# PROCEEDINGS

of the

**2017 Symposium on Information Theory and Signal Processing in the Benelux**

*May 11-12, 2017, Delft University of Technology, Delft, the Netherlands*

http://cas.tudelft.nl/sitb2017

*Richard Heusdens & Jos H. Weber (Editors)*

ISBN 978-94-6186-811-4

The symposium is organized under the auspices of

Werkgemeenschap Informatie- en Communicatietheorie (WIC)

& IEEE Benelux Signal Processing Chapter

and supported by

Gauss Foundation (sponsoring best student paper award)

IEEE Benelux Information Theory Chapter

IEEE Benelux Signal Processing Chapter

Werkgemeenschap Informatie- en Communicatietheorie (WIC)

TU Delft

Delft University of Technology

IEEE

# Towards understanding behavioural biometric classifier performance over time and practice

E. Haasnoot[1] J.S. Barnhoorn[2] L.J Spreeuwers[1]
R.N.J Veldhuis[1] W.B. Verwey[2]
University of Twente
[1]EWI, Group SCS [2]BMS, Group CPE
Drienerlolaan 5, 7522 NB Enschede
e.haasnoot@utwente.nl

**Abstract**

Behavioural biometrics in the context of security and authentication looks at the discriminative features of a user's measurable behaviour. This generally includes timing and location information related to, for example, screen touches and key presses, otherwise known as keystroke dynamics (KD). Research into KD has looked at discriminating features of behavioural patterns of expert typists, which are generally very stable, as well as patterns of novices, which are generally very unstable, if only because of rapid increases in skill level due to practice. The general population, however, at which such authentication solutions are aimed are not expert typists, and quickly move away from being novices, which we found causes significant degradation of biometric recognition performance over time. This is because the biometric data entered at a later stage will increasingly differ from data gathered during enrollment. Accounting for practice effects in KD systems is difficult, as not much is yet known about the way behaviour develops over time in biometrics literature. However, with the advent of open science, not only can we incorporate many new insights from the psychology of (motor) learning, but also re-analyse data sets gathered in this field in a biometric context. In this paper we present initial analyses over a single data set, in which 36 (18 older, 18 young) participants were asked to complete a "password-entry"-like cognitive task, the Discrete Sequence Production task, a significant number of times spread over two sessions. Using out-of-the-box classifiers, we found that biometric recognition performance takes a long time, 75 repeats, to stabilize, as well as hints towards better initial biometric recognition performance for older participants than for younger.

# 1 Introduction

Behavioural biometrics in the context of security and authentication looks at measurable and measured behaviour in order to make classification decisions about someone's identity. It can be said to encompass some of the most basic forms of identification and authentication, for example most people recognize Rembrandts and van Goghs by their distinct painting style. We can listen to a piece of music and say with some degree of accuracy whether it was composed by Bach, Beethoven or Mozart, and whether the performance was by Glenn Gould, Brendel or Poporevich. We can marvel at the abilities of football players Lionel Messi, Cristiano Ronaldo and Arjen Robben on our TVs. Or, between amateur radio operators, recognize other radio operator's "fist" or telegraphic style, even before direct identification is received[9]. All of these examples can be seen as a record of behaviour by expert performers. Some having stood the test of time better than others.

The earliest academic papers on behavioural biometrics focused on the keystroke characteristics or keystroke dynamics (KD), timing and spatial information related to keyboard key presses, of professional typists[9, 23]. Typists are more likely to find themselves amongst the top performers in the skill of typing, each having stable "signature" components to their typing behaviour owing to both mental and physiological factors. As such they show stable patterns in their typing behaviour, by which other typists can (and do) recognize them. Looking at the keystroke characteristics of typists was done both out of convenience (due to relatively easy access to a number of typists) and the likely existence of the before mentioned statistically significant behavioural patterns. [9] results were especially impressive, as they reported 0 instances of false acceptance, and only 2 instances of false rejection, out of 55 comparisons.

Contemporary research into behavioural biometrics focuses more on novices, or unpracticed individuals, whereas experts can be considered extremely practiced individuals. Although biometric recognition performance deteriorated with the change of focus on novices[4], behavioural biometrics is still seeing much interest due to increasingly ubiquitous access to sophisticated sensors (e.g. in smartphones) with feature-level [14, 26, 13, 25], system-level [5, 20] and even framework-level [15] validation studies being published regularly.

A major issue with studies on novices is the majority of the population that uses a specific biometric application is not actually a novice, or at least not for long. User behaviour and patterns change quickly as users practice and learn to use said application. Ideally, validation studies should look at the average use case of the application[19], and thus focus on users with an "intermediate" amount of practice, rather than novices or experts. Setting up studies that take into account time and practice can often be prohibitively expensive in time, money and complexity. Researchers therefore often compromise on study length, in order to keep time investment, expenses and complexity down, but also reducing the generalizability of the validation study. This results in the effects of practice being often acknowledged as important, but studied little[6].

Luckily, there are other areas of research interested in the effects of practice on (motor) learning. Data sets collected in this area are often quite well suited to re-analysis in a biometric context. In order to investigate the important role of practice on biometric recognition performance, we present a re-analysis of the data gathered by [3]. They asked their 36 participants to perform a relatively simple password-entry-like task, the Discrete Sequence Production (DSP) task, a large number of times (864 times in total, per participant). Besides the effects of practice, the nature of this data set also allows us to test how age affects biometric recognition performance. It is not necessarily a given that findings from validation studies on young students generalize well to the whole population, for example to older adults.

Given the large difference in biometric recognition performance between novices and experts, and power law of learning experiments' results of decreasing gains in performance under equal amounts of practice as one becomes more practiced[16], we hypothesize that biometric recognition performance will level-off (that it shows asymptotic behaviour) and that there will be a smaller difference in recognition performance between intermediate and expert users, than between novice and intermediate users. Also, we are interested in finding out whether hitting such a plateau means that behaviour stops changing, thus providing useful information on required lengths of future validation studies.

The contribution of this paper is thus threefold. First, we present a novel way of analysing existing data sets in a biometric context. Second, we investigate the effects of practice are on behavioural biometric systems. third, we shed some light on differences in the effects of practice for different age groups.

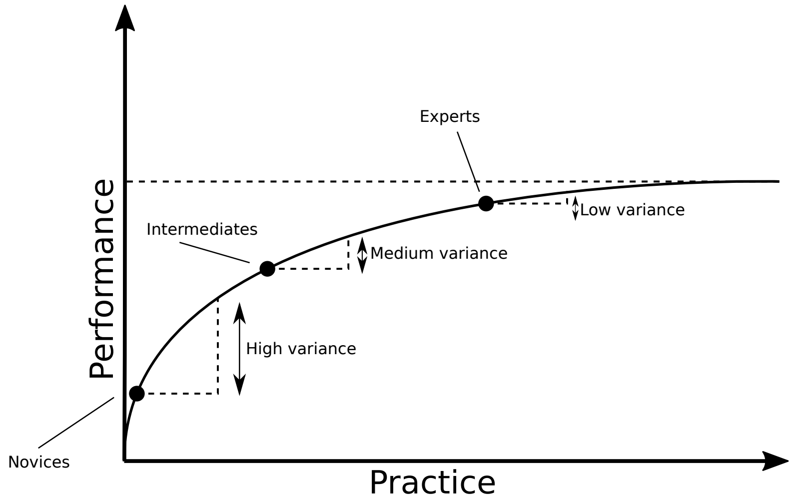


Figure 1: A power law of practice. Different starting points (novice, intermediate, expert) lead to different levels of behavioural variance due to practice.

# 2 Background

In this section we give a short summary of the psychological literature relevant to the problem of practice, and an introduction to the theory behind the Discrete Sequence Production task, which was used previously[3].

## Practice and expertise

The role of practice in skill learning and eventual expert performance is most often studied in a psychological context. [8] proved a landmark paper in this, in that it shifted beliefs about the cause of expert performance away from "natural ability" or talent[10] and gave due respect to deliberate practice. Deliberate practice being "the collection of activities that have been found most effective in improving performance", as opposed to practice as "the collection of activities that are only related to the domain in which one wishes to improve performance"*.

The relation between practice and performance is often depicted in a power law[16]. See Figure 1. Given an equal amount of (deliberate) practice, a high performer will see their performance improve less than a low performer. Such power laws are often criticized as being overly simplistic[7], or only hold for averages[11], but they are some of the simplest forms in which one can show a decreasingly lower performance gain from equal practice and provide a relatively simple way of showing why biometric recognition performance deteriorated when switching from testing on experts to novices, from highly practiced to unpracticed individuals. If we use performance as a biometric feature, novices will show much higher variance in this feature, as they rapidly rise up the curve, than experts or intermediately practiced individuals will, as Figure 1 shows. Higher variance leads to lower classifier performance, as it will be more difficult to separate subjects from each other.

Computational frameworks of learning[22] posit the existence of a fundamental trade-off between expending energy to discover new actions, or to perform actions

*Note that although training and practice quite often refer to the same concept of improving (user) skill/performance by (repetitive) doing in psychological literature, training as a term is also used in context of machine learning and training classifiers. In order to avoid confusion, we will use training exclusively as meaning the training of classifiers, and practice exclusively as improving user skill/performance.

which we estimate to yield the best results. This is known as the exploration vs exploitation dilemma. One can see [8]'s deliberate practice in this light as well. The constant trade-off of attempting to discover more activities that are more effective at improving performance, or to perform activities that are currently found to be most effective in improving performance. Early behavioural variance seems to be positively correlated with eventual relative ranking of a performer among their peers[21], which seems to suggest that strategies with high early exploration fare better than those with high early exploitation. A similar finding is reported in the area of motor learning, an area particularly interesting to behavioural biometrics, where high motor variance seems to promote learning [12].

## Motor Sequence Learning and the Discrete Sequence Production task

Many, if not all, tasks in behavioural biometrics can be seen as sequences of motor actions, or the repeated production of movement. For example, biometric systems that use keystroke dynamics, gait, gaze and handwriting all derive their data from tasks that ask a (willful or not) participant to respectively repeatedly press keys, to walk, to look at things and to write things down. The theory of how we learn such motor sequences is therefore of special interest to, but seems to be ignored in, contemporary behavioural biometric research. Most behavioural biometric tasks ask for relatively simple behaviour, which after enough practice become highly automated (e.g. walking, writing, typing). Motor sequence learning focuses on how we attain such automated behaviour. In other words, how we learn to "rapidly and accurately produce a sequence of movements with limited effort and/or attentional monitoring"[1].

Rather than effortfully initiating and executing each individual movement in a motor sequence, with practice one learns to group movements together in so-called motor chunks. These motor chunks can be, much like individual movements, initiated and executed, but while initialisation takes (slightly) longer, execution takes significantly fewer attentional resources, reducing execution time and errors. These reduced execution time and errors, however, come at a significant decrease in flexibility of movement. Formation of motor chunks can therefore be considered the result of an efficiency computation trade-off[18].

Many paradigms exist to tease apart the mechanisms by which motor chunks are formed, and by which we learn motor sequences. The Discrete Sequence Production task is one such paradigm[24]. In the DSP, participants rest a number of fingers (four to eight) on keys, while viewing an equal number of squares (each corresponding to one key) on a screen in front of them. One by one, the squares will light up, indicating to the participant the order in which the keys on the keyboard need to be pressed. The sequence indicated by the order is generally 3-7 stimuli long, and participants are asked to learn two sequences at the same time. With practice in the order of 500-1000 trials, the DSP turns from cueing individual key presses, to cueing longer sequences of key presses, eventually resulting in a 2-choice task where the first square that's lit up indicates which sequence to press. For a more in-depth explanation of the intricacies of the DSP task, and the theoretical dual processing model behind it, we refer to [1].

The DSP is in essence very similar to repeated entry of a password, with the limitation that the amount of characters one can choose for the password is limited to the amount of fingers involved in the specific implementation of the task. Also, sequences are not very long and characters generally don't repeat in a sequence. Although this limits the possible space of different sequences, the "passwords" are much simpler and it should be much easier develop automated behaviour for them than for regular 8-16 characters long passwords.

# 3 Methods

## Data set

The data set we performed our analyses over was collected by [3] in order to better understand how ageing is related to increased difficulty in developing new motor skills. As such our participants group is split into two equally sized group of 18 individuals. One group consisted of older adults (OA, age = 79 +- 3.5, 13 females) and were recruited via local media. The other group consisted of younger adults (YA, age 21 +- 1.2, 7 females) and were students participating for course credit. All participants were right-handed. Both groups of participants visited the lab on two consecutive days, in which they performed 9 blocks of 48 DSP trials for both a 3-element and 6-element sequence (so 96 trials per block), with short breaks between every full block of DSP trials. In total, each participant performed 864 trials. For each element a reaction time (RT) was recorded, both relative (to the previous RT or start of the trial) and absolute (to the start of the experiment). Errors in a particular sequence meant that the trial would be stopped and the next trial would start. Further RTs were filled with Null values, and an accuracy of 0 was recorded (as opposed to 1 for accurate trials). For more details please refer to [3].

## Analyses

We first cleaned up the data. Missing values in inaccurate trials were filled in with a placeholder value, 1e5 (which was far above what the experimental design would allow). Also, a new column was created, preciseACC, that indicated WHEN the mistake was made, which ranged from 0-1. This allowed us to even include inaccurate trials in our analyses, which would keep the amount of samples per participant the same. We dropped all the 3-key sequence trials, as we felt these were too short to give any results. Because we only have access to key down times (RTs) and no key up times were recorded, we report on performance relative to other points, and do not interpret the absolute performance values.

For our data analyses we used the out-of-the-box random forests (RF) classifier available in the python library scikit-learn[17], with settings set to default. RF classifiers are resistant to outliers, such as those introduced by our placeholder missing value. They also have the added benefit of being somewhat introspectable, as they are based on decision tree classifiers. We measured the performance of our classifier in terms of Equal Error Rates (EER) and employed three different strategies of temporally segmenting our data set, in order to gain a better understanding the development of the biometric recognition performance over practice. These strategies were chosen so that we could compare situations that might arise in behavioural biometric systems. Users might have significant amounts of practice before enrolling (strategy 1), users might use a system for very long after enrolling (strategy 3), and how behavioural patterns continue to change with significant training (the delta between strategy 1 and strategy 2). See Figure 2 for a visual explanation. We compared day 1 and day 2 for each individual strategy, both for all groups together and split between younger and older adults. Because of the anecdotal and exploratory nature of this paper, we report the results in qualitative features of the graphs.

# 4 Results

We present our results of the whole-group analysis in Figure 4. Strategy 1, moving both training and testing segments, results in a recognition performance plateau around an offset of 80  trials for day 1, with day 2 seemingly immediately reaching the plateau.

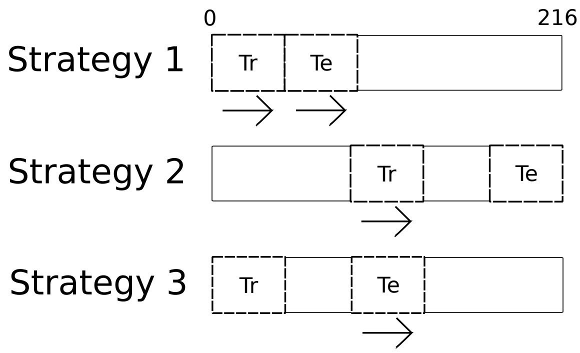


Figure 2: We repeatedly temporally segment our samples into up to 5 distinct blocks. We extract two of these segments, and label one as the training set (Tr) for our classifier, and one as the test set (Te). Segmentation happens differently for every run of the classifier, and we employ three different strategies for picking the segmentation points. Arrows denote which segmentation points we move. 1) Moving both segments - we move both the train and the test segment in lockstep. 2) Moving training segment - move the training segment towards the fixed test segment. 3) Moving test segment - we move the test segment away from the fixed training segment.

Strategy 2, where we move the location of the training segment closer and closer to the location of the test segment results in a significantly lower performance between the beginning of the plateau stage of recognition performance ( 80 samples in for day 1; start of day 2) and the final samples taken on the respective days. Strategy 3, where we move the location of the test segment increasingly further away from the training segment, we find a significant differences between all points in day 1 and day 2, as well as between both the beginning and end of day 1 and day 2.

We perform the same analysis again, Figure 4, but now with our data split into two groups, one with young adults (YA), and one with older adults (OA). For strategy one, interestingly, we find a significant difference between the start of day 1 and day 2 for the OA group, but not for the YA group. Again, no significant difference exists between the end of day 1 and 2, both for OA and YA. For YA, there is however a significant difference between the end of day 1, and the start of day 1. Strategy 2 again shows significant differences for both YA and OA between the training on samples in the beginning of the plateau stage ( 80 samples in for day 1; start of day 2 and the end for both day 1 and day 2. For the OA group in strategy 3, there might not be a significant difference between day 1 and 2, whereas for the YA group the difference is very clear. There is a difference between YA and OA for day 1, but this disappears for day 2.

# 5 Discussion

In this paper, we attempted to answer three questions that are especially relevant to the poorly understood effects of time and practice on the performance of a behavioural biometric recognition system. Specifically, does recognition performance eventually plateau and does this mean that behaviour (or some small sets of patterns in behaviour) stops changing as well and does age have any influence on this?

From plot 1 in Figure 4 we indeed learn that our biometric recognition performance reaches a stage where it plateaus (after $\tilde{8}0$ trials), and that surprisingly there is only a weak positive difference between performance at the start and end of day two. Suggesting that even on relatively simple keystroke tasks, it takes a long time for

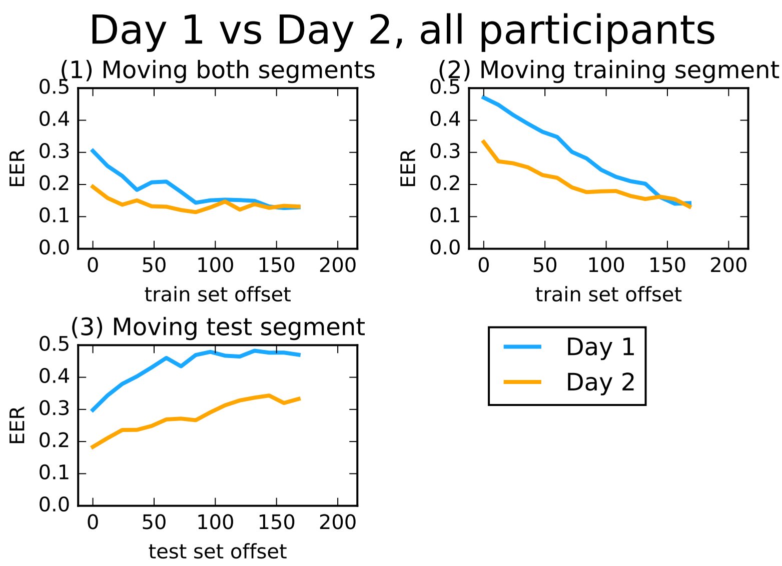


Figure 3: This figure shows the development of recognition performance, expressed in EERs, for the three different strategies for day 1 and day 2

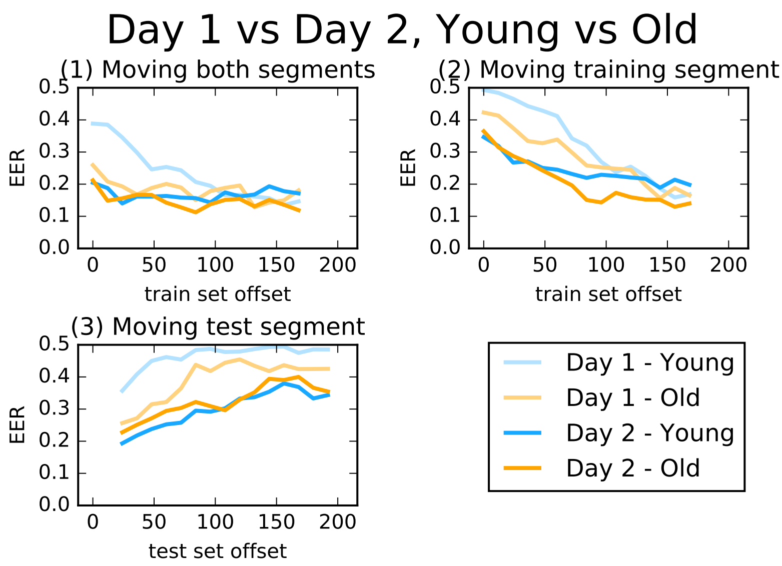


Figure 4: This figure shows an additional split between young and old age groups, as opposed to Figure 4

the recognition performance to plateau, more than is usually allowed in behavioural biometric validation studies. Such validation studies thus likely systematically under-report the eventual performance of their biometric system.

One caveat that needs mentioning is that participants in [3] did many trials of their task in a small amount of sessions (2), whereas a typical use-case for a behavioural biometric system has few trials in a large amount of sessions. Given the same environments, the typical use-case thus has to deal with a lack of task focus, whereas small amounts of longer sessions are more likely to deal with hyper-focus on the task, and fatigue. How these effects compare is unknown in behavioural biometric literature[19].

Taking our moving segments strategy as indicative of when the plateau stage sets in, our moving training segment and moving test segment allowed us to test whether behaviour still changed even after the plateau stage was hit. We find big differences in recognition performance between the start and end of the plateau stage, for both day 1 and 2 in our moving training segment. A similar pattern holds when we train our classifier on the earliest samples, and test on samples increasingly further away. In fact, our performance at the end of the day is no better than random guessing. This means that even though we find our recognition performance to stabilize eventually, underlying patterns are likely to keep changing for a long time.

Related to the timescales over which behaviour has not stabilized, there is evidence of motor chunks, and thus behaviour, changing for a very long time (much longer than our study ran for) in primates and humans alike[2]. As is also shown in[2], in different stages of practice (novice, intermediate, experts), different types of correlations (features) become more and less predictive of chunking structures, to point where some features can lose all of their predictive value. This indicates that we will need to incorporate such long-term behavioural changes into our classifier, and be very particular about training on- and weighting features as time goes by.

Interestingly, the poor performance in early stages of day 1 for strategy 1 can mostly be explained by even poorer performance for YA when compared to the whole group. Recognition performance for OA was relatively stable in strategy 1, compared to the dramatic increase in YA. On the other hand, strategy 2 and 3 show similar results between YA and OA. It is likely that OA will more quickly show acceptable performance when the practice phase is short, where YA showed very low recognition performance for a long time. This suggests that it is likely that validation studies that mainly focus on one of these groups (YA or OA) will need to adjust their length to take these major differences into account.

## Conclusion

In this paper we set out to better understand the effects of time and practice on behavioural biometric recognition performance. First, we presented a novel method of analyzing existing (psychological) data sets in a biometric context. Second, we investigated the relationship between practice and biometric recognition performance. We found that, as hypothesized, recognition performance eventually plateaus. However, we also found that this plateau is hit only after a somewhat large amount of practice, more than what most validation studies allow. Also, we found evidence that behaviour patterns keep changing even after a recognition plateau is hit, which is in line with what is known in motor sequence learning as well[2]. Third, we investigated the difference between in recognition performance trends for young and older adults. We found evidence that recognition performance develops differently between young and older adult groups. With sufficient practice, this does not result in differences in recognition performance, but initial performance for young adults takes much longer to "plateau" than for old adults.

We believe these results warrant some novel weak advisory warnings on the design of validation studies. First, in order to avoid under-reporting performance of a behavioural biometric system, one should allow participants to practice untill they hit such a plateau stage. Second, training a classifier on early samples without dropping any of the earliest samples will result in unnecessarily rapid performance degradation, which can easily be mitigated. Third, different participants of different age groups will need different amounts of practice before we can say behavioural patterns have stabilized enough, with older adults needing less practice than younger adults.

We believe re-analyses such as these of existing data sets provide insights of interest to warrant further similar analyses. Also, based on [2], we suggest a deeper biometric feature-level study about how classifiers learn which features are discriminative of participants, and how the features that are used change. They might provide both interesting insights into the theory between motor sequence learning, and help us to build classifiers that are more robust to changes due to practice.

# References


[1] Abrahamse, E. L., Ruitenberg, M. F., De Kleine, E., and Verwey, W. B. Control of automated behavior: insights from the discrete sequence production task.

[2] Acuna, D. E., Wymbs, N. F., Reynolds, C. A., Picard, N., Turner, R. S., Strick, P. L., Grafton, S. T., and Kording, K. P. Multifaceted aspects of chunking enable robust algorithms. *Journal of neurophysiology 112*, 8 (2014), 1849–1856.

[3] Barnhoorn, J. S., Van Asseldonk, E. H. F., and Verwey, W. B. Differences in chunking behavior between young and older adults diminish with extended practice. Unpublished.

[4] Bergadano, F., Gunetti, D., and Picardi, C. User authentication through keystroke dynamics. *ACM Transactions on Information and System Security (TISSEC) 5*, 4 (2002), 367–397.

[5] Buriro, A., Crispo, B., Delfrari, F., and Wrona, K. Hold and sign: A novel behavioral biometrics for smartphone user authentication. In *Security and Privacy Workshops (SPW), 2016 IEEE* (2016), IEEE, pp. 276–285.

[6] de Wilde, L., Spreeuwers, L., and Veldhuis, R. Exploring how user routine affects the recognition performance of a lock pattern. In *Biometrics Special Interest Group (BIOSIG), 2015 International Conference of the* (2015), IEEE, pp. 1–8.

[7] Donner, Y., and Hardy, J. L. Piecewise power laws in individual learning curves. *Psychonomic bulletin & review 22*, 5 (2015), 1308–1319.

[8] Ericsson, K. A., Krampe, R. T., and Tesch-Römer, C. The role of deliberate practice in the acquisition of expert performance. *Psychological review 100*, 3 (1993), 363.

[9] Gaines, R. S., Lisowski, W., Press, S. J., and Shapiro, N. Authentication by keystroke timing: Some preliminary results. Tech. rep., DTIC Document, 1980.

[10] Galton, F. *Hereditary genius: An inquiry into its laws and consequences*, vol. 27. Macmillan, 1869.

[11] Heathcote, A., Brown, S., and Mewhort, D. J. The power law repealed: The case for an exponential law of practice. *Psychonomic bulletin & review 7*, 2 (2000), 185–207.

[12] Herzfeld, D. J., and Shadmehr, R. Motor variability is not noise, but grist for the learning mill. *nature neuroscience 17*, 2 (2014), 149–150.

[13] Idrus, S. Z. S., Cherrier, E., Rosenberger, C., Mondal, S., and Bours, P. Keystroke dynamics performance enhancement with soft biometrics. In *Identity, Security and Behavior Analysis (ISBA), 2015 IEEE International Conference on* (2015), IEEE, pp. 1–7.

[14] Kambourakis, G., Damopoulos, D., Papamartzivanos, D., and Pavlidakis, E. Introducing touchstroke: keystroke-based authentication system for smartphones. *Security and Communication Networks 9*, 6 (2016), 542–554.

[15] Khan, H., Atwater, A., and Hengartner, U. Itus: an implicit authentication framework for android. In *Proceedings of the 20th annual international conference on Mobile computing and networking* (2014), ACM, pp. 507–518.

[16] Newell, A., Rosenbloom, P. S., and Anderson, J. Mechanisms of skill acquisition and the law of practice. *Cognitive skills and their acquisition 1* (1981), 1–55.

[17] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[18] Ramkumar, P., Acuna, D. E., Berniker, M., Grafton, S. T., Turner, R. S., and Kording, K. P. Chunking as the result of an efficiency computation trade-off. *Nature Communications 7* (2016).

[19] Revett, K. *Behavioral biometrics: a remote access approach.* John Wiley & Sons, 2008.

[20] Sae-Bae, N., Memon, N., Isbister, K., and Ahmed, K. Multitouch gesture-based authentication. *IEEE transactions on information forensics and security 9*, 4 (2014), 568–582.

[21] Stafford, T., and Dewar, M. Tracing the trajectory of skill learning with a very large sample of online game players. *Psychological Science 25*, 2 (2014), 511–518.

[22] Sutton, R. S., and Barto, A. G. *Reinforcement learning: An introduction*, vol. 1. MIT press Cambridge, 1998.

[23] Umphress, D., and Williams, G. Identity verification through keyboard characteristics. *International journal of man-machine studies 23*, 3 (1985), 263–273.

[24] Verwey, W. B. Concatenating familiar movement sequences: The versatile cognitive processor. *Acta psychologica 106*, 1 (2001), 69–95.

[25] Xu, H., Zhou, Y., and Lyu, M. R. Towards continuous and passive authentication via touch biometrics: An experimental study on smartphones. In *Symposium On Usable Privacy and Security, SOUPS* (2014), vol. 14, pp. 187–198.

[26] Zheng, N., Bai, K., Huang, H., and Wang, H. You are how you touch: User verification on smartphones via tapping behaviors. In *Network Protocols (ICNP), 2014 IEEE 22nd International Conference on* (2014), IEEE, pp. 221–232.