Measuring and Assessing Typing Skills in Writing Research

Luuk Van Waes[a], Mariëlle Leijten[a], Jens Roeser[b], Thierry Olive[cd] & Joachim Grabowski[e]

[a] Department of Management, University of Antwerp | Belgium

[b] Department of Psychology, Nottingham Trent University | United Kingdom

[c] Research Centre on Cognition & Learning, CNRS & University of Poitiers | France

[d] Maison des Sciences de l'Homme et de la Société, CNRS & University of Poitiers | France

[e] Institute for Psychology, Leibniz University Hannover | Germany

Abstract: In keyboard writing, typing skills are considered an important prerequisite of proficient text production. We describe the design, implementation, and application of a standardized copy-typing task in order to measure and assess individual typing fluency. A test-retest analysis indicates the instrument's reliability.

While the task has been developed across eleven different languages and the related keyboard layouts, we here refer to a corpus of Dutch copy tasks (N = 1682). Analyses show that copying speed non-linearly varies with age. Bayesian analyses reveal differences in the typing performance and the underlying distributions of inter-key intervals between the different task components (e.g., lexical vs. non-lexical materials; high-frequent vs. low-frequent bigrams).

Based on these findings it is strongly recommended to include copy-task measures in the analysis of keystroke logging data in writing studies. This supports a better comparability and interpretability of keystroke data from more complex or communicatively-embedded writing tasks across individuals. Further potential applications of the copy task for writing research are explained and discussed.

Keywords: copy task; typing skills; writing processes; writing fluency; transcription processes; motor skills

journal of WRITING RESEARCH

Van Waes, L., Leijten, M., Roeser, J. Olive, T., & Grabowski, J. (2021). Measuring and assessing typing skills in writing research. *Journal of Writing Research, 13*(1), 107-153 https://doi.org/10.17239/jowr-2021.13.01.04

Contact: Luuk Van Waes, University of Antwerp, Faculty of Business Economics, Department of Management, Prinsstraat 13, 2000 Antwerp | Belgium - Orcid: 0000-0002-3642-9533 luuk.vanwaes@uantwerpen.be

1. Introduction

Since the introduction of the personal computer in the 1980s, typing has gradually taken a more prominent place in text composition. Both in personal and professional contexts typing has become the most important text production mode (Brandt, 2014). The importance and the impact of typing fluency on writing performance has been demonstrated in many studies (e.g., Aldridge & Fontaine, 2019; Johansson, Wengelin, Johansson, & Holmqvist, 2010; Van Weerdenburg, Tesselhof, & Van der Meijden, 2019; Weigelt-Marom & Weintraub, 2018). These researchers contend that writers with a certain level of keyboarding automatization are more productive, in line with the findings in studies on graphomotor automatization in handwriting (see, e.g., Limpo, Vigário, Rocha, & Graham, 2020). More fluent typing ability allows writers to reduce the cognitive cost that is related to the motor component of text production, freeing up the writer's attention to focus on other writing components (see also Alves, Castro, de Sousa, & Strömqvist, 2007). When typing skills are automatized, motor skills only minimally affect higher cognitive levels of writing processes, allowing for cascading activation of other processes (like planning or revision). All these writing components compete for the same working memory capacity and interact with each other (Kellogg, 2001). Therefore, it is not surprising that younger students (from grades 4, 5 and 6) who participated in a touch-typing course produced narrative texts with a higher quality compared to their non-trained peers (Van Weerdenburg et al., 2019). This finding was also supported by Tate, Warschauer and Kim (2019), who reported a positive effect of keyboarding fluency – and prior computer use – on writing quality for grade 8 students.

As most writers have shifted to digital composition, writing researchers have shifted their methods accordingly. A good example is the use of keystroke logging to observe writing processes. In recent years, keystroke logging has become one of the mainstream research methods in writing studies (Lindgren & Sullivan, 2019). Keystroke logging records every keystroke and mouse click or movement related to the production of the text. These logging data are time coded, making it possible to exactly reconstruct the writing process and analyze the writing-process dynamics from different perspectives (Leijten & Van Waes, 2013). Because keystroke logging is unobtrusive, it is suitable for both ethnographic and experimental writing studies.

Pausing behavior is one of the main cognitive indicators used in keystroke logging research (Wengelin, 2006). Shorter latencies of finger movements between subsequent keys are often referred to as 'interkey intervals' (IKIs) or 'interkey latencies'; longer latencies as 'pauses'; long latencies during the logged process that are not directly related to writing, e.g., answering a phone call, are referred to as 'downtime' (see Figure 1). Latency values below 30 ms could be considered as 'noise' as they normally do not relate to deliberate typing actions. However, the

alignment between keystroke logging data and cognitive models of writing is not straightforward (Galbraith & Baaijen, 2019). Cognitive processes underlying pausing are often quite difficult to interpret (Chukharev-Hudilainen, 2014; Conijn, Roeser, & Van Zaanen, 2019; Wengelin, 2006). One of the reasons is the large variability in writers' typing proficiency (see below). To distinguish keystroke-transition durations that are associated with low levels of activation from those associated with higher levels of activation, researchers use a fixed pause threshold: often 2 seconds, sometimes 500 ms (e.g., Aldridge & Fontaine, 2019; Chukharev-Hudilainen, 2014), or – when, focusing on higher level processes – even 5 or 10 seconds (e.g., Schumacher, Klare, Cronin, & Moses, 1984). However, using thresholds leads consequently to a different 'filtering' of pauses or interkey latencies, which makes comparisons of research results difficult. This is, for instance, shown in fluency analyses using a gradually increasing threshold of 200, 500, 1000 and 2000 ms (Van Waes & Leijten, 2015). In their studies, Medimorec and Risko (2016, 2017) and Allen et al. (2016) explored functions of pauses (or non-scribal periods). They set different thresholds to define discrete time intervals (e.g., 300–999, 1000–1999, and >2000 ms). These approaches bear the problem, however, that such thresholds do not adequately reflect individual differences in typing skills.

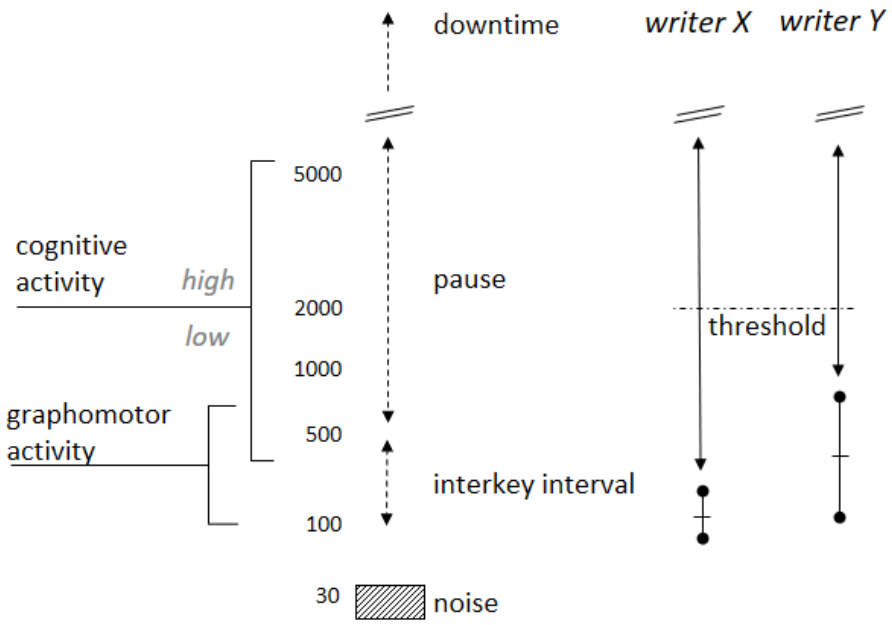

Figure 1: Schematic representation of latency in relation with low- and high-level cognitive writing processes, including the comparison of two writers' cases.

For instance, when one writer's optimal interkey latency is measured around 120 ms (see Figure 1; writer X) and another's is around 340 ms (writer Y), the threshold filtering – whatever measure chosen – affects the pause analysis differently. Especially when we are interested in lower level cognitive activities, we contend that it is important to consider interparticipant differences in typing skill when conducting pause analyses.

Taking these considerations into account, it is surprising that few researchers have developed instruments to adequately measure typing skills (Weigelt-Marom & Weintraub, 2018). A wide variety of tests exist in the context of typing courses, but most of them are limited to rather rough measurements such as words or characters per minute (often combined with an accuracy percentage). Moreover, most of these tests are based on product measurements (final product based on a timed task). As typing competence is a layered concept (Van Waes, Leijten, Mariën, & Engelborghs, 2017; Roeser, De Maeyer, Leijten, & Van Waes, 2021), we explored approaches to define the latency range that underlies the writer's typing competence in more detail.

To our knowledge, one of the few process-oriented studies on keyboard-copying abilities are those by Grabowski and colleagues (Grabowski, 2008; Wallot & Grabowski, 2013) in which they developed a set of three different tasks with varying complexity and cognitive demands. Their copy task consisted of three subtasks. In the first task, participants were shown a printed text with the first line of a well-known German nursery rhyme. They had to copy this line twelve times from memory. The second task consisted of a text with 156 words which had to be copied from paper. Finally, in the third task participants had to write a description of the route from where they live to university. Only the third task required planning components; the first two task were designed to eliminate planning and formulation processes as far as possible.

The copy task introduced in this article includes five components addressing different aspects of typing proficiency. Differences between tasks are based on the degrees of lexicality, bigram frequency, and keyboard spread (left-right, adjacency, repetitiveness; see Section 4.1 and Table 1 for an overview). To standardize this copy task across different languages, parallel forms with corresponding task characteristics were developed for eleven languages.

This paper describes the design principles of our copy task in detail and, more specifically, the choices that have been made in designing a multi-component copy task. First, however, we briefly review the theoretical framework in which we situate typing skills in the context of (cognitive) writing-process research. This is necessary to inform the methodological significance of copy-task assessment for writing process research that we will propose subsequent to the findings of the reported

study. In the second part of this paper, we present results based on a corpus analysis of copy tasks collected from 1682 Dutch participants (i.e. about 1 million keystrokes). We report on different analyses that each address a perspective characterizing an aspect of the underlying corpus. For instance, age-related performance, but also differences between the different components the copy task consists of (e.g., with respect to frequency and lexicality; test-retest comparisons). These analyses aim at contributing to our understanding of how biomechanical constraints and linguistic text characteristics interact during the writing process. Moreover, we used advanced statistical methods that are novel in this domain, i.e. Bayesian statistical inference and in particular mixture modeling. By introducing these techniques, we aim to illustrate the suitability of this type of data analysis in writing research in general, and in copy-task studies in particular. Finally, in the closing discussion section, we discuss some theoretical, methodological, and practical implications of a standardized copy task and explain potential applications in various fields of keystroke logging and writing research.

1.1 Typing skills in the context of text production

Keystroke logging research mainly resorts to measuring interkey intervals or pauses to analyze and further comprehend writing. However, pauses are difficult to interpret because they are multi-determined by the variety of cognitive processes involved during text production.

Producing a text requires several high-level central cognitive processes for planning content, formulating the linguistic surface, and for revision by assessing whether the text fits with the communicative goals, the selected linguistic register, the intended audience, etc. However, also peripheral motor processes are engaged to handwrite or type the sequences of graphemes that constitute the written trace. Once the appropriate graphemes are selected by spelling processes, they are stored into an orthographic working memory and, finally, conveyed to motor execution (Rapp, Purcell, Hillis, Capasso, & Micelli, 2016). When handwriting, motor execution requires three hierarchical steps to determine characteristics of the selected allographs (e.g., uppercase, lowercase, cursive script), the number of strokes forming each allograph, and the execution of the required movement sequences. In contrast to handwriting, typing movements are less complex (Van Galen, 1991). For instance, there is no allograph selection and no related motor programs (except for specific finger combinations to produce, for example, an uppercase letter). Although typing requires simpler movements for reaching and pressing the appropriate keys, it however requires bimanual and finger coordination for pressing key combinations.

Learning to typewrite is therefore different from learning handwriting. In contrast to novice handwriters, novice typists do not learn the motoric parameters of letters (stroke order, direction). To type a word, they encode a first letter, find

the correct key, select a finger to press that key, press the key, and encode the next letter to type, etc. This requires efforts to translate each letter to a keystroke (Yamaguchi & Logan, 2014). In contrast, skilled typists activate multiple keystrokes in parallel while using more than one finger from each hand and press the appropriate keys without looking at the keyboard (Crump & Logan, 2010). Expert typists can type up to 100 words per minutes; typing contests often see typists producing around 150 words per minute (Logan, 2018).

Learning to typewrite requires associations between (1) words and letters, (2) letters and keys, and (3) keys and fingers (Yamaguchi & Logan, 2014). Skilled typing is hierarchically controlled with an outer loop that operates on the word-level, whereas an inner loop operates on the letter- or keystroke-level (Logan & Crump, 2011). This depends on our ability to chunk information. While practicing and acquiring knowledge of words, typists chunk letters depending on their co-occurrence. In skilled typists, chunking occurs at the perceptual level, in memory, and at the motor level (Yamaguchi & Logan, 2014) and scales up units of processing from letters to words, which results in more fluent typing with bursts of key presses. Specifically, in skilled typing, the word context activates sequences of keypresses which are retrieved on-the-fly from long-term memory (Logan, 2018), as evidenced by the sensitivity of keystroke-transition durations to the co-occurrence frequency of letter combinations (i.e., bigrams) and as a function of typing proficiency. This explains why words are typed faster than non-words (Van Waes et al., 2017; Wallot & Grabowski, 2013). It is also worth noting that the intervals between successive keypresses executed with fingers of the same hand are on average slower than when the involved hands alternate (Logan; 2003; Salthouse, 1984).

Skilled typists have poor explicit knowledge of both the locations of the keys on the keyboard, and the finger they use to press a key, which is common in several areas of expertise. Their most frequent errors consist in difficulties in finger coordination and not in miss-aimed movements (Logan, 2018). Proficient or skilled typists (i.e., screen gazers) use both hands and several fingers of each hand fluently – which is called touch-typing – without looking at the key but mainly looking at the computer screen (Johansson et al., 2010) or, in the case of copying, on other adjacent materials. By contrast, novice typists (i.e., keyboard gazers or hunt-and-peck typists) alternate their attention between the keyboard for searching keys, spending less time processing their developing text. As a consequence, this might affect their unfolding mental representation of the text (Haas, 1989; Olive & Passerault, 2012).

Importantly, all writing processes compete for a common and limited working memory (Kellogg, 2001; McCutchen, 2000). Hence, low skills in typing may affect central processes because they occupy working memory resources (Bouriga & Olive, in press; Grabowski, 2010; Mangen, Anda, Oxborough, & Brønnick, 2015; Van Weerdenburg et al., 2019). As a consequence, writers with low transcription skills

encounter difficulties activating high-level writing processes during handwriting and typing. They need to suspend transcription when preparing or revising their texts (the thinking-then-writing strategy; Olive, 2014). Conversely, advanced transcription skills free up working memory resources which can be allotted to planning, formulating, and reviewing the text while transcribing (the thinking-while-writing strategy; Olive, 2014). This simultaneous coordination of central writing processes and transcription has been found in skilled handwriting (Olive & Kellogg, 2002) and in skilled typing (Alves, Castro, & Olive, 2008), resulting in longer transcription periods (Alves et al., 2007).

In sum, a less-developed typing proficiency draws resources away from the central writing processes that are responsible for preparing the text for transcription. This leads to texts of lower quality (Van Weerdenburg et al., 2019) and other difficulties in managing the writing processes. A recent meta-analysis by Feng, Libdener, Ji, and Joshi (2017) confirmed that keyboarding skills are positively associated with the development of writing for a variety of writing measures, supporting the idea that low typing skills negatively contribute to text generation. As a methodological consequence, keystroke logging studies that focus on text production need to control for individual differences in typing skills in order to maintain the focus on planning, formulating, and revising the text. One possibility to address these individual differences is to assess typing skills by asking writers to complete tasks that do not engage central writing processes. Copy tasks have been used in text production research for that purpose (Berninger et al., 1992; Grabowski, 2008). Copy tasks do not require planning, formulation, and revision and hence, central writing processes are controlled. Performance at a copy task is therefore mainly determined by the writer's typing skills (and, if necessary, short-term memory capacity to memorize a portion of the stimulus string while it is copied). Additionally, to target the different factors that influence typing processes, it is necessary to control for the materials to be copied. The copy task we present was specifically designed for that purpose.

1.2 Copy tasks in writing research

The analysis of keystrokes is a well-developed subdiscipline in the domain of biometrics (for a review, see Banerjee & Woodard, 2012). Researchers approach keyboard dynamics using statistical analyses, neural networks, pattern recognition, learning algorithms and search heuristics to describe, identify and classify typing patterns.

Also, several tasks have been developed in the domain of writing research to assess handwriting and typing fluency (Cooper, 1983; Gentner, 1983; Salthouse, 1986). The most well-known tasks are (full) text, sentence, phrase and letter copying. We briefly introduce each of these tasks before introducing the copy-task used in

this paper. This overview functions as a framework for the copy-task presented here because our copy-task draws on aspects of all previously introduced variations.

In text copy tasks, participants are usually instructed to retype a text as precisely and as accurately as possible. This type of task has been used, for instance, by Van Weerdenburg et al. (2019). In their intervention study, the authors used this task as part of a pretest-posttest design in order to assess pupils' typing skills. The study investigated the effects of taking a touch-typing course on spelling and narrative-writing skills. As expected, they found an effect on typing skill, and the experimental group also outperformed the control group with respect to spelling and narrative-writing skills.

A recent example of a sentence copy task can be found in a study by Dhakal and colleagues in which they describe their 136 million keystrokes corpus (Dhakal, Feit, Kristensson & Oulasvirta, 2018). Participants had to copy-type 15 sentences randomly drawn from a set of 1525 sentences taken from two corpora: the Enron mobile email corpus (Vertanen & Kristensson, 2011) and the English gigaword newswire corpus (Graff & Cieri, 2003). The sentences were selected according to a set of criteria (e.g., containing at least 3 words and maximally 70 characters). Participants had to transcribe the sentences directly from the screen, one by one. The study aimed at identifying typing patterns to differentiate between typing profiles. Using unsupervised cluster analysis, the authors were able to divide the participants (N = 168,000) into eight groups based on differences in performance, accuracy, hand and finger use, and rollover (i.e., pressing the next key before the release of the previous one).

In this study the sentence copy task was presented on screen. In other studies, however, texts had to be 'copied' from memory. For instance, in Grabowski's (2008) study participants had to repeat the first sentence of a well-known nursery rhyme. Because these sentences were learned by heart, it is likely that they can easily be recalled with minimal central influences, such as those from reading, and thus making typing skills more likely to directly affect copying. This material facilitated comparison of the typing performance of touch and non-touch typists.

Fewer researchers have used phrases in designing copy tasks. However, an influential study that is frequently used in technical research evaluating text entry methods has been presented by MacKenzie and Soukoreff (2003). They developed a collection of 500 prompts which consisted of combination of noun phrases (NPs) and sentences (without punctuation), 16 to 43 characters in length. For instance, "the back yard of our house", or "this is a very good idea". Utility programs are also provided to compute statistical properties of the phrase set selected (e.g., with respect to word frequency). Their corpus is made available on the internet.

Finally, a typical example of a letter copy task has been developed by Berninger and her colleagues. Their alphabet task is widely used as a copy task in handwriting research, in order to assess children's handwriting skills (Berninger & Rutberg, 1992;

Berninger et al., 1992). In this task, pupils have to write down the alphabet in order using only lowercase manuscript letters. Scoring is based on the number and rate of letters correctly produced in the right order within the first 15s or 60s.This task has been frequently used in research on the development of handwriting because it turned out to be predictive of other writing-related performances.

In the next section, we present a copy task that combines different aspects of the tasks described above.

2. Description of the Inputlog copy task

The copy task presented here was developed as a lean instrument to measure and assess typing skills in a series of well-defined transcription conditions within a limited time. The copy task at hand is a combination of components that consist of a short sentence, three-word combinations and letter clusters. The task is developed for online use and is freely available for researchers as part of Inputlog, a keystroke logging program used in writing and translation studies (Leijten & Van Waes, 2013; Inputlog is freely available for researchers at www.inputlog.net; see Appendix A for more details).

The copy task was first developed in Dutch (Van Waes et al., 2017), then further developed in collaborations within the European Literacy Network, a European Commission COST action. It is now available for eleven languages (Dutch, English, French, German, Italian, Norwegian, Polish, Portuguese, Spanish, Turkish, and Welsh) and three keyboard layouts (AZERTY, QWERTY, and QWERTZ).

To allow for comparisons across languages, we created a set of guiding design principles and implemented them as consistently as possible. Subsequently, however, in the analysis section we will only refer to the Dutch copy-task corpus. Below we briefly describe the Inputlog copy task. For a more detailed (technical) description, see Van Waes, Leijten, Pauwaert, and Van Horenbeeck (2019) and the materials on the Github repository (https://github.com/lvanwaes/Inputlog-Copy-Task).

2.1 Design of the components

The default copy task consists of five components, addressing different levels of lexicality as well as word and bigram frequency (Table 1). Moreover, the location of the keys on the keyboard is a design factor (e.g., left-right; adjacency). Participants are guided through the flow by a sequential interface. They are instructed to copy-type the given prompts as quickly and accurately as possible. They are not required to correct errors. At the start, the participant selects the language for the copy task and completes some background information. A privacy notice is included, in line with the privacy policy of the General Data Protection Regulation (GDPR) of the European Union.

The copy task starts with a brief introduction to instruct and inform the participant about the task, then guides them through the different components. Finally, a brief questionnaire is presented in which the following topics are addressed: handedness, hardware and software used for the test, dominant language, reading or writing difficulties, and familiarity with this task. For the handedness test we chose the reduced Edinburgh handedness test (Oldfield, 1971; Veale, 2014).

The main part of the copy task consists of seven tasks grouped in five components: (1) a tapping task, (2) a sentence copy-task, (3a-c) three-word combinations with high-frequency bigrams (repeated three times for different combinations); (4) a three-Word combination with low-frequency bigrams, and (5) a series of four consonant groups. On average the complete task takes about five to ten minutes.

Table 1. Overview of the five components of the copy task.

Components	**Description**
Tapping task	press the 'd' and 'k' key alternatively during 15 s
Sentence	copy a sentence during 30 s
Word combination HF 1-3	copy a combination of three words 7 times (high-frequent bigrams)
Word combination LF 4	copy a combination of three words 7 times (low-frequent bigrams)
Consonant groups	copy four blocks of six consonants once

Tapping task

The first component, the Tapping task, is a non-lexical task that intends to measure the fastest possible motor speed by pressing two keys with alternating hands (viz. 'd' and 'k', resp. a left-right and right-left hand combination). Participants are asked to type the 'd'-'k' key combination for 15 seconds (Salthouse, 1984). A time circle at the top right corner is used as time indicator. Finger-tapping tasks are commonly used to study the human motor system. They are simple enough to be used to study individuals with neuropathologies affecting the motor system (Witt, Laird, & Meyerand, 2008).

Sentence copy task

The next component, the sentence task, intends to measure the typing skills related to copying a series of short and high-frequent words in sentence context (Pinet, Dubarry, & Alario, 2016; Pinet, Ziegler, & Alario, 2016). For example, “the cat was

sleeping under the apple tree", in the English copy task. Participants are asked to repetitively type this sentence for 30 seconds. Repetitive typing reduces the reading-writing interaction that characterizes most existing copy tasks. As the stimulus is very simple, participants tend to produce the sentence from memory. In contrast with, for instance, nursery rhymes, no proper names and no infrequent (written) words were included. The prompted sentence consists, in all currently available languages, of seven to nine short, high-frequency words.

Three-word combination

In this task, four three-word combinations have to be copy-typed seven times each: the first three three-word combinations target the repetitive production of mainly high-frequent bigrams; the fourth contains low-frequent bigrams. These word combinations allow us to strictly control several bigram characteristics, which is not possible in an open copy task with pre-existing natural texts.

Table 2 gives an example of the presented word combinations (of the English version), together with their characteristics selected from previous studies (Dhakal et al., 2018; John, 1996; Logan & Crump, 2011; Ostry, 1983; Pinet et al., 2016; Salthouse,

Table 2. Examples of the word combination prompts in the English UK (QWERTY) copy task including the controlled characteristics.

	word combination 1	word combination 2	word combination 3	word combination 4
Word 1 (numerical)	four	seven	five	some
Word 2 (adjective)	interesting	wonderful	important	awkward
Word 3 (noun)	questions	surprises	behaviors	zigzags
#characters	24	23	23	18
High-frequent bigrams (HF – e.g., 'nt')	19	18	18	8
Low-frequent bigrams (LF – e.g., 'gz')	0	0	0	4
Left-Left (LL – e.g., 'es')	4	6	1	5
Left-Right (LR – e.g., 'fo')	4	6	2	3
Right-Right (RR – e.g., 'ou')	4	2	5	1
Right-Left (RL – e.g., 'us')	3	4	2	2
Adjacent keys (e.g., 'io')	7	6	3	4
Repetitive keys	0	0	0	0

1986; Wobbrock & Myers, 2006). For instance, Salthouse (1986) has clearly demonstrated that high-frequency bigrams are typed faster than low-frequency bigrams (see also Pinet et al., 2016). Word combinations were controlled for keyboard characteristics because Logan (1999, 2003) has shown that bigrams requiring the coordination of both hands are executed faster than same-hand bigrams. Also, Gentner et al. (1988) showed that keystroke transitions are affected by the keyboard layout (adjacency and same-finger key repetition, cf. also Hess, Mousikou, & Schroeder, 2020).

Consonant groups

The final component, the Consonants task, measures typing skills in another non-lexical context (Grabowski, Weinzierl, & Schmitt, 2010). Participants are asked to copy four blocks of six *consonants* once. The blocks are identical for all languages: tjxggl pgkfkq dtdrtt npwdvf. These strings of consonants do not have a meaning and cannot be phonologically encoded (beyond letter names), which inhibits phonological and semantic chunking.

2.2 Implementation

The copy task is a web-based application (platform independent), developed as a separate JavaScript component. Using the design principles described above, the task has been developed for eleven languages and implemented by considering national keyboard layouts. The stored logging file is an XML-file which can be used for further analysis.

Within the Inputlog analysis component, a default copy-task analysis provides a large variety of perspectives to interpret the logged data. About 900 variables are being provided representing descriptive statistical data for within-bigram transition time (or inter-key interval) at the component level, trial level and related to the manipulated bigram characteristics (frequency, adjacency, repetition, hand combination needed to type the bigram). Apart from means, medians and standard deviations, also geometric means and confidence intervals are presented. Correctness scores indicate response quality. Time and trial filters are applied.

Finally, the output is available as a non-aggregated CSV file containing a full list of the typed bigrams enriched with information about the component (and trial) in which they were composed, their interkey-interval duration, frequency class, and keyboard location.

3. Method

As part of a series of research studies using keystroke logging, we collected a corpus of Dutch copy tasks. The corpus consists of keystroke data of 1682 participants

completing the copy task. We first present some general information to characterize this corpus data set. Next, we explore the test-retest reliability of the copy-task components, before we present an analysis in Bayesian mixed effects models with extension to mixture models. This analysis aims at characterizing the data from each copy-task component and differences between them.

3.1 Participants and age distribution

For the present analysis, we used the Dutch corpus of copy-task data. The data were collected from 1682 participants (1107 females, 495 males, 80 unknowns) aged between 13 and 83 years old (median = 21 years; *SD* = 11.81) in the context of various research projects, research courses and training schools. The age groups are not equally represented because most of these studies took place at secondary schools or universities: 78.94% of the participants are 23 or younger. The age distribution of the copy-task corpus can be found in Figure 2A showing a bimodal, positively skewed distribution. This figure shows that the majority of our participants were younger than 23, a large proportion was between 20 and 30, and the right tail showing a smaller number of participants older than 30.

The distribution of participants' age and their inter-keystroke intervals can be found in Figure 2B. This distribution is illustrated in different colors for each copy-task component illustrating the relationship between the participants' performance on the copy task components and their age. For each component, we observed a nonlinear relationship between age and keystroke latencies (Bosman, 1993; Van Waes et al, 2017).

Handwriting research has shown that automaticity is reached at about the age of fourteen (Berninger & Winn, 2006; Graham & Harris, 2000; Medwell & Wray, 2014). In our keyboarding corpus, typing performance is the fastest at a later stage, viz. between 21 and 30. From the age of 30 onward, typing speed seems to gradually slow down. In the higher age group of this corpus it is important to notice that typing skills might differ considerably. The effect of age on the keystroke transitions at the component level shows that the evolution pattern is more or less consistent for each of the components. Interestingly, all components relate highly similar to each other across lifespan. However, as the lexical components (i.e. Sentence task and Word-combination tasks) are characterized by a slight increase of speed at the initial stage (between the age of 14 and about 25), this trend is less explicit for the non-lexical tasks (Tapping task and Consonants task). This graph suggests different functions for lexical and non-lexical tasks. Lexical tasks show longer IKIs for younger participants, followed by a speed-up that is not seen in the non-lexical tasks. Thus, age-related fluency progression seems to affect the lexical components more than the non-lexical components in our corpus. Note, these initial observations are merely descriptive and not central to this paper but open for further investigations.

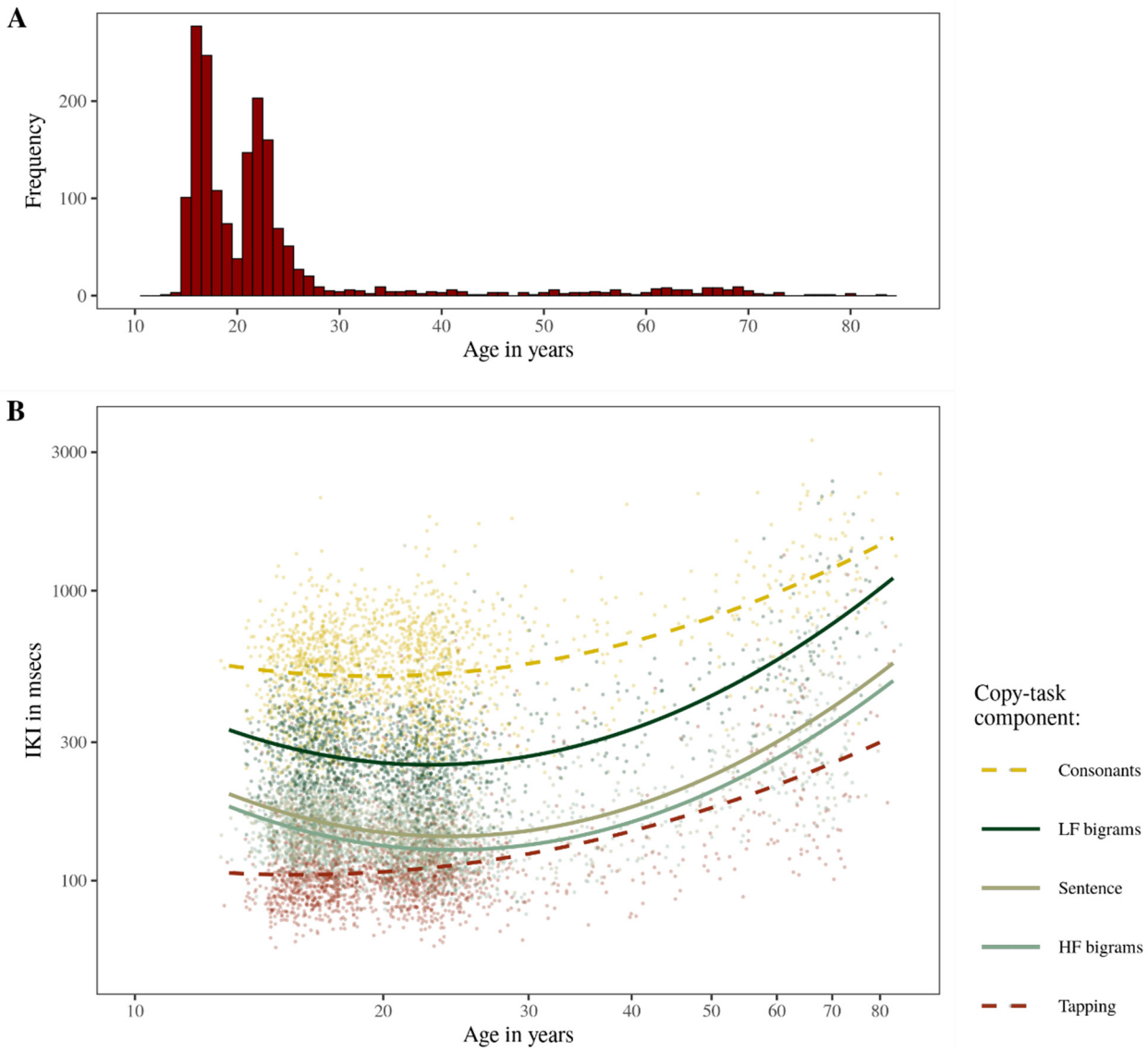

Figure 2: Panel A shows a histogram with the age distribution of the copy-task corpus. Panel B shows the distribution of inter-keystroke intervals and the relationship between age and inter-keystroke intervals (IKIs), both log-scaled. For visualisation data were aggregated on participant-level and IKIs were capped at 3,500 msecs. Lexical tasks were indicated by solid lines and non-lexical tasks were displayed with dashed lines. Age information was available for 1,612 participants.

3.2 Materials

The Dutch copy-task corpus contains 1,447,310 inter-keystroke intervals (per logfile: mean = 700.54; *SD* = 78.56) characterized by a set of variables (see Section 2.2). R-scripts, Stan code and dataset are available on Github (https://github.com/jensroes/Copy-task-analysis) and can be used for reproducing and extending the presented analysis for further investigation involving the copy-task data.

3.3 Data analysis

The aim of this analysis is to provide population parameter-estimates for all copy-task components. This is achieved in two steps: First, we assessed the test-retest reliability of the copy-task data. Second, and importantly, we inferred inter-keystroke intervals for each copy-task component after controlling for variability specific to the sample (i.e. bigrams and participants) and accounting for process disfluencies. The inferred data were then used to determine differences in inter-keystroke intervals between copy-task components. For example, typing differences are specific to frequency (HF vs LF bigrams), lexical information (Tapping vs HF bigrams), syntactic structure (HF bigrams in an adjective-noun combination vs. a simple sentence) or increased cognitive demands (LF bigrams vs. Consonants task).

The inter-keystroke intervals (IKIs), i.e. the latency between two consecutive keystrokes, were analyzed in Bayesian linear mixed-effects models (see, e.g., Gelman & Hill, 2007; Gelman et al., 2014; Kruschke, 2014; Lambert, 2018; McElreath, 2016) using the Stan package in R (Carpenter et al., 2016; Hoffman & Gelman, 2014; Stan Development Team, 2015a, 2015b). These models were extended to a mixture model (Farrell & Lewandowsky, 2018; Vasishth, Chopin, Ryder, & Nicenboim, 2017).[1]

The rationale for using mixture models for the copy-task data can be illustrated with the following example. Baaijen, Galbraith, and de Glopper (2012) used a mixture model approach to analyze pause length in text production. They found that pause length is determined by a combination (i.e. mixture) of three normal distributions (i.e. Gaussians) of which each has a specific share in the data, i.e. mixing proportion. They found that pause durations of 330 ms have a mixing proportion of .65, pauses of 735 ms a proportion of .26, and those with durations of 2697 ms have a proportion of .09. In other words, pausing in text writing is a combination of three processes that are represented in 65%, 26% and 9% of the data, respectively. We can infer these pauses reflect the varying demands of cognitive processes.

As highlighted in the introduction of this paper, keystroke data – as well as many other lower-level tasks – are a combination of different cognitive processes which cascade from higher into lower levels of representation. We extended our analysis to mixture models as they constitute a modeling framework that elegantly maps onto this combination of cognitive processes (Baaijen et al., 2012; Vasishth et al.,

2017). Mixture models allow us, therefore, to detect keystroke-intervals that correspond to different underlying (mental) processes. Instead of removing disfluencies using fixed thresholds, we can use statistical models to estimate the ratio of fluent and disfluent keystroke transitions. This is important as disfluencies are a natural part of the writing process and may reflect higher-level planning. Copy-typing – similar to text production – comprises a mixture of cognitive processes. For example, copy-typing requires visually encoding the target string, buffering a mental representation as well as activating and executing the relevant motor code. In the Tapping task, we would expect only a small amount of data to be subject to higher-level demands while the majority of data will represent fluent typing. In the Consonants task, on the other hand, the data are expected to reflect a mixture of typing execution, visual encoding and mental buffering. Mixture models provide a principled approach to account for keystroke intervals that are reflective for typing execution on the one hand and higher-level planning on the other hand to provide reliable population estimates.

From a statistical viewpoint, mixture models assume that the underlying data generating process is a combination of Gaussian distributions (i.e. normal distributions). In contrast, statistical methods such as linear regression models assume the data represents a single underlying normal distribution. While regression models estimate one population mean and variance from the data, a mixture model with two mixture components would estimate the mean and the variance for two components involved in the underlying data generating process. This is achieved by introducing an additional (latent) model parameter – the mixing proportion – which captures the probability to which data are attributed to either distribution.

In psychological terms, the use of Bayesian mixture-models allows us to match the keystroke data to the assumed underlying combination of cognitive processes, after excluding variance that is subject to participant and bigram-specific variability.[2]

3.4 Data preprocessing

For this analysis, the keystroke data were minimally trimmed. Missing data (0.01%) and inter-keystroke intervals equal to zero (0.02%) were removed as well as non-targeted bigrams (4.25%). The overall median score for correctly typed bigrams across participants was 96.35% (interquartile range (IQR) = 3.03). The overall accuracy was lowest in the low-frequency (LF) bigram task (median = 89.66%, IQR = 14.04) and highest in the Tapping task (median = 98.6%, IQR = 4.28).

The following analysis data were aggregated across repetitions rendering one inter-keystroke interval per letter combination (i.e. bigram) per participant. The analysis was performed on a random subset of 500 participants (i.e. about half a

million keystrokes). We chose a random subset for two reasons. First, our analysis focuses on analyzing data with a minimum amount of aggregation to address known sources of random variance. A realistic estimation of random variation in Bayesian models is computationally difficult. Using a subset is beneficial to reduce the computational power needed. Second, 500 participants is a large sample for the investigation of typing processes. In practice, most researchers will not be able to use a sample as large as the entire copy-task corpus. For the aims of the analysis outlined below, there are no benefits of extending the analysis to the entire corpus. The presented results should be reproducible with any other subset obtained from the corpus. All scripts are available online for further investigations of the entire copy-task corpus.

4. Results

First, we present an overall characterization of the copy-task data in our corpus. Next, we evaluate the copy task's reliability in a test-retest analysis. Finally, we evaluate the distribution characteristics of fluent and disfluent keystroke transitions for each copy-task component and their respective differences.[3]

4.1 Overall interkey latency

Table 3 presents descriptive statistics directly taken from the aggregated Inputlog copy-task analysis. The table shows that the overall correctness score for the data collected is 94.2% indicating that the participants strived at completing the copy tasks very accurately. The overall mean IKI latency – calculated across all copy-task components and restricted to the targeted bigrams only – is 169 ms (*SD* = 84), with an average rate of 392 (*SD* = 96*)* characters per minute.

Table 3: Overall IKI descriptives (in ms) for targeted bigrams (overall and for lexical components only).

	Mean	SD	lower 2.5% CI	higher 2.5% CI
IKI bigrams: overall				
mean	169	84	165	172
median	137	62	134	139
converted logmean (trimmed)	141	64	138	143
characters per minute	392	96	388	396
IKI bigrams: lexical components				
mean	152	75	149	155
median	134	61	132	137
converted logmean (trimmed)	134	62	132	137
characters per minute	438	110	434	442

When we limit the results to the lexical task components, we notice a slight increase of typing speed: the mean IKI is 152 ms (*SD* = 75). The report also shows the trimmed log mean. As pausing data are right-skewed, a geometric mean provides a better representation of a person's typing performance. The values for the log-based means are respectively 141 ms (*SD* = 64) overall and 134 ms (*SD* = 62) for the lexical task components (Sentence and Word combination tasks).

Figure 3 shows boxplots of the inter-keystroke intervals for each of the copy-task components with the individual data as jittered dots. The distributions show both the central tendency and dispersion of the data. The non-lexical tasks demarcate the limits of the inter-keystroke intervals. The purely motoric tapping component shows very short keystroke latencies and the Consonants task involving memory storage and eye-hand coordination shows particularly long keystroke intervals. The lexical components show a clear difference between the LF bigram task on the one side and the HF bigrams and Sentence task on the other side.

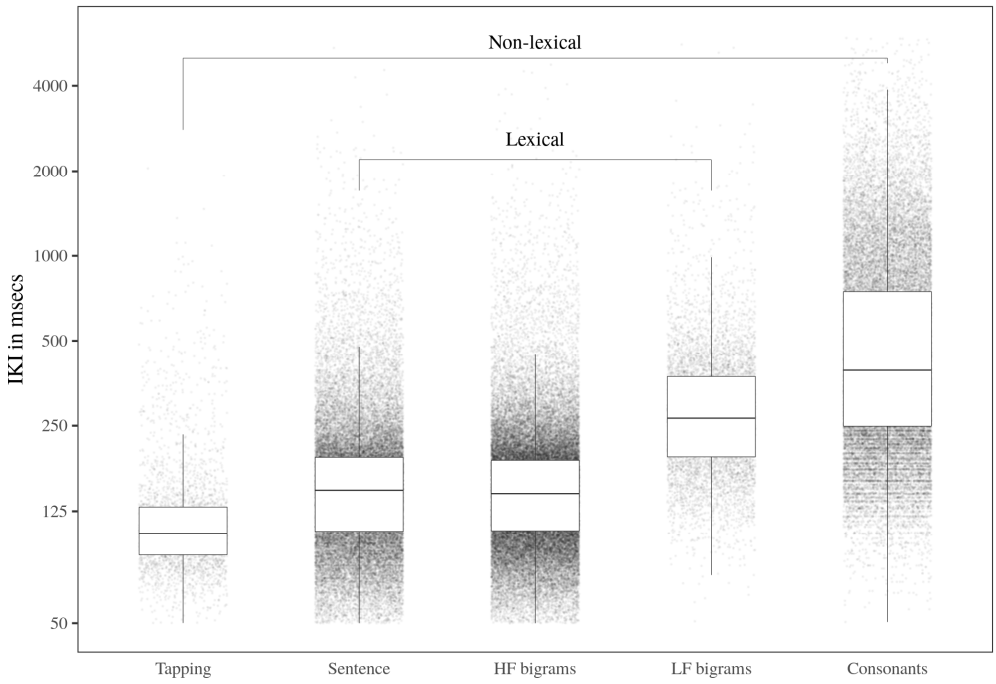

Figure 3: Boxplots of the inter-keystroke intervals (IKI) for each copy-task component of the full copy-task corpus. For visualization purposes, the data were capped at 3,500 ms and shown on a log scale.

In addition to the differences between components shown in Figure 3, there are clear differences in the variance between the components. In particular, the Consonants task shows a wide dispersion. The dark area around the boxplots for the Sentence and HF bigram task illustrates a dense distribution of the inter-keystroke intervals around a center. While most components seem to be associated with a specific variance, the Sentence copy task and the HF bigram task show similar distributions. In other words, the variance of the majority of copy-task components is unequal (i.e. heteroscedasticity).

4.2 Test-retest reliability

Measure stability was assessed in a test-retest design: 245 participants completed the copy task twice with a time-lag interval of at least one week (195 males, 45 females, 5 unknowns; median age = 22 years, range: 15, 73).

We used Bayesian linear mixed-effects models to evaluate whether, first, participants were faster during the completion of the second session, and second, whether we could predict the IKI data from the second session from the IKIs produced in the first session. Models were fitted with the copy-task component as fixed effect and random intercepts for participants and bigrams. We calculated the estimated population mean and the 95% highest posterior density interval (HDPI) expressing the range of values that contains the true parameter value with a probability of 95% (see Appendix B).

First, comparison of the first and second sessions revealed a small speed-up effect of -11 ms for the Tapping component. The HDPI interval indicates that the speed-up effect for the Tapping task has a 95% probability to be between -20 ms and -3 ms. In other words, there is negligible evidence for 0 ms as a plausible parameter value. Similarly, we found a speed-up effect for LF bigrams of -12 ms (95% HPDI: -24, -4) and for the Consonants task of -9 ms (95% HPDI: -15, -4). Neither HPDI included 0 ms as plausible parameter value, thus indicating support for a systematic but small speed-up effect of around 10 ms. The speed-up effect for the Sentence task (-2 ms; 95% HPDI: -5, 0) and HF bigrams (-1 ms; 95% HPDI: -3, 1) has a small magnitude and the HPDIs includes 0 as possible parameter value.

These differences are too small to suggest a strategic change in participants' responses but seem related to task familiarity. The population means for each session are illustrated in Figure 4. Although there is a systematic speed-up in some copy-task components, the effect does not change the overall pattern of the copy-task components.

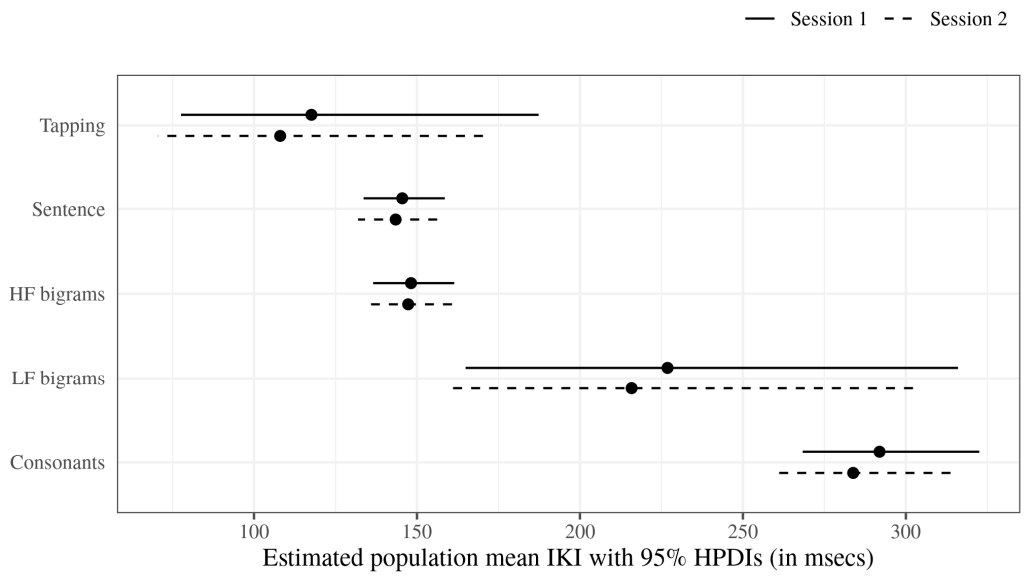

Figure 4: Estimated population mean for each session shown by copy-task component. Dots indicate the most probable parameter value and error bars show 95% HPDIs.

Second, for all copy-task components we found evidence for a positive predictive relationship between IKIs from the first session and the second session. This was supported for the Tapping component (0.34; 95% HPDI: 0.25, 0.41), the Sentence component (0.44; 95% HPDI: 0.41, 0.46), the HF-bigrams task (0.6; 95% HPDI: 0.57, 0.62), the LF-bigrams task (0.49; 95% HPDI: 0.43, 0.55), and the Consonants task (0.3; 95% HPDI: 0.28, 0.32).

These results confirm the reliability of all five copy-task components. The general pattern observed in IKIs was reproduced in a second session in spite of a systematic but small typing advantage in the second session. This can be explained in terms of hesitations that might be related to accommodation to the novelty of the copy task environment. Further we found evidence for a positive predictive relationship between IKIs from the first session and IKIs from the second sessions, supporting a strong test-retest reliability.

4.3 Comparing models of inter-keystroke intervals

In the following analysis, we compared four models fitted to the copy-task data. The aim of these models was to derive population estimates for each copy-task component by accounting for the possibility that the data for each component arise

from a mixture of distributions. Finally, we tested for differences between the parameter estimates of the copy-task components.

First, we fitted a linear mixed-effects model with an intercept term only (a null model). This model was compared to a model with each copy-task component as fixed effect. A comparison of these two models allows us to determine the predictive performance of each copy-task component (see Appendix B). Next, we implemented a model that assumed that each copy-task component has its own variance parameter. This assumption was carried over into the mixture model. In other words, the mixture model was constrained, such that the distribution of longer IKIs has a larger variance. The reason for this was that larger values from reaction-time data in particular (Wagenmakers & Brown, 2007) and human motor behavior in general (Schöner, 2002; Wing & Kristofferson, 1973) are associated with a larger variance.

To compare the fit of the different models, we used leave-one-out cross-validation which allows us to test the predictive ability of models, by penalizing models with more parameters (see Farrell & Lewandowsky, 2018; Lambert, 2018; Lee & Wagenmakers, 2014; McElreath, 2016). We determined the out-of-sample predictive performance via Pareto smoothed importance-sampling leave-one-out cross-validation (Vehtari, Gelman, & Gabry, 2015, 2017). This predictive performance was estimated as the sum of the 'expected log pointwise predictive density' (*elpd*). Model comparisons can be found in Table 4. The difference between the predictive quality of the best fitting model compared to the remaining models was expressed as $\Delta elpd$. A negative difference $\Delta elpd$ indicates that the predictive performance of a model is lower compared to the best fitting model.

Table 4: Predictive performance of four Bayesian models, three Bayesian linear mixed-effects models (BLMM) and one Mixture of Gaussians (MoG) model with two mixture components. The model fit was ordered starting with the model with the highest predictive performance on the top. Differences in model fit assessed as expected log pointwise predictive density $\Delta elpd$ (SE = standard error) are shown with reference to the model with the highest predictive performance.

Model	$\boldsymbol{\Delta elpd}$	$\boldsymbol{SE}$
MoG	0	0
BLMM (unequal variance)	-1,128.49	73.79
BLMM (equal variance)	-4,304.25	118.00
BLMM (intercept-only)	-4,554.04	120.44

The mixture model showed a higher predictive performance than the linear mixed-effects models with an increase in predictive performance of more than $\Delta elpd$= 1,000 compared to the best fitting linear mixed-effects model. In other words, modeling the keystroke intervals as mixture processes largely improved the model fit compared to mixed-effects models. The Bayesian linear mixed-effects model with

unequal variance for each copy-task component rendered a higher predictive performance compared to the equal variance standard linear mixed-effects models. The lowest predictive performance was observed for the intercept-only model. Including the copy-task components as model parameter improved the predictive performance of the model. Allowing varying variance parameters increases the predictive performance of the models as well as the addition of mixture components. In other words, it is important for the statistical model to acknowledge that there is a different variability associated with the intervals of each component.

The estimates for each copy-task component are summarized in Figure 5. For each copy-task component, the most probable parameter estimate (estimated population mean) and 95% HPDIs are shown indicating the range in which the population mean is contained with a 95% probability. Values are shown for each mixture component. For all copy-task components, except HF bigrams, two mixture components were detected, representing fluent and less fluent keystroke transitions. For HF bigrams, however, the centers of the mixture components are almost identical.

We compared the estimates of the Bayesian linear mixed-effects model to the estimates of each mixture component. The parameter estimate of the first mixture component (K_1) is similar to the estimates of the Bayesian linear mixed-effects model in the Tapping task, the Sentence and the HF bigrams task, and the LF bigram task. In the Tapping, Sentence and LF bigram task, the second mixture component (K_2) accounts for 13% of longer keystroke intervals. In the HF bigrams task, both mixture components are consistent with the parameter estimates of the linear mixed-effects models. However, the Consonants task shows shorter IKIs for K_1 and longer estimates for K_2 compared to the linear mixed-effects model. This indicates that the estimates of the linear mixed-effects model were biased. In sum, for the majority of copy-task components, both mixture and linear mixed-effects models provide reasonably similar estimates. However, the linear mixed-effects model provides a misleading estimate for the Consonants task which is better captured as a mixture of two processes.

The estimated population mean of the mixing proportion and their 95% HPDIs are shown in the panel labels of Figure 5 for each copy-task component. This mixing proportion can be understood as the probability to observe relatively long latencies (i.e. disfluencies) which ranges between 0 and 1.

As the model is constrained to be bimodal, we know that the estimated population mean of fluent keystroke transitions is $1 - x$, where x is the mixing proportion displayed in the panel stripes of Figure 5. For the Consonants task, this means that the two distinct distributions appear with a probability of 31% for shorter IKIs (fluent typing) and 69% for longer IKIs (hesitant typing).

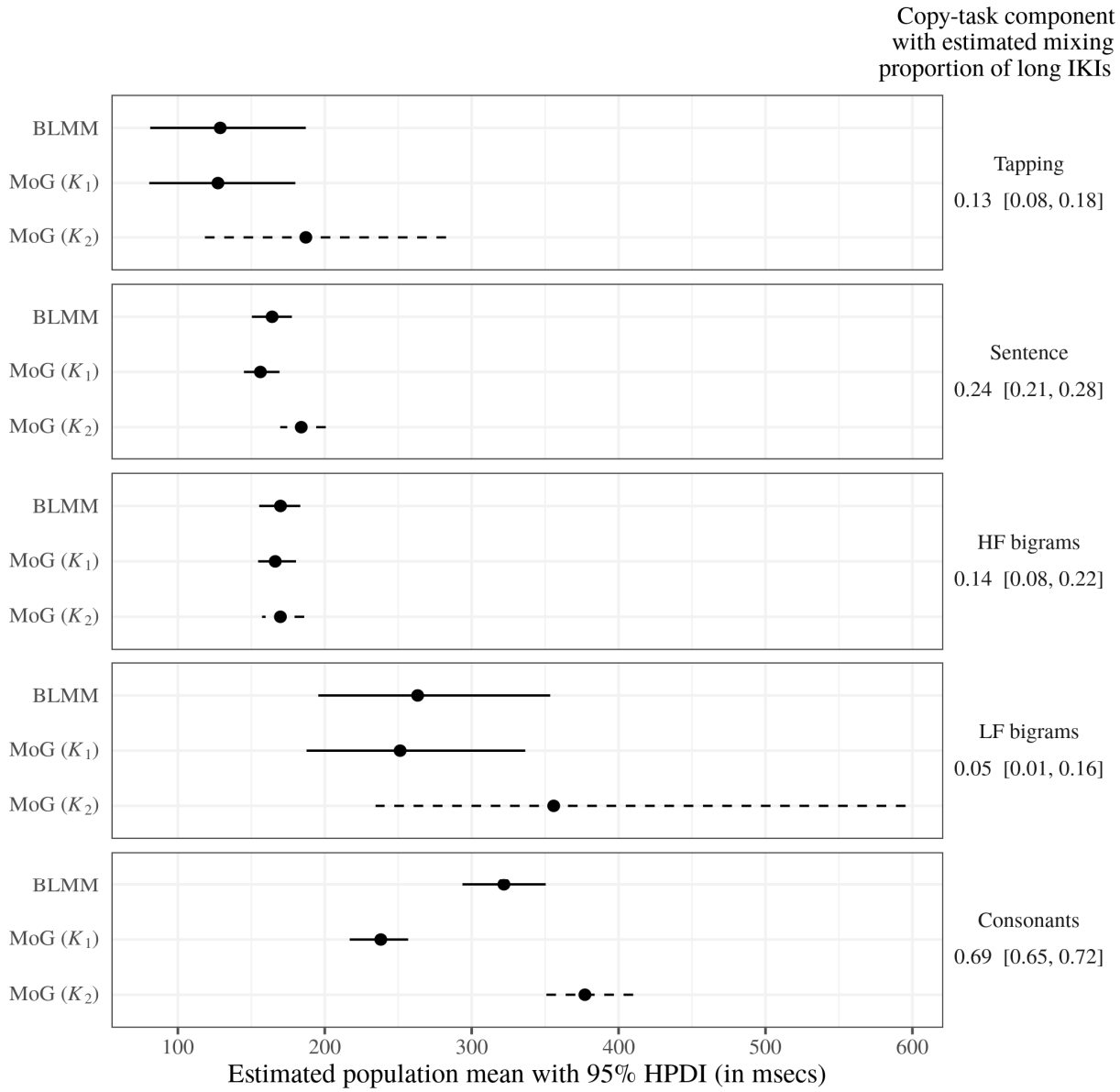

Figure 5: Comparison of parameter estimates of the Bayesian linear mixed-effects model (BLMM) and the mixture model (Mixture of Gaussians [MoG]) showing the estimated population mean and 95% HPDIs. Mixture components are indicated by subscripts. For each copy-task component the estimated population mean of the mixing proportion is shown with 95% HPDIs in the panel strips.

For the Sentence task, longer IKIs have a probability of 24%. The mixing proportion or the remaining copy-task components is relatively small. In other words, the majority of the copy-typing processes is reflected by one mixture component whereas the smaller proportion may be attributed to occasional pausing, disfluencies or possibly slower initial keystrokes due to processing on higher cognitive levels. The HPDIs indicate a larger uncertainty about the parameter estimate in the Tapping task and the LF-bigram task than in the remaining three tasks. This variability is indicating flexibility of IKIs specific to either of these two components rather than an overall variability in relatively fast and slow tasks. This is because the Consonants task shows less variability compared to the distribution of the LF-bigram task, even though the former is cognitively more demanding. Also, as the Tapping task allows fast typing, a slowdown is more likely than in tasks that exhibit a more consistent typing speed, viz. the Sentence task, the HF-bigram task, and the Consonants task.

The differences between copy-task components were assessed in pairwise comparisons. The population estimate for these differences (with 95% HPDI) can be found in Figure 6. The differences are shown for the population estimates of the first mixture component (fluent typing) and the second mixture component (long IKIs). Also, we show the differences in the probability of long IKIs. Numbers above the intervals show $P(\Delta\mu < 0)$, the probability that the posterior difference between the copy-task components is smaller than zero; a value approaching 0 indicates a low probability that the true difference is smaller than 0 and a value approaching 1 indicates a high probability.

As shown in Figure 6 we found differences for fluent IKIs across copy-task components; no notable differences were seen for fluent IKIs in the Consonants task compared to the LF-bigrams task, and Tapping compared to both the Sentence task and the HF-bigrams task. Comparing fluent IKIs for the Sentence task and the HF-bigrams task rendered a very small but consistent difference. Comparing the population estimates for the mixture component of long keystrokes rendered longer disfluencies for the Consonants task compared to the Tapping task, the Sentence task, and the HF-bigrams task, but not compared to LF bigrams. The latter showed longer disfluencies compared to HF-bigrams task. The remaining comparisons show no notable difference. The probability of long IKIs was found to be larger in the Consonants task compared to all other copy-task components. The remaining components show small or no substantial differences.

In summary, in all copy-task components we observe a mixture of two processes in which one is reflected in short and the other in relatively long intervals. Longer intervals can be attributed to inhibitions on higher levels of processing which delays keystroke intervals.

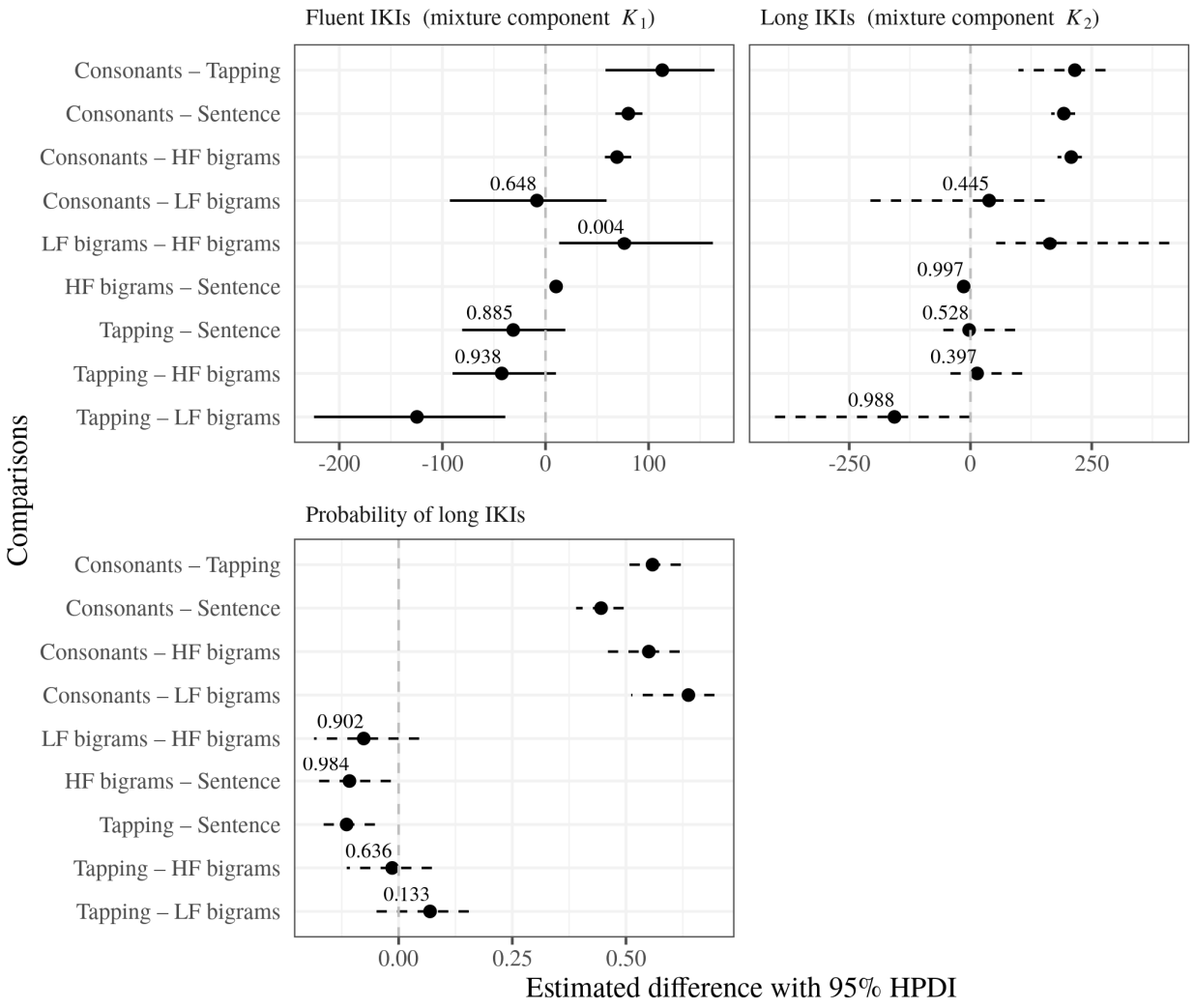

Figure 6: Difference between estimated population IKIs of all copy-task components shown for both mixture components, representing fluent IKIs and long IKIs, and in the lower panel the differences in the mixing proportion of long IKIs. Error bars indicate 95% HPDIs. Numbers in the graph show the posterior probability that the true difference is smaller than zero $P(\Delta\hat{\mu} < 0)$. Values of $P(\Delta\hat{\mu} < 0) < .001$ or $P(\Delta\hat{\mu} < 0) > .999$ were omitted.

Shorter intervals are representative for uninhibited typing execution. After accounting for the possibility that copy-typing underlies a mixture of two processes resulting in fluent IKIs and slow IKIs we made the following observation: fluent copy-typing is less affected by lexical information than it is affected by the frequency of bigrams; disfluent keystroke transitions in copy-typing were primarily associated with difficulty in the absence of lexical information. In particular, we found that the reduced frequency of groups of non-lexical bigrams slowed down the typing speed compared to both purely motoric copy-typing and lexical tasks that involve high frequency bigrams. This is supported by the finding that neither lexical information in low-frequent bigrams (LF bigrams vs Consonants task) nor lexical information in high-frequent bigrams (Tapping task vs HF bigrams and Sentence task) affects the typing speed. Conversely, the frequency of the involved bigrams has a strong impact on the typing speed (Tapping, HF bigrams, Sentence task on the one side and LF bigrams and Consonants on the other side). This is important as it shows that lexical information per se (HF bigrams, Sentence component) does not radically change the typing speed compared to purely motoric typing (Tapping task).

This is, however, not to say that lexical information does not affect copy-typing at all. Indeed, we observed that keystroke transitions in the Consonants task were more frequent compared to the Tapping task, the HF-bigrams task, the Sentence task, and LF-bigrams task. Although these comparisons showed longer pauses in Consonants tasks for the Tapping, HF-bigrams, and the Sentence task, pauses were equally long in the Consonants and the LF-bigrams task. Similarly, pauses were found to be longer in LF-bigrams task compared to HF-bigrams task but not more frequent.

Further, for almost all copy-task components, long intervals were in the minority except for the Consonants task data; while less data were attributed to shorter intervals, they were still systematically shorter than copy-typing in the remaining components. This pattern reflects the cognitive demands involved in typing low-frequency bigrams and encoding, or updating, memory representations of the target bigrams. In other words, for low frequency bigrams, lexical information facilitates copy-typing.

4.4 Summary

We demonstrated that an extension from linear mixed-effects models to mixture models can be used to evaluate typing characteristics of a population for each copy-task component and their respective differences. Mixture-process models allow us to model keystroke intervals as a combination of fluent keystroke transitions and long intervals reflecting inhibition at higher levels of activation. Hence, mixture models show an elegant mapping between typing data and the assumed underlying mental process that feeds into the motor execution. Accounting for this mixture

process, we showed that lexicality does not affect copy-typing as much as frequency information does. Further we proved the test-retest reliability of the copy task. We observed a small but systematic speed-up effect for individuals performing in a second session of the copy task; this difference did not affect the overall pattern across copy-task components.

5. Conclusion and Discussion

5.1 Age and cohort effects

We have already, in the theory section, refered to the potential of copy tasks for the decomposition of the writing process into bio-mechanical, linguistic, and cognitive components (Grabowski, 2008) that determine resulting patterns of keystroke latencies. Over and above these theoretical implications, we observed a non-linear relation between age and keystroke transitions. However, we do not yet know whether a cohort effect has additional influence on the observed typing skill development. Not long ago, it was far from common that almost all people have typing experience; rather, particularly high-education professionals would have secretaries or stenotypists for the manual execution of writing who underwent a special training. Now, whole generations of younger people are in active contact with – virtual or physical – keyboards from their early years on, but they frequently acquired their skills rather autodidactically, life-long remaining "advanced typing laypersons". As a result, there are many different, though functional typing strategies between professional touch-typing and effortful hunt-and-peck typing. This situation may affect the slope of the speed curves towards the higher ages, when these generations become older. Discussing further these age-related differences is beyond the scope of this paper. However, they might provide an interesting avenue for future research. For example, the reproducibility of the observed differences depending on lexicality of the stimulus can be directly explored through analyses of existing copy-task data from languages other than Dutch.

5.2 Possible applications of a standardized copy task

In the introduction, we described the problems associated with pause thresholds that are often crucial for the attribution of different indications when interpreting IKI intervals (cf. Medimorec & Risko, 2017). For the copy task, we can be quite sure about the sources and influences of the observed patterns of temporal typing progression. Since the task is standardized, interindividual differences depend on skill variation and, at best, on different experiences with basic linguistic properties such as letter combination frequencies. Therefore, the individual copy-task results can be used to better interpret the logging protocols of other, more complex

writing tasks in focus. Either, pause thresholds can be more subtly defined according to the individual typing baselines (not only in terms of simple typing speed and basic mechanics, but also differentiated for, e.g., high and low bigram or word frequencies), allowing for a better indication of other, more demanding cognitive processes that are reflected in the temporal typing patterns. Or, the relevant copying parameters are introduced as covariates in the analyses of the respective research questions. Moreover, having standardized copy-task results would also inform the researcher about the skill-related homogeneity or heterogeneity of the participants.

Beyond simple typing-speed measures, there is no simple vademecum for researchers with respect to a general recommendation on which particular task variable to include. We used a Bayesian mixture-models analysis to identify two distributions from the data representing a mixture of fluent and disfluent keystroke transitions (see Roeser et al., 2021, for a detailed discussion). We contend that identifying disfluent keystroke transitions might be helpful to correct estimates for fluent transitions. For instance, when analysing typing data, the estimated length of fluent transitions in the lexical components (e.g. HF-bigrams or the Sentence task) can be used as typing-skill indicator. If the focus is not on low-level processes, disfluent transitions in the LF-bigrams task might be of interest. Data from non-lexical components, the Tapping and the Consonants task, could serve as a more stretched reference for low-level fluency and disfluency. Moreover, they can be considered as a theoretically relevant outcome variable representing inhibition at a higher level of cognitive activation at the level of intra-word transitions. We suggest these possibilities for future research.

Nevertheless, it appears feasible to finally develop a standardized copy task with norms for languages and across age, as soon as data sets of sufficient size exist beyond the Dutch sample presented here. The observed reliabilities of the task components support the assumption that individual typing skills have sufficiently stable characteristics that would not vary from one typing situation to the other. Moreover, as the task is systematically constructed, it shows construct validity and face validity. However, additional external validation can be achieved by comparing our results to other copy tasks, e.g. tasks that are used for the vocational assessment of typing proficiency, and to typical copy-typing results established elsewhere (e.g., Wu & Liu, 2008). Note, however, that the copy task that we introduced produces meaningful results also for typing strategies other than "perfect" touch-typing. It is the individual typing strategy, or proficiency, that can help to understand and analyze more complex writing processes and their respective results. Lower typing skills, as indicated, e.g., by longer inter-keystroke intervals, do not necessarily make the resulting products worse, but they will change the production process and the contexts and resources a writer may need to achieve a sufficient text quality.

With respect to research applications, the standardized development and description of a multilingual copy task can be seen as a first contribution to a classification of writing tasks in general. Research on writing processes uses different, often unique tasks (e.g., instructions, descriptions, narrations, essays) that lead to quite different result patterns with different requirements regarding the identification of interpretable data structures. Having a consistent description and classification of writing tasks, the typical process patterns they evoke, and the relevant factors responsible for the emerging variation, would allow us to run comparable studies that focus on specific problems that are worth to be investigated in greater depth. In general, this can create a more coherent picture of what we know about writing processes from keystroke logging.

There are also further applications that do not refer to writing-process research in a narrow sense. First, the repeated assessment of typing skills via the copy task can indicate not only whether trainees show progress after a typing-skill training but also whether sufficient automatization has been acquired, or improved. For instance, with respect to the representation of language-specific frequency phenomena as indicated by faster copying of words that contain high-frequency bigrams. Mastering a second language may also lead to an increasing differentiation of the observed temporal writing patterns. Further, the copy task can be used for other diagnostic purposes, e.g. in the clinical field (e.g., aphasia, dementia or dyslexia), as soon as we know about the specific interference on typing patterns associated with certain impairments. For example, Van Waes et al. (2017) found first indications that a general cognitive impairment as associated with age has partly different effects on the copy task than a beginning Alzheimer's disease. Testing patients that suffer from aphasia, with or without an accompanying agraphia (Behrns, Hartelius & Wengelin, 2009), may also help to identify some of the particular difficulties of language use that could be addressed by individual neuro-linguistic rehabilitation training.

5.3 Further research

This paper presented an analysis of the Dutch sample of a copy-task corpus. Due to the systematic cross-linguistic construction principles, there will be further data sets available that open at least three novel opportunities: (1) the comparison between the targeted languages with respect to generalized findings such as the dependence of IKIs on bigram frequency; (2) the identification of effects or patterns pertaining to the respective language or group of languages, along with the search for appropriate explanations that may relate to linguistic characteristics, but also, e.g., to the keyboard layout in use or to orthographical conventions such as noun capitalization in German; (3) the correction against language-specific patterns which eliminates single-language effects in order to make existing data cross-linguistically comparable. Nevertheless, there are linguistic characteristics that may

yield unique overall patterns of IKIs. For example, languages differ with respect to the consonant clusters they allow at the onset and the coda of a syllable. If a language like Italian has mostly open syllables without a coda, there are necessarily less different bigrams, with relatively higher frequencies of occurrence, than in a language like German where more complex consonant clusters are possible on both sides of the nucleus (e.g., "Strumpf" [stocking] or "geplantscht" [plashed]) which necessarily leads to a larger variety of bigrams with lower relative frequencies.

Effects in the temporal writing patterns do not only occur with typing, but also with handwriting. For example, bigram-based effects, but also effects of syllable boundaries (which may separate otherwise coherent bigrams) have been observed in handwriting studies as well (Kandel, Peereman, Grosjacques, & Fayol, 2011; Nottbusch, Grimm, Weingarten, & Will, 2005; Sausset, Lambert, Olive, & Larocque, 2012). When compared to handwriting results, a standardized copy task can identify effects across the two writing modalities, indicating that they do not (only) rely on practise and mechanical fluency, but also on linguistic representations above the level of motor execution (Van Galen, 1991), or even on a phonological level. In turn, this may reveal information on how – i.e., through which kinds of writing tasks – to support children during the development of writing fluency (which is obviously not restricted to the mastery of orthography). Still, we do not know much about the principles by which low-level writing skills may transfer from handwriting to typing or vice versa, and how integrated early writing instruction should be didactically conceptualized. It should be clear from our results that it needs more than mere touch-typing training to bring pupils and students closer to the typing proficiency they need to manage the requirements of our modern digital world.

Finally, the use of a standardized copy task can detect differences in the writing convenience across the different keyboard technologies. Desktop keyboards have different pressure points, and more clearly separated keys, than notebook keyboards, while tablet and smartphone keypads allow for, and call for, different movement patterns, namely touching an area rather than pressing a key (Palin et al., 2019). It is not yet predictable whether traditional keyboards will remain the predominant tool for the production of complex texts, involving the entire range of high- and low-level processes described in the theories and models on written language production (for a review of these models, see Alamargot & Chanquoy, 2001).

Notes

1. As Bayesian data analysis is novel in the domain of writing research, we illustrate central theoretical differences in Appendix B, comparing between a frequentist and a Bayesian analysis in linear mixed-effects models.

2. All reported models include random intercepts for participants and bigrams to account for individual differences between both participants' typing ability and differences between individual bigrams (e.g. frequency, hand combination, adjacency). To avoid over-parametrization of the models we did not include random by-participants slopes for copy-task components (see Baayen, Davidson, & Bates, 2008; Bates et al., 2015a), unless stated differently. Further, all models were fitted with weakly informative regulating priors (see Lambert, 2018). Six thousand iterations (3,000 warm-up) were run for 3 Markov chain Monte Carlo chains. Model convergence was tested via the Rubin-Gelman statistic (Gelman & Rubin, 1992), traceplots and cross-validation (Vehtari, Gelman, & Gabry, 2015, 2017).
3. For an example of how this analysis of the copy task can be used as a diagnostic tool, see Appendix D.

Acknowledgements

For the technical development of the Inputlog copy task, we would like to thank Tom Pauwaert, Eric Van Horenbeeck and Sebastian Fierens. For collecting the Dutch corpus of copy tasks, we would like to thank all the researchers that contributed to this corpus. Especially thanks to Nina Vandermeulen (NWO-LIFT project), Lise Paesen and Catherine Meulemans (FWO/BOF project) and master students of Multilingual Professional Communication (University of Antwerp) for making their copy task data available for this study. We also thank Thomas Quinlan for proofreading an earlier version of this manuscript.

Within the context of the European literacy network (ELN) COST Network (IS1401) and EarlyWritePro (Developing methods for understanding early writing through analysis of process disfluencies) the copy task has been developed in different languages. We thank Lise Fontaine (Cardiff University, UK), Mark Torrance (Nottingham University, UK), Thierry Olive (CNRS at University of Poitiers, France), Esther Breuer (University of Cologne, Germany), Olga Witczak (Adam Mickiewicz University, Poland), Teresa Limpo (University of Porto, Portugal), Anna Sala (University of Barcelona, Spain), Åsa Wengelin (University of Gothenburg, Sweden), Victoria Johansson (University of Lund, Sweden), Gulay Tiryakioglu (University of Lyon 2, France), Vibeke Rønneberg (University of Stavanger, Norway), Anne Sætersdal Myklestad (Western Norway University of Applied Sciences, Norway) and Alessandra Rossetti (University of Antwerp, Belgium).

References

Alamargot, D., & Chanquoy, L. (2001). *Through the models of writing*. Dordrecht: Kluwer. https://doi.org/10.1007/978-94-010-0804-4

Aldridge, M., & Fontaine, L. (2019). Using keystroke logging to capture the impact of cognitive complexity and typing fluency on written language production. In K. Sullivan & E. Lindgren

(Eds.), *Observing writing: Insights from keystroke logging and handwriting* (pp. 285–305). Leiden: Brill. https://doi.org/10.1163/9789004392526_014

Allen, L. K., Jacovina, M. E., Dascalu, M., Roscoe, R. D., Kent, K., Likens, A. D., & McNamara, D. S. (2016). ENTER ing the time series SPACE: Uncovering the writing process through keystroke analyses. In *Proceedings of the 9th international conference on educational data mining (EDM)* (pp. 22–29). https://eric.ed.gov/?id=ED592674

Alves, R. A., Castro, S. L., de Sousa, L., & Strömqvist, S. (2007). Influence of typing skill on pause-execution cycles in written composition. In M. Torrance, D. Galbraith, & L. Van Waes (Eds.), *Recent developments in writing-process research* (Vol. 20, pp. 55–65). Dordrecht-Boston-London: Kluwer Academic Press.

Alves, R. A., Castro, S. L., & Olive, T. (2008). Execution and pauses in writing narratives: Processing time, cognitive effort and typing skill. *International Journal of Psychology, 43,* 469–479. https://doi.org/10.1080/00207590701398951

Baaijen, V. M., Galbraith, D., & Glopper, K. de. (2012). Keystroke analysis: Reflections on procedures and measures. *Written Communication*, *29*(3), 246–277. https://doi.org/10.1177/0741088312451108

Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511801686

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Banerjee, S. P., & Woodard, D. L. (2012). Biometric authentication and identification using keystroke dynamics: A survey. *Journal of Pattern Recognition Research, 7*(1), 116–139. https://doi.org/10.13176/11.427

Bates, D. M., Kliegl, R., Vasishth, S., & Baayen, R. H. (2015a). Parsimonious mixed models. *arXiv Preprint arXiv:1506.04967*. https://arxiv.org/abs/1506.04967

Bates, D. M., Mächler, M., Bolker, B., & Walker, S. (2015b). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Behrns, I., Hartelius, L., & Wengelin, Å. (2009). Aphasia and computerised writing aid supported treatment. *Aphasiology, 23*(10), 1276-1294. doi:10.1080/02687030802436892

Berninger, V. W., & Winn, W. D. (2006). Implications of advancements in brain research and technology for writing development, writing instruction, and educational evolution. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 96-114). New York: Guilford Press.

Berninger, V. W., Yates, C., Cartwright, A., Rutberg, J., Remy, E., & Abbott, R. D. (1992). Lower-level developmental skills in beginning writing. *Reading and Writing: An Interdisciplinary Journal, 4*, 257–280. https://doi.org/10.1007/BF01027151

Berninger, V.W., & Rutberg, J. (1992). Relationship of finger function to beginning writing: Application to diagnosis of writing disabilities. *Developmental Medicine & Child Neurology, 34*, 155–172. https://doi.org/10.1111/j.1469-8749.1992.tb14993.x

Bosman, E. A. (1993). Age-related differences in the motoric aspects of transcription typing skill. *Psychology and Aging*, *8*(1), 87–102. https://doi.org/10.1037/0882-7974.8.1.87

Bouriga, S., & Olive, T. (in press). Is typewriting more resources-demanding than handwriting in undergraduates? *Reading and Writing*.

Brandt, D. (2014). *The rise of writing: Redefining mass literacy*: Cambridge University Press.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, *20*. https://doi.org/10.18637/jss.v076.i01

Chukharev-Hudilainen, E. (2014). Pauses in spontaneous written communication: A keystroke logging study. *Journal of Writing Research, 6*(1), 61–84. https://doi.org/10.17239/jowr-2014.06.01.3

Conijn, R., Roeser, J., & van Zaanen, M. (2019). Understanding the keystroke log: the effect of writing task on keystroke features. *Reading and Writing, 32*(9), 2353–2374. http://doi.org/10.1007/s11145-019-09953-8

Cooper, W. E. (Ed.). (1983). *Cognitive aspects of skilled typewriting*. New York: Springer. https://doi.org/10.1007/978-1-4612-5470-6

Crump, M., & Logan, G. (2010). Hierarchical control and skilled typing: Evidence for word-level control over the execution of individual keystrokes. *Journal of Experimental Psychology: Learning, Memory and Cognition, 36*(6), 1369–1380. https://doi.org/10.1037/a0020696.

Dhakal, V., Feit, A. M., Kristensson, P. O., & Oulasvirta, A. (2018). *Observations on typing from 136 million keystrokes.* Paper presented at the Conference on Human Factors in Computing Systems-Proceedings, Montreal, Canada. https://doi.org/10.1145/3173574.3174220

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*(781), 1–17. https://doi.org/10.3389/fpsyg.2014.00781

Farrell, S. & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press. https://doi.org/10.1017/CBO9781316272503

Feng, L., Lindner, A., Ji, X. R., & Joshi, R. M. (2017). The roles of handwriting and keyboarding in writing: a meta-analytic review. *Reading and Writing: an Interdisciplinary Journal*, 1–31. http://doi.org/10.1007/s11145-017-9749-x

Galbraith, D. & Baaijen, V. M. (2019). Aligning keystrokes with cognitive processes in writing. In K. Sullivan & E. Lindgren (Eds.), *Observing writing: Insights from keystroke loggging and handwriting* (pp. 306-325). Leiden: Brill. https://doi.org/10.1163/9789004392526_015

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Chapman; Hall/CRC.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Gelman, A. & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Gentner, D. R. (1983). Keystroke Timing in Transcription Typing. In W. E. Cooper (Ed.), *Cognitive aspects of skilled typewriting* (pp. 95-120). New York: Springer. https://doi.org/10.1007/978-1-4612-5470-6_5

Gentner, D. R., Larochelle, S., & Grudin, J. (1988). Lexical, sublexical, and peripheral effects in skilled typewriting. Cognitive psychology, 20(4), 524-548.

Grabowski, J. (2008). The internal structure of university students' keyboard skills. *Journal of Writing Research, 1*(1), 27–52. https://doi.org/10.17239/jowr-2008.01.01.2

Grabowski, J. (2010). Speaking, writing, and memory span in children: Output modality affects cognitive performance. *International Journal of Psychology*, *45*, 28–39. https://doi.org/10.1080/ 00207590902914051

Grabowski, J., Weinzierl, C., & Schmitt, M. (2010). Second and fourth graders' copying ability: from graphical to linguistic processing. *Journal of Research in Reading, 33*(1), 39--53. https://doi.org/10.1111/j.1467-9817.2009.01431.x

Graff, D., & Cieri, C. (2003). *English gigaword LDC2003T05*. University of Pennsylvania, Linguistic Data Consortium. Retrieved from: https://catalog.ldc.upenn.edu/LDC2003T05

Graham, S., & R. Harris, K. (2000). The role of self-regulation and transcription skills in writing and writing development. *Educational Psychologist*, *35*(1), 3–12. https://doi.org/10.1207/S15326985EP3501_2

Haas, C. (1989). Does the medium make a difference? A study of composing with pen and paper and with a computer. *Human-Computer Interaction, 4*, 149–169. https://doi.org/10.1207/ s15327051hci0402_3

Hess, S., Mousikou, P., & Schroeder, S. (2020). Double-letter processing in developmental and skilled handwriting production: Evidence from kinematics. *Quarterly Journal of Experimental Psychology*, 1747021820908538. https://doi.org/10.1177/1747021820908538

Hoaglin, D. C., & Iglewicz, B. (1987). Fine-tuning some resistant rules for outlier labeling. Journal of the American Statistical Association, 82(400), 1147–1149. https://doi.org/10.1080/01621459. 1987.10478551

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164. https://doi.org/10.3758/s13423-013-0572-3

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Jiménez, J. E & Hernández-Cabrera, J. A. (2019). Transcription skills and written composition in Spanish beginning writers: pen and keyboard modes. *Reading and Writing*, *32*, 1847–1879. https://doi.org/10.1007/s11145-018-9928-4

Johansson, R., Wengelin, Å., Johansson, V., & Holmqvist, K. J. R.(2010). Looking at the keyboard or the monitor: relationship with text production processes. *Reading and Writing, 23*(7), 835851. https://doi.org/10.1007/s11145-009-9189-3

John, B. E. (1996). TYPIST: A theory of performance in skilled typing. *Human-Computer Interaction, 11*(4), 321–355. https://doi.org/10.1207/s15327051hci1104_2

Kandel, S., Peereman, R., Grosjacques, G., & Fayol, M. (2011). For a psycholinguistic model of handwriting production: testing the syllable-bigram controversy. *Journal of Experimental Psychology: Human Perception and Performance, 37*(4), 1310–1322. https://doi.org/10.1037/a0023094

Kellogg, R. T. (2001). Competition for working memory among writing processes. *American Journal of Psychology*, *114*, *175–191*. https://doi.org/10.2307/1423513

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Boston: Academic Press. https://doi.org/10.1016/B978-0-12-405888-0.00008-8

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, *15*(4), 722–752. https://doi.org/10.1177/1094428112457829

Lambert, B. (2018). *A student's guide to Bayesian statistics*. Sage.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication, 30*(3), 358–392. https://doi.org/10.1177/0741088313491692

Limpo, T., Vigário, V., Rocha, R., & Graham, S. (2020). Promoting transcription in third-grade classrooms: Effects on handwriting and spelling skills, composing, and motivation. *Contemporary Educational Psychology, 61*, 101856. https://doi.org/10.1016/j.cedpsych.2020.101856

Liu, Y., Gelman, A., & Zheng, T. (2015). Simulation-efficient shortest probability intervals. *Statistics and Computing*, *25*(4), 809–819. https://doi.org/10.1007/s11222-015-9563-8

Lindgren, E., & Sullivan, K. (Eds.). (2019). *Observing writing: Insights from keystroke logging and handwriting*. Leiden, NL: Brill. https://doi.org/10.1163/9789004392526

Logan, F. A. (1999). Errors in copy typewriting. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 1760–1773. https://doi.org/10.1037//0096-1523.25.6.176

Logan, G. D. (2003). Simon-Type effects: Chronometric evidence for keypress schemata in typewriting. *Journal of Experimental Psychology: Human Perception and Performance, 29*(4), 741–757. https://doi.org/10.1037//0096-1523.25.6.1760

Logan, G. D. (2018). Automatic control: How experts act without thinking. *Psychological Review*, *125*(4), 453–485. http://doi.org/10.1037/rev0000100

Logan, G. D., & Crump, M. (2011). Hierarchical control of cognitive processes: The case for skilled typewriting. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 54, pp. 1–27). Burlington, MA: Academic Press. http://dx.doi.org/10.1016/B978-0-12-385527-5.00001-2

MacKenzie, I. S. & Soukoreff, R. W. (2003). *Phrase sets for evaluating text entry techniques.* Paper presented at the CHI'03 extended abstracts on Human factors in computing systems. https://doi.org/10.1145/765891.765971

Mangen, A., Anda, L. G., Oxborough, G. H., & Brønnick, K. (2015). Handwriting versus keyboardwriting: Effect on word recall. *Journal of Writing Research*, *7*(2), 227–247. https://doi.org/10.17239/jowr-2015.07.02.1

McCutchen, D. (2000). Knowledge, Processing, and working memory: Implications for a theory of writing. *Educational Psychology*, *35*(1), 13–23. https://doi.org/10.1207/S15326985EP3501_3

McElreath, R. (2016). *Statistical rethinking: A bayesian course with examples in R and Stan*. CRC Press.

Medimorec, S., & Risko, E. F. (2016). Effects of disfluency in writing. *British Journal of Psychology, 107*, 625–650. https://doi.org/10.1111/bjop.12177

Medimorec, S., & Risko, E. F. (2017). Pauses in written composition: on the importance of where writers pause. *Journal of Reading and Writing, 30*(6), 1267–1285. https://doi.org/10.1007/s11145-017-9723-7

Medwell, J., & Wray, D. (2014). Handwriting automaticity: The search for performance thresholds. *Language and Education*, *28*(1), 34–51. https://doi.org/10.1080/09500782.2013.763819

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2015). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123. https://doi.org/10.3758/s13423-015-0947-8

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas – Part II. *arXiv Preprint arXiv:1602.00245*. https://doi.org/10.1111/lnc3.12207

Nottbusch, G., Grimm, A., Weingarten, R., & Will, U. (2005). Syllabic structures in typing: Evidence from hearing-impaired writers. *Reading & Writing, 18*, 497–526. https://doi.org/10.1007/s11145-005-3178-y

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia, 9*(1), 97-113.

Olive, T. (2014). Toward a parallel and cascading model of the writing system: A review of research on writing processes coordination. *Journal of Writing Research, 6*(2), 173–194. https://doi.org/10.17239/jowr-2014.06.02.4

Olive, T. & Kellogg, R. T. (2002). Concurrent activation of high- and low-level production processes in written composition. *Memory and Cognition, 30*, 594–600. https://doi.org/10.3758/BF03194960

Olive, T., & Passerault, J. M. (2012). The Visuospatial Dimension of Writing. *Written Communication*, *29*(3), 326–344. https://doi.org/10.1177/0741088312451111

Ostry, D. J. (1983). Determinants of interkey times in typing. In W. E. Cooper (Ed.), *Cognitive aspects of skilled typewriting* (pp. 225–246). New York/Heidelberg/Berlin: Springer. https://doi.org/10.1007/978-1-4612-5470-6_9

Palin, K., Feit, A. M., Kim, S., Kristensson, P. O., & Oulasvirta, A. (2019). *How do people type on mobile devices? Observations from a study with 37,000 volunteers.* Paper presented at the Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services.

Pinet, S., Dubarry, A.-S., & Alario, F.-X. (2016). Response retrieval and motor planning during typing. *Brain and Language, 159*, 74–83. https://doi.org/10.1016/j.bandl.2016.05.012

Pinet, S., Ziegler, J. C., & Alario, F.-X. (2016). Typing is writing: Linguistic properties modulate typing execution. *Journal of Psychonomic bulletin review, 23*(6), 1898–1906. https://doi.org/10.3758/s13423-016-1044-3

Rapp, B., Purcell, J., Hillis, A. E., Capasso, R., & Miceli, G. (2016). Neural bases of orthographic long-term memory and working memory in dysgraphia. *Brain*, *139*(2), 588–604. https://doi.org/10.1093/brain/awv348

Roeser, J., De Maeyer, S., Leijten, M., & Van Waes, L. (2021 - under review). *Modelling typing disfluencies using Bayesian mixture models*. https://doi.org/10.17605/OSF.IO/Y3P4D

Salthouse, T. A. (1984). Effects of age and skill in typing. *Journal of Experimental Psychology, 113*(3), 345–371. https://doi.org/10.1037/0096-3445.113.3.345

Salthouse, T. A. (1986). Perceptual, cognitive, and motoric aspects of transcription typing. *Psychological Bulletin, 99*(3), 303–319. https://doi.org/10.1037/0033-2909.99.3.303

Sausset, S., Lambert, E., Olive, T., & Larocque, D. (2012). Processing of syllables during handwriting: Effects of graphomotor constraints. *The Quarterly Journal of Experimental Psychology*, *65*, 1–9. https://doi.org/10.1080/17470218.2012.715654

Schöner, G. (2002). Timing, clocks, and dynamical systems. *Brain and Cognition*, *48*(1), 31–51. https://doi.org/10.1006/brcg.2001.1302

Schumacher, G. M., Klare, G. R., Cronin, F. C., & Moses, J. D. (1984). Cognitive activities of beginning and advanced college writers: A pausal analysis. *Research in the Teaching of English*, 169–187.

Sorensen, T., Hohenstein, S., & Vasishth, S. (2015). Bayesian linear mixed models using stan: A tutorial for psychologists, linguists, and cognitive scientists. *arXiv Preprint arXiv:1506.06201*. https://doi.org/10.20982/tqmp.12.3.p175

Stan Development Team. (2015a). Stan: A C++ library for probability and sampling. http://mc-stan.org/

Stan Development Team. (2015b). Stan modeling language user's guide and reference manual. http://mc-stan.org/

Tate, T. P., Warschauer, M., & Kim, Y.-S. G. (2019). Learning to compose digitally: The effect of prior computer use and keyboard activity on NAEP writing. *Reading and Writing*, 1–24. https://doi.org/ 10.1007/s11145-019-09940-z

Van Galen, G. P. (1991). Handwriting: Issues for a psychomotor theory. *Human Movement Science*, *10*(2-3), 165–191. http://doi.org/10.1016/0167-9457(91)90003-g

Van Waes, L., & Leijten, M. (2015). Fluency in Writing: A Multidimensional Perspective on Writing Fluency Applied to L1 and L2. *Computers & Composition, 38*, 79–95. https://doi.org/10.1016/j.compcom.2015.09.012

Van Waes, L., Leijten, M., Mariën, P., & Engelborghs, S. (2017). Typing competencies in Alzheimer's disease: An exploration of copy tasks. *Computers in Human Behavior*, *73*, 311–319. https://doi.org/10.1016/j.chb.2017.03.050

Van Waes, L., Leijten, M., Pauwaert, T., & Van Horenbeeck, E. (2019). A multilingual copy task: Measuring typing and motor skills in writing with Inputlog. *Journal of Open Research Software, 7*(1:30), 1–8. https://doi.org/10.5281/zenodo.2908966

Van Weerdenburg, M., Tesselhof, M., & Van der Meijden, H. (2019). Touch-typing for better spelling and narrative-writing skills on the computer. *Journal of Computer Assisted Learning*, *35*(1), 143–152. https://doi.org/10.1111/jcal.12323

Vasishth, S., Chopin, N., Ryder, R., & Nicenboim, B. (2017). Modelling dependency completion in sentence comprehension as a Bayesian hierarchical mixture process: A case study involving Chinese relative clauses. https://arxiv.org/abs/1702.00564

Veale, J. F. (2014). Edinburgh handedness inventory–short form: a revised version based on confirmatory factor analysis. *Laterality: Asymmetries of Body, Brain and Cognition, 19*(2), 164-177. https://doi.org/10.1080/1357650X.2013.783045

Vehtari, A., Gelman, A., & Gabry, J. (2015). Pareto smoothed importance sampling. *arXiv Preprint arXiv:1507.02646*.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vertanen, K., & Kristensson, P. O. (2011). *A versatile dataset for text entry evaluations based on genuine mobile emails.* Paper presented at the 13th International Conference on Human Computer Interaction with Mobile Devices and Services, Stockholm, Sweden. https://doi.org/10.1145/2037373.2037418

Wagenmakers, E.-J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review*, *114*(3), 830–841. https://doi.org/10.1037/0033-295X.114.3.830

Wallot, S., & Grabowski, J. (2013). Typewriting dynamics: What distinguishes simple from complex writing tasks? *Ecological Psychology, 25*(3), 267–280. https://doi.org/10.1080/10407413.2013.810512

Weigelt-Marom, H., & Weintraub, N. (2018). Keyboarding versus handwriting speed of higher education students with and without learning disabilities. *Computers & Education, 117*(C), 132–140. https://doi.org/10.1016/j.compedu.2017.10.008

Wengelin, Å. (2006). Examining pauses in writing: Theories, methods and empirical data. In K. P. H. Sullivan & E. Lindgren (Eds.), *Computer Key-Stroke logging and Writing: Methods and Applications* (Vol. 18, pp. 107–130). Oxford: Elsevier.

Wing, A. M., & Kristofferson, A. B. (1973). Response delays and the timing of discrete motor responses. *Perception & Psychophysics*, *14*(1), 5–12. https://doi.org/10.3758/BF03198607

Witt, S. T., Laird, A. R., & Meyerand, M. E. (2008). Functional neuroimaging correlates of finger-tapping task variations: an ALE meta-analysis. *Neuroimage, 42*(1), 343–356. https://doi.org/10.1016/j.neuroimage.2008.04.025

Wobbrock, J. O., & Myers, B. A. (2006). Analyzing the input stream for character-level errors in unconstrained text entry evaluations. *ACM Transactions on Computer-Human Interaction (TOCHI), 13*(4), 458–489. https://doi.org/10.1145/1188816.1188819

Wu, C., & Liu, Y. (2008). Queuing network modeling of transcription typing. ACM Transactions on Computer-Human Interaction, Vol. 15, No. 1, Article 6. http://doi.org/10.1145/1352782.1352788

Yamaguchi, M., & Logan, G. D. (2014). pushing typists back on the learning curve: Contributions of multiple linguistic units in the acquisition of typing skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 1–20. http://doi.org/10.1037/xlm0000026

Appendix A: Using the Inputlog copy task

The copy task has been developed as part of Inputlog 8. The task itself is accessible via the 'Record component'; the analyses are provided in the 'Analysis component' (see Inputlog manual for more details). Inputlog is made available for researchers on https://www.inputlog.net (see the download section for more information on the installation procedure).

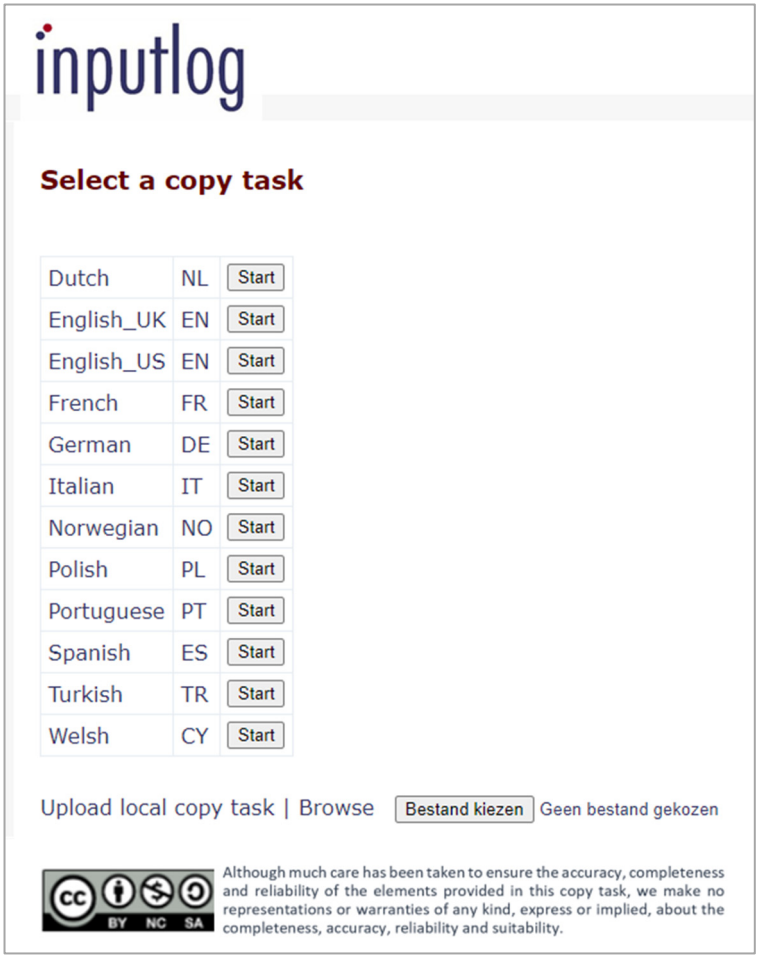

Figure A1: Introductory page of the copy task webtool.

However, as this component is programmed in Javascript, the copy task itself – not the analysis – is also directly accessible as a webtool via https://inputlog.uantwerpen.be/WebSite/copytask/tasks.html (see Figure A1). The

source code is downloadable via GitHub (https://github.com/lvanwaes/Inputlog-Copy-Task). For a more elaborate, technical description, see Van Waes, Leijten, Pauwaert, and Van Horenbeeck (2019).

The default copy task described in this paper is made available in eleven languages. However, if researchers want to translate/transpose the default copy task into another language or want to expand or customize the current copy task, they can use the so called 'Copy task creator', also made available as an integrated tool in Inputlog 8. Moreover, this copy task creator is also made available as an isolated, stand-alone tool (see Github). The copy task creator consists of several building blocks that can be combined into a customized task flow.

Appendix B: Comparing Bayesian and Frequentist linear mixed-effects models

This appendix compares the interpretation of a standard Frequentist linear mixed-effects model analysis and its Bayesian equivalent. In particular, we will see that even though the numeric results of both analyses are relatively similar, their interpretation is fundamentally different.

The advantages of Bayesian data analysis for hypothesis testing are well documented in the literature (Dienes, 2014; Kruschke, 2014; Kruschke, Aguinis, & Joo, 2012; Kruschke & Liddell, 2018; Nicenboim & Vasishth, 2016; Sorensen, Hohenstein, & Vasishth, 2015). Bayesian inference is based on posterior (i.e. statistically inferred) samples which allows a fundamentally different theoretical interpretation than with Frequentist quantities (see for detailed discussions Kruschke, 2014; Lambert, 2018; McElreath, 2016; Nicenboim & Vasishth, 2016). An attractive property of Bayesian statistics is the fact that posterior samples are associated with a probability distribution. In other words, Bayesian inference allows us to determine the most probable population estimate (i.e. *maximum a posteriori*) for each copy-task component or their differences. Along with this estimate, we can calculate the shortest interval containing 95% of the posterior probability mass. This interval is called the 95% Highest Posterior Density Interval (HPDI) and indicates the 95% range that contains the true parameter value. Although frequentist confidence intervals are often mistaken to have similar properties, they do not provide any information about the probability of an inferred parameter value (see Hoekstra, Morey, Rouder, & Wagenmakers, 2014; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2015).

We will illustrate the differences between Bayesian and Frequentist model estimates in a linear mixed-effect model analysis of the copy-task components. The model is fitted on the log-transformed inter-keystroke intervals with copy-task component as fixed effects and random intercepts for participants and bigrams. To evaluate the fit of this model, we fitted an intercept-only model without copy-task component as fixed effect.

The linear mixed-effects model performed in the R package lme4 (Bates et al., 2015b) showed a statistically significant fit ($F(5) = 4{,}152.96$, $p < .001$; AIC = 47,498.07). This model, with copy-task component as fixed effect, rendered a better fit compared to the intercept-only model and was found more informative ($\chi^2_5 = 580.64$, ΔAIC = 572.64). The Bayesian linear mixed effects model with copy-task component as fixed effect rendered a higher predictive performance compared to an intercept-only model (Δelpd = -249.80, $SE = 25.02$). This model comparison provides inference akin to the Frequentist linear mixed-effects model.

The coefficients for each copy-task component estimated by the Frequentist and Bayesian linear mixed effects model ([B]LMM) are shown in Figure B1. The estimate of Bayesian model is the maximum a posteriori, the most probable

parameter estimate, or population mean. The lower and upper bound are the 95% confidence intervals for the Frequentist model and the 95% highest posterior density interval (HPDI) for the Bayesian model. The parameter estimates and the associated intervals show a general difference between the Tapping task, the HF bigram component, and the Sentence task one the one hand, and the LF bigram component and the Consonant copy task on the other hand.

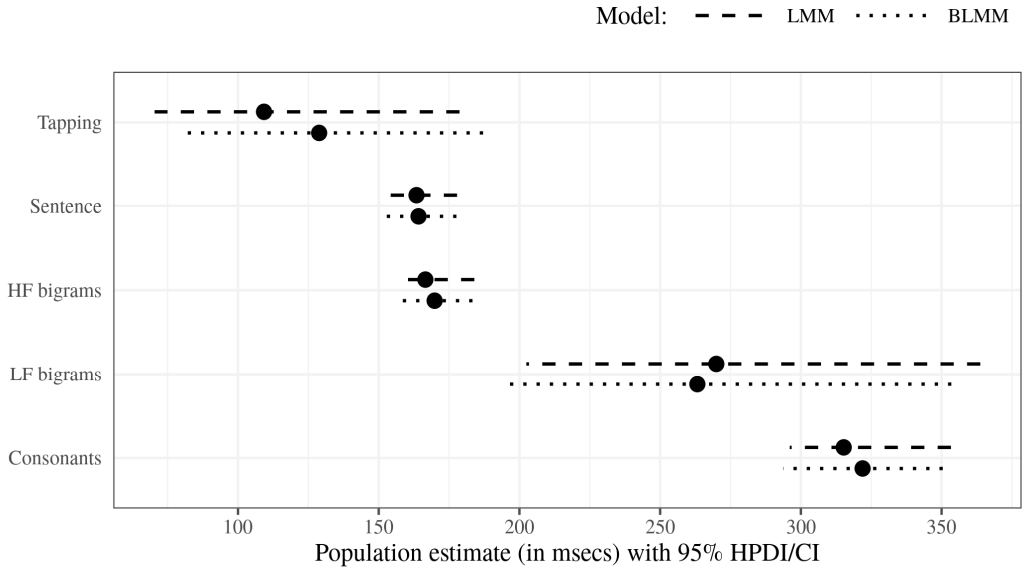

Figure B1: Model estimates of the inter-keystroke intervals (IKI) for each copy-task component estimated in a Bayesian and Frequentist linear mixed effects model ([B]LMM). Dots show the population estimate and intervals show 95% CIs for the LMM and 95% HPDIs for the BLMM.

Overall, the coefficients are numerically very similar. However, the interpretation associated with the two statistical frameworks is crucially different. The estimate of Bayesian model represents the most probable parameter estimate of the unknown population mean. The Frequentist estimate it is not associated with a probability distribution. The same holds for the lower and upper bounds of the error bars. The lower and upper bound for the Frequentist model are the 95% confidence intervals (CI). For the Bayesian model, the interval shows the 95% Highest Posterior Density Interval (HPDI). Although the intervals look very similar the interpretation is different. The HPDI is defined as the shortest interval containing 95% of the posterior probability mass and represents the area in which the largest amount of posterior estimates lie. Bayesian HPDIs, probability/percentile intervals and credible intervals all provide the probability range in which the true parameter value lies with the highest certainty. 95% CIs have a more involved definition and

can be understood to represent the intervals that would contain the true parameter value if we were to repeat an experiment a large, if not infinite, number of times under the same conditions. Therefore, 95% CIs cannot be understood as intervals that contain a true parameter value with a probability of 95%, whereas Bayesian posterior intervals do have this interpretation.

As Bayesian models provide probability distributions of population estimates, we can derive inference directly from the model's posterior, as shown in Figure B2. From these distributions we can calculate the probability of observing keystroke intervals below or above a particular threshold or within a particular range. For example, from the posterior shown in Figure B2 we can determine that in the Tapping task the probability of observing keystroke intervals below 100 ms is 0.16, below 80 ms is 0.02 and below 50 ms is virtually 0. For the Consonants component, the probability of observing keystroke intervals for all of these thresholds is 0. In contrast, the probability to observe keystroke intervals above 350 ms in the Consonants task is 0.03 but 0 in the Tapping task.

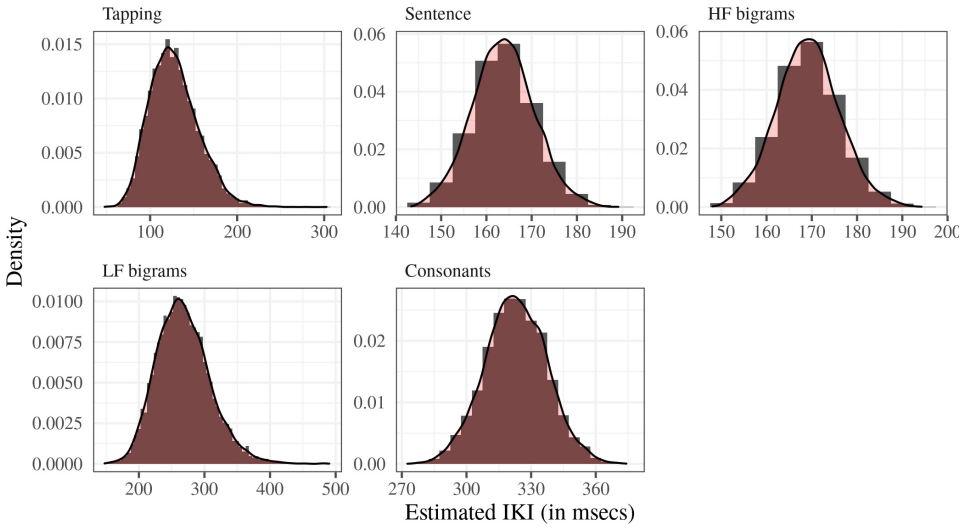

Figure B2: Histograms of the posterior probability distribution of inter-keystroke intervals (IKI) generated from the BLMM.

Appendix C: Data trimming

Deviating interkey intervals were not removed from the analysis for the following reasons. A bottom threshold defined at 30 ms which has been argued to indicate unintentional double strokes or continuous key pressing does not apply because non-targeted bigrams were removed. As can be seen in the left panel of Figure B1, any other threshold on the lower end would need to be arbitrary and is difficult to be justified for the data. As for an upper threshold, following Hoaglin and Iglewicz (1987), we may consider 1827 ms.*

The right panel of Figure C1 illustrates the distribution of data above the determined threshold. This proposal, however, hinges on the assumption that the data follow a normal distribution which we know is not the case. Data are positively skewed which results from the fact that the data are zero bound (Baayen, 2008). In other words, inter-keystroke intervals can be infinitely slow but not faster than or equal to 0 ms. More generally, trimming on the upper end would affect in particular the data of participants that are slow typists (e.g. many elderly participants) and those copy-task components that are more challenging (e.g. the Consonants task) but extreme values in faster components, such as the Tapping task, or in usually fast typists would be disregarded.

Instead of removing those values, we used statistical methods that are capable of accounting for large values (i.e. mixture models presented in the results section of the main text) and that is associating extreme data with a lower posterior probability (see Appendix B).

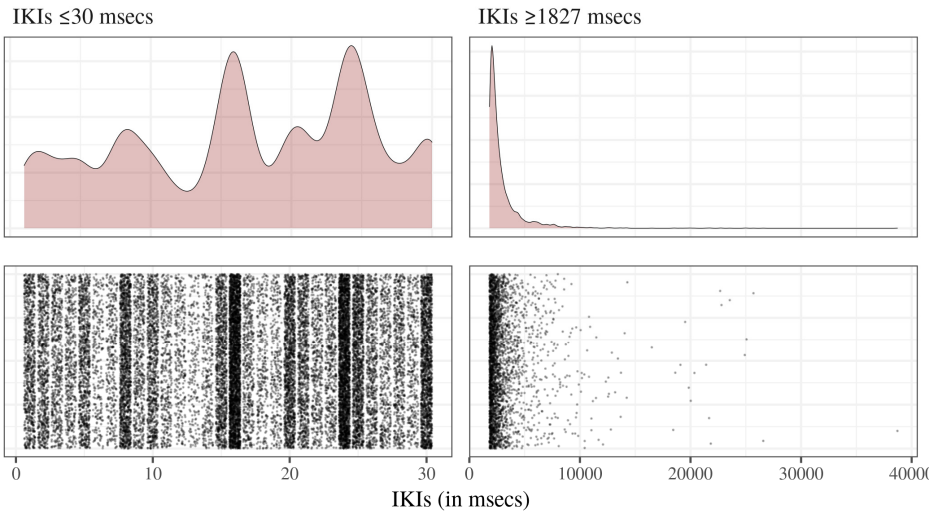

Figure C1: Distribution of inter-keystroke intervals shown on the lower (left panel) and upper (right panel) extreme. Graphs show the density distribution of the data in the top panel and the jittered individual data points on the lower panel.

Note

* Hoaglin and Iglewicz (1987) defined outliers based on the differences between quartile 3 (Q3) and quartile 1 (Q1). This interval is multiplied by factor 2.2. The upper threshold for outliers is then calculated by adding this score to Q3. We applied this formula to the slowest component, i.e. the Consonant task. This resulted in the following calculation: $((Q3 - Q1) \times 2.2) + Q3 = ((772.93 - 293.59) \times 2.2) + 772.93 = 1{,}827.47$.

Appendix D: Copy-task as a diagnostic tool

Mixed-effects models allow us to estimate varying intercepts for each participant. The parameter estimates for each individual can be used to identify those individuals that are relatively fast or slow typists compared to the remainder of the sample. In other words, we can use this analysis to test the typing performance and isolate individuals that vary from the remainder of the sample. Figure D1 shows a caterpillar plot for the deviation estimates for each participant extracted from the mixture model analysis.

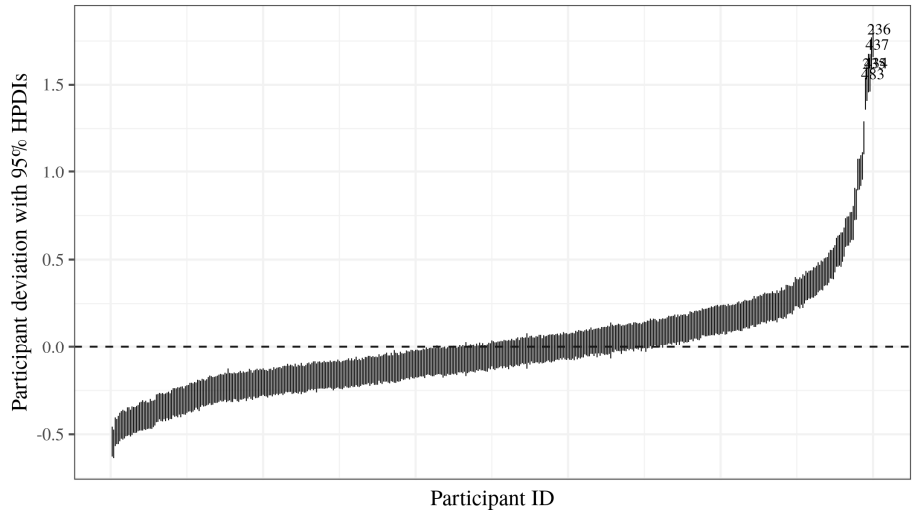

Figure D1: Participant deviation from the model intercept. Each error bar represents the 95% HPDI for one participant. Participants with large deviations (estimated population mean > 3 × SD) were labeled with their participant ID.

For each participant we plotted the 95% HPDI interval showing the deviation from the overall intercept (population mean) marked by the dashed line. As the distribution of these deviations is normal, it allows us to identify those individuals that are relatively slow or fast. In Figure D1 relatively fast individuals are shown on the left side (with shorter IKIs compared to the population mean) and relatively slow individuals are shown on the right (with longer IKIs compared to the population mean). HPDIs that did not fall into the normal distribution of those deviations were labelled with the participants' ID. For example, the slowest typist in the sample shown were participants 236 and 437.

The logic of this analysis can be extended to isolate individuals with difficulty in certain copy-task components, e.g., to identify individuals with specific difficulties on the motor level (Tapping), in lexical typing tasks (HF/LF bigrams, Sentence task),

or tasks that involve eye-hand coordination and a memory component (Consonants task). To illustrate how we can use the copy-task as a diagnostic tool we focused on a group that typically shows more difficulty. We selected the 2.5% oldest participants in the full copy-task corpus (N = 80, 32 males, 48 females; median age = 67 years, range: 60 – 83). We fitted a Bayesian linear mixed-effects model with copy-task component as fixed effect and random intercepts for bigrams and participants with by-participants slopes for each copy-task component. The results of this model can be found in Figure D2. Figure D2 shows a caterpillar plot for the deviation estimates from the overall intercept (shown as dashed line) for each participant plotted against age. Those individuals that showed a relatively slow performance were labeled with their participant ID (similar to Figure D1)

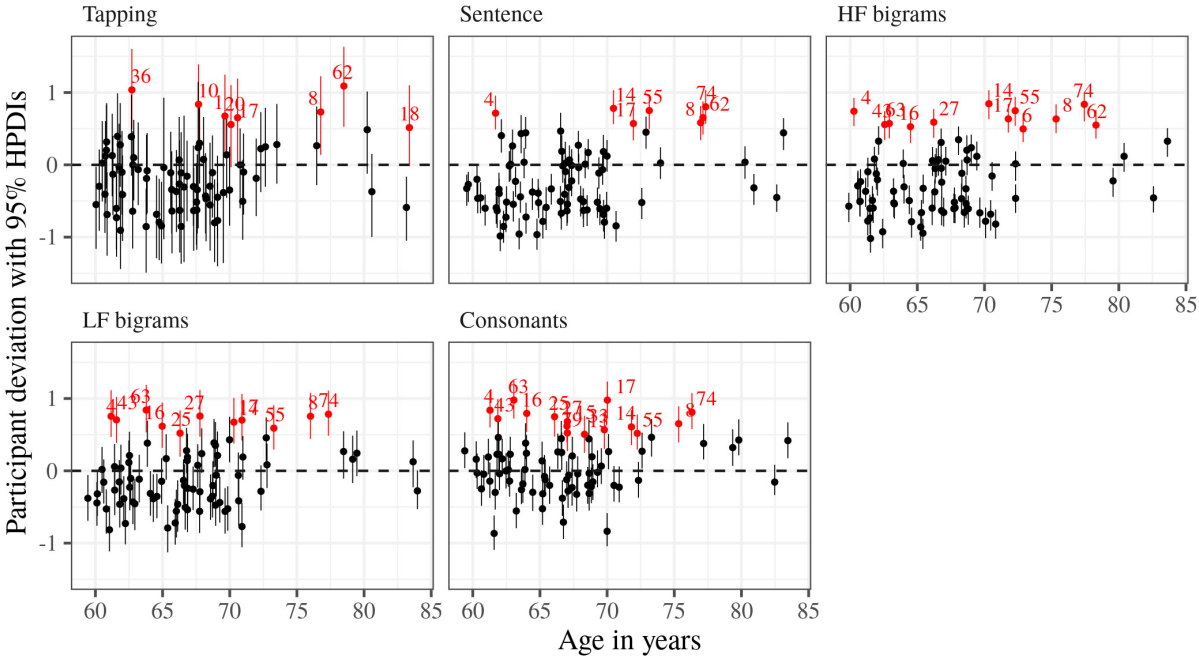

Figure D2: Participant deviation from the overall intercept with 95% HPDIs. Age is shown on the x-axis. For each component, participant IDs are shown for individuals that show slower IKIs compared to the remainder of the sample (estimated population mean > 3 × SD).

For each participant we plotted the 95% HPDI interval showing the deviation from the overall intercept marked by the dashed line. Figure D2 illustrates that participant 74 shows more difficulty in most copy-task components, except the Tapping task. In contrast, participants 10, 18, and 36 show difficulty in the Tapping task but in none

of the other copy-task components. Participant 14 shows difficulty in the Sentence and HF bigrams task only. Participant 6 shows difficulty in the LF bigrams and Consonants task. Participant 8 shows difficulty across all copy-task components. In other words, we can detect participants that have difficulty related to the properties of certain copy-task components.