An Integrated Trial-Level Performance Measure: Combining Accuracy and RT to Express Performance During Learning

Florian Sense
f.sense@rug.nl
Department of Experimental Psychology & Behavioral and Cognitive Neuroscience
University of Groningen, Groningen, The Netherlands

Tiffany Jastrzembski, Michael Krusmark, Siera Martinez
{tiffany.jastrzembski, michael.krusmark.ctr, siera.martinez.ctr}@us.af.mil
Air Force Research Laboratory, Wright-Patterson Air Force Base, Dayton, OH, USA

Hedderik van Rijn
d.h.van.rijn@rug.nl
Department of Experimental Psychology & Behavioral and Cognitive Neuroscience
University of Groningen, Groningen, The Netherlands

Abstract

Memory researchers have studied learning behavior and extracted regularities describing learning and forgetting over time. Early work revealed forgetting curves and the benefits of temporal spacing and testing for learning. Computational models formally implemented these regularities to capture relevant trends over time. As these models improved, they were applied to adaptive learning contexts, where learning profiles could be identified from responses to past learning events to predict and improve future performance. Often times, past performance is expressed as accuracy alone. Here we explore whether a model's predictions can be improved if past performance is expressed by an integrated measure that combines accuracy and response times (RT). We present a simple, data-driven method to combine accuracy and RT on a trial-by-trial basis. This research demonstrates that predictions made using the Predictive Performance Equation improve when past performance is expressed as an integrated measure rather than accuracy alone.

Keywords: Learning; forgetting; cognitive model; accuracy; response time; integrated measure

Introduction

What data from fact learning trials are needed to predict whether a student will know the correct answer some time in the future? Does it help to know how often (and when) the student has previously answered correctly? Or how long it took them to provide the answer?

These questions are at the heart of models that describe learning and forgetting over time. Computational models are often fit to historical data to demonstrate that they can capture relevant behavioral effects exhibited by human learners (e.g. Pavlik & Anderson, 2008; Walsh, Gluck, Gunzelmann, Jastrzembski, Base, et al., 2018). Yet, the strongest test of a model is accurately predicting future performance—especially if predictions are made for each item studied by each student. The Second Language Acquisition Modeling (SLAM) challenge recently posed by Duolingo required such predictions (Settles, Brust, Gustafson, Hagiwara, & Madnani, 2018). Data from a subset of Duolingo users were made available and users submitted model performance predictions as part of a modeling competition..

As in these challenges, adaptive fact-learning systems must decide which features of the available data are taken into account to detect differences in item difficulty and participants' abilities to make accurate predictions. An obvious candidate is accuracy, since it indicates whether the student knew an answer previously. Forgetting then reduces the probability that responses are correct over time. Systems such as Duolingo (Settles & Meeder, 2016) strive to ensure that study repetitions occur *before* knowledge is forgotten.

Yet, if most responses are correct, there is very little information in the responses if only accuracy is considered, making it difficult to optimally adapt to learning and item difficulty. Response times (RT) can provide an additional source of information to differentiate between otherwise identical responses. The basic assumption is that observed RTs correlate with the difficulty of memory retrieval (e.g., Pavlik & Anderson, 2008; Pyc & Rawson, 2009). Indeed, analyses of the models submitted to the SLAM challenge support the view that RTs provide valuable information for predicting later performance (see Table 4 in Settles et al., 2018).

As accuracy and RT are often correlated (e.g., speed-accuracy trade-offs), methods have been proposed to combine them into a single performance metric. A recent suite of simulation studies discusses the merits of seven such integrated performance measures (Vandierendonck, 2017). All these measures, however, are aggregate measures: For example, the mean RT is combined with the average accuracy to express performance per participant, per condition. As this discards all information pertaining to *when* responses are given, these measures are less suited for parametrizing adaptive learning systems.

To our knowledge, there are at least two adaptive fact-learning systems that use both accuracy and RTs on a trial-by-trial level. Adaptive Response-Time-based Sequencing (ARTS; Mettler & Kellman, 2014; Mettler, Massey, & Kellman, 2016) schedules repetitions adaptively by continuously computing priority scores and presenting

the item with the highest priority. If the previous response was incorrect, that item's priority is increased drastically to ensure timely repetition. If the previous response was correct, however, the priority score is a function of the (log-transformed and scaled) RT associated with that response.

The second system is an extension of ACT-R's declarative memory module and uses the associated equations to approximate an item's memory strength (or "activation") through observed RTs (Pavlik & Anderson, 2008). Instead of using priority scores, items are repeated based on their estimated activation, a value that decreases over time (van Rijn, van Maanen, & van Woudenberg, 2009). Note that the observed RT of incorrect responses is replaced by a fixed, long RT, reflecting that it took "too long" for the correct response to be retrieved. These two examples demonstrate that combining information from accuracy and RTs is feasible in practice. Neither system really uses an *integrated* performance measure, however—they both use a transformation of RT that is conditional on accuracy.

Here, we will present an approach to computing an integrated, trial-level performance measure that combines accuracy and RT. Ideally, such a measure is purely data-driven, easy to interpret, computationally simple, and applicable to existing datasets. We are most interested in situations in which item-level data of the learning history are available and the goal is to validly predict future performance.

In the following, we will outline two datasets that we use as a test bed. Both datasets concern learning of paired associates, which provides a context in which the RT reflects relevant memory processes. We will demonstrate how our trial-level integrated performance measure can be computed for such data. Lastly, we describe how these integrated "Readiness" scores can be used as input to a computational model (the Predictive Performance Equation, Walsh, Gluck, Gunzelmann, Jastrzembski, Base, et al., 2018) to generate predictions based on past performance.

The central focus of this research seeks to explore whether use of this "Readiness score improves model predictions compared to scores that do not integrate accuracy and latency.

Methods

Datasets

We leverage two existing datasets, labelled WSU and TopiCS, to explore the idea of an integrated, trial-level performance measure. Each dataset consists of a study and a test phase. Trial-level information for response accuracy and RT is available for both datasets but they vary drastically in the structuring of the study phase and in the time between study and test. Importantly, accuracy during study is very high in both datasets (85.9% in WSU and 89.8% in TopiCS).

Washington State University (WSU) data WSU data is part of an (as of yet) unpublished multi-day fatigue study. Participants spent four days in a sleep lab and completed a battery of tests throughout that period. Here, we will focus on the paired-associates learning data of 36 participants who were not withheld any sleep (the control group). Fifty-one nonsensical line drawings—e.g.,  —were used as cues and participants learned two-digit numbers—e.g., "79"—as a response. Each paired associated was repeated 20 times according to different presentation schedules.

Table 1: Number of repetitions of an item at each test moment depending on the schedule in the WSU data. RI = retention interval between the last encounter of an item and the test.

	Study phase					Test phase
Schedule	9am	1pm	3pm	7pm	9pm	9am
Spaced	4	4	4	4	4	2 (36h RI)
Massed early	20	·	·	·	·	2 (48h RI)
Massed late	·	·	·	·	20	2 (36h RI)

We will focus on three schedules that distributed the 20 repetitions across a single day. Table 1 shows that the *spaced* schedule distributed the 20 repetitions equally among five study periods throughout the day (four repetitions each), while the *massed* schedules presented each item 20 times either *early* or *late* in the day. The test phase featured two repetitions for each paired associate, and the retention interval (RI; i.e., the temporal space between the last encounter of an item and the test) depended on which study schedule the paired associate was assigned to. Each participant studied with all schedules and encountered three unique paired associate per schedule, resulting in 6,156 observations from the study phase that were used to make predictions for the 648 observations from the test phase.

The recorded RT corresponds to the first key press. If participants did not respond within 6 seconds, the trial was recorded as incorrect (with RT set to 6 sec, hence the spike in the lower right panel of Figure 1A).

TopiCS data The TopiCS data were taken from Sense, Behrens, Meijer, and van Rijn (2016), published in Topics in Cognitive Science. Participants completed three sessions of two blocks each (six total). In each block, material was studied for 20 minutes using an adaptive fact-learning system (van Rijn et al., 2009), followed by a five-minute distractor task (Tetris), followed by a test of the studied material. Here, we will only use the first block of each session. In each of these three blocks, participants studied Swahili-English vocabulary word pairs. Each Swahili block featured 25 unique paired associates.

A total of 50,665 responses are available from 67 participants. Since the introduction and repetition schedules of items during study were governed by an adaptive model, these data do not have the controlled temporal structure of the WSU data: The number of repetitions as well as their timing varied between items and participants. The test was the same for everyone, however. After a five-minute delay, participants

were tested on all 25 potential Swahili cues at the end of each block and accuracy was recorded (4,965 observations).

The study phase was entirely self-paced and RTs correspond to the first key press recorded after a Swahili word appeared on screen. The RT distributions, split by accuracy, are shown in Figure 1A.

Computing integrated "Readiness" scores

The goal is to derive a trial-level, quasi-continuous performance metric from accuracy and RT data. This "Readiness" value can take any value between 0 and 1. Values closer to 1 correspond to a correct, fast response. There are two versions of the metric: For R_0, all incorrect responses are treated equally and set to 0. For R_c, incorrect responses are transformed such that faster incorrect responses are more severely penalized (i.e., closer to 0) than slow incorrect responses (cf. Klinkenberg, Straatemeier, & Van der Maas, 2011). The term "Readiness" is used because values close to the 1 indicate that a response was readily available, resulting in a fast RT. Overall, the higher the "Readiness" value, the better the performance. In the following, we will detail how R_0 and R_c are computed from behavioral data.

Figure 1A depicts the distribution of RTs for correct (top panels) and incorrect (bottom panels) responses from the two datasets. For a more precise depiction of the data, the axes across the panels vary and only RTs faster than 15*s* are shown for the TopiCS data (99% of all observations). Since the vast majority of responses were correct, there are fewer RTs for the incorrect responses in the bottom panels.

Figure 1B makes the mapping from observed RTs to the probability of a correct response explicit: In both datasets, the log-transformed RTs (in ms) are strong predictors of accuracy, such that slower RTs reduce the probability of a correct response. The exact mapping differs in the two datasets: Responses are generally faster in the WSU data and time out after 6 seconds. The mapping is expressed as the two coefficients estimated by a simple logistic regression, which is $\beta_0 + \beta_1 \cdot \log(\text{RT})$. For the WSU data, β_0 is 24.26 and β_1 is -3.09. For the TopiCS data, β_0 is 12.99 and β_1 is -1.39. All four coefficients differ significantly from 0 with $p < 0.001$.

The relationship shown in Figure 1B provides the quantitative basis for the "Readiness" metrics. The mapping provided by the logistic regression allows an unbound performance metric (RT) to be transformed to a continuous metric with range [0, 1].

For the first metric, R_0, all correct responses are transformed using the mapping provided by the logistic regression coefficients. *Incorrect* responses are treated as performance of 0 (as with accuracy; hence the subscript 0). Using this approach, a correct response given quickly is considered "more correct" than a correct response given after longer deliberation, which is in line with behavioral data and theoretical assumptions (Pavlik & Anderson, 2008). Numerically, R_0 is computed by taking the inverse logit (L^{-1}) of the regression formula shown above, using log-transformed RTs (in *ms*) when accuracy (A) is 1:

$$R_0 = \left\{ \begin{array}{ll} A = 0: & 0 \\ A = 1: & L^{-1}(\beta_0 + \beta_1 \cdot \log(\text{RT})) \end{array} \right\} \quad (1)$$

The second "Readiness" metric, R_c, assumes that latencies for incorrect responses are informative too. Specifically, the assumption is that a fast incorrect response is *worse* than an incorrect response given after longer deliberation, a notion also present in other learning systems (e.g., Math Garden—an adaptive, online arithmetic-learning environment used by many schools in the Netherlands—formalized the same idea in the "high speed, high stakes" scoring rule; Klinkenberg et al., 2011, see section 2.3.3. and Fig. 2 specifically). Numerically, this is formalized by using the same approach as for correct responses but then subtracting 1 and taking the absolute value[1]:

$$R_c = \left\{ \begin{array}{ll} A = 0: & |L^{-1}(\beta_0 + \beta_1 \cdot \log(\text{RT})) - 1| \\ A = 1: & L^{-1}(\beta_0 + \beta_1 \cdot \log(\text{RT})) \end{array} \right\} \quad (2)$$

For example, a correct response with an RT of 1,834*ms* would result in the same R_0 and R_c scores but they would depend on the dataset the response was observed in. In the WSU data, the "Readiness" score would be 0.739 ($L^{-1}(24.26 - 3.09 \cdot \log(1,834))$ but in the TopiCS data it would be higher ($L^{-1}(12.99 - 1.39 \cdot \log(1,834) = 0.927)$ because RTs are generally longer, which results in a different mapping (cf. Figure 1A and B). If the RT is the same but associated with an *incorrect* response, the R_0 score is simply 0 (see Eq. 1). The R_c score, on the other hand, would be 0.261 (i.e., $|0.739 - 1|$) in the WSU and 0.073 (i.e., $|0.927 - 1|$) in the TopiCS data (see Eq. 2).

Figure 1C gives an overview of the R_c values computed in the two datasets, split again by accuracy and dataset. Note that the y-axes differ due to the unequal number of observations. For both datasets, the correct responses (top panels) mostly have values between 0.75 and 1. The incorrect responses (bottom panels) are more spread across the range for the R_c metric[2]. For the TopiCS data, most incorrect responses have R_c values between 0 and 0.5 but mostly values are <0.25. In the WSU data, the values are spread more widely. The distributions of correct and incorrect responses barely overlap within a dataset, though (note the small numbers on the y-axis for incorrect reponses from the WSU data). For the R_0 metric, all values corresponding to incorrect responses are simply 0 (cf. Eq. 1).

Taken together, the approach outlined here has multiple advantages. A binary and an unbound performance metric (accuracy and RT) are combined into an integrated, trial-level measure that is continuous and bound between 0 and 1. Importantly, a trial-level performance metric preserves

[1] This could be thought of as flipping the mapping in Figure 1B along the horizontal axis at 0.5.

[2] For the WSU data, timed-out observations with RTs of 6*s* (N = 35) were transformed to the fastest observed RT (391*ms*) to make them "very wrong". If these observations are simply dropped, none of the reported results change qualitatively.

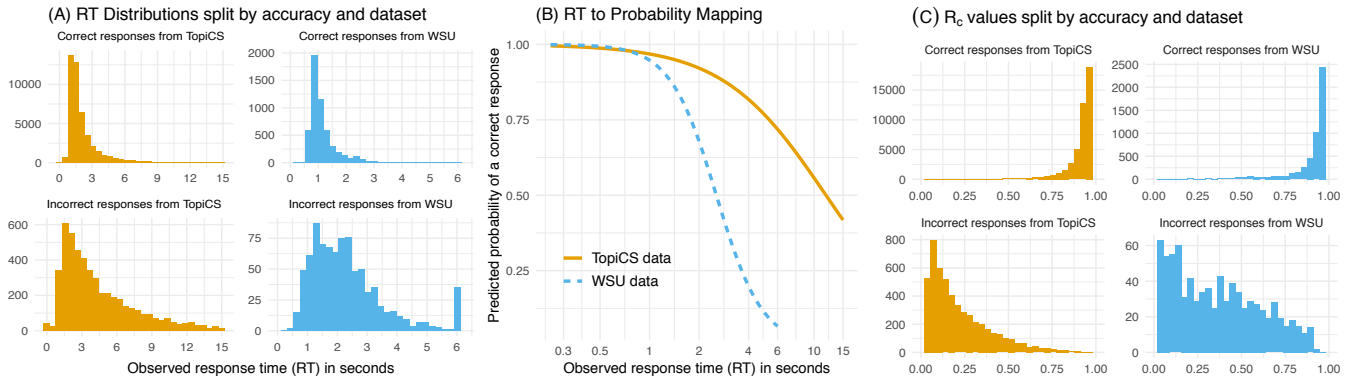

Figure 1: A: The observed RT in seconds split by accuracy and dataset; B: Logistic regression lines showing the mapping from observed RTs in seconds to the probability of a correct response; C: The continuous "Readiness" (R_c) values computed using the mapping in B, split by accuracy and dataset. In all plots, color indicates the dataset. See text for additional details.

information about the timing of individual encounters that would be lost if one simply computed, for example, the mean accuracy during study. The approach is also computationally extremely simple and makes minimal theoretical assumptions that are easily checked: Are longer RTs associated with a lower probability of giving an accurate answer? If the logistic regression's slope coefficient (β_1) does not differ significantly from 0, computing "Readiness" values is probably not sensible. Visual checks akin to Figure 1B and C also provide easy sensibility checks. Finally, the interpretation of "Readiness" values is straightforward: Higher values indicate better performance and values at the boundaries indicate faster responses.

Predictive Performance Equation (PPE)

To explore whether the "Readiness" scores are useful, we will explore their utility as input to the Predictive Performance Equation (PPE), a computational model developed to capture individual differences in learning and forgetting (for an extended description of the model see Walsh, Gluck, Gunzelmann, Jastrzembski, Base, et al., 2018). If the scores expressed meaningful individual differences, a computational model should more closely mimic a participant's learning and forgetting process than when other scores are used. The end result would be more accurate predictions of future performance based on past performance.

In two recent studies, PPE was compared to other models to test "the theoretical adequacy and applied potential of computational models" more generally (Walsh, Gluck, Gunzelmann, Jastrzembski, & Krusmark, 2018) and to shed light on "the mechanisms underlying the spacing effect in learning" specifically (Walsh, Gluck, Gunzelmann, Jastrzembski, Base, et al., 2018). Due to space constraints, we will keep the current description of the model mechanics brief and refer the interested reader to those papers for a detailed overview.

The PPE component we are ultimately interested in is the predicted performance, P, which is a logistic function of activation (M) that has two free parameters, τ and s:

$$P = \frac{1}{1 + \exp(\frac{\tau - M}{s})} \tag{3}$$

The activation M is the product of learning and forgetting, expressed as $N^{0.1} \cdot T^{-d}$. The learning term increases exponentially as a function of the number of repetitions (N) and the forgetting term decreases exponentially. The latter has two components: The elapsed time (T) is the weighted sum of the time since each previous repetition (see Eq. 3 and 4 in Walsh, Gluck, Gunzelmann, Jastrzembski, & Krusmark, 2018) and the decay rate (d), which has free intercept (b) and slope (m) parameters and is a function of the lag between consecutive repetitions:

$$d = b + m \cdot \left(\frac{1}{n-1} \cdot \sum_{j=1}^{n-1} \frac{1}{\ln(\text{lag} + e)} \right) \tag{4}$$

Model fitting In the form outlined above, PPE has four free parameters (b, m, τ, and s) and requires two pieces of information to be fit: The time point of each repetition (to compute T and d) and the observed performance at each time point. The model is agnostic with regards to what the performance metric represents and only requires it to fall in the range of [0, 1], as the "Readiness" measure provides. The best-fitting parameters are found by minimizing the error between the supplied performance metric and the predicted performance P (see Eq. 3) produced by a given combination of the free parameters. The error is defined as the summed squared error between the performance metric and P across the data available for each unique participant-item combination.

Here, we only vary the performance metric that is used during model fitting, using either accuracy, R_0, and R_c. All other factors—free parameters, allowed parameter ranges[3],

[3]The ranges for the free parameters are b = [0, 0.5], m = [0, 0.5],

and timing-related information—are held constant.

Results

For both datasets, we determined the best-fitting PPE parameters for each participant-item combination, and then computed item-level predicted performance, P, on the test. The model fit will be evaluated for the predictions—i.e., comparing predicted P with recorded accuracy—rather than fit to the study data because we consider the ability to predict future performance given historical data most relevant.

For each dataset, PPE was fit three times, using the three performance metrics outlined above: Accuracy, which is binary; R_0, which is continuous for correct responses but all incorrect responses are 0 (see Eq. 1); and R_c, which is continuous for both correct and incorrect responses (see Eq. 2 and Figure 1C). The main results of the comparison are presented in Table 2, which lists two model fit statistics for each performance metric. Also included is a baseline, which simply predicts that all responses during the test are correct.

In both datasets, performance on the test is expressed as accuracy. PPE, on the other hand, predicts the probability of a correct response. To evaluate the model predictions, we use fit statistics commonly used when evaluating performance in binary classification problems. The fit statistics are: (1) The area under the receiver operating characteristic curve (*AUC*), which can be interpreted as the probability of a randomly drawn correct response outranking (i.e., having a higher P value) a randomly drawn incorrect response. Note that the baseline condition, in which all responses are predicted to be correct, would result in an AUC of 50%. (2) *Log loss*, expressing the accuracy of a classifier by penalizing inaccurate classifications. The AUC measure can range from .5 to 1, higher values are better. Log loss is unbound and lower values are better. In Table 2, the best-performing metric is highlighted in bold for each fit statistic.

Table 2: Fit statistics for predictions made in the two datasets. The baseline predicts that all responses on the test are correct.

Dataset	Statistic	Baseline	Accuracy	R_0	R_c
WSU	AUC	0.500	0.687	0.768	**0.769**
	Log loss	19.635	6.170	1.900	**1.141**
TopiCS	AUC	0.500	0.679	0.712	**0.755**
	Log loss	1.298	3.110	1.701	**0.638**

Table 2 shows the fit statistics for the 648 predictions made in the WSU data. All three measures outperform the baseline. Of these, using accuracy as performance measure scores lowest, and there is no clear difference between the two "Readiness" scores. This impression is confirmed by statistical comparison of the AUC values, which tests the null hypothesis that the difference between two AUCs is 0 against the alternative hypothesis that it is not (DeLong, DeLong, & Clarke-Pearson, 1988). The tests yield significant differences between the accuracy- and R_0-based AUCs (z = -4.449; $p < 0.001$) and accuracy- and R_c-based AUCs (z = -3.850; $p < 0.001$) but not between R_0- and R_c-based AUCs (z = -0.063; $p = 0.950$). The log loss is very high for the baseline because the actual accuracy on the test was only 45.5%, resulting in a high penalty.

In the TopiCS data, on the other hand, the observed accuracy on the test was extremely high: 96.4% of the 4,965 responses were correct. Thus, the all-correct baseline gets less than 4% of the predictions wrong, resulting in a relatively low log loss value. Only the R_c score yields predictions that result in a lower log loss value than the baseline. Regarding the AUC values, all predictions derived from the computational model outperform the baseline. The statistical test for the comparison of the accuracy- and R_0-based AUC is inconclusive (z = -1.451; $p = 0.147$), while the R_c-based predictions are significantly better than both the accuracy- (z = -2.809; $p = 0.005$) and R_0-based predictions (z = -2.148; $p = 0.032$).

Discussion

Here, we explored the predictive power of an integrated performance measure that combines accuracy and RT information. Unlike aggregate measures (see Vandierendonck, 2017, for an overview), the "Readiness" scores presented here are computed for each observation individually. Using two datasets, we demonstrated how "Readiness" scores are computed. The practical utility of the resulting integrated performance measures was demonstrated by fitting a computational model to past performance in order to predict future performance. Statistical analyses reveal evidence that predictions are more accurate when past performance was expressed as a "Readiness" score rather than accuracy.

We present two variations of the "Readiness" score that differ in how they treat incorrect responses. The R_0 score regards all incorrect responses equally, setting them to 0 (analogously to accuracy). The continuous score, R_c, scales both correct *and* incorrect responses (see Figure 1C) such that fast incorrect responses are considered worse than slow incorrect responses. Both versions express performance as scores between 0 and 1, with higher values indicating better performance. For R_c, scores closer to either boundary correspond to responses that were given quickly.

Whether R_0 or R_c should be preferred—or whether either should be used—depends on the context and the assumptions the researcher can make, especially regarding incorrect responses. Since the "Readiness" scores are based on empirical data, the data can provide an immediate check. If the slope of the logistic regression model that provides the mapping (cf. Figure 1B) does not significantly differ from 0, the crucial assumption that observed RTs and accuracy are associated is violated and "Readiness" scores are probably

τ = [0, 1], and s = [0, 0.1].

not meaningful. As discussed in the Methods section, visualizations such as those in Figure 1B and C can also inform the researcher's choice. In a very large dataset, for example, the logistic regression model might have a significant but very small slope coefficient, resulting in a mapping (cf. Figure 1B) for which even very slow RTs result in near-ceiling performance, which would in turn yield R_c values that are quasi-equivalent to accuracy.

Exploring to which extent "Readiness" scores could be a useful expression of past performance in different contexts would be a logical extension of the current work, which presents an initial exploration of the idea in two relatively small datasets. This first exploration is promising, however, given that even though both datasets differed in a number of important aspects, the "Readiness" measure outperformed accuracy in both. Most importantly, the retention intervals differed dramatically (five minutes in the TopiCS data and 36–48 hours in the WSU data), which meant that test performance was near-perfect in the TopiCS data and lower than 50% in the WSU data. Another possible extension of the current work would be to investigate the utility of "Readiness" scores in computational models other than PPE.

In conclusion, we present a simple, data-driven way to combine accuracy and response time information into an integrated, trial-level performance measure that we call "Readiness." This approach makes minimal assumptions that are easy to check and resulting performance scores are easy to interpret. This research demonstrates that a single computational model can capture the general learning and forgetting patterns observed across two very diverse sets of paired associate learning data, and that the model's predictive validity is enhanced when past performance is expressed in terms of an integrated "Readiness measure, rather than use of simple accuracy alone.

Acknowledgements

FS was supported by L3 Technologies through the Air Force Research Laboratory at Wright-Patterson Air Force Base.

References

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, *44*(3), 837–845. doi: 10.2307/2531595

Klinkenberg, S., Straatemeier, M., & Van der Maas, H. L. J. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, *57*(2), 1813–1824. doi: 10.1016/j.compedu.2011.02.003

Mettler, E., & Kellman, P. J. (2014). Adaptive response-time-based category sequencing in perceptual learning. *Vision Research*, *99*, 111–123. doi: 10.1016/j.visres.2013.12.009

Mettler, E., Massey, C. M., & Kellman, P. J. (2016). A Comparison of Adaptive and Fixed Schedules of Practice. *Journal of Experimental Psychology: General*, *145*(7), 897–917. doi: 10.1037/xge0000170

Pavlik, P. I., & Anderson, J. R. (2008). Using a Model to Compute the Optimal Schedule of Practice. *Journal of experimental psychology. Applied*, *14*(2), 101–117. doi: 10.1037/1076-898X.14.2.101

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447. doi: 10.1016/j.jml.2009.01.004

Sense, F., Behrens, F., Meijer, R. R., & van Rijn, H. (2016). An Individual's Rate of Forgetting Is Stable Over Time but Differs Across Materials. *Topics in Cognitive Science*, *8*(1), 305–321. doi: 10.1111/tops.12183

Settles, B., Brust, C., Gustafson, E., Hagiwara, M., & Madnani, N. (2018). Second Language Acquisition Modeling. In *Thirteenth workshop on innovative use of nlp for building educational applications* (pp. 56–65).

Settles, B., & Meeder, B. (2016). A Trainable Spaced Repetition Model for Language Learning. *Association for Computational Linguistic (ACL)*, 1848–1858.

Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior Research Methods*, *49*, 653–673. doi: 10.3758/s13428-016-0721-5

van Rijn, H., van Maanen, L., & van Woudenberg, M. (2009). Passing the Test: Improving Learning Gains by Balancing Spacing and Testing Effects. In *Proceedings of the 9th international conference on cognitive modeling* (pp. 110–115). Manchester, UK.

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., Base, F., Myung, J. I., … Zhou, R. (2018). Mechanisms Underlying the Spacing Effect in Learning: A Comparison of Three Computational Models. *Journal of Experimental Psychology: General*, *147*(9), 1325–1348. doi: 10.1037/xge0000416

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., & Krusmark, M. (2018). Evaluating the Theoretic Adequacy and Applied Potential of Computational Models of the Spacing Effect. *Cognitive Science*, *42*, 644–691. doi: 10.1111/cogs.12602