USING INVARIANCE TO MODEL PRACTICE, FORGETTING, AND SPACING EFFECTS

Martin Riopel[1], Pierre Chastenay[1], Gabrielle Fortin-Clément[1], Patrice Potvin[1], Steve Masson[1], and Patrick Charland[1]

[1]Université du Québec à Montréal (CANADA)

Abstract

Practice plays an important role in e-learning, yet it is still difficult to find a consensual and usable mathematical model to describe the shapes of learning, forgetting, and spacing curves [1]. Studies on distributed practice date back to Ebbinghaus (1885), and hundreds more have been conducted on the related effects [2]. One of these effects is the law of practice [3], which links higher success rates and shorter response times to the number of practice trials. Another is the forgetting curve [4], which links a lower observed probability of retention to a longer period of time before retrieval. A third and more abstract phenomenon is the spacing effect, which suggests that breaking up practice over time produces more learning than if the sessions are grouped together.

This study uses three scale-invariant hypotheses to determine the mathematical properties of the empirical curves. The first hypothesis, composition invariance, states that, for each subject, simultaneous success at two equivalent and valid tasks within a certain context will always produce a new valid, though possibly rescaled, task within the same context. The second hypothesis, temporal invariance, states that slowing a sequence of tasks does not change the mathematical properties of the empirical curve for correctness, apart from a single scaling parameter. Similarly, the third hypothesis, ordinal invariance, states that when counting successive occurrences of identical tasks, one is free to choose what constitutes a unit (a single task or group of tasks) without changing the mathematical properties of the empirical curve for correctness.

An online PHP/HTML application was developed for this study. Members of professional and amateur astronomy associations were invited to participate via social media. Subjects were shown an image of one of the 88 constellations and asked to choose the correct name from three options. They were then given three seconds to study the correct answer. The questions were repeated several times in a within-session randomized block design with different scales. Correctness of answers and response times were recorded. A total of 325 participants completed the sequence; this consisted of three blocks of 136 items, with each block taking approximately 15 minutes to complete. The three scale-invariance hypotheses were tested in a two-step procedure. A scale parameter was first chosen to make each pair of related curves as similar as possible by minimizing the chi-squared statistic. The same statistic was then used to test whether the first curve could serve as a model for the second. Preliminary results show that the appropriate scale parameter can be used to make all related curves statistically equivalent. Lastly, the hypotheses were used to deduce certain properties of the curves observed within this context.

Keywords: Astronomy, Practice, Forgetting, Spacing, Learning Theory.

1 INTRODUCTION

When students prepare for a test, they have several study sessions. They usually ask themselves: for how long, how often, and exactly when should they study to get the best results? It is difficult to know the answers because of the number of variables involved. But what if students were to use an online application that had access to all pertinent variables? Could the application propose the optimal sequence of study sessions? The answer is still no, because research on distributed practice has yet to reach definitive conclusions. More precisely, even if practice plays an important role in e-learning, it is difficult to find a consensual and usable mathematical model to describe the shapes of the learning, forgetting, and spacing curves relating to distributed practice. As pointed out by [1], “existing theories cannot convincingly account for the different empirical findings, and predictions about the optimal point in time to review material are hard to derive” (p. 78); furthermore, “future research on the distributed practice effect should focus on validating existing theories as well as on developing new ones to put this strong effect on strong theoretical grounds” (p. 79).

Studies on distributed practice date back to Ebbinghaus (1885), and hundreds more have been conducted on the related effects [2]. One of these effects is the law of practice [3], which links higher success rates and shorter response times to the number of practice trials. Another is the forgetting curve [4], which links a lower observed probability of retention to a longer period of time before retrieval. A third and more abstract phenomenon is the spacing effect, which suggests that breaking up practice over time produces more learning than if the sessions are grouped together. According to [5], the spacing effect is directly related to the lag effect (longer spacing is better) and the testing effect (testing is better than studying alone), and it can be confounded with at least five spacing-like phenomena—namely, the recency effect, rehearsal effect, zero-sum effect, deficient-processing effect, and list-strength effect. It is understandably difficult to produce a single model that encompasses all of these effects.

One way to construct such models is to base them on known cognitive processes, such as contextual variability [6, 7, 8], memory accessibility [9, 10], and the study-phase retrieval hypothesis [11, 12, 13]. This first family of general models tends to be conceptually strong, but unable to specify the mathematical shapes of the observed curves. Another approach is to use the observed empirical curves as a starting point, as with the adaptive control of thought-rational (ACT-R) model [14] and the multiscale context model (MCM) [15]. These models are usually consistent with real data, perhaps because they are based on empirical results; however, they do not provide strong theoretical arguments for the precise shape of the curves. In this paper, we will attempt a third approach, the search for invariance, to construct a more general model. If validated, the general invariance properties could be used to deduce the precise mathematical form of the curves.

For the purposes of this study, three hypotheses were chosen to determine the mathematical properties of the empirical curves. The first hypothesis, composition invariance, states that, for each subject, simultaneous success at two equivalent and valid tasks within a certain context always produces a new valid, though possibly rescaled, task within the same context. The second hypothesis, temporal invariance, states that slowing a sequence of tasks does not change the mathematical properties of the empirical curve for correctness, apart from a single scaling parameter. Similarly, the third hypothesis, ordinal invariance, states that when counting successive occurrences of identical tasks, one is free to choose what constitutes a unit (a single task or group of tasks) without changing the mathematical properties of the empirical curve for correctness. This study sets out to validate these three hypotheses.

2 METHODOLOGY

An online PHP/HTML application was developed for the study and is available at http://riopel.uqam.ca/astro. Members of professional and amateur astronomy associations were invited to participate via social media. Subjects were shown an image of one of the 88 constellations and asked to choose the correct name from three options. They were then given three seconds to study the correct answer (displayed in green on the same screen). The questions were repeated several times in a within-session block design with different scales. A sample question is shown in Fig. 1.

Constellations were randomized for each participant. Correctness of answers and response times were recorded. Before beginning, participants were asked to indicate their gender, age, language (English or French), and self-declared level of prior knowledge about constellations (beginner, intermediate, advanced, or expert). The complete sequence consisted of three blocks of 136 items, each block taking approximately 15 minutes to complete.

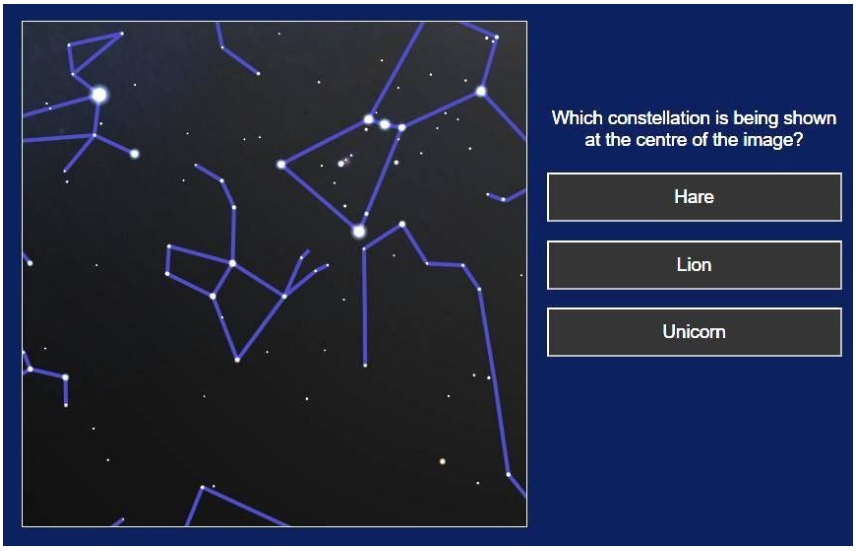

Figure 1. Sample constellation identification question

The three scale-invariance hypotheses were tested in a two-step procedure. A scale parameter was first chosen to make each pair of related curves as similar as possible by minimizing the chi-squared statistic. The same statistic was then used to test whether the first curve could serve as a model for the second. This paper presents new results for this ongoing experiment (for which some results were introduced in [16]).

3 RESULTS

3.1 Participants

A total of 325 subjects completed the 408 items. Table 1 summarizes their descriptive statistics.

Table 1. Descriptive statistics for the participants

Gender	%	Language	%	Age	%	Level	%
Male	58	English	42	5 to 24	9	Beginner	48
Female	40	French	58	25 to 44	24	Intermediate	40
Other	2			45 to 64	48	Advanced	10
				65 to 84	19	Expert	2
TOTAL	325	Subjects					

3.2 Testing invariance hypotheses with correctness

Fig. 2 presents the data used to test the composition invariance and temporal invariance hypotheses. Solid lines represent the reference sequence for each effect (practice, forgetting, and spacing). Dotted lines are scaled versions of the solid line, minimizing the chi-squared statistic for the three other cases. Slower curves (square and rhombus points) test the temporal scale invariance for the faster curves (round and triangle points). Lower curves (triangle and rhombus points) test the composition scale invariance (combined success for equivalent tasks) for the corresponding higher curves (round and square points). It can be seen that the scaled versions are almost always consistent with the data within the 2σ error bars. Each subject contributed at least four times to each point for a minimum of 1,300 observations per point for the higher curves and 650 for the lower curves.

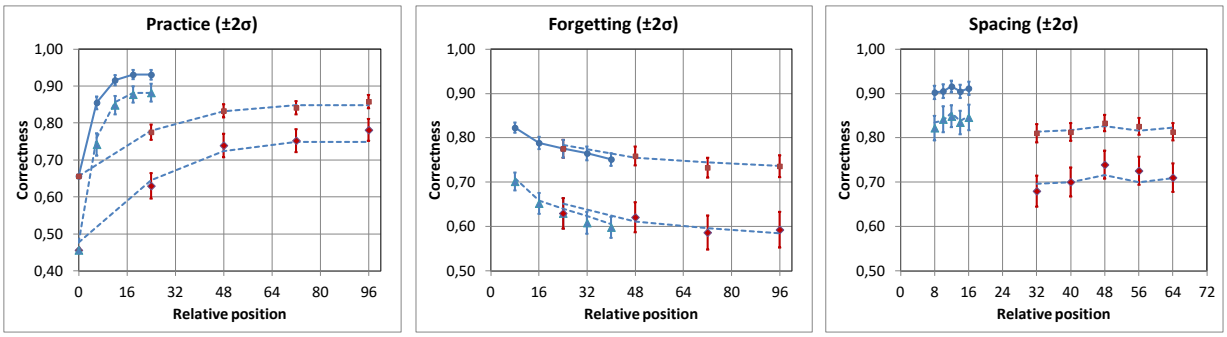

Figure 2. Probability of success as a function of position relating to the first appearance in sequence for different effects and different scales

Subsequent chi-squared tests related to Fig. 2 for the three effects and two scales (composition and temporal) all showed that the null hypotheses cannot be rejected for any of the nine possible scaled associations, where χ^2 (3 to 4) = 0.55 to 6.06 and p = .20 to .97. This outcome confirms their statistical equivalence with respect to this experiment—a conclusion that is reinforced when the test is done globally for Fig. 2: χ^2 (44) = 28.97, p > .99.

Fig. 3 uses the same convention as Fig. 2, but with a sequence of two consecutive questions for the same constellation as the unit task. This is required in order to test the ordinal invariance. The same reference sequences as Fig. 2 are used, and, once again, the scaled versions almost always fit the data within the 2σ error bars.

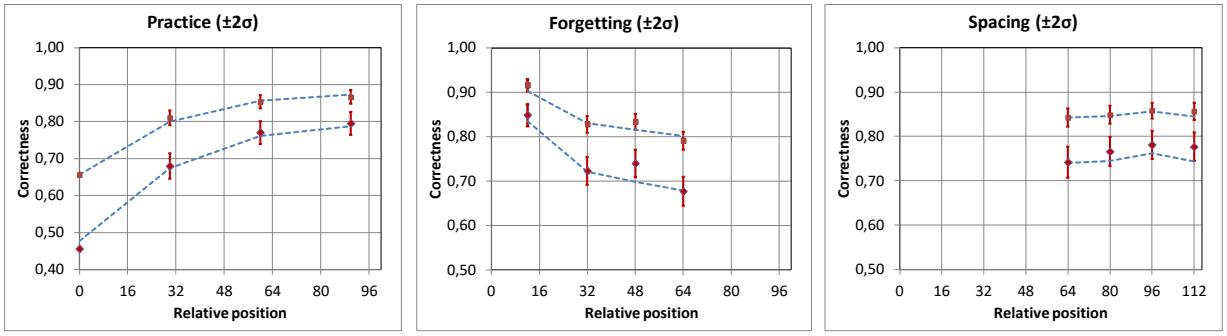

Figure 3. Probability of success as a function of relative position when two consecutive questions are counted as one

Subsequent chi-squared tests for the three effects and the ordinal scale all showed that the null hypotheses cannot be rejected for any of the six possible scaled associations, where χ^2 (4) = 1.59 to 6.95 and p = .08 to .67. This outcome confirms their statistical equivalence with respect to this experiment—a conclusion that is reinforced when the test is done globally for Fig. 3: χ^2 (23) = 21.23, p > .92.

These results show that, at least within the context of this experiment, the invariance hypotheses can be used to link many curves to a given effect before knowing the exact shape for that effect. It can be seen, for example, that invariance is appropriate for correctness as an increasing function for practice, a decreasing function for forgetting, and a convex function for spacing. There is no indication that this general procedure could not be used for any other shape.

3.3 Testing consequences of composition invariance

A first consequence of composition invariance is the uniquely defined shape of the probability function for succeeding at a task. This can be understood with the formalism of the item response theory. According to [17], latent-trait models are usually used with tests in which performance (P) for an item (i) is a function of an ability parameter (θ) that ranges from − ∞ to + ∞. In these models, the cumulative probability distribution P_i (θ) is called the item response function. A popular item response function is the one-parameter logistic (or Rasch) model,

$$P_i = 1/\{1+\exp[-a\,(\theta - b_i)]\}\ ,$$

where a is a constant discrimination parameter and b_i is the item difficulty. If two such items are independent, then the probability of succeeding at both is the product of their response functions ($P_1 \cdot P_2$). The composition invariance within this context is verified if this product can also be described by the same model (P_3). It can be shown that this is not the case for either the logistic model or the normal ogive model, another popular item response function.

Finding a cumulative distribution function for which it is possible to obtain $P_1 \cdot P_2 = P_3$ requires applying the statistical extreme value theorem. Similar to the central limit theorem for normal distribution, it states that the only possible stable classes for the product of identical cumulative distributions (or maxima) are the Gumbel, Fréchet, and Weibull functions [18]. Of these three, only one has a range of $-\infty$ to $+\infty$. Therefore, the only cumulative distribution that shows the composition invariance property with an appropriate range is the Gumbel distribution,

$$P_i = \exp\{-\exp[-a\,(\theta - b_i)]\}\ ,$$

where a and b_i have the same values as defined in the one-parameter logistic model. This skewed response function is not common, but it has been used with some success in the past [19]. It shows that ability is unchanged when two or more equivalent tasks are combined to form a valid but more difficult third task. The discrimination parameter is also unaffected by this transformation.

In the case of empirical data where θ varies, it can be more convenient to set the following values, without loss of generality:

- unit discrimination parameter: $a = 1$
- initial correctness: $P_0 = \exp[-\exp(b_i - \theta_0)]$
- learning: $\Delta\theta = \theta - \theta_0 = -\log\,[\log(P) / \log(P_0)]$

Correctness and learning are thus related by $P = P_0^{\exp(-\Delta\theta)}$, and only the scaling parameter P_0 changes to P_0^N when independent items are combined. This consequence was validated for all nine relevant pairs of empirical curves in Fig. 2 and 3: χ^2 (3 to 4) = 0.55 to 6.06, p = .18 to .97. This key relationship between correctness and learning is scale-invariant and can be used to uniquely define a unit of measure for learning.

A second consequence of composition invariance is the explicit relationship between correctness and response time, which can be deduced through the simple observation that, for a given subject, the total response time (T_3) for a combination of successful items is the sum of individual response times ($T_1 + T_2$). Since the corresponding combined correctness (P_3) for independent items is the product of the individual correctness values ($P_1 \cdot P_2$), the general relationship for response time may be expressed as

$$T_i = T_{min} - S \cdot \log\,(P_i)\ ,$$

where S is a constant parameter and T_{min} is the shortest possible response time corresponding to the highest possible correctness. Fig. 4 validates this relationship using empirical data from the same experiment. The lower line represents individual response times for each sequence, while the higher line corresponds to the total response times for the combined success of two equivalent sequences. The extreme left point of each line corresponds to the first time the constellations are presented. The other points correspond to each subsequent time that the same constellation was presented within a particular sequence. One can see that the observed relations are parallel and strongly linear (R^2 = .94 and .95), consistent with composition invariance.

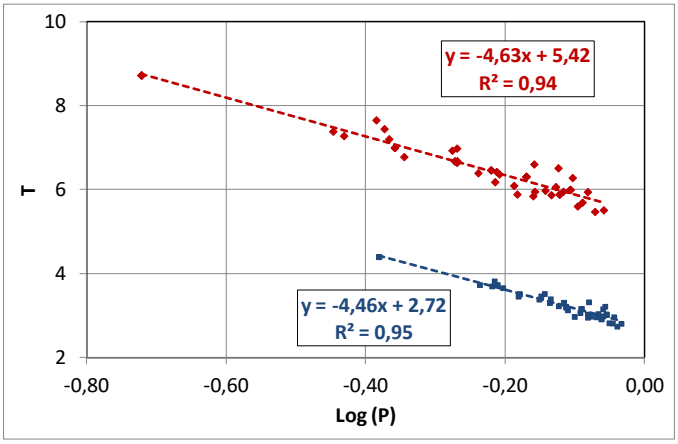

Figure 4. Response time as a function of logarithm of correctness

The equations above can be used to express response time as a general function of ability:

$$T_i = T_{min} + S \cdot \exp[-a\,(\theta - b_i)] = T_{min} - S \cdot \exp(-\Delta\theta) \cdot \log P_0.$$

If N equiprobable choices are proposed, $P_0 = 1/N$ and the new equation (proposed by Riopel) is an extension of the Hick-Hyman law beyond novel performance ($\Delta\theta = 0$ or upper-left point), answering a key challenge raised by [20]. We believe that this is an important result for the field of learning sciences, as it gives a more general scale-invariant explanation for a robust and demonstrable law that is usually deduced from information theory applied to the context of the human brain.

It is also interesting to note that if error on ability (θ) is distributed normally, as might be expected for a single subject in a perfect experiment, the distribution of response time follows the lognormal model of [21].

3.4 Testing invariance hypotheses with response time

The relationship between response time and correctness opens the way for a second validation of the three scale-invariant hypotheses that were tested in Fig. 2, this time using response times (when successful) instead of correctness. This data is presented in Fig. 5. As before, solid lines represent the reference sequence for each effect (practice, forgetting, and spacing), and dotted lines are scaled versions of the solid line, minimizing the chi-squared statistic for the three other cases. Slower curves (square and rhombus points) test the temporal scale invariance for the faster curves (round and triangle points). Lower curves (triangle and rhombus points) test the composition scale invariance (combined success for equivalent tasks) for the corresponding higher curves (round and square points). It can be seen that, even with smaller error bars, the scaled versions are almost always consistent with the data within the 2σ error bars.

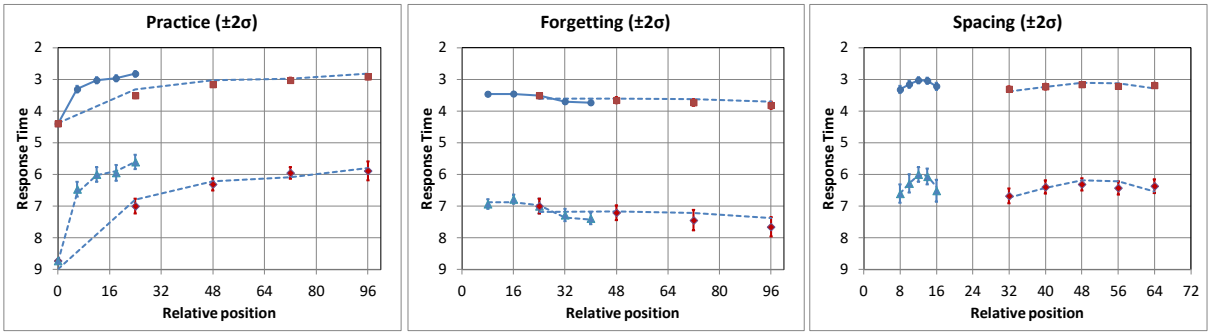

Figure 5. Response time for successful identification as a function of position relative to first appearance in sequence for different effects and different scales

Once again, these results show that, at least within the context of this experiment, invariance hypotheses can be used to connect various curves for a given effect (practice, forgetting, or spacing) before making an assumption about the exact shape for that effect or the data collected (correctness

or response time). To date, there is no indication that this general procedure could not be used for other shapes or pertinent data.

4 CONCLUSIONS

The results for this ongoing experiment largely confirm that the general scale-invariance hypotheses can be applied to the spacing, forgetting, and distributed practice effects. They also propose a scale-invariant unit of measure for learning and a demonstrable extension of the robust Hick-Hyman law. We believe that this unifying proposition is a significant achievement for the field of learning sciences, as the shapes of the curves describing these effects can be varied and unrelated. However, future work is still needed to test the applicability of these results with greater precision and within broader contexts. Using the general hypotheses of this experiment to deduce the precise mathematical form of the curves is another interesting avenue that is explored in [22].

ACKNOWLEDGEMENTS

The authors wish to thank the board of directors of the Canadian Astronomical Society (CASCA), the Royal Astronomical Society of Canada (RASC), the Fédération des astronomes amateurs du Québec (FAAQ), and the Astronomical League for inviting their members to participate in this research project. The authors also wish to thank Salim Sader and Julien Mercier for discussions about the ideas presented in this paper and Reza Shams Latifi for his help with the English version of the online application.

REFERENCES

[1] C. E. Küpper-Tetzel, “Understanding the Distributed Practice Effect : Strong Effects on Weak Theoretical Grounds”, *Zeitschrift fr Psychologie*, vol. 222, no. 2, pp. 71–81, 2014.

[2] N. J. Cepeda, H. Pashler, E. Vul, J.T. Wixted and D. Rohrer, “Distributed practice in verbal recall tasks: A review and quantitative synthesis”, *Psychological Bulletin*, vol. 132, pp. 354–380, 2006.

[3] A. Heathcote, S. Brown and D.J.K. Mewhort, “The power law repealed: The case for an exponential law of practice”, *Psychonomic Bulletin & Review*, vol. 7 no. 2, pp. 185-207, 2000.

[4] L. Averell and A. Heathcote, “The form of forgetting curve and the fate of memories”. *Journal of Mathematical Psychology*, vol. 55, no. 1, pp. 25–35, 2011.

[5] P. F. Delaney, P. P. Verkoeijen and A. Spirgel, “Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature”, In *The psychology of learning and motivation: Advances in research and theory* (B. H. Ross Ed.), pp. 63–147, London, UK: Academic Press, 2010.

[6] W. K. Estes, “Statistical theory of distributional phenomena in learning”, *Psychological Review*, vol. 62, pp. 369–377, 1955

[7] A. M. Glenberg, “Component-levels theory of the effects of spacing of repetitions on recall and recognition”, *Memory & Cognition*, vol. 7, pp. 95–112, 1979.

[8] J. G. Raaijmakers, “Spacing and repetition effects in human memory: Application of the SAM model”, *Cognitive Science: A Multidisciplinary Journal,* vol. 27, pp. 431–452, 2003.

[9] R. A. Schmidt and R. A. Bjork, “New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training”, *Psychological Science,* vol. 3, pp. 207–217, 1992.

[10] W. B. Whitten and R. A. Bjork, “Learning from tests: Effects of spacing”, *Journal of Verbal Learning and Verbal Behavior,* vol. 16, pp. 465–478, 1977.

[11] S. J. Thios and P. R. D’Agostino, “Effects of repetition as a function of study-phase retrieval”, *Journal of Verbal Learning & Verbal Behavior,* vol. 15, pp. 529–536, 1976.

[12] F. S. Bellezza and D. R. Young, “Chunking of repeated events in memory”, *Journal of Experimental Psychology: Learning, Memory, and Cognition,* vol. 15, pp. 990–997, 1989.

[13] P. P. J. L. Verkoeijen, R. M. J. P. Rikers and H. G. Schmidt, “Detrimental influence of contextual change on spacing effects in free recall”, *Journal of Experimental Psychology: Learning, Memory, and Cognition,* vol. 30*, pp.* 796– 800, 2004.

[14] J. R. Anderson, J. M., Fincham and S. Douglass, “Practice and retention: A unifying analysis”, *Journal of Experimental Psychology: Learning, Memory, and Cognition,* vol. 25, pp. 1120–1136, 1999.

[15] M.C. Mozer, H. Pashler, N. J. Cepeda, N. J., R. Lindsey and E. Vul , “Predicting the optimal spacing of study: A multiscale context model of memory”. In *Advances in neural information processing systems* (Y. Bengio *et al.* Eds.), pp. 1321–1329, La Jolla, CA: NIPS Foundation, 2009.

[16] M. Riopel, P. Chastenay, G. Fortin-Clément, P. Potvin, S. Masson and P. Charland, “Using invariance to model practice, forgetting, and spacing effects: the constellations’ case”, *ESERA 2017 conference* (21st-25th August, Dublin, Ireland), 2017.

[17] J. Allen and W. Yen, *Introduction to measurement theory*, Long Grove: Illinois: Waveland Press, 2002.

[18] B. Basrak, “Fisher-Tippett Theorem”. In *International Encyclopedia of Statistical Science* (M. Lovric Ed.), pp. 525-526, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[19] H. Goldstein, “Dimensionality, bias, independence and measurement scale problems in latent trait test score models”, *British Journal of mathematical and statistical psychology,* vol. 33, no. 2, pp. 234-246, 1980.

[20] S. C. Seow, “Information Theoretic Models of HCI: A Comparison of the Hick–Hyman Law and Fitts' Law”, *Human–Computer Interaction,* Vol. 20, no. 3, pp. 315–352, 2005.

[21] W. J. Van der Linden, “A Lognormal Model for Response Times on Test Items”, *Journal of Educational and Behavioral Statistics,* vol. 31, no. 2, pp. 181–204, 2006.

[22] M. Riopel, “Practice and forgetting curves deduced from scale invariance”, *EDULEARN17 conference* (3rd-5th July, Barcelona, Spain), 2017.