Identifying writing tasks using sequences of keystrokes

Rianne Conijn m.a.conijn@tilburguniversity.edu
Menno van Zaanen mvzaanen@tilburguniversity.edu
Department of Communication and Information Sciences, Tilburg University, The Netherlands

Keywords: keystroke analysis, classification, writing processes, writer identification

Abstract

The sequences of keystrokes that are generated when writing texts contain information about the writer as well as the writing task and cognitive aspects of the writing process. Much research has been conducted in the area of writer identification. However, research on the analysis of writing processes based on sequences of keystrokes has received only a limited amount of attention. Therefore, in this study we try to identify properties of keystrokes that indicate cognitive load of the writing process. Moreover, we investigate the influence of these properties on the classification of texts written during two different writing tasks: copying a text and free-form generation of text. We show that we can identify properties that allow for the correct classification of writing tasks, which at the same time do not describe writer-specific characteristics. However, some properties are the result of an interaction between the typing characteristics of the writer and the writing task.

1. Introduction

Students' activities in online learning systems can provide useful information about their learning behavior. Educational data mining focuses on the use of data from learners and their context to better understand how students learn, to improve educational outcomes, and to gain insight into and explain educational phenomena (Romero & Ventura, 2013). Data can be collected from different sources, such as online learning systems, student administration, and questionnaires. This results in data from multiple contexts, over different time periods, ranging from low to high granularity. In this study, we analyze fine-grained data: keystroke data from a writing task.

In the literature, two different goals can be distinguished in the analyses of keystroke data: authentication or identification of writers, and determination of writing processes. Keystrokes have mainly been used for the former (Longi et al., 2015; Karnan et al., 2011). In the field of educational data mining, the authentication and identification of writers is used, for example, for authentication in online exams (Gunetti & Picardi, 2005) or for the identification of programmers (Longi et al., 2015). These studies mainly focus on the typing or motor processes, since these are considered unique per person. The majority of studies focus on statistical properties, such as mean, standard deviation, and Euclidean distance of three attributes: keystroke duration, keystroke latency, and digraph duration (Karnan et al., 2011). These features can be used to identify and authenticate writers to a large extent, with accuracies up to 99% (Tappert et al., 2009). These high accuracies show that keystroke logs contain much information that denotes writer-specific characteristics.

Yet, keystroke data also includes other information, denoting the writing process itself. The determination of these writing processes has received less attention. This might be due to the fact that keystrokes are not clear measures of the underlying writing processes (Baaijen et al., 2012). The data need to be preprocessed and analyzed in a way such that it provides meaningful information to be used by students and teachers for improving learning and teaching. Therefore, this study explores the writing processes derived from students' keystrokes.

Some studies already investigated the possibilities of determining writing processes using keystrokes. Baaijen et al. (2012) analyzed keystroke data from 80 participants during a 30-minute writing task. The relation between pauses, bursts, and revisions were analyzed. Using these features, text production could be distinguished from revisions. Revision bursts were shorter than new text production bursts. In another writing task, keystroke data from 44 students during a 10-minute essay was collected to determine emotional states (Bixler & D'Mello, 2013). Four feature sets were used: total time, keystroke verbosity (number of keys and backspaces), keystroke latency, and number of pauses (categorized by length). All feature sets combined could classify boredom versus engagement with an accuracy of 87%. Keystroke data have also been analyzed in programming tasks, to determine performance. Thomas et al. (2005) analyzed keystroke data from 38 experienced programmers and 141 novices in a programming task. Keystroke latencies and key types were found related to performance. Key latencies (within and before or after a word) were found negatively correlated with performance. Additionally, it was found that experienced programmers used more browsing keys and were faster in pressing those.

These studies show that keystrokes do not only differ due to writer-specific characteristics (which is used in authentication and identification), but also because of differences in revisions and text production, emotional states, and level of experience. Whereas the differences in writer-specific properties may be due to physical differences and differences in typing style, the differences in writing properties are expected to come from differences in cognitive processes required. Indeed, keystroke duration and keystroke latencies are often seen as an indicator of cognitive load (Leijten & Van Waes, 2013). As different tasks lead to differences in cognitive load, we may find these differences using different writing tasks. However, existing studies do not compare differences in keystrokes between tasks. Therefore, in the current study, the writing processes in two different tasks are compared: writing a free-form text versus a fixed text (copying a text). Here we assume that writing a free-from text requires a different cognitive load than writing a fixed text, resulting in differences in the keystroke data.

Having knowledge of the cognitive load while producing a text may provide useful information, for example, for teachers. Currently, teachers often only have access to the final writing product for evaluation purposes. This does not provide insight in what students did during the writing process. Insight in students' writing behavior or cognitive load during an assignment may trigger the teacher to further investigate this behavior and adapt the task or provide personalized feedback on the writing process.

To identify properties of keystrokes that indicate the cognitive load of the writing process, an open dataset is used, which has been used for writer identification. In a previous study, it was already shown that keystroke data differed between free-form and fixed text (Tappert et al., 2009). However, these differences were not made explicit nor evaluated. Therefore, in the current study, we analyze which features differ within the keystrokes of free-form versus fixed text using three different feature sets. As an evaluation, the differences found between fixed and free-form text are used to classify text as being either fixed or free-form text. This is done using all possible combinations of the different feature groups, to determine which feature group is most useful for the classification. At the same time, since we are not interested in the writer-specific information, the properties should not allow for an accurate identification of the actual writer.

2. Method

2.1. Data

Data used in the current experiments has been taken from the Villani keystroke dataset (Tappert et al., 2009; Monaco et al., 2012). The Villani keystroke dataset consists of keystroke data collected from 142 participants in an experimental setting. Participants were free to chose to copy a fable, a fixed text of 652 characters, or to type a free-form text, an email of at least 650 characters. Participants could copy the fable multiple times and could also type multiple free-form texts. Since typing the texts was not mandatory, not all participants typed both a free-form text and a fixed text. In total, 36 participants typed both at least one fixed text and one free-form text, resulting in keystroke data of 338 fixed texts and 416 free-form texts. The other 106 participants only wrote either free-form or fixed texts, resulting in a further 21 fixed texts and 808 free-form texts. The keystroke data consisted of timestamps for each key press and key release and the corresponding key code. More information about the dataset and the collection of the dataset can be found in Tappert et al. (2009). In this research, we only use the data of participants who created both text types.

2.2. Data processing

First, for all keystrokes, the type of key was derived: letter, number, browse key (e.g., LEFT, HOME), punctuation key, correction key (BACKSPACE, DELETE), and other (e.g., F12). The time between a key release and the subsequent key press (keystroke latency or key pause time) was calculated. Thereafter, the type of pause between the key was derived using the key types. Six pause types were identified: pause before word (after SPACE, before letter or number), within word (between letters or numbers), after word (after letter or number, before SPACE), before correction, within correction, and after correction.

Lastly, words were identified as the letters and numbers between two SPACE keys. For all words the word length (number of letters and numbers) and the word time was calculated (time from key press of the first character until time of the key release of the last character). For further analyses on the word length and word time, only words without corrections were included. The use of corrections within a word would have a significant influence on the time of typing. Additionally, since a BACKSPACE or DELETE key can be used to remove multiple characters, it is hard to determine word length if corrections are made within the word.

Figure 1 shows the measurement of the timing of the different types of pauses. Given that the writer types the words "the book" with two incorrect letters after the SPACE key ("do"), which are corrected using two BACKSPACEs, the key presses and releases per key are illustrated in the second row. The following rows each depict which periods between key releases and key presses are measured for that type of pause. For instance, the pause before the word "the" and the pause between the SPACE key and the letter "d" are counted as pause before word type (third row). The last row indicates the timing used to compute the word length. In this case, the word "the" is identified (which has a length of three characters). The word "book" is not used, as it contains corrections.

After data enrichment, the three groups of features were computed: pause times, corrections, and word length. For all six different types of pause times (see Figure 1), the normalized average pause times were calculated by dividing the average pause time of each type by the overall average pause time over all types. Additionally, the normalized standard deviations of the pause times were calculated. In total, this resulted in 12 different features related to pause time. For the corrections, two features were calculated: the total number of corrections and the percentage of words with corrections. Lastly, four features related to word length were computed: the average time and standard deviation for short words (having less than four characters) and the average time and standard deviation for long words (consisting of between 9 and 13 characters). All four features were normalized using the average time and standard deviation of all words.

Obviously, the keystroke sequences contain much more

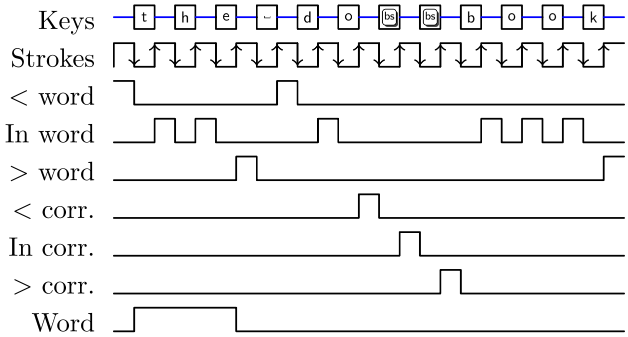

Figure 1. Measurement of timing of pauses of "the book" with two corrections using the backspace key (bs). "<" means before, ">" means after, "corr." stands for correction. The last row indicates the time of the word "the" (of length three characters). The word "book" is not used, as it contains corrections.

information than what we extracted here. The selection of these features was made in order to reveal as little as possible about the actual text being typed. Actual key code information, for instance, is not used as that information should be quite consistent between the fixed texts.

For the statistical analyses and training of the models, only data from the 36 participants who typed both fixed and free-form texts were included. From these, four texts were excluded, because they consisted of less than five words and inspection showed that these texts were random key strokes. Thereafter, 750 texts (338 fixed and 412 free-form) remained for analyses.

2.3. Analyses

To identify the relationship between keystroke information and cognitive load in writing tasks, two types of analyses were used: statistical analyses and model evaluation. First, paired t-tests were conducted to determine whether differences were found between the features in the fixed and free-form texts of participants.

Thereafter, support vector machines were trained to classify texts as being fixed or free-form. Support vector machines were trained for all combinations of the three feature groups (pause times, corrections, and word length), resulting in a total of seven models. The radial base function was used as kernel ('svmRadial' from the 'caret' package in R (Kuhn, 2016)). The data was trained using 10-fold cross-validation. Grid search was used during training (with a tuning part held aside from the testing) to optimize the parameters σ and cost. The average accuracy were calculated as performance measures. Since the groups were not equally distributed, the average κ was also calculated. The

Table 1. Descriptive statistics and paired t-tests of features in fixed and free-form text (N = 36). *=$p < .05$, **=$p < .01$, ***=$p < .001$.

Feature	Fixed text M	Fixed text $S.D.$	Free-form text M	Free-form text $S.D.$
Total time ***	376	100	432	129
# Keys **	703	54.6	749	57.5
# Corrections ***	21.7	13.6	35.3	16.4
% Words corrected ***	0.08	0.05	0.13	0.06
Average pause time before word **	1.17	0.18	1.25	0.18
S.D. pause time time before word *	1.04	0.25	1.16	0.28
Average pause time within word ***	0.86	0.09	0.78	0.07
S.D. pause time within word ***	0.60	0.21	0.40	0.15
Average pause time after word **	0.91	0.15	0.84	0.17
S.D. pause time after word **	0.79	0.29	0.61	0.30
Average pause time before correction	2.49	0.76	2.40	1.52
S.D. pause time before correction *	1.50	0.55	1.22	0.51
Average pause time within correction	0.95	0.61	0.82	0.16
S.D. pause time within correction	0.48	0.30	0.44	0.24
Average pause time after correction	1.75	1.60	2.16	3.29
S.D. pause time after correction	1.14	0.29	1.03	0.30
Average short word time **	0.52	0.05	0.48	0.09
S.D. short word time	0.33	0.09	0.32	0.11
Average long word time	2.87	0.26	2.98	0.71
S.D. long word time	1.14	0.23	1.16	0.28

κ corrects for random guessing, by comparing the observed accuracy with the expected accuracy (chance):

$$\kappa = \frac{observed\ accuracy - expected\ accuracy}{1 - expected\ accuracy}$$

Additionally, a one-way ANOVA with Tukey post-hoc test was used to determine whether the models differed significantly in accuracy.

Lastly, since we focus on the writing process, the learned models should preferably not be able to classify personal writer-specific characteristics. Thus, the learned model should perform really badly when classifying writers. Therefore, as an additional evaluation support vector machines were trained to classify the writers. The best σ and cost values from the models classifying fixed versus free-form text were used. Again, the average accuracy and κs were calculated using the same folds in 10-fold cross-validation.

3. Features measuring cognitive load

Paired t-tests were used to determine which features differed significantly between fixed and free-form text created by the same writer. This is assumed to provide insight in which features are indicative of cognitive load. The results can be found in Table 1. Note that we use both the mean as well as the standard deviation (S.D.) within a text as features (and both types of features have their own standard deviations per document). In the table, the descriptive statistics of these features can also be found.

It was found that fixed texts consisted of significantly fewer keystrokes compared to free-form texts (703 versus 749). Although the fixed text consisted of 652 characters, the mean number of keystrokes was 703. This can partly be explained by the fact that sometimes multiple keys are needed to type one character (e.g., SHIFT + character to type a capital letter). Additionally, this can indicate that the participants made typos and fixed those, requiring BACKSPACE or DELETE keystrokes. Indeed, it was shown that corrections were made in 740 of the 750 sessions. The free-form texts contained more corrections and a higher percentage of words with at least one correction, compared to the fixed texts. Lastly, the participants were faster in typing the fixed text compared to the free-form text. All these findings provide some evidence that typing the free-form text requires a different cognitive load.

Several features were analyzed to determine where significant differences in pause duration between the text types were found. Specifically, we investigated the differences between the pauses before, after, and within words and corrections. Since the free-form and fixed texts differed in total length and time, timing of key pauses were normalized based on the average time per key pause in that session. It was found that writers spend more of their writing time on pauses before a word and less time on pauses within a word or after a word in free-form text, compared to fixed texts. Additionally, the standard deviation of pauses within and after a word was significantly lower for free-form texts compared to fixed texts. For the pauses before words, the opposite was found: when typing free-form text, a larger proportion of pause time was spent before a word, compared to fixed texts. Moreover, the standard deviations were larger. This may be because the writer will need to think (longer) about which word to type in free-form text, which is not needed for fixed texts.

For the key pause times before, after, and within corrections, no significant differences were found between free-form and fixed texts. The only exception is the standard deviation of key pause time before corrections: free-form texts lead to a larger standard deviation for key pause time before corrections compared to fixed texts.

When comparing the average word time between the two types of text, participants were faster in typing short words (consisting of less than four letters) in free-form text compared to fixed text. This indicates that in free-form text, of all words, less time is devoted to short words, compared to fixed text. No significant differences were found between fixed and free-form texts for the average word time for long words (8–13 letters). Additionally, no significant differences were found in the standard deviation of time on short and long words between the text types.

4. Model evaluation

4.1. Classifying fixed versus free-form text

To measure the effect of the different groups of features, we trained support vector machines and measured how well they could distinguish between a fixed and a free-form text. The models were trained using all combinations of three different feature groups: average and standard deviations of the pause times (Pauses); correction information (Correction); and average and standard deviation of short and long word typing time (Words). The accuracies and κs of all seven models for classifying fixed and free-form text can be found in Table 2.

Table 2. Accuracies and κs of support vector machine models on the different feature groups that classify fixed versus free-form text.

Feature group	Fixed vs. free Accuracy	κ
Pauses	0.739	0.465
Correction	0.689	0.368
Words	0.687	0.370
Pauses, Correction	0.763	0.513
Correction, Words	0.767	0.522
Pauses, Words	0.739	0.465
Pauses, Correction, Words	0.781	0.551

Table 3. Accuracies and κs of support vector machine models on the different feature groups that classify writers (36 classes).

Feature group	Writer Accuracy	κ
Pauses	0.248	0.223
Correction	0.091	0.065
Words	0.073	0.046
Pauses, Correction	0.311	0.287
Correction, Words	0.121	0.095
Pauses, Words	0.239	0.215
Pauses, Correction, Words	0.291	0.267

The results show that all feature groups are useful for the classification of fixed versus free-form text. The feature group of the key pause times led to best accuracy (73.9% with a κ of 0.465, or approximately 47% above chance), compared to the other individual feature groups. Not surprisingly, the combination of all feature groups yielded the overall highest accuracy: 78.1% with a κ of 0.551. A one-way ANOVA showed that the seven models differed significantly in accuracy ($F(6, 63) = 4.728, p < .001$). Using two feature groups was always better than using only correction features or word length features. However, the combination of all feature groups did not lead to a significantly higher accuracy than the pause time features alone. Thus, the word length and the corrections features did not have much additional value next to the pause time features for classifying fixed versus free-form text.

4.2. Classifying writers

To determine whether the learned models did not include any writer-specific characteristics, models with the same settings as the models that classify text types were trained and tested to classify writers. The results of these experiments can be found in Table 3. Similarly to the models classifying fixed versus free-format text, the key pause time features led to a higher accuracy (24.8% and $\kappa = 0.223$) than the correction and word length features. The correction and word length features resulted in the lowest accuracies: 9.1% and

7.3%, respectively. The model with both correction and pause time features led to the highest accuracy: 31.1% with a κ of 0.267. Although this is a reasonably low accuracy, the model clearly outperforms the model based on chance for the 36-class classification ($1/36 = 2.8\%$). This means that some writer-specific characteristics are encoded within the feature groups that are used in these models, which in this case is unwanted.

A one-way ANOVA showed that the seven models differed significantly in accuracy ($F(6, 63) = 37.43, p < .001$). Interestingly, the accuracies for the corrections and words feature groups combined are significantly lower than the accuracy of all models that include pause time features. This indicates that the corrections and word length features include fewer writer-specific characteristics than the pause time features.

5. Discussion

Keystroke data include both writer-specific information and information about the writing processes. In this study, we focused on the writing processes and aimed to identify properties of keystrokes that indicate the cognitive load of the writing process. In order to do this, keystrokes of two different writing tasks were analyzed, which are assumed to differ in cognitive load: copying a text (fixed text) and writing a free-form text.

Our first analysis showed that several features extracted from the keystroke data differed significantly between the fixed and free-form texts of a writer. These findings support previous work which showed that keystrokes differ for different (types of) text entered (Gunetti & Picardi, 2005; Tappert et al., 2009). As an extension, we also identified which features differed and how these differed. When typing free-form text, the pauses before a word were longer, while the pauses within or after a word were shorter, compared to typing fixed text. This might indicate that the participants were thinking about the next word to type in the free-form text before they typed the word, while writers in the fixed text situation could immediately copy it as it was provided for them. Thus, differences in cognitive load may be identified in the pauses before words.

As an evaluation, we showed that the differences in keystroke information can be used to classify fixed and free-form text. Using a support vector machine, the key pause time features (which measure time spent between key releases and key presses) were found to lead to the highest accuracies for the text identification task. Adding the corrections and word length features did not lead to significantly higher accuracies, showing that the word length and corrections features do not add much information in addition to the pause times. When all feature groups were included, 78.1% accuracy was reached. Although this accuracy is reasonably high, it also shows there is still some room for improvement. Especially considering that classifying writers, being a more complex classification problem with more classes, have shown accuracies up to 99% (Tappert et al., 2009).

Since we aimed to identify features related to the writing process, we wanted to exclude writer-specific information. In other words, the models should perform badly when classifying the writer. To test this, we tried to classify the writers with the same settings as used in the writing task classification for the support vector machine models. The lowest accuracy, while using information from the keystroke sequences were found when using word length features only (7.3%). This corresponds to an accuracy of 68.7% on the text type classification task. The highest accuracy (31.1%) was obtained with both pause time and correction features (corresponding to 76.3% accuracy on the text type classification task, which is close to the highest accuracy on that task: 78.1%). Even though the accuracies on writer classification are higher than chance, it is much lower than the 90%–99% accuracy reached in other studies (Longi et al., 2015; Tappert et al., 2009). Thus, the feature groups that have been extracted, actually contain mostly information related to the writing task and not to the writer-specific characteristics.

Interestingly, especially the corrections and word length features showed low accuracies on classifying writers. Thus, these feature groups contained little information about individual typing characteristics. Adding additional information to improve the quality of the text type classification task, also increases accuracy of the writer classification task. For example, if we add the key pause features, the accuracy of the text type task increases, but the writer identification accuracies also increase. In other words, key pause properties contain useful information for the text type classification task, but also contain information that allows for the identification of the writer, which is unwanted in this case.

There are at least three directions for future work. Firstly, future work could try to improve the accuracy on task classification, while not improving the accuracy on writer identification. Additional features or feature groups could be identified, such as bursts ending in a pause (P-bursts) or ending in a revision (R-bursts) (Baaijen et al., 2012). In addition, the key pause feature group seems to contain useful information, but these features should be modified in order to remove any writer-specific properties. Alternatively, other machine learning algorithms, such as neural networks, may be tried to achieve higher accuracies.

Secondly, a wider range of writing tasks could be considered. For example, semi-fixed tasks with specific task descriptions (e.g., writing a sorting algorithm) can be investigated to determine whether differences between tasks that are more similar can also be distinguished. In this way, we may identify which tasks require more cognitive load and in which properties of the process of typing this effort can be found. This information can be used to improve the writing task instruction or to provide feedback on the writing process to the learner during the writing task (see also Poncin et al., 2011; Kiesmueller et al., 2010).

Lastly, this study assumed that the differences in keystrokes provide an indication of cognitive load. However, we did not actually measure the cognitive load. Future work could explicitly measure the cognitive load during the task, for example by using a secondary task, or a questionnaire (Paas et al., 2003). In this way, the problem could be approached as a regression problem rather than a classification task.

6. Conclusion

To conclude, this research has shown that keystroke data can be used to identify differences in writing tasks, which we believe require different cognitive load. Additionally, we showed which feature groups (key pause, correction, and word length) have an influence on the performance. In particular, the word length and correction feature groups led to a good performance on the writing task classification task and a low performance on the writer identification task. The key pause feature group increases the performance on the text classification task, but the performance on the writer classification task also increases (which is unwanted, as the key pause features also include writer-specific properties).

Having insight in these features which identify writing processes and cognitive load can be useful for improving learning and teaching. For example, in this way, teachers can also get insight in the writing process, instead of only the product of writing. This can be useful for adapting the course materials or providing (personalized) feedback.

References

Baaijen, V. M., Galbraith, D., & de Glopper, K. (2012). Keystroke analysis: Reflections on procedures and measures. *Written Communication*, *29*, 246–277.

Bixler, R., & D'Mello, S. (2013). Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits. *Proceedings of the 2013 international conference on Intelligent user interfaces* (pp. 225–234).

Gunetti, D., & Picardi, C. (2005). Keystroke analysis of free text. *ACM Transactions on Information and System Security (TISSEC)*, *8*, 312–347.

Karnan, M., Akila, M., & Krishnaraj, N. (2011). Biometric personal authentication using keystroke dynamics: A review. *Applied Soft Computing*, *11*, 1565–1573.

Kiesmueller, U., Sossalla, S., Brinda, T., & Riedhammer, K. (2010). Online identification of learner problem solving strategies using pattern recognition methods. *Proceedings of the fifteenth annual conference on Innovation and technology in computer science education* (pp. 274–278).

Kuhn, M. (2016). *caret: Classification and regression training*. R package version 6.0-73.

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using inputlog to analyze and visualize writing processes. *Written Communication*, *30*, 358–392.

Longi, K., Leinonen, J., Nygren, H., Salmi, J., Klami, A., & Vihavainen, A. (2015). Identification of programmers from typing patterns. *Proceedings of the 15th Koli Calling Conference on Computing Education Research* (pp. 60–67).

Monaco, J. V., Bakelman, N., Cha, S.-H., & Tappert, C. C. (2012). Developing a keystroke biometric system for continual authentication of computer users. *Intelligence and Security Informatics Conference (EISIC), 2012 European* (pp. 210–216).

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist*, *38*, 63–71.

Poncin, W., Serebrenik, A., & van den Brand, M. (2011). Mining student capstone projects with frasr and prom. *Proceedings of the ACM international* *conference companion on Object oriented programming systems languages and applications companion* (pp. 87–96).

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*, 12–27.

Tappert, C. C., Villani, M., & Cha, S.-H. (2009). Keystroke biometric identification and authentication on long-text input. *Behavioral biometrics for human identification: Intelligent applications*, 342–367.

Thomas, R. C., Karahasanovic, A., & Kennedy, G. E. (2005). An investigation into keystroke latency metrics as an indicator of programming performance. *Proceedings of the 7th Australasian conference on Computing education-Volume 42* (pp. 127–134).