural-network models that learn by error correction, learning requires a comparison of the actual output to the desired output. When the target (or output) response is presented simultaneously with the cue (or input), as in the present study-only condition, learning is less efficient, because the presence of the target interferes with the estimation of error critical for learning (see Mozer, Howe, & Pashler, 2004, for an instantiation of such a model that accounts for the benefit of testing with feedback over studying only for paired-associate learning). In other words, forcing the learning system to make a prediction (via testing) and then receiving corrective feedback produces a more reliable error signal, which leads to better tuning of the system (i.e., learning). This account of the testing effect, we believe, provides a plausible explanation of our findings in the domain of function learning. Of course, confirming this account with formal model simulations is beyond the scope of this article, and we leave it up to future research.

Author Note This work was supported by the Institute of Education Sciences (U.S. Department of Education, Grants R305H020061 and R305H040108 to H.P.), the National Science Foundation (Grant BCS-0720375, H.P., principal investigator [PI], and Grant SBE-0542013, G. W. Cottrell, PI), and a collaborative activity award from the J. S. McDonnell Foundation. We acknowledge the contributions of the following individuals: David Cun programmed the experiments, and Steven Rocha and Harrison Wong assisted with the data analysis.

References

Abbott, E. E. (1909). On the analysis of the factors of recall in the learning process. *Psychological Monographs, 11,* 159–177.

Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior, 8,* 463–470.

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Busemeyer, J. R., Byun, E., DeLosh, E., & McDaniel, M. A. (1997). Learning functional relations based on experience with input–output pairs by humans and artificial neural networks. In K. Lamberts & D. R. Shanks (Eds.), *Knowledge, concepts, and categories* (pp. 405–435). Hove, U.K.: Psychology Press.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1563–1569.

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name-learning. *Applied Cognitive Psychology, 19,* 619–636.

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review, 14,* 474–478. doi:10.3758/BF03194092

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20,* 633–642.

Carroll, J. D. (1963). *Functional learning: The learning of continuous functional mappings relating stimulus and response continua*. Princeton, NJ: Educational Testing Service.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 968–986.

Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological Review, 62,* 255–262.

Hammond, K. R., & Stewart, T. R. (Eds.). (2001). *The essential Brunswik: Beginnings, explications, applications*. New York: Oxford University Press.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin, 57,* 116–131.

Kang, S. H. K. (2010). Enhancing visuospatial learning: The benefit of retrieval practice. *Memory & Cognition, 38,* 1009–1017.

Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus–response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 811–836.

Lachman, R., & Laughery, K. R. (1968). Is a test trial a training trial in free recall learning? *Journal of Experimental Psychology, 76,* 40–50.

McDaniel, M. A., & Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: Comparison of rule and associative based models. *Psychonomic Bulletin & Review, 12,* 24–42.

McDaniel, M. A., Dimperio, E., Griego, J. A., & Busemeyer, J. R. (2009). Predicting transfer performance: A comparison of competing function learning models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 173–195.

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology, 16,* 192–201.

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 370–384.

Mozer, M. C., Howe, M., & Pashler, H. (2004). Using testing to enhance learning: A comparison of two hypotheses. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the twenty-sixth annual conference of the cognitive science society* (pp. 975–980). Mahwah, NJ: Erlbaum.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316.

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181–210. doi:10.1111/j.1745-6916.2006.00012.x

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 233–239.

Wagenaar, W. A., & Sagaria, S. D. (1975). Misperception of exponential growth. *Perception & Psychophysics, 18,* 416–422.