Reading and Writing (2024) 37:329–357
https://doi.org/10.1007/s11145-022-10284-4

Constructing theoretically informed measures of pause duration in experimentally manipulated writing

Sophie Hall[1] · **Veerle M. Baaijen**[2] · **David Galbraith**[1]

Accepted: 7 March 2022 / Published online: 18 April 2022

Abstract

This paper argues that traditional threshold-based approaches to the analysis of pauses in writing fail to capture the complexity of the cognitive processes involved in text production. It proposes that, to capture these processes, pause analysis should focus on the transition times between linearly produced units of text. Following a review of some of the problematic features of traditional pause analysis, the paper is divided into two sections. These are designed to demonstrate: (i) how to isolate relevant transitions within a text and calculate their durations; and (ii) the use of mixture modelling to identify structure within the distributions of pauses at different locations. The paper uses a set of keystroke logs collected from 32 university students writing argumentative texts about current affairs topics to demonstrate these methods. In the first section, it defines how pauses are calculated using a reproducible framework, explains the distinction between linear and non-linear text transitions, and explains how relevant sections of text are identified. It provides Excel scripts for automatically identifying relevant pauses and calculating their duration. The second section applies mixture modelling to linear transitions at sentence, sub sentence, between-word and within-word boundaries for each participant. It concludes that these transitions cannot be characterised by a single distribution of "cognitive" pauses. It proposes, further, that transitions between words should be characterised by a three-component distribution reflecting lexical, supra-lexical and reflective processes, while transitions at other text locations can be modelled by two-component distributions distinguishing between fluent and less fluent or more reflective processing. The paper concludes by recommending that, rather than imposing fixed thresholds to distinguish processes, researchers should instead impose a common set of theoretically informed distributions on the data and estimate how the parameters of these distributions vary for different individuals and under different conditions.

✉ Sophie Hall
s.m.hall@soton.ac.uk

[1] Southampton Education School, University of Southampton, Building 32, Southampton SO17 1BJ, UK

[2] Center for Language and Cognition Groningen, University of Groningen, Oude Kijk in 't Jatstraat 26, 9712 EK Groningen, The Netherlands

Keywords Pause analysis · Writing processes · Keystroke analysis

Introduction

The immediate attraction of keystroke logging is that it provides a moment-by-moment record of the writer's output as it is produced. It provides us with a record, not just of how often a writer pauses for thought, but also how those pauses are distributed during the course of writing. The analysis of pauses to try to make inferences about the processes involved in writing has consistently been used as a method in writing research from its early days (e.g. Chanquoy et al., 1990; Matsuhashi, 1981; Schilperoord, 2001). Some of the relatively well established findings from such research include: (i) a relationship between fluency and writing quality (Alves & Limpo, 2015; Alves et al., 2008; Chenoweth & Hayes, 2001; Connelly et al., 2006; Medimorec & Risko, 2016; Medimorec et al., 2017; Olive et al., 2009); (ii) a relationship between the frequency and duration of pauses and units of the text—typically pauses are more frequent and longer at more global text boundaries, increasing as one moves up from within-word boundaries, through boundaries between words and sentences to paragraph boundaries (Alamargot et al., 2007; Baaijen et al., 2012; Medimorec & Risko, 2017; Spelman Miller, 2000; Wengelin et al, 2009); and (iii) a relationship between the complexity of the writing task and the frequency and duration of pauses (Beauvais et al., 2011; Medimorec & Risko, 2017; Van Hell et al., 2008). An important qualification here is that these relationships—particularly those between fluency and writing quality—may be strongly moderated by the age and experience of the writers. This is analogous to the way that the relationship between decoding skill and reading performance varies with age: typically, reading comprehension is strongly determined by decoding skill in the early years, but this relationship declines and is replaced by higher level inferential skills as readers develop (Cain, 2010). It remains an open question how fluency—and pausing more generally—relates to writing performance when writing goes beyond knowledge-telling (Bereiter & Scardamalia, 1986) and involves more complex problem-solving within and across a number of drafts.

Research in this field has been dominated by a focus on "cognitive" pauses – pauses above a certain threshold which are assumed to represent higher-level reflective thought—and their relationship with relatively undifferentiated features of the text as a whole. Schilperoord (2001) has questioned whether such an approach can be used to investigate the underlying cognitive processes involved:

> if the writing process that one wants to study is characterized by intensive problem-solving and massive editing on the part of the writer, with actual language production being but one aspect of these processes, it [may] prove impossible to relate pause data to ongoing processes (Schilperoord, 2001, p. 67).

We have characterised this as a problem of alignment (Baaijen et al., 2012; Galbraith & Baaijen, 2019). In this paper, after discussing some of the theoretical and

methodological issues in addressing this problem, we focus on two specific areas. First, we describe the procedures we use to align calculations of pause durations more directly with underlying cognitive processes. We focus particularly on the need to construct reproducible and transparent measures of pause durations, and provide the Microsoft Excel scripts that we used to do this. Second, we demonstrate how mixture modelling (McLachlan & Peel, 2000) can be used to identify structure within the distributions of pause times at different text boundaries and evaluate the extent to which such models provide a better fit to the data at these locations.

Contrasting approaches to pause analysis: from the thinking behind the text to text production

The overwhelming emphasis on the analysis of cognitive pauses in research on writing is partly a consequence of the theoretical assumptions made by the classic cognitive models of the writing process (Bereiter & Scardamalia, 1987; Flower & Hayes, 1980; Hayes, 2012). These models focus on the reflective processes carried out during planning and revision and tend to assume that text production is a relatively passive process responsible for translating the output of higher-level conceptual processes into words. A natural implication is that one should focus on identifying such instances of reflective thought and assess how these vary depending on characteristics of the writer (their age, for example), of different text boundaries (between sentences or within words, for example) and different writing tasks (argumentative or narrative, for example). One of the virtues of this approach is that there is no need to examine every transition between keys. One can simply identify whether the transition time between two keystrokes exceeds a defined threshold and classify these as instances of pauses during writing. One can then examine how these reflective events are distributed within the text or use retrospective verbal protocols to analyse the contents of these episodes.

There are three main difficulties with this approach. First, pauses are largely defined in terms of keystrokes rather than by their underlying cognitive or linguistic function. For example, Inputlog (Leijten & Van Waes, 2013) classifies all the transitions related to full stops (before < FULL STOP >, after < FULL STOP >, and after < SPACE > key) as between sentence transitions. Frequently, it is unclear whether researchers aggregate these into an overall pause time or treat these as independent instances of potential pauses (Baaijen et al., 2012). This clearly affects the number of pauses identified in a log. Second, there is no agreed threshold for deciding between cognitive and non-cognitive pauses (Baaijen et al., 2012; Chenu et al., 2014). What is more, this arguably may vary for different writers. Typically, a conservative approach is taken, and a relatively high threshold is used—for example two seconds—and it is assumed that, although any individual pause might simply reflect a momentary distraction or other off-task thought, these pauses, as a class, reflect higher level reflective thought (Leijten & Van Waes, 2013). Third, by definition, the use of a threshold means that below-threshold processes are ignored, and hence that as a method this cannot provide much information about the more transient processes involved in formulating thought in language (Baaijen et al., 2012;

Chenu et al., 2014). In itself, this is not necessarily an issue, particularly if this is regarded as a relatively autonomous but passive component of the writing process. However, it is much more problematic when text production is treated as a more active process of content generation and the distinction between higher level thought and the formulation of thought in language is less distinct.

The dual-process model (Galbraith, 2009; Galbraith & Baaijen, 2018) claims that there are two distinctively different types of processes involved in writing. The first type corresponds to the reflective thought emphasized by the classic cognitive models of writing, and varies between a knowledge-telling process, in which content is retrieved directly from episodic memory in accordance with its associative structure, and a knowledge-transforming process, in which content retrieval is guided by, and evaluated and reorganised in order to satisfy, communicative goals. The second type corresponds to the translation component of classical models, but the dual-process model characterises this as an active knowledge-constituting process in its own right, in which content is synthesized according to constraints within semantic memory, rather than as a passive process of translating knowledge retrieved from episodic memory. For present purposes, the key feature of the model is that text production is assumed to vary in the processes it involves, and in the effects that it has, depending on how it interfaces with higher level planning processes. This calls into question whether the underlying processes will be reflected in measures based simply on cognitive pauses above a certain threshold and highlights that pause durations may have very different interpretations depending on the writing context within which they occur.

To briefly illustrate this, consider some of the findings of a recent study by Baaijen and Galbraith (2018). This study was carried out with undergraduate students who were asked to write argumentative essays about a current affairs topic under either outline or synthetic planning conditions. In the outline planning condition, participants created hierarchically organised plans for their essays prior to writing. In the synthetic planning condition, participants were instructed to write down a single sentence summing up their overall response to the essay topic. However, they were not allowed to write an explicit plan for their essays. The participants were asked to provide subjective ratings of their understanding of the topic before and after writing and the quality of the final texts was rated by two independent markers. Keystroke logs were collected during writing and two composite measures of writing processes were constructed: (i) Global linearity, which distinguished between the linear and non-linear production of text, and (ii) Sentence production, which distinguished between spontaneous and controlled sentence production, and which included, among other indicators, measures of the mean duration of linear transitions between sentences and words. Crucially, for present purposes, these "pause" measures were defined in terms of whether they reflected linear transitions between units rather than whether they exceeded a threshold.

Two features of the findings are important in the present context. First, the relationships between the sentence production variable and both text quality and the development of understanding varied depending on the type of planning carried out in advance of writing. Spontaneous sentence production (short pauses between units of text production and high revision levels at the leading edge) was associated with increased

understanding when it followed synthetic planning but not outline planning. Similarly, text quality was unrelated to the sentence production variable when it followed outline planning but positively related to controlled sentence production (longer pauses at grammatical boundaries and between words and short bursts of text production) when it followed synthetic planning. Note that these results were observed in combination with revision of the global structure of the text. The critical methodological implication of these findings is that, although the sentence production measure did show relationships with both the development of understanding and the quality of the text, it only did so when the form of advance planning was manipulated. Under natural conditions, without this experimental manipulation, there would have been no apparent relationship between the sentence production measure and either outcome measure. Second, to capture the variation in how participants produced their text, it was crucial to estimate all the linear transitions between the units of text. If the authors had counted the number of pauses above a threshold (say 2 s), and then had only included these pauses within their analysis, there would have only been a categorical distinction between the presence or absence of "cognitive" pauses, rather than the more nuanced measure of controlled versus spontaneous sentence production.

The first point that we want to emphasize, therefore, is that in order to analyse keystroke logs productively, we need to go beyond simple measures of fluency or counting frequencies of threshold-determined cognitive pauses. We also need to employ these measures in theoretically motivated experimental designs—it is tempting with the large data sets made available from keystroke logging to imagine that one might simply rely on data mining to reveal, bottom-up, the underlying patterns in the process.

Constructing reproducible measures of pauses in text production

For present purposes, the main implication we draw from Baaijen and Galbraith's (2018) findings is the value of assessing all the transitions between units of text rather than focussing only on "cognitive pauses" above a given threshold. Doing so, however, is not a straightforward matter, particularly if one is concerned with identifying the underlying cognitive processes involved (Baaijen et al., 2012; Galbraith & Baaijen, 2019). We briefly summarize some of the main issues below. In addition, partly because of the detail required for full transparency and the problems of providing these details in length-limited articles, it is often unclear precisely how different keystroke measures have been operationalised. Given the growing awareness over the past decade or so of the need for transparency and reproducibility in research practices (Aarts et al., 2015; Munafò et al., 2017) this is particularly problematic. In the interests of greater transparency, we provide scripts for all the analyses that we report.

Defining pauses

There is wide variation in the literature in the thresholds used to define pauses. Furthermore, it is often hard to identify precisely how pauses have been measured. In handwriting, the pause duration between words is simply the time between the end of the final letter of a word and the beginning of the initial letter of the next word.

For keyboarding, however, the equivalent transition has two components: the time between the typing of the final letter and the space bar, and the time between the typing of the space bar and the initial letter of the next word. Baaijen et al. (2012) argued that these two intervals should be combined to form a single measure of the transition time between two words, on the grounds that this interval is more representative of the underlying cognitive process than either of its two sub-components [see Conijn et al., (2019) for evidence that measures of such cognitively-informed intervals are more sensitive to task differences than neutral measures of the separate interkey intervals which constitute the transition between words]. Similar issues arise with calculating pauses at sentence and paragraph boundaries. In both these cases, Baaijen et al. (2012) argue that a composite time should be calculated. It is worth noting here that, although this may be less of an explicit problem for approaches that search for cognitive pauses above a certain threshold, and then locate where these occur within the text, it nevertheless remains an issue. Researchers who identify transition times before and after the < SPACE > bar as separate instances will systematically underestimate the frequency of pauses compared to researchers who treat pauses between words as the sum of the two transition times, despite using identical thresholds. Typically, however, the procedures involved in calculating these measures are not clearly specified, and it is impossible to work out how pauses have been operationalised in the analysis. In the analyses that follow, we provide explicit definitions for each transition time along with Excel scripts for automatically calculating their duration.

The context within which pauses occur: linear and event transitions

Baaijen et al. (2012) argued further that, in order to gain more accurate estimates of the characteristics of text production and how these vary between individuals, one needed to distinguish full text production carried out at the leading edge of the text from other forms—such as text produced as part of revision, inserted as part of explicit planning or as an overall title for the text. Thus, rather than taking an undifferentiated keystroke log and simply identifying the pauses above a certain threshold, one should first differentiate between forward text production and other forms. This depends to a certain extent on the purposes of the research, and we should stress that this is not a judgement about the relative importance of the different components of the writing process: a lengthy pause during revision may be an important indicator of a writer's processing. The point is that such pauses are not an indicator of the sentence production process, and hence should not be included as an indicator of that component of the writing process, though other indicators of revision are included in the composite scale used by Baaijen and Galbraith (2018).

Given an initial sorting of the keystroke log into different functional elements, the key distinction that Baaijen et al. (2012) made was between *linear* and *event* transitions. Linear transitions occur when the movement between two units is uninterrupted. Note that, for them, the < SPACE > bar would not constitute an interruption; indeed, they also excluded some other low level mechanical elements. Event transitions, by contrast, occur whenever the transition between units is interrupted by some other activity (e.g., a revision or movement away from the leading edge to

insert new text). This distinction serves two functions. First, it provides a measure in its own right. Important elements of the global linearity measure used by Baaijen and Galbraith (2018) consisted of the percentage of linear transitions between units of text (and note that, in their study, global non-linearity—i.e. a higher proportion of events among other indicators—was associated with higher text quality and greater development of the writer's subjective understanding). Second, it isolates the transitions over which estimates of pause time are calculated. Baaijen et al. (2012) argued that, although individual instances of such pauses might be a consequence of a miscellany of processes, when aggregated across a text, they provided indicators of the planning time typically devoted to different components of the writing process. In common with many other researchers, they found that these systematically varied in duration depending on the units between which they occurred (between paragraphs, sub-sentences, sentences, words, and within words).

Emergent structure within pause distributions: mixture modelling

The factors that we have considered so far involve top-down decisions about pause analysis. Baaijen et al. (2012) also suggested that mixture modelling (McLachlan & Peel, 2000) could be used as a strategy to explore the distribution of linear pauses at different locations within a keystroke log (see Chenu et al., 2014; Hird & Kirsner, 2010; Kirsner, Dunn & Hird, 2005; Little et al., 2013, for similar applications to the analysis of pauses in handwriting and in speech). Mixture modelling is essentially a form of cluster analysis. Using a bottom-up, data-driven approach, mixture modelling allows the researcher to evaluate how many sub-components can be distinguished within a distribution. These components can all be the same type of distribution, or they can vary. A top-down approach to mixture modelling can also be used, in which the researcher constrains the number of components within the model to test whether the data fits that number of theorised distributions appropriately. Baaijen et al. (2012) found that, even after log-transformations of positively skewed pause data, there was evidence of distinctions between the types of pause at different locations within the text. For example, using mixture modelling, they found that linear pauses between words could be sorted into three components, which they hypothesised might represent three underlying processes: word retrieval, phrase planning and higher-level reflective processing. Furthermore, these distinctions were apparent at intervals well below the thresholds customarily used in pause analyses.

More recent research using mixture modelling (Almond, Deane, Quinlan, Wagner, & Sydorenko, 2012; Guo et al., 2018; Roeser et al., 2021; Van Waes et al., 2021) has varied in whether it analyses pauses at different boundaries separately, or aggregated across the whole text, but has typically modelled these as mixtures of two log-normal distributions. Almond et al. (2012) suggest that shorter pauses may reflect the mechanics of typing, whilst longer pauses capture deeper processes such as spelling, word choice and critical thinking. Guo et al. (2018) suggest that most keystroke pauses represent processes relating to the fluency of word-finding, spelling and keyboard skills, but that longer pauses, particularly at the between-word, -sentence, and -paragraph level might represent more complex processes such

as sentence-level planning. Research by Van Waes et al. (2021) and Roeser et al. (2021), using copy tasks and focussing more specifically on typing skill, has presented robust evidence that inter-key intervals in such tasks can be treated as a mixture of two distributions representing fluencies and disfluencies.

In this paper, we will use mixture modelling to examine the presence of such distributions at different text locations in a new sample of keystroke logs of free text production and explicitly test whether these provide a better fit to the data than single lognormal distributions. Given Baaijen et al.'s (2012) finding that linear transitions between words can be modelled as a mixture of three lognormal distributions, we will also compare the relative goodness of fit of two- and three-component distributions at these text locations.

Aims

Our overall aim in this paper is to describe the procedures we used to identify and analyse pauses extracted from keystroke logs. We focus specifically on the analysis of pauses related to the text production component of writing. In doing so, we follow the same rationale as Baaijen et al. (2012) and Baaijen and Galbraith (2018) but focus exclusively on pauses rather than on the broader scale which they constructed to quantify variation in text production processes. Furthermore, we focus on discussing the issues that arise in analysing pauses for this purpose rather than on presenting the results of inferential tests of hypotheses.

Our first aim is to introduce a framework for defining and calculating pause times that reflect text production processes rather than other components of the writing process. In order to promote reproducibility and transparency for these procedures, we provide scripts for the automated calculation of these pause times, and evaluate the extent to which such automatic processes need to be complemented by manual identification of intervals.

Second, we demonstrate how these pause durations can be isolated by (i) separating first-draft text-production from other types of text production (titles, explicit planning and post-draft revision) and (ii) distinguishing between linear and non-linear intervals between units of text.

Finally, having established how these pauses can be identified we demonstrate how Gaussian mixture models can be used to identify further distinctions between pauses. We examine, first, whether multicomponent models provide a better fit to the data than single distributions and, second, the possible interpretations that can be given to these distributions.

The data set

The data which we will use to illustrate these methods and procedures were collected as part of a wider project investigating the effects of writing beliefs and planning strategies on writing outcomes. We collected keystroke data using Inputlog Version 8 (Leijten & Van Waes, 2013). The keystroke data were from a timed

essay-writing experiment, which 32 university students have so far completed (all under face-to-face conditions). The participants all had English as a first language and were asked to write an argumentative essay in 30 min discussing the pros and cons of either veganism or social media. They were randomly assigned to one of two conditions.

In the outline planning with complete drafting condition, participants were asked to type a well-organised essay with accurate spelling. Prior to writing their essay, they were given 5 min to create a handwritten organised essay outline which indicated their opinion, main ideas and the order that they were to go in. They were allowed to refer to this plan while writing the essay.

In the synthetic condition, participants were asked to write a rough draft of an essay. Prior to writing, the participants were given 5 min to work out what they thought about the essay question and were instructed to sum this up in a hand-written single sentence. Importantly, they were not allowed to make a structured written plan to inform their essay writing. However, they were allowed to refer to their sentence as often as they liked throughout writing. Throughout the 30 min of writing, they were instructed to discuss the question with themselves, writing down their thoughts as they occurred spontaneously, then forming a conclusion. They were told not to worry about how well-formed their texts were, instead focusing directly on expressing their thoughts as they unfolded. Additionally, they were told not to worry about spelling.

This design replicated most features of the study by Baaijen and Galbraith (2018) with the important exception that in the previous study both planning groups were instructed to write well-formed texts, whereas in the present study the participants in the synthetic planning condition were instructed to write a rough draft.

A reproducible framework for calculating pause durations

Ideally, the calculation of pause durations would be completely automated. Typically, keystroke logging programs, such as Inputlog (Leijten & Van Waes, 2013) and Scriptlog (Andersson et al., 2006), provide a specification of each keystroke, the duration of a keypress and the duration of the intervals between each key press. We will focus on Inputlog here.

As our paper is demonstrational, rather than providing a full-scale analysis of our keystroke data, we focus on linear pauses at different pause boundaries (within-words, between-words, between-subsentences, and between-sentences), and describe the steps that we took to conceptualise, identify, calculate, and analyse these types of linear pauses. However, the method we used to develop our framework can be applied to other writing process features as well.

To start with, we set out clear conceptual definitions for how we defined linear pauses. The purpose of these explicit definitions was to help make our methods of coding transparent and reproducible. Table 1 exhibits our conceptual definitions.

Whilst the definitions presented in Table 1 show the conceptual underpinnings of how we describe linear pauses, they do not explain how we calculated linear pauses. Thus, we also decided on a framework for calculating linear pause

durations based on the collective time of the pause between and after the pause marker. Table 2 gives an example of how we calculated the between-sentence pause times, but our complete linear pause calculation framework is available in our supplementary materials (https://osf.io/r53h2/).

To calculate the pause durations identified in Table 2, we created Inputlog general analysis files for each individual participant's keystroke log data. These files give information about the start and end time of key presses, the length of time a key is held down for (action time), and the empty time between key presses (pause time). Inputlog saves these files in xml format, which makes them easy to read into multiple different programs. We are aware that many writing researchers use Microsoft Excel to prepare and code their data, before analysing with other statistical software, and so to keep in line with this, we exported the xml files into Microsoft Excel. It is important to note here that Windows and Macintosh versions of Excel read xml data differently, which can result in a different layout of the data. We mention this because to calculate when linear pauses occurred in each participant's data, and the length of their linear pause times, we scripted a series of VBA macros within Excel, based on the rules in our linear pause calculation framework. These macros work on data that is structured in a specific way. More information about the requirements for the macros and the scripts for the macros themselves can be found in our supplementary materials (https://osf.io/r53h2/).

Table 1 Our conceptual definitions of linear pauses

Keystroke feature	Conceptual definition
Linear pause	A clean, 'pure' transition between keystrokes. That is, there is only forward progression and no instances of revision, mouse movements, or insertion and edits away from the leading edge
Linear within-word pause	The time between the letters pressed at the within-word level, when there are no instances of revision, mouse movements, or insertion and edits away from the leading edge
Linear between-word pause	The time between the end of a word and the beginning of the next word when there are no instances of revision, mouse movements, or insertion and edits away from the leading edge
Linear sub-sentence pause	The time between the end of a word that is followed by a comma, and the start of the next word that is preceded by the same comma, when there are no instances of revision, mouse movements, or insertion and edits away from the leading edge
Linear between-sentence pause	The time between the end of a sentence and the beginning of the next sentence, when there are no instances of revision, mouse movements, or insertion and edits away from the leading edge
Non-linear event	A transition between keystrokes (at the within-word, between-word, sub-sentence, or between-sentence boundary) that is interrupted by a revision or movement to another location within the text

Table 2 A framework for the calculation of linear between sentence pause-times

Keystroke feature	Keystroke rule	Calculation of pause time
Linear between-sentence pause	**4 KEYSTROKE COMBINATIONS** <LETTER>/<NUMBER>/<QUOTES>/<APOSTROPHE> <FULL STOP> <SPACE> <LETTER>/<NUMBER>/<APOSTROPHE> **5 KEYSTROKE COMBINATIONS** <LETTER>/<NUMBER>/<QUOTES>/<APOSTROPHE> <FULL STOP> <SPACE> <COMBINATION KEY> [b] <LETTER>/<NUMBER>/<QUOTES>/<APOSTROPHE> OR <LETTER>/<NUMBER>/<QUOTES>/<APOSTROPHE> <FULL STOP> <SPACE> <SPACE> <LETTER>/<NUMBER>/<QUOTES> **6 KEYSTROKE COMBINATIONS** <LETTER>/<NUMBER>/<QUOTES>/<APOSTROPHE> <FULL STOP> <SPACE> <SPACE> <COMBINATION KEY> <LETTER>/<NUMBER>/<QUOTES>/<APOSTROPHE> **5 KEYSTROKE COMBINATION WHEN FIRST SENTENCE ENDS IN ? OR !** <LETTER>/<NUMBER>/<QUOTES> <COMBINATION KEY> <?> OR <!> <SPACE> <LETTER>/<NUMBER>/<APOSTROPHE> **6 KEYSTROKE COMBINATIONS (? OR !)** <LETTER>/<NUMBER>/<QUOTES>/<APOSTROPHE> <COMBINATION KEY> <?> OR <!> <SPACE> <COMBINATION KEY <LETTER>/<NUMBER>/<QUOTES>/<APOSTROPHE> OR <LETTER>/<NUMBER>/<QUOTES>/<APOSTROPHE> <COMBINATION KEY> <?> OR <!> <SPACE> <SPACE> <LETTER>/<NUMBER>/<QUOTES> **7 KEYSTROKE COMBINATIONS (? OR !)** <LETTER>/<NUMBER>/<QUOTES>/<APOSTROPHE> <COMBINATION KEY> <?> OR <!> <SPACE> <SPACE> <COMBINATION KEY> <LETTER>/<NUMBER>/<QUOTES>/<APOSTROPHE>	Difference between the start time of the last character [a] in the previous sentence and the start time of the first character in the new sentence

The keystroke combinations are based on a QWERTY keyboard on a Windows computer

[a] 'character' is a letter, number, quotation marks or an apostrophe

[b] the combination key can fall anywhere between the full stop and first character of the next sentence. This applies to all given keystroke combinations

Linear versus non-linear transitions between units of text

After an initial run of our macros (which we describe in more detail later in this section), we assessed the extent to which these accurately distinguished linear and non-linear transitions between units of text. We illustrate this analysis for the sentence boundary. This distinction is key to constructing a valid measure of the pause durations which we have argued can be taken as indicators of planning during text production and is also a key indicator of the global linearity measure used by Baaijen and Galbraith (2018). However, it is not a distinction that is made automatically by Inputlog. Inputlog classifies the combination of keystrokes surrounding a < FULL STOP > individually as between (before and after) sentence pauses but leaves it to the researcher to decide whether to sum this collection of pauses to reflect an overall transition time. Furthermore, it does not distinguish between linear and non-linear transitions: the transition time before the capital letter indicating a new sentence is classified as a before-sentence pause regardless of whether it directly follows the < SPACE > key or occurs following an extended excursion elsewhere in the text. This means that researchers who simply sum the duration of the transitions classified by Inputlog as before or after sentences will dramatically underestimate the average amount of time elapsing between the ending of one sentence and the beginning of the next. Alternatively, researchers who, like us, calculate the transition time from the final letter of the previous sentence to the initial letter of the succeeding sentence will create an extremely amorphous indicator if they make no distinction between linear and non-linear transitions between sentences.

To demonstrate the importance of this distinction, we inspected the number of linear between-sentence pauses versus the number of non-linear sentence pauses. The number of linear pauses were identified via our macros, whereas the number of non-linear pauses were identified manually due to the large variety of ways they could be formed. This revealed that, in the synthetic condition, the mean percentage of linear sentence pauses out of all sentence pauses was 55.09% (SD = 18.64). In the outline condition, 55.43% (SD = 19.76) were linear out of all sentence pauses. In our view, it is essential to distinguish between linear and non-linear transitions between units of text, not just between sentences as we have demonstrated here, but also between other units of text. Doing so enables us both to calculate pause times which are more likely to reflect the underlying cognitive processes involved in formulating text and to create more global indicators of the linearity of the writing process.

It is important to stress here that non-linear boundaries are not just something to be filtered out in order to create purer measures of pause duration but also that they reflect multiple processes in their own right. For example, in our data, we found instances where writers moved away from the leading edge to insert new text, but also sometimes just moved away from the leading edge to scroll back through previously written text. Sometimes, a non-linear between-sentence event involved the writer staying at the leading edge and ending the sentence but then deleting the end of the sentence and replacing it with new content. Similarly, we saw participants who ended a sentence but then deleted their < FULL STOP > and replaced it with a < COMMA > . This brings into question whether a non-linear pause such as this should be described as sentence level or sub-sentence level.

The necessity of combining automated coding of keystrokes with manual inspection

The benefit of using macros to identify linear pause locations and linear pause times is that the analysis is automated, which makes the overall analytical process standardised and fast. Additionally, the code can be made openly available so that other researchers can implement the same analysis or edit the code to perform a different variation of the analysis. The development of our macros was an iterative process. We developed code and tested the code on our participants' keystroke logs based on our linear pause calculation framework. We then manually checked all participants' data to ensure that the macros were correctly identifying pauses as we defined them. This screening enabled us to identify potential linear pauses that we had not identified in the first version of our calculation framework and its associated macros. It also enabled us to identify cases which were accurately excluded but which called into question our definition of the boundary between linear and non-linear transitions. We present examples of each below.

An example of a linear between-sentence pause that was only identified after manual screening of the data is shown in Fig. 1.

The participant shown in Fig. 1 has come to the end of a sentence and linearly progresses straight onto the next sentence. However, rather than using the standard <FULL STOP>, <SPACE>, <LSHIFT>, <LETTER> sequence, they have pressed the <LSHIFT> key directly after the <FULL STOP>. This would have not been identified as a linear between-sentence pause within our original macros and, interestingly, was not picked up as a sentence transition by Inputlog either, but rather as a word-level transition. This happened for multiple other participants. Thus, if we had relied on our original macros or the automated pause output from Inputlog, we would have not included several linear between-sentence pause cases in our further analysis, which could have affected our mixture-model results.

Manually screening the keystroke data allowed us to identify cases such as the one described here. Based on these identifications, we were then able to improve the macros from their original state. Ultimately, the development of these macros (which is still an ongoing process) meant that we could automate some of our analysis and pick up features not identified in the Inputlog output. However, macros cannot be relied on exclusively to analyse pause data. Macros and scripts in any programming language are rule based so identification of keystroke features via rule-based scripts alone necessarily means that only features conforming to those rules are identified. However, there are complicated keystroke combinations which conceptually meet the definition of a writing feature, such as a linear pause, but for which it is hard to write a general script. There are also boundary cases where individual decisions must be made. An example of this found in our keystroke data follows.

For Participant 15 in the outline condition, we observed 32 sentence transitions. However, only two of them were linear between-sentence pauses (as calculated by our macros). Further manual inspection of their data revealed that in most cases where the participant was ending a sentence, they were pressing the <SPACE> key, followed by the <BACKSPACE> key before pressing

output	startTime	endTime	actionTime	pauseTime	pauseLocationFull
i	569063	569123	60	136	BEFORE WORDS
t	569123	569212	89	60	WITHIN WORDS
.	569282	569362	80	159	AFTER WORDS
LSHIFT	569411	569560	149	129	COMBINATION KEY
SPACE	569451	569521	70	40	AFTER WORDS
T	569501	569610	109	50	BEFORE WORDS
h	569591	569670	79	90	WITHIN WORDS
i	569739	569809	70	148	WITHIN WORDS
s	569839	569890	51	100	WITHIN WORDS

Fig. 1 An example of a linear between-sentence pause that was not picked up in our original macros and was also not identified by Inputlog due to the non-standard order of keys pressed at this boundary

the < FULL STOP > key. Due to the short duration of these key presses, it was apparent that the < SPACE > and < BACKSPACE > were being pressed out of habit, and so rather than reflecting a conscious revision, we concluded that these were just automatic key presses, reflecting the typing motor-skills of the participant.

This led to a discussion about whether to change our conceptual definition of a linear between-sentence pause to include minor revisions (as in Baaijen & Galbraith, 2018; Baaijen et al, 2012). For this paper, we decided against this because we would have had to identify cut-off points for the number of backspaces that could be included within a linear-between sentence pause (and indeed any other linear pause type), and the duration of these key presses. We decided to postpone this until we had examined a larger data set. It may, in general, be a decision best left to individual researchers in specific contexts.

We think that the provision of comprehensive open-source, adaptable scripts for the coding of keystroke data would provide an important and useful tool for writing researchers, both in terms of speed and reproducibility. We note also that, although our scripts go some way to doing this, they are a work-in-progress and do not provide comprehensive coding for keystroke data. In particular, they are restricted to the analysis of linear pauses, and do not attempt to analyse the wide range of other processes involved in writing. Additionally, our focus was to make sure that our code worked, rather than it being succinct. Thus, we expect that our macros can be substantially shortened to make them run more efficiently. If users wish to adapt our macros so that they are more concise, we welcome them to do so but ask that they make their code openly accessible, as we have, so other researchers may also use their scripts. Finally, if a writing researcher wishes to use automated scripts to code keystroke data, it is still especially important to manually screen their keystroke logs because the high level of typing variation between participants means that scripts will not necessarily pick up all features that they are designed to identify.

In Fig. 2, we show an example output from running our work-in-progress scripts on a participant's keystroke log file using Windows Microsoft Excel 2016.

Isolating first-draft text production

After the final run of our macros on the keystroke log data, we set about isolating first draft text production from other types of text production. Firstly, we excluded titles, as, in line with Baaijen et al. (2012) and as explained in the introduction, we think they do not reflect the same underlying processes as general text production in essay writing. For the same reason, we also wanted to isolate any explicit pre-planning that the writer may have produced in their Inputlog word documents. In this dataset, there were no instances of explicit planning. We think that this is because the participants in the outline condition were told explicitly to plan on paper prior to writing their essays, and the synthetic group were told to write spontaneously, rather than in a pre-planned fashion.

Finally, we separated the text produced in an initial draft, versus the text produced in any post-draft revision. We define post-draft revision as text that is produced after the writer starts to close their essay. Closing of an essay is made apparent through contextual factors, such as the participant writing 'to conclude', or 'finally'. Importantly, post-draft revisions are often made using a "top-to-bottom" strategy. That is, the writer moves away from the leading edge and then starts to make edits to or insertions in their essay, roughly from the beginning of their text, and then working systematically towards the end of their text.

Of course, post-draft revision classification relies on an element of subjectivity. For example, the contextual clues one uses to identify when a writer is drawing their first essay draft to a close will vary dependent on the writer. In addition, the extent to which writers make post draft edits will differ. After participants have ended their first draft, they may start to make their post-draft, top-to-bottom edits from the first line of their text product, or they may start making their edits further into the text product. To overcome issues that may have risen due to the subjective nature of the post-draft revision classification, our research team worked collaboratively in deciding which participants had post draft revisions present in their keystroke logs, so that we could discuss any ambiguities that may have led to differences in our categorisations. Additionally, using the Inputlog 8 playback feature to watch the writer producing text in real time helped to clarify whether the participant had instances of post-draft revision or not, because cases where the writer moved away from the leading edge to start making top-to-bottom edits and insertions could clearly be seen.

In our data, approximately 50% of participants in both conditions made post-draft revisions. This meant that if we had analysed the linear pause data in the keystroke logs without isolating the first-draft text production, our results would have aggregated linear pauses across several types of text production which reflect different types of writing processes.

After applying all of the methods outlined within this section, we had isolated 'pure', first-draft, linearly produced text from other types of text production. We had also identified linearly produced within-word, between-word, subsentence and between-sentence pauses within the isolated text, based on our conceptual and calculation pause frameworks. All further analyses described in this paper were conducted on these final isolated keystroke files, which we converted from xlsm to csv format for analysis in R.

output	startTime	startClock	endTime	endClock	actionTime	pauseTime	pauseLocation	pauseLocationFull	LWW	LWWTime	LBW	LBWTime	LBS	LBSTime
v	470606	00:07:50.606	470665	00:07:50.665	59	139	1	WITHIN WORDS						
e	470755	00:07:50.755	470805	00:07:50.805	50	149	1	WITHIN WORDS	LWW	149				
g	470824	00:07:50.824	470874	00:07:50.874	50	69	1	WITHIN WORDS	LWW	69				
a	470934	00:07:50.934	470993	00:07:50.993	59	110	1	WITHIN WORDS	LWW	110				
n	471006	00:07:51.006	471073	00:07:51.073	67	72	1	WITHIN WORDS	LWW	72				
i	471182	00:07:51.182	471242	00:07:51.242	60	176	1	WITHIN WORDS	LWW	176				
s	471262	00:07:51.262	471311	00:07:51.311	49	80	1	WITHIN WORDS	LWW	80				
m	471351	00:07:51.351	471401	00:07:51.401	50	89	1	WITHIN WORDS	LWW	89	TRANSITION			
SPACE	471470	00:07:51.470	471520	00:07:51.520	50	119	3	AFTER WORDS			LBW	2165		
i	473516	00:07:53.516	473576	00:07:53.576	60	2046	2	BEFORE WORDS			TRANSITION			
s	473616	00:07:53.616	473675	00:07:53.675	59	100	1	WITHIN WORDS	LWW	100	TRANSITION			
SPACE	473745	00:07:53.745	473795	00:07:53.795	50	129	3	AFTER WORDS			LBW	5483		
a	479099	00:07:59.099	479158	00:07:59.158	59	5354	2	BEFORE WORDS			TRANSITION			
d	479278	00:07:59.278	479327	00:07:59.327	49	179	1	WITHIN WORDS	LWW	179				
v	479436	00:07:59.436	479476	00:07:59.476	40	158	1	WITHIN WORDS	LWW	158				
a	479605	00:07:59.605	479665	00:07:59.665	60	169	1	WITHIN WORDS	LWW	169				
n	479645	00:07:59.645	479725	00:07:59.725	80	40	1	WITHIN WORDS	LWW	40				
g	479826	00:07:59.826	479863	00:07:59.863	37	181	1	WITHIN WORDS	LWW	181				
t	479836	00:07:59.836	480013	00:08:00.013	177	10	1	WITHIN WORDS	LWW	10				
a	480013	00:08:00.013	480062	00:08:00.062	49	177	1	WITHIN WORDS	LWW	177				
e	480142	00:08:00.142	480192	00:08:00.192	50	129	1	WITHIN WORDS	LWW	129				
g	480201	00:08:00.201	480271	00:08:00.271	70	59	1	WITHIN WORDS	LWW	59				
o	480499	00:08:00.499	480599	00:08:00.599	100	298	1	WITHIN WORDS	LWW	298				
BACK	480817	00:08:00.817	480857	00:08:00.857	40	318	11	REVISION						
BACK	480940	00:08:00.940	480966	00:08:00.966	26	123	11	REVISION						
BACK	481036	00:08:01.036	481076	00:08:01.076	40	96	11	REVISION						
g	481086	00:08:01.086	481165	00:08:01.165	79	50	1	WITHIN WORDS						
e	481264	00:08:01.264	481324	00:08:01.324	60	178	1	WITHIN WORDS	LWW	178				
o	481325	00:08:01.325	481384	00:08:01.384	59	61	1	WITHIN WORDS	LWW	61				
u	481453	00:08:01.453	481503	00:08:01.503	50	128	1	WITHIN WORDS	LWW	128				
s	481553	00:08:01.553	481602	00:08:01.602	49	100	1	WITHIN WORDS	LWW	100			TRANSITION	
.	481632	00:08:01.632	481692	00:08:01.692	60	79	5	AFTER SENTENCES					LBS	11462
SPACE	481771	00:08:01.771	481831	00:08:01.831	60	139	4	BEFORE SENTENCES					TRANSITION	
CAPS LOCK	490601	00:08:10.601	490671	00:08:10.671	70	8830	15	UNKNOWN					TRANSITION	
V	493015	00:08:13.015	493085	00:08:13.085	70	2414	2	BEFORE WORDS					TRANSITION	

Fig. 2 An example of the macro outputs on a participant's Inputlog general analysis file, where LWW = Linear within-word pause, LBW = linear between-word pause, LBS = linear between sentence pause, and TRANSITION identifies cells involved in the calculation of the pause time. The highlighted numbers are the calculated pause times in milliseconds

Single Gaussian distribution models versus multi-component Gaussian distribution models

In what follows, we construct Gaussian mixture models (GMM) using the expectation–maximization algorithm (EM; McLachlan & Peel, 2000) to investigate whether the linear pause data have an underlying structure that is better represented by multiple Gaussian distributions rather than single Gaussian distributions. The EM algorithm provides maximum likelihood estimation for data that have an underlying latent variable structure (Do & Batzoglou, 2008). The EM algorithm first estimates a latent variable for each of the values within the dataset (e.g., for each linear between-sentence pause time in a single participant). The algorithm then optimises the parameters for those underlying variables in the form of a Gaussian distribution. The process is iterated until an appropriate set of latent values that fits the data is achieved, alongside maximum likelihood. This process is explained in greater detail in Little et al. (2013) and Martinez and Martinez (2002, p.296). We used the R package Mclust version 5.4.7 (Fraley et al., 2020) to do this, and provide the scripts for the analysis in our supplementary materials (https://osf.io/r53h2/).

We present this analysis in three parts: (i) We describe the initial steps involved in preparing the data and illustrate the need for mixture modelling, using the distribution of between-word pauses as an example; (ii) We then describe the process of evaluating the relative fit of a series of mixture models to the data; (iii) Having established the set of distributions to represent the data at each pause location, we present estimates of the parameters of these distributions and their mixing proportions.

Initial preparation of the data and the need for mixture modelling

To illustrate this analysis, we focus on the between-word pause durations. Given that all these distributions were extremely positively skewed, we first carried out natural log transformations. Figures 3 and 4 show the distributions of these log-transformed pause durations for each participant in the outline (Fig. 3) and synthetic (Fig. 4) planning conditions.

Figures 3 and 4 illustrate two important features. First, although log transformations did reduce skew, they did not eliminate it for any of the participants. In some cases (e.g. S017, third along on the top row of Fig. 3), one might characterise this simply as a skewed distribution. However, in the majority of cases (e.g. S008, on the left of the bottom row in Fig. 3), the distributions are clearly multi-modal and appear similar to the three-component distribution of linear between-word pause durations described by Baaijen et al. (2012). Visual inspection of pause durations at the other text boundaries also suggested a multi-modal pattern.

Second, there are some clear outliers representing extremely short pauses between words, For example, participant S045, whose between-word pause durations are plotted at the bottom right of Fig. 3, showed a pause duration of 29 ms ($\log_n 3.38$) for one between word transition. Similarly, participant O037, at the right hand end of row 3 in Fig. 4, showed two extremely short between-word transitions of 30 and 40 ms. Inspection of these cases showed that these reflected presumably accidental space-bar presses in the middle of a word. Given that these did not reflect genuine transitions between words, they were excluded from further analysis. There were similar, though more frequent, outliers for the within-word transitions—transition times of 1, 8 and 10 ms for example—which reflected accidental "joint" key presses. In principle, these could be identified manually. However, given the relatively high frequency of such errors for within-word transitions, it is more practicable to use a threshold to exclude such very brief transitions from analysis. Van Waes et al. (2021), for example, used a threshold of 30 ms in their study of typing skills. For present purposes, we used a common threshold of 50 ms, applied at all text locations, to exclude such accidental transitions from the analysis. Inspection of the between-word transitions in our sample showed that all those, relatively rare, transitions below this threshold reflected accidental presses of the space bar within words.

Evaluating the relative fit of mixture models

To test explicitly whether adding mixture components truly provided a better fit than a single Gaussian distribution model, we formally evaluated the relative goodness of fit of single Gaussian distributions compared to the two- and three-component GMMs suggested by previous research (for each participant's data at each of the pause locations). Hence, we used the EM algorithm to fit multiple GMMs to each participant's log-transformed linear pause data at each of the text boundaries for all transitions above 50 ms. The decision to fit a maximum of three Gaussian distributions to each participants' data was based on the research that we reviewed earlier (Almond et al., 2012; Baaijen et al., 2012; Guo et al., 2018; Roeser et al., 2021; Van

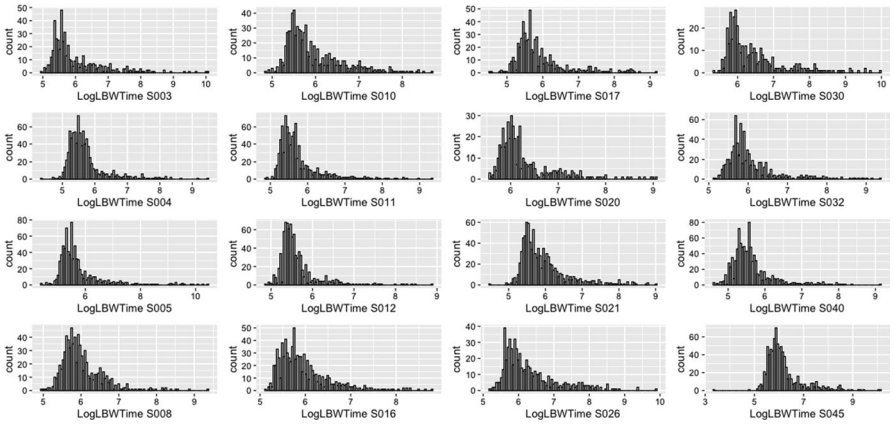

Fig. 3 Histograms showing the log-transformed distributions of linear between-word pause times for each of the participants in the synthetic condition, where LogLBWTime = the natural log of the linear between word pause times for a given participant, "S" = synthetic condition and the "0XX" = the participant number

Waes et al., 2021). To compare the fit of the models, we used the Bayesian Information Criterion (BIC), where a lower value indicates a better fitting model. The advantage of using BIC compared to other goodness-of-fit criteria is that it reduces the likelihood of overfit. The more parameters, the more BIC penalises the model. Hence, BIC usually favours the most parsimonious model (Berchtold, 2010).

Table 3 shows the percentage of participants in each condition whose data were best fit by a given number of components. Thus, the first line of Table 3 shows that, for the within-word boundary, a 1-component model did not provide the best fit for any of the participants. By contrast, a 2-component model fitted the within-word data best for 75% of the participants in both the synthetic condition and the outline condition (12 participants in each condition). The third row shows that a 3-component model fitted the date best for 25% of participants in both the synthetic and outline conditions (4 participants in each case).

The first important finding of this analysis was that, as can be seen in Table 3, a single distribution fitted the data best in an extremely small minority of cases. Indeed, at the within-word boundary, none of the participants' data were best fitted by a single distribution. Even at the sentence boundary, where one might perhaps expect that linear transitions more uniformly reflected a single category of reflective thought or content planning, and even after the data had been log-transformed, approaching 70% of the participants in both planning conditions showed evidence of more than a single component. This is a strong indicator that the processes taking place at different text boundaries are heterogeneous and are not well captured by a single indicator of being "above threshold". We turn now to consider the potential sources of the distributions at each text boundary, focussing first on a detailed consideration of the between-word transitions.

The linear transitions between words showed that a three-component model fitted best for the majority of participants (75%, 12 participants) within the synthetic

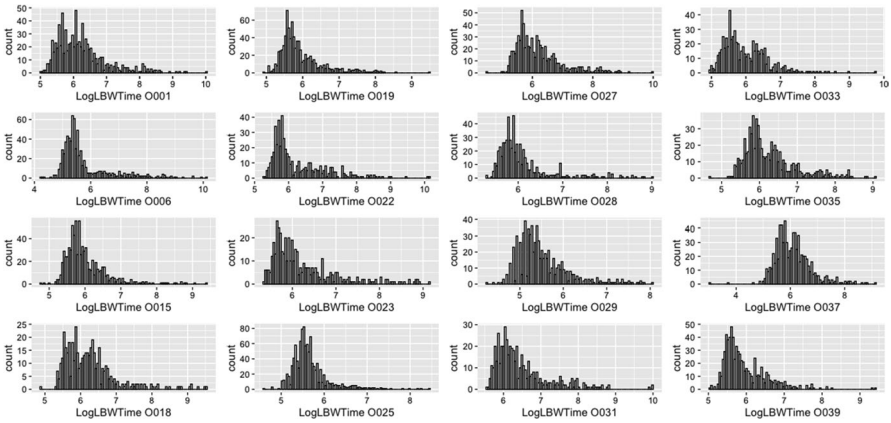

Fig. 4 Histograms showing the log-transformed distributions of linear between-word pause times for each of the participants in the outline condition, where LogLBWTime = the natural log of the linear between word pause times for a given participant, "O" = outline condition and the "0XX" = the participant number

planning condition and 69% (11 participants) within the outline planning condition. This replicates, and indeed provides stronger evidence for, Baaijen et al.'s (2012) finding, which found that the distribution of linear between-word pauses was best fit by the 3-component distribution for 58% of their sample. Furthermore, several participants for whom 2-component mixtures were best-fitting showed very small differences (BIC differences < 5) between 2- and 3-component models. We decided therefore to impose the three-component distribution on the linear between-word transitions for all the participants. Inspection of the plots in Figs. 3 and 4 shows that many individuals have a characteristic pattern of a normal distribution on the left-hand side, a long tail on the right-hand side, and a more or less well-defined normal distribution in between. What seems to vary is how well-defined the middle distribution is. Imposing a common 3-component model enables the participants to be compared in terms of the mixing proportions of the three distributions and varying parameters of those distributions.

Baaijen et al. (2012) suggested that the overall pattern represented a distinction between word retrieval processes (the left-hand distribution), phrase boundary planning (the middle distribution), and reflective thought (the long tail). We would modify that here to suggest a more general, and perhaps vaguer, characterisation of the middle distribution. We propose that the left-hand distribution represents variation in automatic lexical processes and reflects factors such as word frequency. The right-hand distribution represents more reflective thought, including the evaluation of content that is already written, and conceptually planning what to say next when revision is required. The middle distribution reflects sub-structural planning of units within the text, and variations in the presence of this distribution may be related to the duration of between-sentence pauses being more present following brief intersentential pauses. Essentially, the proposal is that this distribution reflects supra-lexical processes within sentence production but is not full formulation of novel content

Table 3 The number of components for the best fitting models and the number/percentage of participants these models were assigned to in the synthetic and outline conditions (and the range of BIC values for these models)

Linear pause boundary	Synthetic condition Best fitting number of components counts and percentages (BIC)	Outline condition Best fitting number of components counts and percentages (BIC)
Within-word	1–0 (0%)	1–0 (0%)
	2–12 (75%)	2–12 (75%)
	3–4 (25%)	3–4 (25%)
	(1845.504 to 5426.658)	(2043.966 to 4806.208)
Between-word	1–0 (0%)	1–0 (0%)
	2–4 (25%)	2–5 (31%)
	3–12 (75%)	3–11 (69%)
	(593.8311 to 1899.715)	(447.4839 to 1925.832)
Sub-sentence	1–2 (13%)	1–4 (25%)
	2–10 (63%)	2–7 (44%)
	3–1 (6%)	3–4 (25%)
	NA–3 (19%)	NA–1 (6%)
	(0.992 to 64.975)	(4.974 to 100.131)
Between-sentence	1–4 (25%)	1–4 (25%)
	2–7 (44%)	2–8 (50%)
	3–4 (25%)	3–2 (12.5%)
	NA–1 (6%)	NA–2 (12.5%)
	(6.198 to 70.619)	(2.949 to 85.096)

NA* indicates where model fit could not be assessed accurately for certain participants because there were too few pause observations

nor evaluation and revision of sentence content (see Roeser et al. (2019) for a discussion of some of the issues here).

For the within-word pauses, the 2-component model fitted the data for the majority of cases in both conditions. This strongly supports previous research assuming 2-component models (Almond et al., 2012; Guo et al., 2018; Roeser et al., 2021; Van Waes et al., 2021) and suggests that this applies just as much to free text production as it does to the copy tasks used by Roeser et al. (2021) and Van Waes et al. (2021). Typically, these two distributions are assumed to reflect differences between fluent and non-fluent typing processes.

Finally, we want to make a few brief comments about the pauses between subsentences and sentences. As Table 3 shows, the main point we would emphasize about these relatively longer pauses, many of which would be captured by threshold-defined measures, is that, although some participants' data would be best-fit by a single lognormal distribution, for both sub-sentence and sentence boundaries, most participants show best fits for 2- and 3-component distributions. This suggests the possibility that they reflect a range of different processes. For example, a

2-component distribution for a between-sentence boundary might indicate a distinction between relatively linear linguistic planning of the proposed next sentence and more reflective thought about the next piece of content. The difficulty is that for the half-hour long texts we are considering here, there are a relatively small number of these boundaries, so it is questionable how meaningful mixture modelling is for these data. Nevertheless, for illustrative purposes, we will assume a 2-component model in the next section, where we estimate the parameters of these distributions and their mixing proportions.

The parameters of the assumed mixture components and their mixing proportions

The virtue of imposing the three-component distribution for the linear transitions between words and two-component distribution for the transitions at other locations is that it enables researchers to compare the properties of the distributions across participants and conditions. These include the proportion of pauses falling within each of the three distributions and the mean and standard deviation of each distribution.

Table 4 shows the mean duration of the transitions between units at different locations in the text for the participants. Thus, individuals typically paused for around 270 ms ± 51 ms for linear transitions between words for the first component, which we argue reflects "normal" lexical retrieval processes. This is remarkably similar to the estimate of 270 ms found by Baaijen et al. (2012) for this component. Other comparisons with their findings are not possible because, in their illustrative analysis of their data they either didn't report comparable estimates or fitted different mixture models at other text locations. Note, finally, that Baaijen et al. (2012) suggest that the "long tail" of pauses seen in their between-word data should not be treated as a normal distribution but rather as a miscellaneous set of reflective thoughts, and that the cut-off point should be defined as the right-hand boundary of the "middle distribution" (they suggested 3 standard deviations above the mean of the middle distribution as the cut-off point). They estimated this as around $\log_n$ 7.43 (1,686 ms). In the present sample, the equivalent threshold is 1426 ms (sd = 309 ms). These are well below the threshold of 2 s usually used to identify "cognitive" pauses. In effect, then, we recommend modelling between-word transitions as a mixture of two normal distributions and a count variable of "reflective thinking" estimated above a threshold varying for differ individuals. Comparisons with other findings are not possible because there has been no other research examining free text production which has modelled linear transitions between units in terms of mixture models.

The second important feature of the data is the mixing proportions of the different components. It is noticeable that for these linear transitions, the overwhelming majority of within-word transitions (around 95%) are fluent. This would suggest that the participants were relatively skilled typists/keyboarders. But is it is important to remember that these data only reflect linear transitions: the other feature that we have not analysed here is percentage of non-linear transitions. At higher-level text locations, there is a much more even spread of the different component processes,

suggesting that they reflect meaningful distinctions between processes. Furthermore, the standard deviations for the mixing proportions indicate that there is substantial variation between individuals in the relative proportion of the different component processes.

Finally, a noticeable feature of these data is that there is relatively little difference between outline and synthetic planning conditions at lower levels of text production (within- and between- words) but more evidence of potential differences for higher levels of text production. Thus, at sub-sentence and sentence locations, transitions for equivalent components are typically longer for the synthetic condition than the outline condition. Similarly, the mixing proportions show some evidence of being different depending on type of advance planning, with higher proportions of relatively fluent transitions in the outline planning condition than in the synthetic planning condition. We should stress that there is considerable individual variation and, with this relatively small sample, these modest effect sizes are not significantly different. Nevertheless, we do take this an indication that this form of analysis has the potential for revealing difference in writing processes under different conditions.

General discussion

Our overall aim in this paper has been to advocate the analysis of transition times between keystrokes rather than of "cognitive" pauses defined relative to a threshold. We argue that this extends analysis beyond the higher-level reflective thought involved in writing and enables researchers to examine the less explicit processes involved in text production. However, this is not simply a matter of changing the unit of analysis. In order to isolate text production processes from the other components of the writing process, pause analysis has to be restricted to the sections of the text produced as part of the forward progression of the text. The key distinction here is between *linear* transitions between units of text and *event* transitions between units of text. In aggregate, we argue that *linear* transitions reflect the characteristic features, for a given text, of an individual's text production process. The scripts that we have provided are designed to provide transparent and reproducible procedures for identifying these linear transitions at different text boundaries and for calculating their duration.

The first observation that we want to make about this process is the necessity of complementing it with manual screening of the keystroke logs combined with visual inspection of playbacks. This is not just a matter of checking how comprehensive the search has been and modifying the scripts accordingly, as in the example we gave of a misidentified sentence transition (the atypical timing of the < LSHIFT > key press). It is fundamental to the identification of appropriate sections of text production and to the definition of that as text production. Three aspects of this are particularly important and are open to debate.

First, there is the question of which parts of the log count as "text production". We think it is relatively uncontroversial to exclude titles and episodes of explicit planning or note making. However, it can be more problematic to decide on what counts as forward text production. If a writer returns to an earlier section of text

Table 4 Mean duration (and SD) of pauses for each mixture component at different text locations within the different planning conditions, along with the mean proportion of pauses (and SD) falling within each mixture component

Text location	Writing condition	Mixture component	Mean duration (sd)		Mixing proportion (sd)
Within words	Outline	Component 1	139.38	(26.51)	.95 (.05)
		Component 2	462.58	(125.85)	.05 (.05)
	Synthetic	Component 1	136.90	(25.16)	.94 (.05)
		Component 2	459.50	(154.66)	.06 (.05)
Between words	Outline	Component 1	266.83	(52.00)	.43 (.15)
		Component 2	400.55	(99.87)	.40 (.10)
		Component 3	1294.83	(598.98)	.17 (.08)
	Synthetic	Component 1	274.55	(50.26)	.40 (.08)
		Component 2	443.31	(113.71)	.42 (.08)
		Component 3	1259.00	(405.34)	.17 (.09)
Sub sentences	Outline	Component 1	785.43	(382.27)	.66 (.24)
		Component 2	4614.83	(4254.92)	.34 (.24)
	Synthetic	Component 1	864.32	(367.24)	.60 (.20)
		Component 2	4478.09	(7218.43)	.40 (.20)
Sentences	Outline	Component 1	1175.52	(486.80)	.69 (.20)
		Component 2	5527.74	(3628.67)	.31 (.20)
	Synthetic	Component 1	1558.16	(1087.45)	.55 (.17)
		Component 2	7756.81	(7927.44)	.45 (.17)

and inserts a brief phrase here and there within a paragraph, we would classify this as a non-linear event. However, if a writer returns to an earlier section of the text and inserts several paragraphs linearly, we would count these as linear transitions. The higher-level non-linearity would be reflected in a separate measure (e.g., global linearity as in Baaijen & Galbraith, 2018). There is clearly room for debate here and, indeed, scope for empirical research about whether parts of the text that differ in this way show different characteristics.

Second, there are always idiosyncratic cases, which can only be identified through manual checking. The example we gave of the writer who habitually pressed the < SPACE > key before deleting it and inserting a < FULL STOP > is a case in point. For the analyses in this paper, we left these, unaltered, as non-linear transitions and did not include them in our estimates of linear pause durations. However, it is important to note that this is rather a conservative decision, and that it may have important consequences. Under this decision, this participant would score very low on a measure of linearity of sentence production; taking the opposite decision, would change this instantly into a very high score for linearity. To a certain extent, this may be compensated by the use of composite measures of global linearity, as in Baaijen and Galbraith's (2018) study, which included six separate indicators. Indeed, a strong argument for using such composite measures is precisely that they are less sensitive to the vagaries of single indicators.

Nevertheless, such cases can have potentially large, but hidden, influences on analysis. Manual checking is necessary if these cases are to be identified.

Finally, it is important to exclude brief transition times between key presses, which, though apparently linear, and hence not identified by our scripts, nevertheless represent errors such as, for example, simultaneously pressing keys. In this paper, we have used a threshold of 50 ms to exclude such transition times from the analysis. Inspection of these below-threshold pauses within our data showed that they all corresponded to such errors rather than genuine transitions. However, we did not carry out a systematic analysis of every transition above this threshold to verify the validity of this threshold, though this is, in principle, possible. Further research is needed to establish the optimum threshold here and whether this varies for different populations and in different contexts.

Overall, then, although we think that, in the interests of both economy of effort and *reproducibility*, it is valuable to provide automated scripts for analysis, we would also emphasize the need for this to be complemented by manual coding and checking. The key element here is that the decision-making process should be *transparent*.

Given these procedures for calculating the duration of linear transitions that reflect text production processes, we turn now to the mixture modelling of these durations. The first important finding here was that, for both within-word and between-word transitions, there was no evidence that a single distribution fitted these data. Furthermore, although there was some evidence that, for some participants, a single lognormal distribution fitted the data best for both the sub-sentence and between-sentence transitions, the majority of the participants' data showed better fits for multicomponent distributions. We take this as conclusive evidence that counting "cognitive" pauses above an arbitrarily given threshold fails to capture the range of processes occurring during text production, and that this applies even to those pauses appearing above threshold. This is an important demonstration of the value of analysing transition times between units rather than searching for threshold-defined pauses. We recommend that, rather than imposing a threshold distinguishing between "cognitive" and "non-cognitive" pauses, researchers should instead impose a common set of distributions on the data, and estimate how the parameters of these distributions vary between individuals and as a function of independent variables.

Beyond establishing empirically-defined thresholds for cognitive pauses, mixture-modelling has the potentially more illuminating function of enabling us to identify potential processes taking place below such thresholds, and of testing theoretical models of these processes. The first important finding here was the relatively strong evidence for a three-component structure in the linear between-word data. This finding replicates with a new sample the findings of Baaijen et al. (2012), who suggested that linear transitions between words have a three-component structure: a mixture of two normal distributions combined with a long tail consisting of a miscellany of above-threshold pauses. Given this replication, it is important to carry out further research testing the hypothesis that these distributions reflect a combination of processes relating to automated lexical processes, higher level supra-lexical planning, and reflective thought. Second, we also replicated previous research by van Waes et al. (2021) and Roeser et al. (2021) indicating that a mixture of two

components—reflecting a distinction between fluent and disfluent typing—fits the within-words data better than a single lognormal distribution does. The distributions at other locations have a less clear structure. In the case of sub-sentence and sentence boundaries, there are probably too few data points for the distributions to be fitted reliably: longer episodes of text production are needed to explore these distributions.

An important methodological consideration arising here (and raised insightfully by one of the reviewers of the article) is the question of whether the data should be analysed in terms of the number of distributions that best fit an individual's data or, instead, in terms of the varying parameters of a common distribution. For example, should we characterise the distribution of between-word pauses as a varying property of writers—some writers show two distributions, perhaps reflecting two underlying processes, other writers show three distributions, perhaps reflecting three underlying processes—or, instead, in terms of a common three-distribution model, with individual writers employing different mixtures of these common processes. The reviewer advocated fitting a common set of distributions across all writers, on the grounds that the sequence of cognitive operations involved (for example, in sentence production) are common across individuals. We agree with this argument, and would in general advocate fitting a single set of distributions after comparing the fit of different mixture models and identifying the mixture that best fits the majority of the data. However, we do not assume that there are necessarily a common set of processes—or, more precisely, that there will necessarily be a direct match between underlying cognitive processes and the number of distributions of pauses. For example, we have observed in some unpublished data collected from second language writers that they typically show evidence for two rather than three distributions of between word pauses. We think, therefore, that variations in the number of distributions may reflect genuine differences in the distribution of processes (see also our cautionary words below about the nested nature of these data) and that this is a relevant analytic feature. That said, for comparison across writers, we would advocate fitting a single set of distribution across all writers, with variations represented by the varying parameters of these distributions and the mixture proportions rather than the number of distributions.

In sum, the overall strategy that we would recommend is a combination of a bottom-up, data-driven, approach and a, top-down, theory-driven approach. It is bottom-up insofar as we fit a series of models and evaluate their relative goodness of fit using purely statistical criteria (BIC). But we have restricted our analyses to mixtures of three distributions or less because these have been suggested by previous research and have the potential to be theoretically interpretable. Models with a higher number of components can be fitted and do occasionally provide better fits for specific writers but these are relatively infrequent and lack any clear interpretation. Generally, we would recommend fitting mixtures up to one level beyond those assumed by current theory as a test of how well current theory fits the data. The analysis is also top-down insofar as we decide, a priori, which boundaries are to be analysed separately, and, on the basis of theory, which set of distributions ultimately to fit to represent variations across writers. Thus, for between-word pauses, the original observation of a three-component distribution of pauses (Baaijen et al.,

2012) has been replicated here, and we advocate that comparisons across writers be made using the parameters of these distributions. Given the hypothesis that the "middle" distribution represents supra-lexical processes, we propose that this could be tested by analysing the pauses at clause, or perhaps phrase, boundaries as a distinct boundary. We would expect that these pauses should typically be shorter than between sentence pauses but longer than the remaining between-word pauses and, further, that once these are removed from the sample of between-word pauses, the best-fitting between-word distribution should be a mixture of two rather than three-distributions.

These findings demonstrate the value of mixture modelling as a method of identifying further structure within linear transitions between units of text. It is important to note, however, that, for demonstration purposes, we have treated the pause distributions at the different text locations as if they were independent. In reality, pause durations at higher-level boundaries such as sentences are clearly likely to influence the duration of the pauses between words within those sentences. To carry out inferential testing of the effects of advanced planning on these nested units of pause distributions in the future, we will use multilevel mixture modelling to analyse a larger data set (Asparouhov & Muthén, 2008; Muthén & Asparouhov, 2009).

In conclusion, we want to emphasize again the importance of combining these observational methods with experimental manipulations designed to assess the relationships between these process measures and outcome variables such as the development of understanding and text quality (Baaijen & Galbraith, 2018). We note, also, the importance of combining the pause measures that we have discussed in this paper with other indicators of both text production processes and higher-level planning and revision processes (Baaijen & Galbraith, 2018; Baaijen, et al, 2012; Galbraith & Baaijen, 2019).

Funding This work was supported through funding from the ESRC South Coast Doctoral Training Partnership.

Declarations

Conflict of interest The authors declare that there are no conflicts of interest.

Data availability The Excel macros and R scripts used to analyse the data in this paper have been made publicly available and can be found here (https://osf.io/r53h2/).

Ethical Approval This research received ethical approval from the University of Southampton Ethics Committee.

References

Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., Babel, M., Bahník, Š, Baranski, E., Barnett-Cowan, M., Bartmess, E., Beer, J., Bell, R., Bentley, H., Beyan, L., Binion, G., Borsboom, D., Bosch, A., Bosco, F. A.,...Zuni, K. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Alamargot, D., Dansac, C., Chesnet, D., & Fayol, M. (2007). Parallel processing before and after pauses: A combined analysis of graphomotor and eye movements during procedural text production. In M. Torrance, L. van Waes, & D. Galbraith (Eds.), *Writing and cognition: Research and applications* (pp. 13–29). Elsevier.

Almond, R., Deane, P., Quinlan, T., Wagner, M., & Sydorenko, T. (2012). *A preliminary analysis of keystroke log data from a timed writing task* (Research Report No. RR-12–23). Educational Testing Service.

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*, 969–979. https://doi.org/10.1080/00207590701398951

Alves, R. A., & Limpo, T. (2015). Progress in written language bursts, pauses, transcription, and written composition across schooling. *Scientific Studies of Reading, 19*, 374–391. https://doi.org/10.1080/10888438.2015.1059838

Andersson, B., Dahl, J., Holmqvist, K., Holsanova, J., Johansson, V., & Karlsson, H. (2006). Combining keystroke logging with eye tracking. In L. Van Waes, M. Leiten, & C. M. Neuwirth (Eds.), *Writing and Digital Media* (pp. 166–172). Elsevier.

Asparouhov, T., & Muthén, B. (2008). Multilevel mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 27–51). Information Age Publishing Inc.

Baaijen, V. M., & Galbraith, D. (2018). Discovery through writing: Relationships with writing processes and text quality. *Cognition and Instruction*. https://doi.org/10.1080/07370008.2018.1456431

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

Beauvais, C., Olive, T., & Passerault, J. M. (2011). Why are some texts good and others not? Relationship between text quality and management of the writing processes. *Journal of Educational Psychology, 103*, 415–428. https://doi.org/10.1037/a0022545

Berchtold, A. (2010). Sequence analysis and transition models. In M. D. Breed & J. Moore (Eds.), *Encyclopedia of Animal Behavior* (pp. 139–145). Academic Press.

Bereiter, C., & Scardamalia, M. (1986). Educational relevance of the study of expertise. *Interchange, 17*(2), 10–19. https://doi.org/10.1007/BF01807464

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Erlbaum.

Cain, K. (2010). *Reading development and difficulties*. Wiley.

Chanquoy, L., Foulin, J., & Michel, F. (1990). Temporal management of short text writing by children and adults. *Cahiers De Psychologie Cognitive/current Psychology of Cognition., 10*, 513–540.

Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in writing generating text in L1 and L2. *Written Communication, 18*, 80–98. https://doi.org/10.1177/0741088301018001004

Chenu, F., Pellegrino, F., Jisa, H., & Fayol, M. (2014). Interword and intraword pause threshold in writing. *Frontiers in Psychology, 5*, 182. https://doi.org/10.3389/fpsyg.2014.00182

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. https://doi.org/10.1007/s11145-019-09953-8

Connelly, V., Campbell, S., MacLean, M., & Barnes, J. (2006). Contribution of lower order skills to the written composition of college students with and without dyslexia. *Developmental Neuropsychology, 29*, 175–196. https://doi.org/10.1207/s15326942dn2901_9

Do, C. B., & Batzoglou, S. (2008). What is the expectation maximization algorithm? *Nature Biotechnology, 26*(8), 897–899. https://doi.org/10.1038/nbt1406

Flower, L., & Hayes, J. R. (1980). The cognition of discovery: Defining a rhetorical problem. *College Composition and Communication, 31*(1), 21. https://doi.org/10.2307/356630

Fraley, C., Raftery, A. E., Scrucca, L., Murphy, T. B., & Fop, M. (2020). Package "mclust" Title Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation. https://mclust-org.github.io/mclust/

Galbraith, D. (2009). Writing as discovery. *British Journal of Educational Psychology Monograph Series, II, 6*, 5–26. https://doi.org/10.1080/07370008.2018.1456431

Galbraith, D., & Baaijen, V. M. (2018). The work of writing: Raiding the inarticulate. *Educational Psychologist*. https://doi.org/10.1080/00461520.2018.1505515

Galbraith, D., & Baaijen, V. M. (2019). Aligning keystrokes with cognitive processes in writing. In E. Lindgren & K. P. H. Sullivan (Eds.), *Observing writing Insights from keystroke logging and handwriting* (Vol. 38, pp. 306–325). Koninklijke Brill.

Guo, H., Deane, P. D., van Rijn, P. W., Zhang, M., & Bennett, R. E. (2018). Modeling basic writing processes from keystroke logs. *Journal of Educational Measurement, 55*(2), 194–216. https://doi.org/10.1111/jedm.12172

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication, 29*(3), 369–388. https://doi.org/10.1177/0741088312451260

Hird, K., & Kirsner, K. (2010). Objective measurement of fluency in natural language production: A dynamic systems approach. *Journal of Neurolinguistics, 23*, 518–530. https://doi.org/10.1016/j.jneuroling.2010.03.001

Kirsner, K., Dunn, J., & Hird, K. (2005). *Language productions: A complex dynamic system with a chronometric footprint*. In: Paper presented at the 2005 International Conference on Computational Science, Atlanta, GA.

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

Little, D. R., Oehmen, R., Dunn, J., Hird, K., & Kirsner, K. (2013). Fluency profiling system: An automated system for analyzing the temporal properties of speech. *Behavior Research Methods, 45*(1), 191–202. https://doi.org/10.3758/s13428-012-0222-0

Martinez, W. L., & Martinez, A. R. (2002). *Computational statistics handbook using MATLAB*. Chapman & Hall/CRCe.

Matsuhashi, A. (1981). Pausing and planning: The tempo of written discourse production. *Research in the Teaching of English, 15*, 113–134.

McLachlan, G. J., & Peel, D. (2000). *Finite Mixture Models*. Wiley.

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., Young, T. P., & Risko, E. F. (2017). Disfluency effects on lexical selection. *Cognition, 18*, 28–32. https://doi.org/10.1016/j.cognition.2016.10.008

Munafò, M., Nosek, B., Bishop, D., Button, K., Chambers, C., Percie Du Sert, N., Simonsohn, U., Wagenmakers, E., Ware, J., & Ioannidis, J. (2017) A manifesto for reproducible science. *Nature Human Behaviour 1*(1), 0021. https://doi.org/10.1038/s41562-016-0021

Muthén, B., & Asparouhov, T. (2009). Growth mixture modeling: Analysis with non-Gaussian random effects. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 143–165). Chapman Hall/CRC Press.

Olive, T., Alves, R. A., & Castro, S. L. (2009). Cognitive processes in writing during pause and execution periods. *European Journal of Cognitive Psychology, 21*, 758–785. https://doi.org/10.1080/09541440802079850

Roeser, J., De Maeyer, S., Leijten, M., & Van Waes, L. (2021). Modelling typing disfluencies as finite mixture process. *Reading and Writing*. https://doi.org/10.1007/s11145-021-10203-z

Roeser, J., Torrance, M., & Baguley, T. (2019). Advance planning in written and spoken sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(11), 1983–2009. https://doi.org/10.1037/xlm0000685

Schilperoord, J. (2001). On the cognitive status of pauses in discourse production. In T. Olive & C. M. Levy (Eds.), *Contemporary tools and techniques for studying writing* (pp. 59–85). Springer.

Spelman Miller, K. (2000). Academic writers on-line: Investigating pausing in the production of text. *Language Teaching Research, 4*(2), 123–148. https://doi.org/10.1177/136216880000400203

Van Hell, J. G., Verhoeven, L., & Van Beijsterveldt, L. M. (2008). Pause time patterns in writing narrative and expository texts by children and adults. *Discourse Processes, 45*, 406–427. https://doi.org/10.1080/01638530802070080

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.

Wengelin, Å., Torrance, M., Holmqvist, K., Simpson, S., Galbraith, D., Johansson, V., & Johansson, R. (2009). Combined eyetracking and keystroke-logging methods for studying cognitive processes in text production. *Behavior Research Methods, 41*(2), 337–351. https://doi.org/10.3758/BRM.41.2.337