Why Immediate Repetition Is Good for Short-Term Learning Results but Bad For Long-Term Learning: Explanation Based on Decision Theory

Laxman Bokati[1], Julio Urenda[2,3],
Olga Kosheleva[4], and Vladik Kreinovich[1,3]
[1]Computational Science Program
[2]Department of Mathematical Sciences
[3]Department of Computer Science
[4] Department of Teacher Education
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
lbokati@miners.utep.edu, jcurenda@utep.edu,
olgak@utep.edu, vladik@utep.edu

Abstract

It is well known that repetition enhances learning; the question is: when is a good time for this repetition? Several experiments have shown that immediate repetition of the topic leads to better performance on the resulting test than a repetition after some time. Recent experiments showed, however, that while immediate repetition leads to better results on the test, it leads to much worse performance in the long term, i.e., several years after the material have been studied. In this paper, we use decision theory to provide a possible explanation for this unexpected phenomenon.

1 Formulation of the Problem: How to Explain Recent Observations Comparing Long-Term Results of Immediate and Delayed Repetition

Repetitions are important for learning. A natural idea to make students better understand and better learn the material is to repeat this material – the more times we repeat, the better the learning results.

This repetition can be explicit – e.g., when we go over the material once again before the test. This repetition can ne implicit – e.g., when we give the students a scheduled quiz on the topic, so that they repeat the material themselves when preparing for this quiz.

When should we repeat? The number of repetitions is limited by the available time. Once the number of repetitions is fixed, it is necessary to decide when should we have a repetition:

- shall we have it immediately after the students have studied the material, or
- shall we have it after some time after this studying, i.e., after we have studied something else.

What was the recommendation until recently. Experiments have shown that repeating the material almost immediately after the corresponding topic was first studied – e.g., by giving a quiz on this topic – enhances the knowledge of this topic that the students have after the class as a whole. This enhancement was much larger than when a similar quiz – reinforcing the students' knowledge – was given after a certain period of time after studying the topic.

New data seems to promote the opposite recommendation. This idea has been successfully used by many instructors. However, a recent series of experiments has made many researchers doubting this widely spread strategy. Specifically, these experiments show that (see, e.g., [1] and references therein):

- while immediate repetition indeed enhances the amount of short-term (e.g., semester-wide) learning more than a later repetition,
- from the viewpoint of long-term learning – what the student will be able to recall in a few years (when he or she will start using this knowledge to solve real-life problems) – the result is opposite: delayed repetitions lead to much better long-term learning than the currently-fashionable immediate ones.

Why? The above empirical result is somewhat unexpected, so how can we explain it? We have partially explained the advantages of *interleaving* – a time interval between the study and the repetition – from the general geometric approach; see, e.g., [3, 5]. However, this explanation does not cover the difference between short-term and long-term memories.

So how can we explain this observed phenomenon? We can simply follow the newer recommendations, kind of arguing that human psychology is difficult, has many weird features, so we should trust whatever the specialists tell us. This may sound reasonable at first glance, but the fact that we have followed this path in the past and came up with what seems now to be wrong recommendation – this fact encourages us to take a pause, and first try to understand the observed phenomenon, and only follow it if it makes sense.

This is exactly the purpose of this paper: to provide a reasonable explanation for the observed phenomenon.

2 Main Idea Behind Our Explanation: Using Decision Theory

Main idea: using decision theory. Our memory is limited in size. We cannot memorize everything that is happening to us. Thus, our brain needs to decide what to store in a short-term memory, what to store in a long-term memory, and what not to store at all.

How can we make this decision? There is a whole area of research called *decision theory* that describes how we make decisions – or, to be more precise, how a rational person should make decisions.

Usually, this theory is applied to conscientious decisions, i.e., decisions that we make after some deliberations. However, it is reasonable to apply it also to decisions that we make on subconscious level – e.g., to decisions on what to remember and what not to remember: indeed, these decisions should also be made rationally.

Decision theory: a brief reminder. To show how utility theory can be applied to our situation, let us briefly recall the main ideas and formulas behind decision theory; for details, see, e.g., [2, 4, 6, 7].

To make a reasonable decision, we need to know the person's preferences. To describe these preferences, decision theory uses the following notion of *utility*. Let us denote possible alternatives by $A_1, \ldots, A_n$. To describe our preference between alternatives in precise terms, let us select two extreme situations:

- a very good situation A_+ which is, according to the user, much better than any of the available alternatives A_i, and
- a very bad situation A_- which is, according to the user, much worse than any of the available alternatives A_i.

Then, for each real number p from the interval $[0, 1]$, we can form a lottery – that we will denote by $L(p)$ – in which:

- we get the very good situation A_+ with probability p and
- we get the very bad situation A_- with the remaining probability $1 - p$.

Clearly, the larger the probability p, the more chances that we will get the very good situation. So, if $p < p'$, then $L(p')$ is better than $L(p)$.

Let us first consider the extreme cases $p = 1$ and $p = 0$.

- When $p = 1$, the lottery $L(p) = L(1)$ coincides with the very good situation A_+ and is, thus, better than any of the alternatives A_i; we will denote this by $A_i < L(1)$.
- When $p = 0$, the lottery $L(p) = L(0)$ coincides with the very bad situation A_- and is, thus, worse than any of the alternatives A_i: $L(0) < A_i$.

For all other possible probability values $p \in (0,1)$, for each i, the selection between the alternative A_i and the lottery $L(p)$ is not pre-determined: the decision maker will have to select between A_i and $L(p)$. As a result of this selection, we have:

- either $A_i < L(p)$,
- or $L(p) < A_i$,
- or the case when to the decision maker, the alternatives A_i and $L(p)$ are equivalent; we will denote this by $A_i \sim L(p)$.

Here:

- If $A_i < L(p)$ and $p < p'$, then $A_i < L(p')$.
- Similarly, if $L(p) < A_i$ and $p' < p$, then $L(p') < A_i$.

Based on these two properties, one can prove that for the probability $u_i \stackrel{\text{def}}{=} \sup\{p : L(p) < A_i\}$:

- we have $L(p) < A_i$ for all $p < u_i$ and
- we have $A_i < L(p)$ for all $p > u_i$.

This "threshold" value u_i is called the *utility* of the alternative A_i.

For every $\varepsilon > 0$, no matter how small it is, we have $L(u_i - \varepsilon) < A_i < L(u_i + \varepsilon)$. In this sense, we can say that the alternative A_i is equivalent to the lottery $L(u_i)$. We will denote this new notion of equivalence by $\equiv$: $A_i \equiv L(u_i)$. Because of this equivalence, if $u_i < u_j$, this means that $A_i < A_j$. So, we should always select an alternative with the largest possible value of utility.

This works well if we know exactly what alternative we will get. In practice, when we perform an action, we may end up in different situations – i.e., with different alternatives. For example, we may have alternatives of being wet without an umbrella and being dry with an extra weight of an umbrella to carry, but when we decide whether to take the umbrella or not, we do not know for sure whether it will rain or not, so we cannot get the exact alternative. In such situations, instead of knowing the exact alternative A_i, we usually know the probability p_i of encountering each alternative A_i when the corresponding action if performed. If we know several actions like thus, which action should we select?

Each alternative A_i is equivalent to a lottery $L(u_i)$ in which we get the very good alternative A_+ with probability u_i and the very bad alternative A_- with the remaining probability $1 - u_i$. Thus, the analyzed action is equivalent to a two-stage lottery in which:

- first, we select one of the alternatives A_i with probability p_i, and then
- depending on which alternative A_i we selected on the first stage, we select A_+ or A_- with probabilities, correspondingly, u_i and $1 - u_i$.

As a result of this two-stage lottery, we end up either with A_+ or with A_-. The probability of getting A_+ can be computed by using the formula of full probability, as $u = \sum_i p_i \cdot u_i$. So, the analyzed action is equivalent to getting A_+ with probability u and A_- with the remaining probability $1 - u$. By definition of utility, this means that the utility of the action is equal to u.

The above formula for u is exactly the formula for the expected value of the utility. Thus, we conclude that the utility of an action is equal to the expected value of the utility corresponding to this action.

Let us apply this to learning. If we learn the material, we spend some resources on storing it in memory. If we do not learn the material, we may lose some utility next tome when this material will be needed. So, whether we store the material in memory depends on for which of the two possible actions – to learn or not to learn – utility is larger (or equivalently, losses are smaller). Let us describe this idea in detail.

3 So When Do We Learn: Analysis of the Problem and the Resulting Explanation

Notations. To formalize the above idea, let us introduce some notations.

- Let m denote the losses (= negative utility) needed to store a piece of material in the corresponding memory (short-term or long-term).
- Let L denote losses that occur when we need this material but do not have it in our memory.
- Finally, let p denote our estimate of the probability that this material will be needed in the corresponding time interval (short-term time interval for short-term memory or long-term time interval for long-term memory).

If we learn, we have loss m. If we do not learn, then the expected loss is equal to $p \cdot L$. We learn the material if the second loss of larger, i.e., if $p \cdot L > m$, i.e., equivalently, if $p > m/L$.

Comment. Sometimes, students underestimate the usefulness of the studied material, i.e., underestimate the value L. In this case, L is low, so the ratio m/L is high, and for most probability estimates p, learning does not make sense. This unfortunate situation can be easily repaired if we explain, to the students, how important this knowledge can be – and thus, make sure that they estimate the potential loss L correctly.

Discussion. For different pieces of the studied material, we have different ratios m/L. These ratios do not depend on the learning technique. As we will show later, the estimated probability p may differ for different learning techniques. So, if one technique consistently leads to higher values p, this means that, in general, that for more pieces of material we will have $p > m/L$ and thus, more pieces of material will be learned. So, to compare two different learning techniques, we need to compare the corresponding probability estimates p.

Let us formulate the problem of estimating the corresponding probability p in precise terms.

Towards a precise formulation of the probability estimation problem. In the absence of other information, to estimate the probability that this material will be needed in the future, the only information that our brain can use is that there were two moments of time at which we needed this material in the past:

- the moment t_1 when the material was first studied, and
- the moment t_2 when the material was repeated.

In the immediate repetition case, the moment t_2 was close to t_1, so the difference $t_2 - t_1$ was small. In the delayed repetition case, the difference $t_2 - t_1$ is larger.

Based on this information, the brain has to estimate the probability that there will be another moment of time during some future time interval. How can we do that?

Let us first consider a deterministic version of this problem. Before we start solving the actual probability-related problem, let us consider the following simplified deterministic version of this problem:

- we know the times $t_1 < t_2$ when the material was needed;
- we need to predict the next time t_3 when the material will be needed.

We can reformulate this problem in more general terms:

- we observed some event at moments t_1 and $t_2 > t_1$;
- based on this information, we want to predict the moment t_3 at which the same event will be observed again.

In other words, we need to have a function $t_3 = F(t_1, t_2) > t_2$ that produces the desired estimate.

What are the reasonable properties of this prediction function? The numerical value of the moment of time depends on what unit we use to measure time – e.g., hours, days, or months. It also depends on what starting point we choose for measuring time. We can measures it from Year 0 or – following Muslim or Buddhist calendars – from some other date.

If we replace the original measuring unit with the new one which is a times smaller, then all numerical values will multiply by a: $t \to t' = a \cdot t$. For example, if we replace seconds with milliseconds, all numerical values will multiply by 1000, so, e.g., 2 sec will become 2000 msec. Similarly, if we replace the original starting point with the new one which is b units earlier, then the value b will be added to all numerical values: $t \to t' = t + b$. It is reasonable to require that the resulting prediction t_3 not depend on the choice of the unit and on the choice of the starting point. Thus, we arrive at the following definitions.

Definition 1. *We say that a function* $F(t_1, t_2)$ *is* scale-invariant *if for every* t_1*,* t_2*,* t_3*, and* $a > 0$*, if* $t_3 = F(t_1, t_2)$*, then for* $t'_i = a \cdot t_i$*, we get* $t'_2 = F(t'_1, t'_2)$*.*

Definition 2. *We say that a function* $F(t_1, t_2)$ *is* shift-invariant *if for every* t_1*,* t_2*,* t_3*, and* b*, if* $t_3 = F(t_1, t_2)$*, then for* $t'_i = t_i + b$*, we get* $t'_2 = F(t'_1, t'_2)$*.*

Proposition 1. *A function* $F(t_1, t_2) > t_2$ *is scale- and shift-invariant if and only if it has the form* $F(t_1, t_2) = t_2 + \alpha \cdot (t_2 - t_1)$ *for some* $\alpha > 0$*.*

Proof. Let us denote $\alpha \stackrel{\text{def}}{=} F(-1, 0)$. Since $F(t_1, t_2) > t_2$, we have $\alpha > 0$. Let $t_1 < t_2$, then, due to scale-invariance with $a = t_2 - t_1 > 0$, the equality $F(-1, 0) = \alpha$ implies that $F(t_1 - t_2, 0) = \alpha \cdot (t_2 - t_1)$. Now, shift-invariance with $b = t_2$ implies that $F(t_1, t_2) = t_2 + \alpha \cdot (t_2 - t_1)$. The proposition is proven.

Discussion. Many physical processes are reversible: if we have a sequence of three events occurring at moments $t_1 < t_2 < t_3$, then we can also have a sequence of events at times $-t_3 < -t_2 < -t_1$. It is therefore reasonable to require that:

- if our prediction works for the first sequence, i.e., if, based on t_1 and t_2, we predict t_3,
- then our prediction should work for the second sequence as well, i.e. based on $-t_3$ and $-t_2$, we should predict the moment $-t_1$.

Let us describe this requirement in precise terms.

Definition 3. *We say that a function* $F(t_1, t_2)$ *is* reversible *if for every* t_1*,* t_2*. and* t_3*, the equality* $F(t_1, t_2) = t_3$ *implies that* $F(-t_3, -t_2) = -t_1$*.*

Proposition 2. *The only scale- and shift-invariant reversible function* $F(t_1, t_2)$ *is the function* $F(t_1, t_2) = t_2 + (t_2 - t_1)$*.*

Comment. In other words, if we encounter two events separated by the time interval $t_2 - t_1$, then the natural prediction is that the next such event will happen after exactly the same time interval.

Proof. In view of Proposition 1, all we need to do is to show that for a reversible function we have $\alpha = 1$. Indeed, for $t_1 = -1$ and $t_2 = 0$, we get $t_3 = \alpha$. Then, due Proposition 1, we have $F(-t_3, -t_2) = F(-\alpha, 0) = 0 + \alpha \cdot (0 - (-\alpha)) = \alpha^2$. The requirement that this value should be equal to $-t_1 = 1$ implies that $\alpha^2 = 1$, i.e., due to the fact that $\alpha > 0$, that $\alpha = 1$. The proposition is proven.

From simplified deterministic case to the desired probabilistic case. In practice, we cannot predict the actual time t_3 of the next occurrence, we can only predict the *probability* of different times t_3. Usually, the corresponding uncertainty is caused by a joint effect of many different independent factors. It is known that in such situations, the resulting probability distribution is close to Gaussian – this is the essence of the Central Limit Theorem which explains the ubiquity of Gaussian distributions; see, e.g., [8]. It is therefore reasonable to conclude that the distribution for t_3 is Gaussian, with some mean μ and standard deviation σ.

There is a minor problem with this conclusion; namely:

- Gaussian distribution has non-zero probability density for all possible real values, while
- we want to have only values $t_3 > t_2$.

This can be taken into account if we recall that in practice, values outside a certain $k\sigma$-interval $[\mu - k \cdot \sigma, \mu + k \cdot \sigma]$ have so little probability that they are considered to be impossible. Depending on how low we want this probability to be, we can take $k = 3$, or $k = 6$, or some other value k. So, it is reasonable to assume that the lower endpoint of this interval corresponds to t_2, i.e., that $\mu - k \cdot \sigma = t_2$. Hence, for given t_1 and t_2, once we know μ, we can determine σ. Thus, to find the corresponding distribution, it is sufficient to find the corresponding value μ.

As this mean value μ, it is reasonable to take the result of the deterministic prediction, i.e., $\mu = t_2 + (t_2 - t_1)$. In this case, from the above formula relating μ and σ, we conclude that $\sigma = (t_2 - t_1)/k$.

Finally, an explanation. Now we are ready to explain the observed phenomenon.

In the case of immediate repetition, when the difference $t_2 - t_1$ is small, most of the probability – close to 1 – is located is the small vicinity of t_1, namely in the $k\sigma$ interval which now takes the form $[t_2, t_2 + 2(t_2 - t_1)]$. Thus, in this case, we have:

- (almost highest possible) probability $p \approx 1$ that the next occurrence will have in the short-term time interval and
- close to 0 probability that it will happen in the long-term time interval.

Not surprisingly, in this case, we get:

- a better short-term learning than for other learning strategies, but
- we get much worse long-term learning.

In contrast, in the case of delayed repetition, when the difference $t_2 - t_1$ is large, the interval $[t_2, t + 2(t_2 - t_1)]$ of possible values t_3 spreads over long-term times as well. Thus, here:

- the probability p to be in the short-time interval is smaller than the value ≈ 1 corresponding to immediate repetition, but
- the probability to be in the long-term interval is larger that the value ≈ 0 corresponding to immediate repetition.

As a result, for this learning strategy:

- we get worse short-term learning but
- we get much better long-term learning,

exactly as empirically observed.

Acknowledgments

This work was supported in part by the US National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science) and HRD-1242122 (Cyber-ShARE Center of Excellence).

References

[1] D. Epstein, *Range: Why Generalists Triumph in the Specialized World*, Riverhead Books, New York, 2019.

[2] P. C. Fishburn, *Utility Theory for Decision Making*, John Wiley & Sons Inc., New York, 1969.

[3] O. Kosheleva and K. Villaverde, *How Interval and Fuzzy Techniques Can Improve Teaching*, Springer Verlag, Cham, Switzerland, 2018

[4] V. Kreinovich, "Decision making under interval uncertainty (and beyond)", In: P. Guo and W. Pedrycz (eds.), *Human-Centric Decision-Making Models for Social Sciences*, Springer Verlag, 2014, pp. 163–193.

[5] O. Lerma, O. Kosheleva, and V. Kreinovich, "Interleaving Enhances Learning: A Possible Geometric Explanation", *Geombinatorics*, 2015, Vol. 24, No. 3, pp. 135–139.

[6] R. D. Luce and R. Raiffa, *Games and Decisions: Introduction and Critical Survey*, Dover, New York, 1989.

[7] H. Raiffa, *Decision Analysis*, Addison-Wesley, Reading, Massachusetts, 1970.

[8] D. J. Sheskin, *Handbook of Parametric and Non-Parametric Statistical Procedures*, Chapman & Hall/CRC, London, UK, 2011.