Revisiting the "strong but wrong" effect

LANGDON E. LONGSTRETH

University of Southern California, Los Angeles, California 90007

In the 1930s, J. M. Stephens found that strong verbal associations called wrong are more likely to be changed than weak verbal associations called wrong, relative to items with no feedback. The present study produced a much larger "strong but wrong" effect, apparently as a result of defining "strong" in terms of meanings presumably stored in long-term memory. It was also found that such items are followed by high recall of feedback on the next trial, and that, indeed, when feedback recall is statistically controlled, the effect disappears. An explanation of why feedback is better recalled for "strong but wrong" items is offered, and several predictions are generated, some of which are tested (and confirmed) from the present data.

Presumably, if one is fairly "sure" of the correct response in a given situation, it might take some "convincing" to be shown otherwise. Yet there are data suggesting the opposite. In the 1930s, J. M. Stephens reported a series of studies concerned with "strong" verbal associations called "wrong" in the context of a verbal learning task (Stephens, 1934; Stephens & Baer, 1937, 1944; Stephens, Baer, & Stinnett, 1940).¹ He found that such "strong but wrong" responses are more likely to be abandoned than "weak but wrong" responses. To the best of the author's knowledge, his unexpected findings have not been followed up, nor are they even acknowledged in the psychological literature. The present study attempts to reproduce the phenomenon as well as to identify some of its necessary conditions.

In Stephens' original experiment, subjects were given three trials on a standard Spanish-to-English vocabulary test. Each item consisted of one Spanish word (stimulus) and five English words, one of which was the equivalent of the Spanish word. On the first trial, no feedback was given. On the second trial, feedback was given ("right," "wrong," or blank), and on the third trial, no feedback. If the choice on Trial 2 was the same as on Trial 1, that item was called a "strong" S-R pair. If the choice on the second trial was different, the item was called "weak." The dependent variable was repetition of Trial 2 responses on Trial 3 as a function of feedback and S-R strength. Percent repetition was as shown in Table 1.

Items with no feedback (blank) were used as a reference point for measuring the effects of right and wrong feedback. Thus, "wrong" feedback of strong responses resulted in 14% less repetition than no feedback (65-51), while "wrong" feedback of weak responses resulted in a 6% reduction (36-30). Thus, relative to no feedback, more strong responses than weak responses were abandoned following a single nonreinforcement. Subsequent experiments not only replicated this result, but produced it at a much larger magnitude (differences of around 30%).

Table 1

Feedback	S-R Strength	
	Strong	Weak
Right	86	59
Wrong	51	30
Blank	65	36

Stephens suspected that rehearsal was the crucial factor. He reasoned that the subject was "surprised" when a strong response was called wrong, and made a special effort not to be "surprised" again; e.g., the response and associated feedback were rehearsed more frequently or differently than other items. He found support for this interpretation in three subsequent studies. In one, a longer spontaneous delay was found after "strong but wrong" items than after other items, suggesting more time spent in rehearsal of these items (Stephens, Baer, & Stinnett, 1940). In another, rehearsal time was reduced to a minimum, and the effect disappeared (Stephens & Baer, 1937). Rehearsal time was not manipulated in this study, however, leaving open the possibility that the phenomenon was simply not replicated. In a third study, rehearsal time was manipulated, the result being that the phenomenon appeared at all intervals, including the minimum (Stephens & Baer, 1944).

There the matter seems to stand. It appears to be a reliable effect, theoretically interesting, yet not well mapped out in terms of the experimental conditions responsible for it. Following Stephens' lead that rehearsal might be involved, the present experiment explored the effects of two variables known to be related to rehearsal activities. One is whether the subject is warned or unwarned prior to the learning trial that he will be tested later. Previous research shows clearly that learning is better under warned conditions when there is time for rehearsal, but not when there is not (Longstreth, 1970, 1971; Nuttin & Greenwald, 1968). The second variable is recall of feedback: more rehearsal should lead to better recall

Table 2
Percentage Repeated Responses (RR) and
Correct Feedback Recall (FR)

Feed-back	Warned		Unwarned	
	Weak RR (FR)	Strong RR (FR)	Weak RR (FR)	Strong RR (FR)
Right	75 (58)	81 (72)	64 (44)	75 (69)
SD	31 (34)	35 (28)	30 (33)	23 (32)
Wrong	28 (47)	16 (86)	53 (47)	14 (86)
SD	28 (33)	46 (38)	34 (28)	22 (24)

of what it is that is rehearsed, which is assumed to include feedback in the case of "strong but wrong" items. This assumption was tested by requiring the subject to recall feedback from the previous trial as well as to choose the correct response.

METHOD

Subjects

Thirty-six college students were randomly divided into two equal groups, warned and unwarned. All said they had no knowledge of German or Latin, the two languages that were used to obtain response alternatives.

Materials and Procedure

The task was presented as one of translation to the subject. The goal was to choose one of two response-alternatives that was "semantically more similar" to the stimulus word. The stimulus word, in English, and two response alternatives, in German or Latin, were typed on 4 x 6 in. cards, which were turned over one at a time by the experimenter. Each card had a stimulus word on the left (e.g., ANNOY) and two response alternatives on the right, with the number 1 or the number 2 behind each word (e.g., KRANGE 1, LAUNE 2). The subject was instructed to name the number of the word he chose rather than the word itself in order to avoid pronunciation difficulties. Following his statement of "one" or "two," feedback was administered ("right" or "wrong") and the next card turned over about 4 sec later. These times were not precise, but were based on a natural rhythm that developed as a result of past experience using this particular procedure (e.g., Longstreth, 1970, 1971).

Three lists were administered, the first two constituting practice lists (A and B) that also served the purpose of illustrating to the subject that the lists were indeed presented just once. Subjects in the warned condition were then told that the next list (X) would be presented twice. The entire sequence for these subjects was thus A--B--test warning--X (learning)--X (test). The unwarned condition was the same except that the warning came *after* the first administration of List X: A--B--X (learning)--test warning--X (test). For both sequences, a 1-min delay intervened between the two presentations of List X.

As each item in List X was presented on the second trial, the subject was asked first to indicate if he had been right or wrong the first time or, if he could not remember, to so indicate. Then he was asked to choose the correct alternative. Thus, two responses were recorded for each item: recall of feedback (right, wrong, uncertain) and response alternative (1 or 2).

Test List (X)

The test list consisted of 12 items, which experience had shown was too long for perfect learning in one trial, but short enough to permit some learning. Half the items had response alternatives neither of which bore any semantic relationship to the stimulus word, the words having been selected from German and Latin dictionaries purely on the basis of length. It was assumed that

associative strength between the stimulus word and each response word was about the same for these items, an assumption that was tested on the learning trial. Two such items were: WEALTHY: PLATZ (1), KOHLE (2) and HAT: BRECHEN (1), AUFGABE (2). These are called the "weak" S-R items. The other six items had one response alternative that was related and one that was not. For example, DIE: WUN (1), MOR (2) and ANIMAL: CARNUM (1), PISCUS (2). These are called the "strong" S-R items, one response alternative presumably being stronger than the other.

The semantically related word also had a higher background frequency count, as determined by Kučera and Francis' (1967) one-million-plus corpus of printed words. The average frequency of the semantically related words (and morphologically similar words) was 18.5, as compared to 5.3 for the unrelated words. For weak items, the counts were very similar: 3.7 and 2.5.

For half the strong and weak items, "right" feedback followed the choice on the first trial. For remaining items, the feedback was "wrong." Thus there were four strength-feedback combinations of three items each: weak-right, weak-wrong, strong-right, and strong-wrong. For half the subjects in each of the warned-unwarned conditions, the feedback was reversed for each item, so that feedback was completely counterbalanced with respect to items. Order of presentation was randomized for each subject, with the restriction that no more than two strong or weak items could occur in succession.

RESULTS

Response distributions on the first trial were examined first in order to confirm the assumption that weak and strong items differed in frequency of selection of the two alternatives. For weak items, the distributions were .50, .53, .64, .75, .76, and .78. For strong items, the distributions were .78, .92, .97, .97, 1.00, and 1.00, with the semantically related alternative chosen more frequently in all cases. Thus, relative strengths of the two alternatives differed widely for weak and strong items. In order to maximize this difference, the two items in each set departing most from the desired distribution (.50 or 1.00) were not included in remaining analyses (inclusion of these items does not change the nature of the results in any essential way). The mean distribution proportions of the remaining four items in each set were .61 (weak items) and .98 (strong items), with no overlap between the two distributions.

The main data are presented in Table 2: percent repeated responses and percent correct feedback recall on Trial 2. The first thing to be observed is that the "strong but wrong" effect is in full evidence, appearing in both the warned and unwarned conditions: the responses to these items were changed more frequently (repeated less frequently) than were responses to "weak but wrong" items. In fact, the correct response to these items was even more frequent than the correct response to "strong and right" items, a result that surely must prove enigmatic to the reinforcement learning theorist. The mean difference between weak and strong "wrong" items is 25.5%, warned and unwarned combined, related $t(35) = 3.4, p < .01$.

Turning next to recall of feedback, Table 2 indicates that recall accuracy is best for these same

strong-wrong items, and, in particular, is *much* better than for weak-wrong items, related $t(35) = 4.2$, $p < .001$. In order to determine if the "strong but wrong" effect could occur in the absence of such recall differences, only items with correct recall were compared. The percentage of repeated responses for weak-wrong items is 8, as compared to zero for strong-wrong items. This 8% difference is significantly less than the 25.5% difference obtained when feedback recall is not controlled, related $t(28) = 2.1$, $p < .05$, and is not itself significant. This suggests that feedback recall may indeed be involved in the "strong but wrong" effect, although there is some possibility of a ceiling effect in the preceding internal analysis.

Table 2 also reveals the anticipated warned-unwarned effect: warned subjects repeated more "right" responses and changed more "wrong" responses, resulting in a greater percentage of correct responses, 78% to 68%, $F(1,34) = 4.1$, $p < .05$. The item effect is also significant: strong items were responded to correctly more often than weak items, with mean percentages of 81 and 65, respectively, $F(1,34) = 15.4$, $p < .001$. This difference is due mainly to the "strong but wrong" effect, although no interactions in Table 2 were significant.

When feedback recall was subjected to an omnibus analysis of variance, one significant effect emerged: the item effect, indicating better feedback recall of strong items (78%) than of weak items (49%), $F(1,34) = 45.9$, $p < .001$. While most of this effect was also due to the strong-wrong items, the interaction was not significant.

Of particular interest is the lack of a warned-unwarned effect on feedback recall, which did produce a significant difference in number of correct responses. How is it that warning subjects of an impending test resulted in better test performance but not in better recall of response outcome on the learning trial? Precisely the same results have appeared in three previous studies (Longstreth, 1971). We interpreted it then to mean that, in a two-choice situation, recall of both feedback and the correct response is redundant. Perhaps, then, warned subjects engage in more rehearsal of the *correct response*, but not the feedback.

There is an internal analysis that makes a fairly sharp test of this notion possible. That is, we would expect to find better performance by warned subjects even on those items for which the feedback could not be recalled, or was recalled incorrectly. Table 3 presents the relevant data. It confirms the prediction for feedback incorrectly recalled, but not for feedback not recalled. In the former case there is significantly better performance by warned subjects, but in the latter case there is (nonsignificantly) better performance by unwarned subjects. Thus our explanation is left hanging.

Table 3
Percentage Correct Responses as a Function of Feedback Recall

Feedback Recall	Warned	Unwarned	σ_d
Correct	97	94	45
Not recalled	53	65	28
Incorrect	29	2*	41
Total	78	68*	15

*Significant at $p < .05$.

DISCUSSION

The data are unequivocal so far as the "strong but wrong" phenomenon is concerned: when one of two choices is "obviously" the correct one, being informed that it is incorrect leads to a higher probability of choosing the alternative on the next trial than any other strength-feedback combination. These data actually constitute a stronger demonstration of the effect than reported by Stephens, who found it only in relation to items with no feedback. If it were to be defined as it was in the present study, in terms of absolute scores, Stephens did not demonstrate it at all.

One possible reason for its stronger appearance in the present study is that it involved stronger "strong" response alternatives. In Stephens' study, a strong response was defined simply in terms of a single repetition. Ignoring the fact that there was no attempt to relate it to the subject's store of knowledge—to his "world schema"—there was one chance in five that it occurred by chance. In the present case, the subject's presumed background knowledge was relied upon exclusively, and was double-checked against response distributions. The near-unanimous choice of the strong alternative for four of the six items confirms the logic of this approach.

Devising such items, however, is delicate. If the "strong" alternative is so embedded in the subject's cognitive structure that its denial would require a major revision of his store of organized knowledge, then negative feedback may simply impugn the experimenter. In the present case, the experimenter was prepared if the subject questioned his accuracy in, say, calling MOR the wrong response to DIE: he simply replied that although MOR was semantically related to DIE, so was WUN, only more so. Since the subject presumably knew no German or Latin in a formal sense, he was not in a position to question the experimenter's authority, and did not. Most subjects, however, did express some surprise when a "strong but wrong" event occurred.

There are definite indications that the phenomenon is at least partly determined by rehearsal activities: recall of feedback for "strong but wrong" items was far superior to that for any other type of item. Furthermore, when recall of feedback was statistically

controlled, the effect was reduced to a trivial and nonsignificant level. If one thus explains the effect in terms of rehearsal, a more basic question then arises, namely why does such differential rehearsal occur?

We suggest that the "strong" responses in the present case were stored in long-term memory; e.g., they were known prior to the first trial (as evidenced by response distributions). The "weak" responses, on the other hand, were unknown (also as evidenced by response distributions). If they were learned, it was *as a result* of the first trial, and, due to the short time allowed for rehearsal (about 4 sec/item), were probably maintained in short-term memory (perhaps some at the tail end, however, made some penetration into long-term structure during the 1-min interval between trials). On "strong" items, therefore, there was no point in rehearsing the correct response following a "right" outcome, since it was already known. On "wrong" outcomes, it was also not necessary to rehearse the correct response: the subject could simply rehearse the feedback (wrong) which, if recalled, would allow him to deduce the correct response, since the choice on Trial 1 was always the "strong" alternative (98% of the time). Thus the six "strong" items could be categorized into two groups, one tagged "right" and one tagged "wrong," and this would allow reconstruction of the correct response for all six items. The "weak" items could not be successfully categorized in this fashion, because feedback recall did not imply one response alternative any more than the other. Thus it was necessary to rehearse the correct response itself for these six items.

This argument, which explains why feedback was superior for the "strong but wrong" items, has three implications, two of which can be checked by internal analyses. First, it predicts that "strong and right" items should also elicit high feedback recall, since they too were collapsible into a single category on that basis. Table 2 confirms this prediction. Second, it predicts that weak items should be harder to learn than strong items across the board, independent of feedback. This prediction is also confirmed in

Table 2. Finally, it predicts that the correct response on strong-wrong items is not as subject to free recall as the correct response on weak-wrong items or on weak-right items, since presumably the correct response on these items was not rehearsed and was not stored in long-term memory. Unfortunately, free recall data were not foreseen to be of particular relevance and were not collected.

REFERENCES

- KUČERA, H., & FRANCIS, W. N. *Computational analysis of present-day American English*. Providence: Brown University Press, 1967.
- LONGSTRETH, L. E. Tests of the law of effect using open and closed tasks. *Journal of Experimental Psychology*, 1970, **84**, 53-57.
- LONGSTRETH, L. E. Relationship between response learning and recall of feedback in tests of the law of effect. *Journal of Experimental Psychology*, 1971, **90**, 149-151.
- NUTTIN, J., & GREENWALD, A. G. *Reward and punishment in human learning*. New York: Academic Press, 1968.
- STEPHENS, J. M. Further notes on punishment and reward. *Journal of Genetic Psychology*, 1934, **44**, 464-472.
- STEPHENS, J. M., & BAER, J. A. The influence of punishment on learning when the opportunity for inner repetition is reduced. *Journal of Genetic Psychology*, 1937, **51**, 209-217.
- STEPHENS, J. M., & BAER, J. A. Factors influencing the efficiency of punishment and reward: The opportunity for immediate review and special instructions regarding the expected role of punishment. *Journal of Genetic Psychology*, 1944, **65**, 53-66.
- STEPHENS, J. M., BAER, J. A., & STINNETT, B. J. Spontaneous delay following the application of punishment and reward to strong and weak connections. *Journal of Genetic Psychology*, 1940, **57**, 279-288.

NOTE

1. The author would like to thank Géry d'Ydewalle of the University of Leuven, Belgium, for introducing him to this literature, and for providing sabbatical facilities for the author while this study was analyzed and written.

(Received for publication January 6, 1975;
revision accepted September 8, 1975.)