Journal of Experimental Psychology:
Learning, Memory, and Cognition
1985, Vol. 11, No. 4, 812-816

0278-7393/85/$00.75

Observations

On Comparing Rates of Forgetting: Comment on Loftus (1985)

Norman J. Slamecka
University of Toronto, Toronto, Ontario Canada

Loftus' (1985) proposal for comparing relative forgetting rates by evaluating the horizontal interaction, rather than by way of the standard vertical interaction, is critically scrutinized. It is maintained that the proposed method entails an unavoidable confounding with the ages of the lists being compared, such that one list is measured at an earlier point in its forgetting curve, whereas the other is measured at a later point. This prevents an unequivocal assessment of the effect of the treatment variable of interest. The vertical interaction appears to be free of artifact. It is concluded that although the horizontal interaction has the advantage of circumventing the scaling problem, this is more than outweighed by the disadvantage of the list-age confound, leaving it less attractive, on balance, than the customary vertical interaction test.

This is a comment on the recent article in this journal by Loftus (1985) entitled, "Evaluating Forgetting Curves." In that article, he proposed an alternative method for comparing rates of forgetting on the part of groups that have been exposed to different levels of some independently manipulated variable, such as the number of original study trials. His article was in response to Slamecka and McElree's (1983) conclusion that variations in the degree of original learning did not affect the subsequent course of normal forgetting, but that they only influenced the intercept levels or heights of the respective retention curves. Loftus made use of this variable for illustrative convenience only, because his intent was to suggest a procedure applicable to the evaluation of relative forgetting rates in general, whatever the basis of difference in treatment between the groups. A brief recapitulation of the substance of his remarks on certain selected aspects of interest will be presented, together with my own considered reactions to the matter in each instance.

Vertical Parallelism

Slamecka and McElree used what can fairly be called the *customary* procedure for estimating whether groups differ in their rates of forgetting. By customary I mean only that it is the method that has been utilized in research on comparative forgetting for at least the past forty years. It is the familiar test for the significance of the interaction between the retention interval and the treatment variable. In other words, the test establishes whether the slopes of the respective retention functions, as measured between the same delay intervals, differ from one another. These slopes represent the loss rates associated with each level of the independent variable, and if they do not reliably differ, then the forgetting is said to be comparable. Loftus pointed out, correctly, that this approach is tantamount to a determination of whether the curves display vertical parallelism. That is, it asks whether the obtained retention level difference between the groups, as measured at the same point in time, maintains itself at a constant value across the various retention intervals. If that difference is the same at every duration of delay interval available, then there is no interaction and both rates of forgetting are functionally equivalent. In graphical form this situation is portrayed as the vertical distance between the groups, and is plotted in Figure 1 for a hypothetical but typical data pattern that exhibits vertical parallelism between high and low degree-of-learning groups, as indicated by the equal lengths of the two vertical dashed lines.

Forgetting rates can also be compared when there is only a single retention interval involved. In this case it is necessary to equate the groups on their terminal acquisition performance levels, or more precisely, on what those levels would be if a final acquisition test had been given. Then, either the respective delayed performances can be compared di-

Preparation of this paper was supported by Operating Grant A7663 from the Natural Sciences and Engineering Research Council of Canada to the author. The assistance of Lilly Katsaiti is gratefully acknowledged.

Requests for reprints should be sent to Norman J. Slamecka, Department of Psychology, University of Toronto, Ontario, Canada M5S 1A1.

rectly or, alternatively, the difference scores between original and delayed performances of each group can be so compared. In either case this still translates into a test of vertical parallelism in the sense that the vertical difference between groups is initially fixed at zero, so that any greater-than-zero difference later on is indicative of differential forgetting. The methodologies for effecting comparisons of forgetting rates from an assured common acquisition level have been worked out in detail by Underwood (1964), and are no doubt familiar to most who work in the field. In passing, it should be recognized that the variable of degree of original learning is unique in this regard, because by its very nature it must produce an initial intercept effect. A claim of different degrees of learning in the absence of a corroborating intercept effect would be hard to defend as an acceptable experimental finding. The Hellyer (1962) graph, in Loftus' Figure 4, shows no intercept difference at the shortest interval, and gives rise to a suspicion that the overall interaction was produced through a confounding with the measurement ceiling at that point.

Our utilization of the standard interaction test simply reflected the currently conventional practice when dealing with the question of comparative rates of forgetting, and its relentlessly consistent null outcomes for our data persuaded us toward the conclusion that forgetting is invariant with the degree of original learning. Loftus noted that, "This conclusion follows if 'forgetting' is operationally defined to be the slope of the forgetting function between any two delay intervals" (1985, p. 397). Indeed it does follow as stated, because we explicitly and deliberately identified the rate of forgetting with the slope of the empirical retention functon (see Slamecka & McElree, 1983, pp. 384 & 394).

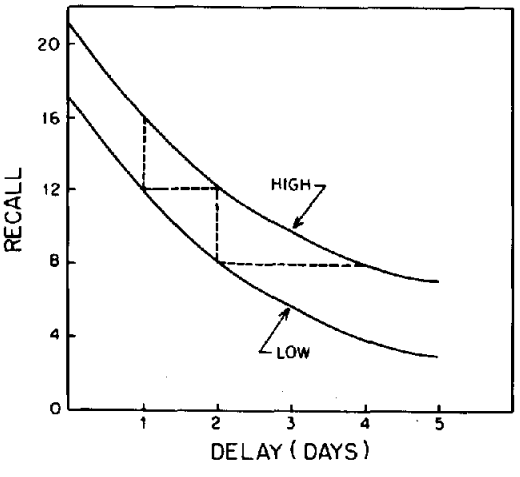

Figure 1. Hypothetical forgetting curves for groups having high and low degrees of original learning. (The vertical lines show vertical parallelism, and the horizontal lines show horizontal nonparallelism.)

It is appropriate to expand somewhat on my view of the forgetting phenomenon, in light of the following assertion made by Loftus: "A data evaluation method should be founded on some reasonably explicit model, or models, of the phenomena being evaluated" (1985, p. 403). If by this he means that one must already have a preexisting theory about something in order even to describe it, then I do not share that position. I do not share it because it seems to place an unnecessary limitation on one's freedom of experimental inquiry. We took forgetting to be literally an empirical occurrence, describable through conventional measurement of performance loss rates. It is not some theoretical entity or hypothesized process, but is a publicly observable manifestation of behavior subject to direct experimental analysis. Although the learning versus performance distinction is well established, with learning being a construct underlying performance gains, I see no need as yet for a corresponding "forgetting versus performance" distinction where forgetting is another construct underlying performance losses. By this view, no psychological theory or cognitive model is a precondition for investigating forgetting or for comparing the relative rates thereof, and therefore a search for the variables of which it is a function can proceed in the absence of a formal model of forgetting.

Theory enters the picture at the point where one postulates some real or metaphorical process that is responsible for forgetting. But, the assumed properties of this process exist at a different level from the objectively visible properties of forgetting itself. Research on the latter, in the form of systematic data collection, can go ahead without waiting for the former. It is likely that even the earliest facts gathered about a phenomenon will always remain useful, whereas early theoretical formulations seldom survive intact. In our own work we did not speculate about underlying processes. This restraint seemed advisable because the independent variables that control the rate of normal forgetting are not yet isolated, and theorizing would have too slim an empirical base from which to begin.

It should be acknowledged that these remarks stem from a position on a continuum of scientific style or preference. At one end are those who first construct a specific psychological model, and then do experiments only to test it. At the other end are those who first do experiments to reveal functional relations, and then perhaps offer a model. Loftus apparently speaks more from the theory-driven side, and I from the data-driven side. If there are any other yet unspoken assumptions on which we differ, I am unaware of them.

Further to our view of forgetting, Loftus said "It isn't clear, however, that this definition will ultimately prove to be the most useful in illuminating the processes that underlie forgetting" (1985, p. 397). But, the ultimate usefulness of a definition would not seem to be a matter of proof, but again a matter of taste, verging almost upon the arbitrary. It is clear that if a phenomenon is redescribed by changing its operational definition, then conclusions that are drawn about it will also be forced to change in at least some instances. Although true, that fact affords no evidential basis for preferring one operational definition of differential forgetting rates over the other. A logical positivist would maintain that a definition is simply an explication of how a given term is to be used, and the sole criterion it must meet is that of clarity. It bears no additional burden of illuminating any underlying processes, because only the findings of well-designed experiments can hope to accomplish that. According to this purist position, definitions are strictly neutral. However, perhaps there is some room for maneuver if one widens the view by also recognizing the demands of the context in which definitions are to be applied. When it comes to comparing forgetting rates, there must be appropriate consideration of what can be called the canons of measurement and the canons of experimental logic. Given that there is no disagreement about the latter, perhaps they will provide the means for saying more about the respective merits of applying these competing definitions. This will be taken up in the sections to follow.

Horizontal Parallelism

Rather than having us continue to compare up and down, Loftus suggests that we compare left and right. Instead of our using the old test of vertical parallelism, he advocates that we use the new test of horizontal parallelism. Rather than our evaluating the vertical interaction, we should evaluate the horizontal interaction. Instead of asking whether the performance difference that exists between groups at one retention interval remains the same at all retention intervals, we are to ask whether the retention interval difference at which performances are equal for both groups remains the same for all equal performance levels. If this latter quantity is indeed constant, there is horizontal parallelism. If it is not, there is an interaction, or horizontal nonparallelism.

The retention functions of Figure 1, which were used to illustrate vertical parallelism, are simultaneously illustrative of horizontal nonparallelism. That is, the two horizontal dashed lines are of unequal length, and that captures the essence of the idea. To put this interesting notion in yet another perspective, the test of horizontal parallelism amounts to a determination of whether it takes the two groups the same amount of time to drop from one common performance level to a lower common performance level. For example, from Figure 1 we can ask how long it takes for each group to drop from a retention level of, say 12 items to one of 8 items. As the figure shows, it takes Group High 2 days to do that, whereas Group Low does it in a single day. Because these elapsed times are different there is horizontal nonparallelism, and because Group High is the one that takes longer to fall from 12 to 8, it is inferred that Group High has a slower rate of forgetting than Group Low. This conclusion is in obvious contradiction to the one called for by the familiar vertical parallelism test, namely, that the groups have the same loss rate. In a sense, the horizontal test turns the vertical one upside down. A vertical analysis uses durations as an independent variable, and performances as the dependent variable, whereas a horizontal test uses performances as an independent variable, and durations as the dependent variable. The former approach keeps durations constant and measures performance differences, whereas the latter approach keeps performances constant and measures duration differences.

In general, given the reality of negatively accelerated forgetting functions, whenever the curves have different intercepts the geometry of the situation dictates that vertical parallelism entails horizontal nonparrallelism (as in Figure 1), and that horizontal parallelism entails vertical nonparralelism (as in Loftus, 1985, Figure 2a), thereby assuring the inevitability of conflicting statements from the two tests about the presence of differential forgetting. Even when there is no parallelism in either direction, there can still be instances of diametrically opposed conclusions, being drawn about which group has the greater, or lesser, loss rate (as in Loftus, 1985, Figure 2b). That is the heart of the issue and it constitutes an intellectually intriguing situation. In any such matter it is worthwhile to take a closer look at the details of each competing proposal, in order to uncover some point of difference that might give one view a significant advantage over the other. I shall discuss two such points. They happen to stand in a trade-off relation, such that the first favors the vertical test, whereas the second favors the horizontal test. One of them speaks to a principle of experimental observation and the other, also mentioned by Loftus (1985), to a principle of measurement. They both converge upon the problem of interpretability. The first is discussed later.

Any experiment that seeks to discover the influence of some treatment variable on the rate of subsequent forgetting must use two independent variables, namely, the treatment variable of interest as well as the delay interval. (If there is only a single retention interval, then terminal acquisition levels of the groups must be equated.) A main effect of

retention interval is a precondition which establishes the presence of forgetting. Then, the usual procedure is to determine the focal variable's influence by evaluating the (vertical) interaction term. This reveals whether the focal variable has affected the slopes of the loss functions, using comparisons where the levels of the retention intervals are held constant between the groups. Identical delay intervals are mandatory for this comparison in order to attain a clean estimate of the effect of the main variable. Nonetheless, one might still feel that a confounding exists with the variable of degree of learning. Group Low has learned the relatively easy items, by definition, whereas Group High has learned items of a greater average difficulty. If easy and hard items are forgotten at different rates, then the High–Low comparisons are illicit. This question was previously addressed by Slamecka and McElree (1983, pp. 388 & 390), so only a brief recapitulation will be given here. First, the literature shows that with degrees of original learning equated, deliberate manipulations of item difficulty leave forgetting rates essentially unaffected (for a review see Keppel, 1968). Second, the High and Low groups in two independent experiments by Slamecka and McElree were also scored on easy items only, these being defined as items recalled at the Immediate interval by Low subjects. The pattern of results was unchanged from that given by total scores, indicating no confounding with item selection.

In contrast, when it comes to the horizontal interaction there is a decided source of confounding, one so intrinsic to the very application of the test that it is difficult to see how it could be eliminated. The problem is that the *ages* of the lists are necessarily uncontrolled whenever the horizontal test is applied. In order to see this situation more clearly, consider Figure 1 again. The horizontal interaction test requires that both group's performances begin to be monitored from some arbitrary common level, say 12 items. That is where the confounding takes place. The groups already differ on the value of the treatment variable (degree of learning), which is appropriate, but now they are also made to differ on the age of the list at the start of measurement, which is inappropriate because it is one difference too many. In the example, Group Low's measurement starts at Day 1, whereas Group High's does not start until Day 2. It is no defense to assert, because the base measurement of both groups begins at a point showing equal absolute performance levels, that they are on the same footing in all important respects aside from the value of the treatment variable. This is simply not so. The key consideration is the fact of negative acceleration of forgetting curves. As a list gets older, its rate of item loss gets progressively slower. For present purposes this is the import of the first of Jost's (1897, cited in Woodworth, 1938, p. 58) two famous laws. In Woodworth's lucid version of it ". . . a young lesson momentarily at the same retention level as an old one is on a steeper part of the curve and doomed to decline more rapidly" (1938, p. 59). Age itself must be acknowledged as a factor in loss rates, and it should always be held constant in order to arrive at a legitimate assessment of the influence of the focal variable. Although it is not essential to the present argument, which is basically formal rather than theoretical, one can view an older list as having endured a more extended item-attrition experience than a younger list, with the consequence that its remaining items are hardier and more resistant to further loss than are those of the younger list. In Figure 1, the 12 items of Group High have all survived the rigors of a 2-day forgetting process, whereas the 12 items of Group Low are the result of a less protracted, and less severe, 1-day forgetting process. Measuring the subsequent fates of both lists from an arbitrary level of equal absolute performances automatically introduces this artifact.

In effect, the horizontal parallelism analysis involves the comparison of a younger segment from one curve with an older segment from another curve. The particular statistical outcome obtained will, of course, depend on the combined effects of the treatment variable and the ages of the list segments, but because it is based on an inherently confounded observation, it would seem to be irremediably uninterpretable. Therefore, by a criterion of freedom from observational confounding, I would put the score at Verticals 1, Horizontals 0.

The Scaling Question

Loftus has rightly reminded us that, "a nonordinal interaction can be made to appear or disappear at will by applying suitable nonlinear transformations to the dependent variable" (1985, p. 403; also see 1978). Slamecka and McElree's raw data consistently and repeatedly described vertically parallel forgetting functions. However, if the scores would have been subjected to any of a number of order-preserving transformations, such as square root or logarithmic, they would no longer have described parallel lines. Such transformations are perfectly legitimate if one makes the assumption that, at best, the measurements conserve only the ordinal relations between objective findings and some presumed underlying scale belonging to a hypothetical construct that could be labeled *retention.* There may be a monotonic transformation that relates the two, but its form is, of course, always unknown. A specific mapping relation can always be proposed for theoretical purposes, but it can never be actually known. Given the ordinality-only assumption, it follows that our obtained interaction terms were not interpretable at the level of an underlying scale. This was

previously acknowledged by Slamecka and McElree (1983, p. 394), who argued on behalf of the informational value of the raw data as such. Those data clearly show that, over time, the same absolute number of list items is lost from performance, regardless of the degree of original learning. What they do not necessarily also establish is that the same pattern holds true for the hypothetical construct of retention. This latter uncertainty stems from the experimental requirement that the groups always start from different intercept levels, properly reflecting the effect of different degrees of submastery learning. With most any other variable it would be possible to equate the starting points, but with this one it is not, and the limitation must be accepted.

On the other hand, the interpretation of a horizontal interaction can be said to apply to an underlying scale, if one puts all other considerations aside. The test neatly circumvents the realistic problem of different starting points by simply chopping off the tops of both curves so that they are now at the same height. However, this boldly Procrustean solution is imposed at a very steep price, namely, the introduction of the list-age artifact discussed in the preceding section. Of what net value is an interpretable interaction if it is based on confounded comparisons? Nonetheless, by a sole criterion of relevance to an underlying scale, I would put the score at Horizontals 1, Verticals 0. Combined with the first score, that makes the final outcome a tie.

But, rather than letting the matter subside into such an indecisive and depressingly unsatisfying balance of trade-offs, I feel an obligation to inject my personal tiebreaker vote. It is based on the conviction that the vast majority of potential independent variables to be examined for their effects on forgetting can readily have their retention levels begin at the same intercept point on the forgetting curve. For that vast majority of uses, therefore, the vertical interaction is completely free of scaling uncertainties, and is fully informative. In contrast, the horizontal interaction always bears the onus of introducing and of having to defend its inherent list-age confounding, regardless of starting points. I know of no resolution to that problem. Therefore, on those grounds, I am persuaded to come down in favor of the greater utility of the vertical test.

References

Hellyer, S. (1962). Frequency of stimulus presentation and short-term decrement in recall. *Journal of Experimental Psychology, 64,* 650.

Keppel, G. (1968). Retroactive and proactive inhibition. In T. R. Dixon & D. L. Horton (Eds.), *Verbal behavior and general behavior theory.* Englewood Cliffs, NJ: Prentice Hall.

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition, 6,* 312–319.

Loftus, G. R. (1985). Evaluating forgetting curves. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 397–406.

Slamecka, N. J., & McElree, B. (1983). Normal forgetting of verbal lists as a function of their degree of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 384–397.

Underwood, B. J. (1964). Degree of learning and the measurement of forgetting. *Journal of Verbal Learning and Verbal Behavior 3,* 112–129.

Woodworth, R. S. (1938). *Experimental Psychology.* New York: Holt.

Received November 12, 1984
Revision received February 13, 1985 ■