Combinations of simple mechanisms explain diverse strategies in the free-hand writing of memorised sentences.

Peter C-H. Cheng & Erlijn van Genuchten
Email: p.c.h.cheng@sussex.ac.uk
Department of Informatics, University of Sussex Brighton, BN1 9QJ, UK

Abstract
Individual differences in the strategies that control sequential behaviour were investigated in an experiment in which participants memorised sentences and then wrote them by hand, in a non-cursive style. Thirty-two participants each wrote eight sentences, which had hierarchical structures with five levels. The dataset included over 31 thousand letters. Despite the deliberately constrained nature of the task and stimuli, 23 patterns of behaviour were identified from the durations of pauses that occurred before the inscription of letters at four chunk levels, spanning letters, word, phrases and sentences. A critical path task analytic model, Graphical Production of Memorised Sentences (GPoMS), shows that the control of writing relies on cues that continuously switch between motor actions and chunk retrievals in a just-in-time fashion at the level of letter information. GPoMS explains the individual differences in terms of variants of a motor production mechanism and variants of a chunk retrieval mechanism, which involve varying degrees of parallelism between cognitive actions and motor actions. A graphical technique for constructing GPoMS models was developed that enabled the estimation of on-going working memory demands during production.

Keywords:
chunk hierarchy; control of sequential behaviour; writing from memory; individual differences; pause analysis; task analysis; critical path model; consistency principles.

9 Feb 2018
(In press *Cognitive Science*)
Introduction

The experiment and model reported here investigate patterns of human behaviour in the task of freehand writing of memorised sentences, a routine task-oriented sequential behaviour. The primary purpose of the study is to examine how basic mechanisms can, in combination, produce diverse strategies across individuals in an apparently simple task that lacks many task characteristics typically associated with the occurrence of alternative strategies. This investigation uses differences between individuals’ behaviours with similar levels of competence as they perform the same task in order to probe the nature of strategies, which contrasts with previous studies that aggregate over participants to obtain a canonical model, or that compare performance across variants of tasks, or that examine participants with different performance levels. As the study addresses hand written production it contributes to our understanding of the pen-based behaviours that is a somewhat neglected aspect of human behaviour compared to, say, interaction with computer interfaces.

Sources of multiple task strategies

Understanding the sources of control of human behaviour is a fundamental issue for cognitive science. It has been well-established for some time that the strategies used on tasks constitute one key to explaining behaviour (Newell & Simon, 1972; Newell, 1990). Explanations have considered the extent to which behaviour can be attributed to characteristics of the cognitive architecture or the particular nature of strategies used on the task. In order to make causal claims about whether, and if so how, fundamental architectural characteristics determine behaviour, it is critical first to understand the strategies applied to the task. Howes & Young (1997) called this the architecture-strategy credit assignment problem. More recently, it has become clear that explanations of behaviour must encompass more than the characteristics of the cognitive architecture and task strategies. For example, in the context of interactive behaviour with technology, Payne and Howes (2013) have formulated a four component framework in which explanations of behaviour include strategies, ecology, mechanisms and utility. They demonstrate the validity of the framework by explaining a broad range of phenomena from HCI. For a particular task, utility is what a person finds of value in what they do, which may be represented by the desire to minimise time, to maximised accuracy, or to achieve a high task relevant performance score, for instance. The mechanisms of cognitive architecture include procedures by which information is mentally processed, which are limited by the mind’s available means and capacities to receive, store, retrieve, transform and transmit information. Ecology takes the form of the constraints imposed by users’ interaction and experience of their external environment that will, for example, include the modality, structure and temporal characteristics of the stimuli encountered. Payne and Howes contend that the mechanism, ecology and utility components are essential to explanations of behaviour, because the space of strategies for a task will be unbounded if mechanism, ecology and utility are not specified. However, this in turn means that the strategy component has a crucial central role in the framework, because the effects on behaviour of the other components are made apparent through their impact on the space of strategies. In other words, mechanism, ecology and utility by their very nature cannot directly interact with each other, but determine behaviours through their interplay in the medium of strategies. Thus, understanding the circumstance in which different strategies are exhibited and the factors that determine their form is an important challenge in studies on the underlying sources of the control of behaviour.

This paper takes up this challenge by identifying diverse strategies in freehand writing; a somewhat neglected domain. The existence of many strategies is surprising because the specific task considered is highly constrained in many ways that might, on current accounts, suggest relatively little opportunity for diversity. The experimental task was to write memorised sets of sentences in a non-cursive style. The strategies were discovered by exploiting the differences between individual’s
writing behaviours as a natural source of variability, which is in contrast to approaches that manipulate task factors to create separate experimental conditions over which sub-groups of participants’ data are aggregated. A model has been developed, using a novel graphical tool, to explain the diverse strategies in terms of combinations of two simple mechanisms and principles of motor production and cognitive consistency, which are described below.

Alternative strategies may occur at many levels in different types of tasks with different characteristic timescales. At time scales of a minute or more, alternative strategies exist to manage *discretionary task interleaving* (Payne, Duggan & Neth, 2007). In such situations, people work on two distinct tasks that have each separate task environments and goals, such that their effort on the tasks cannot be concurrent, so one must decide when to switch between tasks. Payne et al. (2007) found that alternative strategies are determined by the rate of return and by within task sub-goal completion. In dual task situations, two tasks occur concurrently, with rapid switching between the tasks at time scales from 10s and down to 1s. For example, Brumby, Howes, & Salvucci (2007), and Janssen, Brumby & Garnet (2012) investigated the strategies underpinning the dual-task of steering a car whilst also dialling numbers on a mobile phone. Participants’ strategies were sensitive to the relative pre-assigned priority given to the tasks, such that they tended to switch more frequently between dialling and steering, when higher performance on the latter task was required. Importantly, the choice of break points was associated with cognitive chunk boundaries rather than superficially salient but otherwise arbitrary boundaries. The potential for substantial variety in strategies follows from Salvucci & Taatgen’s (2009) theory of threaded cognition. Their theory provides a parsimonious explanation of how a serial procedural resource can be integrated successfully with perceptual, motor and declarative memory resources in dual tasks using a small number of simple mechanisms. To support their theory, Salvucci and Taatgen explain a variety of phenomena. One example is dual-choice tasks that require, separately, specific manual responses to particular visual stimuli and specific verbal responses to particular aural stimuli (e.g., Schumacher et al., 1999). For novices there is a delay in response to the concurrent stimuli (the psychological refractory period) but this disappears with extended practice. Salvucci & Taatgen’s (2009) modelling of dual tasks shows that the reduced performance of novices is not merely due to the lack of automation of the two sets of stimulus-response mappings, as per highly practiced participants, but that interaction of the two streams of processing contributes to the delays. This highlights the idea that sources and explanations of alternative strategies may be both complex and subtle.

Alternative strategies are not only manifest in multi-task situations but are found in individual tasks. The importance of understanding task strategies is emphasised by Gray & Boehm-Davis (2000) who conducted studies using button pressing selection tasks. They found that even within a relatively simple environment, participants switched between two different strategies in successive parts of a trial even though switching saved just 150ms. They conclude that small time differences are sufficient to prompt humans to make strategic changes (Gray & Boehm-Davis, 2000). Strategic variation in task composition has been identified in a broad range of tasks (Gray, Sims, Fu and Schoelles (2006), and Salvucci, Taatgen and Kushleyeva (2006). For instance, Smith et al.’s (2008) task involved the transcription of seven digit phone numbers, on a computer, which participants typically broke down into two groups that were separately encoded and typed. More interestingly, however, their model of task performance explains the choice of strategy in terms of a time saving (utility) that was a consequence of the opportunity to perceive and encode the second chunk in parallel with the motor actions of typing the first chunk.

The operations constituting the processing of hierarchically encoded information may occur in succession or simultaneously, so consideration must be given to both the well-practised automated sub-skills that may occur in parallel with the conscious serially controlled sub-tasks (Schneider & Shiffren,
Although the distinction between serial and parallel processing is often a reasonable idealisation for many analyses, it is likely that the cognitive system cascades processes so that sequential processes locally overlap (McClelland, 1979). Whether operations occur serially or in parallel may depend on their ordering and duration. Some processes, such as positioning a finger at a key and pressing that key necessarily occur serially. Other processes, such as retrieving the next chunk of information and pressing a key may occur in parallel. Hence, executing the same operator serially or in parallel may change the critical path of a process, thereby determine the overall duration of the process (Gray & Boehm-Davis, 2000). This is an important factor in freehand writing.

Models are needed to understand how strategies arising from the complex relations of the task environment and cognitive architecture yield rich patterns of behaviour. Studies that use task-analytic models have been effective in exploring individual differences in strategy choice. This approach was used by John (1996) to examine alternative strategies in a typing task under alternative interpretations of the same task instructions. By formally defining cognitive architectures with fixed capabilities computational models have provided a sound basis upon which to conduct studies of strategies in many domains (e.g., Anderson, 1998, Newell, 1990, Meyer & Kieras, 1997). Unfortunately, as observed by Kieras & Meyer (2000), task strategy often serves as a degree of freedom in modelling, with researchers varying the assumed strategy in order to achieve a good match between model prediction and empirical data. In addition, modellers regularly make informal assumptions about these strategies but do not fully justify them. Thus, the assumed strategy commonly serves as a free parameter to maximise the model fit to data rather than to genuinely explore what credit should be attributed to the strategies or to the cognitive architecture. Howes, Lewis and Vera’s (2009), cognitively bounded rational analysis is an approach that avoids the problem of arbitrary strategy choice in model building by deliberately exploring the space of plausible task strategies for a given specification of a cognitive architecture and a given task environment. The subjective utilities of the strategies are computed using a chosen quantitative utility function, which is an operationalization of some principle of bounded rationality (Howes et al., 2009). The strategies that have the highest utility are selected for comparison to the empirical data. The use of a formal language to describe the architecture and task environment, and the specification of a utility function to drive the computation of the utilities of each strategy, provide a greater degree of rigor and objectivity.

Clearly, there is ample scope for the occurrence of multiple strategies in complex task environments, which demands modelling to adequately explain. Given the importance of strategies in explanations of the control of behaviour it is interesting to explore the boundaries of when alternative strategies may be manifest. Do alternative strategies exist in tasks that are simpler than those previously studied? The existence of alternative strategies in simple tasks is an interesting and important issue, because it lowers the threshold of complexity at which explanations of behaviour must be devised not just in terms of generic psychological factors, but also in relation to strategies (or more generally in terms of Payne and Howes’s (2013) mechanism, ecology, utility and strategy framework). In other words, Newell’s (1973) injunction not to aggregate over task strategies may be applicable to a wider range of tasks than previously envisaged, spanning simple tasks that have not previously been expected to exhibit diverse strategies.

The task in the present experiment simply involved the writing of memorised sentences such as:

You just signed up for a trip, from your favourite society, because you like visiting different places. You paid with some money, which you got from your mum, because you did shopping for her. You have never been to Holland, so you would like to visit Amsterdam, and have a great time. (1 paragraph, 3 sentences, 9 phrases, 52 words, 224 letters)
The sentences are written non-cursively and without punctuation marks. As we will see below, major individual differences occur in the length of the pauses between the writing of successive letters but not in the time to actually scribe letters (pen-off versus pen-on paper durations). What might be the sources of strategic differences that cause such variation in pause lengths? Consider two, of the many, alternative ways that we might manage the process of preparing to write the \( f \) in \( from \) in the second phrase of the first sentence, above. (a) A long pause may occur between writing \( p \) (in \( trip \)) and the \( f \), because we choose to physically lift the pen from the paper, then whilst it is hovering we retrieve the next phrase from memory, select the first word of the phrase (\( from \)), select the first letter (\( f \)), prepare to write the letter by retrieving its motor program, and finally physically move the pen into position to start the first stroke of \( f \). This is a wholly serial sequence of actions. (b) Alternatively, a shorter pause may occur, because we retrieve the next phrase from memory and focus upon \( from \) whilst inscribing \( p \). Then when \( p \) is complete, we select \( f \) and retrieve its motor program, whilst simultaneously moving the pen to its next starting position. In this sequence cognitive actions run in parallel with motor actions, including an overlap between the processing of the second phrase before the completion of the first phrase. This paper identifies the mechanisms that control this task and explores how alternative combinations of them lead to substantial individual differences.

To study individual differences at this level, the experiment was designed to avoid six major sources of strategic variability, which are recognised in the literature, that could mask our phenomenon of interest. First, the occurrence of strategies associated with the coordination of motor or verbal output with the input and encoding of sensory information was avoided by having participants memorise the target sentences so that no reading of text was required during the production parts of trials. This contrasts with transcription tasks (e.g., Cramp & Logan, 2010; Smith et al., 2008) and display interaction tasks (e.g., Gray & Boehm-Davis, 2000; Janssen et al., 2012), in which stimuli must be perceived and encoded alongside production. Second, the occurrence of alternative strategies associated with switches of visual attention and eye movements between different parts of the external task environment were also avoided by writing the sentences from memory (cf., Janssen et al., 2012; Smith et al., 2008). So, during the production parts of the trials, participants could focus just on the pen. Third, by using simple English sentences with a predetermined hierarchical structure the opportunity for different decompositions of each task stimulus was reduced (cf., Smith et al., 2008). The hierarchical structure was emphasised to participants during the memorization phase of the trials. Fourth, the participants were well-educated adults, each with substantially more than ten years of experience in the handwriting of sentences. Thus, opportunities for strategy variations due to inherent differences task expertise were minimized (cf., John, 1996). Fifth, variation to do different levels of declarative processing of the basic task instructions is likely to be minimal (cf., John, 1996). Sixth, the participants were told not to be concerned about mistakes and to simply continue writing in such cases, which eliminates the possibilities of strategies occurring that relate to the checking of errors (cf., Smith et al., 2008).

The next sub-section considers the methodological developments needed to implement the task.

**Methodology development**

Typically, individual differences in cognitive experiments are treated as a nuisance to be minimised by careful sampling and mitigated through the aggregation of data over groups and the elimination of outliers. The motivation to specifically use individual differences to give empirical leverage in the investigation alternative strategies arose directly from the observation that the range of magnitudes across participants was substantial (van Genuchten and Cheng, 2010). By embracing the individual differences, it is feasible to investigate existence and sources of alternative strategies despite the design
of the task deliberately avoiding, or holding constant, the known sources of multiple strategies. The grain size of comparison in the experiment is at the level of the individual, rather than at the level of sub-populations, with the aim to coherently and parsimoniously explain all the differences in the observed behaviours across individuals using combinations of basic mechanisms. In other words, the identified individual differences are a source of variability that constitutes a ‘natural experiment’ and each difference serves to reduce the space of possible strategies to account for all the observed patterns of behaviour.

The hierarchical processing has been long studied across a wide variety of domains. Some examples include: the verbalization of memorised groups of letter sequences (Mclean and Gregg, 1967; Johnson, 1970); verbalisation of lists of words (Buschke, 1976); replacing chess pieces in position on a chessboard (Chase & Simon, 1973); copying electronic circuit diagrams (Egan & Schwartz, 1979); drawing complex geometric diagrams (Obaidella & Cheng, 2009; Roller & Cheng, 2014); copying mathematical equations (Cheng & Rojas-Anaya, 2007; Cheng, 2014). In the present experiment, the use of natural language sentences as stimuli permitted three innovations (van Genuchten and Cheng, 2010). First, stimuli with four meaningful hierarchical levels were created, whereas previous studies typically have considered just two levels (e.g., Longan & Crump, 2010) or more rarely three (Cheng & Rojas-Anaya, 2008). Specifically, the four stimuli levels are: sentence (S), phrase (P), word (W) and letter (L) (plus a stroke (St) level that is of no theoretical interest here). The levels are illustrated in Fig. 1, where the leaves at the branch terminals represent the inscription of the letters and the subscripts are labels that distinguish items locally, relative to their own branches (e.g., L3 is the third letter of a word).

Fig. 1. Goal hierarchy and processing order for memorised sentence writing task.

The second innovation is the use of relatively simple sentences and phrases, and common words to impose a natural but strong organization of chunks across participants. Following Johnson (1970), it is taken that the processing of the hierarchy involves the traversal of that structure in a depth first manner, so the presumed goal structure of the task can be represented by the hierarchy in Fig. 1, where the numbers on each branch show the order of processing of the sub-goals. Consistent with this interpretation, Van Genuchten and Cheng (2010) found significant differences in pause durations for each level of the hierarchy: the pause duration before starting a sentence is longer than the pause before starting a phrase, which is longer than the pause before a word, which is longer than the pause before a letter. Thus, in this task alternative strategies must be modelled by means other than variations in task goal structures. In addition, a model should be able to predict the actual pause durations for each chunk.
level across all the observed strategies, which is a substantial challenge, but if done successfully would argue strongly for the validity of the model.

The third innovation is the potential to estimate the instantaneous WM demands as the task progresses, because the participants are expert in key aspects of the task. The participants are educated adults from an academic setting, so it is assumed that motor actions and cognitive operators associated with positioning and moving the pen will essentially be automatic, so WM demands arise from navigating and preparation of the sentences for production; that is the processing of the chunk hierarchy in Fig. 1. So, a new visual tool is introduced below for the construction of task analytic models that shows how WM demands differ during task execution under alternative task strategies.

Although the sentences have four main hierarchical levels, they have otherwise been kept simple, with two or three short phrases and high frequency words, in order to minimize the potential difference of interpretation or impacts of levels of word familiarity, which might be sources of individual difference. Once a participant has accurately memorised a stimulus, it is written (inscribed) on paper that is placed on a graphics tablet to record the times when the pen is off the paper, and times and positions of the pen in contact with the paper. A further innovation that aims to minimize variability among participants at the level of motor actions is a response sheet with small regularly spaced squares into which individual letters are inscribed. The aim of this response grid is to minimize the temporal impact of variable spacing between letters and peculiarities of any individual’s cursive writing style. Further, Landy & Goldstone (2007) showed that the distance between inscribed elements in unconstrained handwriting is larger for inter-chunk elements than intra-chunk elements. Although the increase in distance is small, the regular response grid aims to eliminate any such influence.

In order to avoid perceptual and encoding processes, the sentences are written from memory. Methods to induce given chunk structures in memory have included: structured serial recitation (e.g., Mclean and Gregg, 1967; Cheng & Rojas-Anaya, 2007); repeated stimulus copying (Obaidella & Cheng, 2009); schematic overview of item structure (Cheng and Roja-Anaya, 2008); repetition of the drawing of a final solution diagram following problem solving (e.g., Roller & Cheng, 2014). In the present experiment, each set of sentences was written on a card, with punctuation used to emphasise the hierarchical structure, including breaks between sentences (full stop / space / capital letter), phrases (commas) and words (spaces). Participants’ recited the sentences verbally until they could repeat the sentence perfectly a sufficient number of times.

To investigate alternative strategies this study adopts the Graphical Protocol Analysis (GPA) method that uses the pause durations which occur between inscribed characters or drawn elements. In handwriting, a pause duration (or pause) is the time between finishing a letter and starting the next; i.e., \( \text{pause}_{\text{letter}} = \text{time}_{\text{pen-down-current-letter}} - \text{time}_{\text{pen-up-previous-letter}} \). Inscription time represents the time needed to inscribe the stroke (or strokes) of the letter; i.e., \( \text{inscription}_{\text{letter}} = \text{time}_{\text{pen-up-current-letter}} - \text{time}_{\text{pen-down-current-letter}} \). Like the long established use of pause durations in other contexts (e.g., Mclean and Gregg, 1967; Chase & Simon, 1973), our GPA studies have shown that pauses in graphical production provides a meaningful Temporal Chunk Signal (TCS) that reflects the organization of chunks in memory. When chunks are induced in participants using structured stimuli, the measured pause durations reflect the induced structure, both for handwritten letters (Cheng & Rojas-Anaya, 2005, 2008) and the drawing of geometric diagrams (Cheng, McFadzean, and Copeland, 2001; Obaidellah & Cheng, 2009; Roller & Cheng, 2014). Further, we have shown that the TCS can be used to assess mathematical competence (Cheng & Rojas-Anaya, 2007; Cheng, 2014) and language competence, both in children’s first language (Van Genuchten, Cheng, Leseman & Messer, 2009) and in adults’ second language (Zulfliki, 2013).
Writing as human Behaviour

Handwriting, and writing more generally, is an important skill that is commonly valued across cultures and educational systems. Therefore, substantial time and resources are devoted to teaching this skill and to investigating its nature. For example, Van Galen (1991) identified the multiple levels of processing involved, ranging from basic motor actions to the conceptual processing of ideas. Both basic motor actions of underlying handwriting (for a review, see Teulings, 1996) and conceptual processing in terms of text composition (e.g., Bereiter & Scardamalia, 1987; Hayes & Flowers, 1981) have been extensively studied. However, handwriting at an intermediate level between the motor and the conceptual, the level addressed in the present experiment, is a form of task-oriented routine cognitive behaviour that has been rather neglected. This neglect may be due to the relative ease of studying typewriting at this level. Transcription typing in which participants immediately reproduce what they are reading or hearing is particularly well-studied. For example, Salthouse (1986) provides a detailed review of 29 phenomena associated with this task, including processing, parsing, translating and executing text components. Of the 19 (plus one other) phenomena that John (1996) found within scope to model, just one explored sources of multiple strategies. In a task involving the pressing of a special key in response to seeing unexpected capital letters, John identified three different strategies. One source of strategy difference was an ambiguity in the task instruction and the other source was in relation to participants’ choice of precise order in which to make key presses, which depended on whether there was parallel perception of the capital letter whilst other letters were being processed by cognitive and motor resources, or a serial perceptual, cognitive and motor process just for the capital letter. Such alternative explanations of multiple strategies are not feasible for our handwritten task, because there are no task ambiguities, nor any opportunity for perceptual processes to run in parallel with cognitive or motor processes or both.

Crump and Logan (2010) proposed and tested a hierarchical model of transcription typing that consists of an inner-loop to process the production of letters within words and an outer-loop to acquire words from the stimuli and to prepare them for output. They propose that words are passed from the outer-loop to the inner-loop, and that the inner-loop operates largely independently from the outer-loop. Crump and Logan’s model was not intended to address the low level strategic relation between letter and words, so it is not obvious how it could be extended to include the higher phrase and sentence levels in our experimental task.

One clear difference between the typing and handwriting is the time required to hand write letters. The durations of inscription actions are relatively long compared to key presses in typing, and generally longer than motor actions found in the tasks of previous studies of sequential behaviour, such as mouse clicks, button presses, or the verbalisations of words or letters (e.g., Buschke, 1976; McLean and Gregg, 1967). As millisecond differences in the duration of individual operators may, as previously described, result in the manifestation of alternative strategies for a given task (Gray & Boehm-Davis, 2000), the relatively long duration of inscribing a letter provides a relatively large window in which cognitive operators might (or might not) be selected. As will be seen, the duration of motor actions in handwriting is critical to the explanation of the different strategies.

This paper has two main parts. The first part presents the experiment on the task of handwriting memorised sentences in which diverse patterns of behaviour are found. The second part presents variants of a task analytic model that explains these patterns.
Experiment: handwriting memorised sentences

Method

Participants
The participants were 32 adults, 19 female and 13 male, between 18 and 33 years old ($M = 22.99$ years, $SD = 3.98$ years). They were all native English speakers recruited from among people working or studying at a large university in the UK.

Materials
The eight stimuli were paragraphs consisting of simple English sentences with hierarchical structures that were emphasized with punctuation to ensure that every participant possessed the same chunk organization. An example was given above and here is another:

We like swimming, in the pool next door. You like to cycle, to towns far away. They like to play football, on the top of the hill. As they play all day, they should eat enough. (4 sentences, 8 phrases, 36 words, 133 letters.)

All the stimuli consisted of three or four sentences, containing six to nine phrases and 32 to 54 words. In total, the eight stimuli consisted of 24 sentences, 59 phrases, 298 words and 1148 letters.

A specially written computer program was used to record position and timing of each pen touch and lift on the graphics tablet. The temporal resolution of the timing information was better than 10ms.

Procedure
Each trial consisted of memorising and writing one stimulus. The order of presentation of the stimuli was randomized for each participant. Participants were allowed to apply any strategy and take as long as needed for rehearsing the stimulus until they were able to verbally recite the stimulus twice without error. All sentences were then written on a piece of A4 paper in landscape orientation attached to a Wacom Intuos 2 graphics tablet using an inking pen. A grid of rectangles was printed on the paper (width: 6mm, height: 8mm, spacing: 2mm). Participants were trained to write one letter in each rectangle and to lift their pen from the paper between each rectangle: a short period gave ample practice. Participants were instructed to fill all rectangles within a row before continuing to the next row, splitting words as necessary. Participants took between 45 and 75 minutes to complete the experiment and they received £10 for their time.

Pause durations between the last letter of a line and the first of the next were eliminated from the data. Participants were told not to include any punctuation in their responses and no one produced any. Participants were trained to write a hash symbol (#) at the start of each stimulus following the procedure of Cheng & Rojas-Anaya (2006), in order to ensure that the process of writing was underway before the first stimulus letter was produced.

Results
For each letter, its pause and inscription duration were computed. Another specially written program was used to identify letters based upon the expected position and separation of strokes. This program took spelling errors and omitted words into account, but where it failed, the identity of letters was manually corrected. Given the number of letters in the stimuli and the number of participants a dataset of 36,736 letters was expected but 31,551 were actually written by participants, which corresponds to an omission rate of 14.1%, which were primarily due to missing phrases. Knowing the structure of each stimulus a chunk level was assigned to each letter: 1=letter, 2=word, 3=phrases, 4=sentence (and 0 for breaks between successive pen strokes within a letter).
The results are presented in two parts. The first provides evidence that the pause durations are the primary source of variability associated with chunk levels. The second provides measures and observations to describe the chunk levels and the rich variety of individual differences.

Fig. 2. Box and whisker plots of (A) pause durations and (B) Inscript times across the four chunk levels. The boxes give the medians and quartiles, the whiskers are 5th and 95th percentiles, the wide ticks are data points between the 95th and 99th percentiles, and the narrow ticks are data comprise the highest 1% of values.

Overall pause and inscription durations

Fig. 2A and B shows pause and inscription times across the four chunk levels for all the data. There are 23326, 6747, 929 and 549 data points for chunks levels 1 to 4, respectively. Substantial variability of pauses occurs at each chunk level, Fig. 2A, with greater variability at higher levels, but it is clear that pause duration increases across the levels, with medians of 281, 360, 453, 781 ms. Median pauses at successive chunk levels differ by approximately 100ms, or more, which are substantial differences in the field of cognition. Comparatively, the variability and duration of inscriptions are more uniform across the chunk levels, Fig. 2B, but there is some increase of median inscription durations with chunk level: 313, 359, 375 and 406ms. The differences between successive levels are below 50ms.

Chunk level appears to be a substantial factor in determining the overall temporal patterns in the data. To test whether the duration of pauses is best explained by the chunk levels, multilevel mixed models were analysed (Pituch & Stevens, 2016). The dependent variable was pause duration and various combinations of chunk level, letter identity, and word length were examined. The best model included fixed effects for all three variables and random effects for all three variables with a resulting -2 log likelihood (-2LL) measure of 481104 (on the basis of 45 parameters: 1 for the intercept, 3 for chunk levels, 24 for letter identities, 13 for word length, 3 variance components, and 1 residual). The tests for aggregated fix effects were significant for both the overall intercept (F(1, 45)=216, p<.001) and chunk level (F(1, 45)=69.1, p<.001), but these were not significant for the letters (F(23,
Further, the estimates of fixed effects for the four chunk levels are, respectively, 392, 444, 699 and 1257 ms, which are all significant (letter, $t(92.3)=13.5$, $p<.001$; word, $t(92.3)=11.4$, $p<.001$; phrase, $t(101.0)=7.6$, $p<.001$; sentence $t(510.7)=13.2$, $p<.001$). These values are greater than the median pauses given above, because of the positive skew of the distributions (see Fig. 2A). However, just two of the fixed effects parameters for the letters were significant (at $p<.05$) and all parameters were close to the overall intercept, with no divergence greater than 50 ms. Similarly, none of the fixed effects for word length were significant ($p>.25$, in all cases) and all the parameters were close to the overall intercept, within 62 ms, with the exception of word length 13 (divergence of 225 ms), which is not especially meaningful given the rarity of the long words in the dataset.

Further support for chunk levels as the major determiner of pause duration is seen in models that exclude selected factors. When both letter identity and word length are omitted, so just chunk level is considered, the $-2\log L$ increases to 481154 (on the basis of 7 parameters). This difference is at the margins of significance as the change in $-2\log L$ is 50, the difference in number of parameters (change in df) is 38, and given that the critical $\chi^2(35)=49.80$. In contrast, when the chunk level is omitted, so letter identity and word length are modelled, the $-2\log L$ is 484023 (for 41 parameters), a difference of 2919, which is significant, $\chi^2(4)=13.28$, $p<.01$. In summary, the fit of the model for pause duration degrades dramatically when chunk level is not present, but is relatively indifferent to the effects of letter identity and word length.

From the specific pattern of estimates of fixed effects and the quality of the models with and without chunk levels, it is clear that chunk level is the major factor that determines pause values beyond individual participant variability. Van Genuchten and Cheng (2010) showed using multilevel modelling that each of the mean pause durations for the five predicted levels for this dataset were significantly different. The mean pause durations between items on each level are: strokes=90 ms; letters=273 ms; words=374 ms; phrases=567 ms; sentences=1134 ms. This is consistent with the hierarchical chunk structure proposed in Fig. 1.

The average inscription times and their variability is of some relevance to the modelling below. The mean of the participants’ median inscription times is 361 ms (SD=66 ms). Overall, the median inscription time was 329 ms (SD = 63 ms) and the mean non-parametric skew was 0.059 (SD = 0.12), and the full range (longest minus shortest) of median inscription times across all the participants is 203 ms (excluding P4, see below). However, the median and the shape of the distribution of inscription times varies substantially between letters and between participants. Considering differences between letters, letters with simpler forms, such as ‘c’, ‘l’, ‘o’ and ‘v’, have smaller medians (e.g., median inscription time of ‘c’ is 200 ms), whereas letters with more complex forms, such as a, e, i and r have larger medians (e.g., median inscription time for an ‘a’ is 400 ms). The median inscription time varies gradually across participants from 234 ms (P2) to 437 ms (P30) and then jumps to 555 ms for one participant, P4, who wrote in block capitals that appeared particularly neat, consistent with the unusually long inscription times.
Fig. 3. Individual participant box plots for pauses at letter, word, phrase and sentence chunk levels (left to right, dark to light boxes). Participants grouped into bands of pauses with centres of (A) 200 ms, (B) 300 ms, and (C) 400 ms. The boxes give the medians and quartiles, the whiskers are 5th and 95th percentiles and the ticks are data points data comprise the lowest and highest 5% of values.
Patterns of individual differences

The mean of the median pause durations across all participants was 374ms (SD=103ms). The median pause duration varies gradually across participants from the shortest at 156ms (P6) to the longest at 438ms (P31), and then jumps to 516ms for participant P2. There is no obvious reason that explains why P2’s pauses are unusually large at all levels (but see below). Figs. 3A to C show box plots for each participant; the grouping and order of participants in each plot relates to their classification according to the model below. Observations concerning the overall magnitude of pauses across chunk levels include:

Obs A. In general, the duration of pauses increases with chunk level. Of the 96 possible transitions between successive chunk levels (32 participants X 3 chunk level increments) just four decreases in pauses occur. The participants within each of the graph Figs. 3A-C are arranged approximately in order of pause duration for successive chunk levels.

Obs B. In general, the variability of pause duration increases with chunk level; the interquartile range tends to be greater towards the right of each participant’s set of four bars.

Obs C. The mean and median of participants’ median Letter pauses are 271 (SD=78.2) and 266ms, respectively, with first and third quartiles of 219 and 320ms, respectively, and a skewness of 0.416.

Obs D. The mean and median of participants’ median Word pauses are 372 (SD=110) and 375ms, respectively, with first and third quartiles of 285 and 417.5ms, respectively, and a skewness of 1.485.

Obs E. The mean and median of participants’ median Phrase pauses were 475 (SD=168) and 449.5ms, respectively, with first and third quartiles of 361.5 and 558.25ms, respectively, and a skewness of 1.344.

Obs F. The mean and median of participants’ median Sentence pauses were 779 (SD=364) and 722.5ms, respectively, with first and third quartiles of 447.75 and 1088ms, respectively, and a skewness of 0.511. A histogram (not shown here) of these pauses reveals three clear clusters of pauses with centres at approximately 400, 900 and 1500ms.

Patterns of pauses for individual participants, or subgroups, can be observed in the charts, including:

Obs G. There are substantial individual differences among the participants; e.g., compare individuals on the left and right sides of Figs. 3A-C.

Obs H. For some individuals their letter, word and phrase pauses are relatively constant: for P1, P13, P15 and P25 their range of pauses across the levels is below 100ms for at least the first letter, word and phrase levels. However, the typical absolute value of their pauses differs: P13 ≈200ms, P1 and P25 ≈300ms; P15 ≈400ms.

Obs I. In some cases, pauses appear to increase in an approximately linear fashion; P16, P19 over all four levels; P26, P29, P9 and P8 for the first three chunk levels; P5 for the last three chunk levels. The increases in the pauses with levels are increments of ≈100ms in most instances.

Obs J. Unlike the contrast between the short and long pauses for the letter and word levels, the longest pauses for sentences are quite out of proportion both in terms of magnitude and variability in relation to the short pauses. Eighteen participants have median pauses of at least 750ms. There are also four similar cases for phrase levels pauses, with median pauses of at least 650ms.
Obs K. Participant P2 is unique in possessing exceptionally long pauses at each level: longest at letter level and word level (453 and 797 ms) and the second longest at phrase and sentence levels (1039 and 1532 ms). P2’s inscriptions times fall well within the typical range for participants.

Coherently explaining this set of twelve observations presents a substantial theoretical challenge for any model of writing production.

**Experiment Discussion**

The experimental task was designed to avoid or to minimise potential sources of individual differences, in various ways: the well-educated adult participants were all experts in writing sentences; they were native speakers of English; they were trained to criterion on the stimuli; the chunk structure was fixed and emphasised during training; pen travel distances between inscribed letters was constrained by the uniform response grid; participants were told to ignore any errors they made; interleaving of production with stimulus perception and encoding was avoided through the prior memorization of each set of sentences. Nevertheless, substantial individual differences were found among participants (Figs. 3A-C). Although motor actions took more time in the present task compared to actions in other tasks studied in the literature, such as typewriting (e.g., John, 1996), the major difference between individuals was not in the inscription times, but the duration of pauses between inscriptions. Indeed, much variation in inscription duration can be attributed to the letter shape, with more complex forms taking longer to write. In contrast, differences among participants’ pauses were substantial, with approximately 300 ms separating the shortest and longest median pause durations. As the absolute mean of the participants’ median of all pauses is itself just over 300 ms, it is improbable that such differences are due mainly to simple variability in the speed of elementary cognitive operations, individual or in combination, because their time scales are small and would tend overall to produce some regular distribution rather than distinct patterns or sets of cases seen in Obs H to L. Thus, variations at a strategic level appear necessary to explain the full range of differences in pause durations.

The mixed model analysis showed that chunk levels significantly determined the duration of participants’ pauses and that the particular letters and the length of the words written were minor secondary effects. To what extent can the chunk hierarchy (Fig. 1) explain the observed differences apparent at different levels and across participants? At a global level pause duration (Obs A) increases with chunk level, which is consistent with the increasing amount of processing associated with greater depths in the chunk hierarchy. This explanation is also compatible with the approximately linear increase of pauses across chunk levels observed for some participants (Obs H). However, the chunk hierarchy does not simply explain the increase in the range of pauses with chunk levels (Obs B) nor account for the disproportionally long sentence pauses of some participants (Obs K). Although, such an explanation might account for the overall magnitude of the distribution of pauses (Obs C-F), given assumptions about various process parameters, it would not directly explain the range of the pauses by predicting minimum and maximum magnitudes for each chunk level. Further, Obs I is inconsistent with a hierarchical processing depth explanation, because the pauses for this subset of participants do not change appreciably across chunk levels. This observation is a particular challenge, because the pauses are not only relatively constant for each participant but the overall magnitude of the pauses vary between them. For some individuals sentence level pauses are similar to their lower level chunk level patterns, but in many cases sentence level pauses exhibit a major divergence (Obs F) (and to a lesser extent phrase level pauses, Obs E). The magnitude of such large jumps is not explicable just in terms of the additional of one level in the chunk hierarchy.
To investigate the source of these individual differences, the next section describes the construction and examination of a task analytic model.

**Task analytic models**

Graphical Production of Memorized Sentences (GPoMS) theory aims to model the experimental data from individuals writing sentences from memory. The writing task gives a strict ordering of chunk production (Fig. 1) and a strict ordering of motor actions, thus the major source of variations must arise from the alternative ways in which sub-sequences of cognitive and motor operators are interleaved. Hence, three questions are to be addressed by GPoMS. (1) What is the set of mechanisms (microstrategies) that enables the interleaving of the cognitive and motor actions? (2) How can the interaction of the mechanisms accommodate the processing across the chunk levels given in Fig. 1 in a fashion consistent with the overall magnitudes and general patterns of measured pauses (Obs A-F)? (3) How can the interaction of the mechanisms explain the observed individual differences among the participants (Obs G-L)? For the sake of parsimony, GPoMS should propose a small number of mechanisms to account for the pause duration patterns across all chunk levels for all participants.

Of the various approaches to task analysis (e.g., Crandall, Klein & Hoffman, 2006), and in particular the GOMS family (John & Kieras, 1996), the CPM-GOMS approach (Vera, John, Remington, Matessa, & Freed, 2005) is chosen here, because it facilitates the exploration of how different control mechanisms affect the ordering and timing of operations in terms of different critical paths for participants. The variants of the GPoMS model were initially explored using SANLab (Patton & Gray, 2010), which is a convenient tool for constructing critical path models of cognitive processes.

This part of the paper has six subsections. The first details the assumptions and constraints underpinning GPoMS models and the second subsection describes a visual tool for constructing the models. The third introduces components of the models and the fourth explains how they are combined into complete models. The fifth subsection evaluates the fit to the models to the experimental data and the final section is a discussion of the scope and limitations of GPoMS.

**Theoretical assumptions and modelling constraints**

The GPoMS theory consists of a collection of theoretical assumptions in the form of 18 constraints.

C1. The processing of sentences navigates the goal hierarchy depicted in Fig. 1 in a depth first fashion. A word is viewed as a chunk consisting of letters. The word chunk holds information about the letters and their positions in the word. In a similar fashion, at higher levels in the stimuli, phrases and sentences are retrieval structures for words and phrases, respectively. Note also that a chunk for the whole stimulus must exist, which contains information about the order of the sentences.

C2. The completion of one operation may serve as a trigger (cue) for the initiation of the next operation. Thus, it is assumed that when a motor action or cognitive operation triggers another motor action or cognitive operation, there is no overlap or delay between the end of the first and the start of the second. Although cognitive operations may in general be cascade-like (McClelland, 1979), for simplicity we adopt the practice from GOMS modelling by assuming that the onset and end of operations are discrete.

C3. Both motor actions and cognitive operations can be triggered by motor actions or cognitive operations.
C4. Cognitive operations that are not triggers for motor actions may occur in parallel with those motor actions; for example, chunk retrieval into WM may occur at the same time as positioning the pen to begin an inscription of a letter.

C5. As the writing is non-cursive, it is assumed that motor programmes for individual letters are executed, rather than motor programmes for whole words.

C6. As the task is one of written production, the sub-goal of making inscriptions will be ever-present, so it is assumed that whenever a letter is retrieved, its associated motor program is immediately retrieved and its production then begins immediately.

C7. To avoid interference, it is assumed that the retrieval of a letter and its motor program does not occur during the inscription of another letter. (This assumption is specifically investigated below.)

C8. Motor programs for inscribing letters and for jumping between rectangles are assumed to be automatic. Although writing letters in rectangles needs more precision than unencumbered production, writing in boxes and jumping between boxes are common activities (e.g., on forms) and participants were given a pre-trial practice. So, it is taken that their execution is automatic, does not demand additional cognitive resources and that it is possible for cognitive operations to occur in parallel with inscriptions or jumps, notwithstanding the serial processing conditions above.

C9. The retrieval time of a chunk from memory is taken to be 100 ms (Collins & Quillian), which is near the middle of the 25-170 ms range accepted in the literature (e.g., Card, Moran & Newell, 1983).

C10. As the stimuli has a predetermined structure that is well memorised prior to production, it is assumed that chunks at each level are retrieved in order and this determines the order of production without cognitive operations that select elements for production. (The consequences of violating this assumption are discussed below.)

C11. To obtain the time for initiating and moving the pen, irrespective of any other cognitive operations, the shortest pause duration serves as a sensible estimate. So, the 1st percentile of pauses across all participants was calculated, which is 94 ms, so the default time for a pen move motor action is set to 100 ms.

C12. The time in C11 is used both for a whole continuous jump between two rectangles without stopping and for movements of the pen off the paper to a resting position or from rest back on to the paper. Given the small distances between the rectangles, the largest part of the movement time is likely to be associated with initiating the movement, so the difference in transit time between a whole jump and smaller movements, to or from rest somewhere between the rectangles, is likely to be small.

C13. As both the mean and median inscription times are approximately 300 ms (see above) a representative inscription time of 300 ms was chosen.

C14. In ideal experimental circumstances WM can hold 7±2 chunks (Miller, 1956). However, in rich dynamic task environments, about 4 chunks is a more realistic estimate (Cowan, 2001), because some chunks are needed to encode task sub-goal information. WM demands above 4 chunks will be considered to be large in GPoMS.

C15. This constraint concerns durations over which chunks must remain available in WM for processing. Chunks become active, enter WM, as the chunk hierarchy is processed in accordance to the depth first constraint, C1. It is assumed that they remain active whilst the
information they contain is needed. Motor programs for writing a letter must be available until the completion of the inscription. A letter chunk will remain active until its motor program is complete. A word chunk will remain active until its last letter has been retrieved. For example, in the chunk hierarchy in Fig. 1, when processing reaches letter L₃, at step 16, word W₂ is no longer required. The same holds for phrase and sentence chunks and the retrieval of their last word or sentence, respectively.

C16. *Production consistency* principle. It is hypothesised that an individual will employ the same strategy across all chunk levels for controlling the pen movement between successive letters.

C17. *Cognitive consistency* principle. It is hypothesised that individuals will for the sake of simplicity attempt to be consistent in their control of cognitive actions. GPoMS operationalises the principle in terms of minimising the use of different cues across chunk levels.

C18 *Participant consistency principle*. It is hypothesised that each participant will use a fixed set of strategies, one strategy for each chunk level, without switching within or between stimuli. Put another way, GPoMS models individual participants by aggregating data over stimuli.

See below for further discussion of both C16 and C17. The 18 modelling constraints define a space of potential GPoMS models. The models should predict the observed magnitude and range of pauses across chunk level, and the patterns of pauses across individuals, Obs A to K.

**Visualization tool**

A diagrammatic approach to depicting GPoMS models has been developed to aid comprehension of the patterns of behaviour predicted by the theory, which extends previous graphical tools for presenting task analytic models (John & Kieras, 1996; Vera, 2005). Fig. 4 shows GPoMS components, and Figs. 5 and 6 shows complete models to be described in detail below. This subsection explains how the elements of the diagrams encode the theoretical assumptions of GPoMS models.

- Time progresses from left to right.
- The grey boxes in the figures represent cognitive operations.
- The white boxes represent motor actions.
- The width of the boxes represents the duration of the operation/action (a scale line is provided). (The box height is constant but arbitrary.)
- The sequence of cognitive operations follows the depth first transition of the process hierarchy depicted in Fig. 1 (constraint C1).
- Processing of successive cognitive elements in the same sub-branch in Fig. 1 are offset by half a box height. Transitions between different sub-branches is shown by an offset of a full box height.
- All St(imulus), S(entence), P(hrase), W(ord) and L etter boxes stand for retrieval of a chunk of that type, so their width is 100ms (C9)
- The subscripts of the letters indicate the item being processed relative to its parent item; for example, in Fig 5, the rightmost L₂ is the second letter of W₂, which in turn is the second word of phrase P₂, which is the second phase of sentence S₁.
- M boxes represents the retrieval of the motor programs for letters so their widths are 100ms (C9)
- The I(nscription) box represents the inscription of a letter with a width of 300ms (C13).
- The J(ump) box represents the move of the pen between inscriptions, with a width of 100ms (C11).
• The triggers are represented either by arrows connecting the boxes or by immediately adjacent boxes (touching side-by-side or corner-to-corner) (C2).
• An arrow pointing vertically downwards indicates a cognitive operation triggering a motor action (C3). An arrow pointing upwards indicates a motor action triggering a cognitive operation (C3).
• An action that is a necessary precursor to another, but not a trigger, is shown by a horizontal or diagonal arrow between the two boxes representing the actions.
• The critical path is given by a sequence of adjacent boxes with no horizontal gaps, but which may be vertically displaced and connected by a vertical arrow. A GPoMS diagram is not well-formed unless there is at least one critical path.
• Double-head arrows represent pause durations.
• The grey WM-line attached to the top-right corner of each cognitive element indicates how long it must remain active in WM (C15). For example, in Fig. 5 the second WM-line from the top shows that the first sentence chunk (S₁) must be available in WM until it last phrase chunk (P₂) has been retrieved.

The minimum WM demand at a given instant is estimated by counting the number of grey boxes and WM-lines present at that time. For example, in Fig. 5 the dotted line near the centre crosses the WM-line for S₁, S₁, L₂, M₂ and P₂, for a count of five elements.

The format of the diagrams provides a tool for exploring the consequences of the theoretical constraints on the models, and to make predictions about the duration of pauses and the demands on WM.

Model components
A GPoMS model proposes a pattern of cognitive and motor actions that aim to predict the pauses at each chunk level for particular participants. Each model combines two types of mechanisms: (1) move-patterns and (2) chunk-schedules. Move-patterns are ways the pen may move, or jump, between successive rectangles. Chunk-schedules sequence chunk retrievals in coordination with jumps and inscriptions. Figs. 4A to D show the possible model components generated by different combinations of the move-patterns and chunk-schedules for the production of one letter. The inscription of the letter of interest is to the right of each diagram and the preceding letter inscription is to the left. For example, in Fig. 4A-L₁, I₁W₁L₂ is the target inscription associated with the 200ms pause generated by the retrieval of L₁ and M₁ (C6), and to the left M₁ is a prerequisite the retrieval of L₂, but I₁W₁L₁ is the actual trigger for L₂’s retrieval. In Fig. 4D-S₁A, both W₁ and I₁W₁L₁ are prerequisites to the L₁, and either could be its retrieval trigger. Figs. 4A, B, C and D address the processing at the letter, word, phrase and sentence levels, respectively. The first letter of a component’s label in Fig. 4 is a letter identifying the respective chunk level: L, W, P and S. Letter components, Fig. 4A-L₁ to L₃, process letters within a word (e.g., Fig. 1 steps 6-7), but not the first letter. Word components, Fig. 4B-W₁A to W₃B, process the first letter of the second and subsequent words within a phrase (e.g., Fig. 1 steps 11-13). Phase components, Fig. 4C-P₁A to P₃C, process the first letter of the first word of the second and subsequent phrases within a sentence (e.g., Fig. 1 steps 23-26). Sentence components, Fig. 4D-S₁A to S₃D, process the first letter of the first word of the first phrase of a sentence (e.g., Fig. 1 steps 1-5).
Fig. 4. GPoMS model components at four chunk levels: (A) letter, (B) word, (C) phrase and (D) sentence

The move-pattern mechanisms give three ways in which the pen may be moved between inscriptions. These correspond to the columns in Fig. 4, and the number in each component label identifies the move-pattern. (1) A jump is triggered by the completion of an inscription and the pen is immediately positioned at the next rectangle in preparation for the following inscription. (2) The pen waits at the rectangle of the inscription just completed until retrieval of the motor information for the next letter, which serves as the trigger to make a jump immediately preceding the inscription. (3) This pattern combines the previous two mechanisms with an initial jump away from the previous rectangle, a dwell whilst information about the next letter is retrieved, and a final jump to the next rectangle to begin the inscription. The three patterns represent behaviours in which a jump is considered to be (1) the last physical action of producing a letter, (2) the first physical production action of a letter, or (3) two separate physical actions that are each associated with a different letter inscription. These patterns implement some of the theoretical assumptions. Constraint C6 prescribes that a letter’s motor program should immediately follow the retrieval of that letter, thus L and M boxes in all components occur in pairs. Constraint C7 prohibits the co-occurrence of the retrieval of one letter whilst another is being inscribed, so the retrieval of a subsequent letter always occurs after the completion of the inscription of a preceding letter. An inscription cannot begin until information about the shape of a letter is available, so the inscription must follow the letter and motor retrieval. In component L1 (Fig. 4A-L1), the end of the inscription is the trigger for both the letter retrieval and the jump, which occur in parallel to give a
predicted pause of 200ms. In component L2 the letter retrieval occurs in series with the jump, so the pause is longer at 300ms. In component L3 there are two jumps so the total pause is 400ms. These patterns are repeated at all the other chunk levels.

Chunk-schedules are the second type of mechanism. The number of alternative schedules increases with the number of chunk levels being processed. In Figs. 4B to D, the third symbol of the component label is a letter designating the schedule. Consider chunk-schedule of component W1A, this is similar to component L1 but the word chunk W2 is retrieved in parallel with the inscription of the last letter of the previous word (which happens to be three letters long in Fig. 4B-W1A), so the predicted pause does not change and is 200ms (C4). In component W1B, W2’s retrieval is triggered by the end of the prior inscription, so the predicted pause is lengthened to 300ms, and W2 also becomes the trigger for the retrieval of its first letter, L1 (C3). The schedules for the other move-patterns W2A/B and W3A/B are similar extensions of move-patterns L2 and L3. In W2A and W3A, the trigger for word W2 is the beginning of the inscription I_W1L3, however, completion of the retrieval of M3 could also be its trigger with no effect on the pause associated with inscription I_W2L1. The overall pattern for the word components is repeated for the phrase and sentence components, Figs. 4C and 4D, but with three and four chunk-schedules for each move-pattern, because of the increased number of chunk levels. In the phrase components P1A, P2A, and P3A, the word and phrase (and sentence) chunk retrievals occur in parallel with the inscription, so the pause before the inscriptions remains the same as at the letter level. At the other extreme, phrase components P3A, P3B and P3C, word and phrase chunk retrievals follow the inscription, so the pause duration is 200ms greater across the move-patterns. Thus, the full range of pauses for all phrases is 200 to 600ms. The pattern is similar for sentence chunk retrievals, Fig. 4D, which results in a 200 to 700ms range for sentence pauses.

![Diagram](image)

Fig. 5. GPoMS “Model-1” composed of similar component across levels L2-W2A-P2A-S2A.

**Composing ideal models**

The components in Fig. 4 may be composed in linear sequences to make predictions. Each model variant consists of a fixed set of components, one for each level, in accord with constraint C18. For example, consider two models that make the same prediction, Figs. 5 and 6. Model-1, is composed of components L2-W2A-P2A-S2A which all have 300ms pauses. Model-2 has components L2-W1B-P2A-S1B, which also all have 300ms pauses. All the components in Model-1 have the same move-pattern (‘2’), and the word, phrase and sentence chunk-schedules are also the same (‘A’). In model-2
move-pattern and chunk-schedules vary. Overall, there are 1944 (3*6*9*12) potential models. Now, assume that the two models run over an imaginary stimulus consisting of two sentences, which are each comprised of two phrases, which are each two words long, which are each two letters long. Fig. 5 and Fig. 6 show how the two models, respectively, process such a stimulus, but only covering the first sentence and the start of the second sentence. Model-1 is simpler and as all its components have the same move-pattern and the same chunk-schedule mechanisms throughout, it conforms to the production consistency principle (C16) and the cognitive consistency principle (C17). The same pattern of jumps and inscriptions repeats regularly along the bottom of Fig. 5 with each letter retrieval cued by the completion of the preceding letter inscription. The trigger is the same for the retrieval of the sentence, phrase or word chunks, ‘n’ in Fig. 5; specifically, the completion of the jump just before the start of the inscription of the previous letter. In contrast, the behaviour in Model-2 is complex. The pattern of motor actions changes throughout with inscriptions sometimes immediately preceded or following a jump, or in some cases occurring in isolation from a jump. The triggers for chunks are not consistent, sometimes at the end of an inscription (‘n’ in Fig. 6) or at the beginning of an inscription (‘m’), so that the retrievals occur both serially or concurrently with the preceding inscription. Thus, Model 2 violates the production consistency principle (C16) and the cognitive consistency principle (C17), despite having the same overt pattern of pauses.

In both models the critical path (see above) switches between the cognitive and the motor actions in the production of every letter. Cognitive and motor actions each serve as triggers for both types of actions. In Figs. 5 and Fig. 6 the WM-lines to show the minimum time that each item must remain in memory, according to C15, and the numbers along the top of the diagram gives the WM-demand. For both Model-1 and 2, the WM demand rapidly increases as the task starts (Figs. 5 and 6, left) and mostly fluctuates between 4 and 5 chunks, but as the first sentence is complete the demand diminishes to 3 chunks before raising again at the next sentence.

To produce models for individual participants, the median pauses for all chunk levels of all participants were sorted into separate 100ms bins with central values of 200, 300, 400, 500, 600 and 700ms. Multiples of 100ms were chosen because the elementary move (C7) and retrieval times (C9) are 100ms and the representative inscription duration was 300ms. Pauses up to 250ms were grouped in the 200ms bin, pauses between 250ms and 350ms were classified as 300ms, and so forth. This classification of pauses smooths the data whilst preserving its overall structure at a theoretically
meaningful level of granularity suitable to match with the values predicted by the models. Table 1 (left) summarises the classification of participants. Pauses larger than 750ms are marked ‘X’ to indicate they are beyond the range of any component. Each line in Table 1 is a particular set of pauses and some include multiple participants, for a total of 23 distinct patterns. For example, participants 16 and 19 are the same with pauses of 200, 300, 300 and 400ms across the four chunk levels. Participants 31 and 2 are exceptional as they have no matching components for most, or all, of the chunk levels.

Table 1. Classification of observed median pauses and best fit models characteristics.

<table>
<thead>
<tr>
<th>Participants’ pause band classification</th>
<th>Feasible models</th>
<th>Best models</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Letter</td>
<td>Word</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>13</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>11</td>
<td>300</td>
<td>300</td>
</tr>
<tr>
<td>22</td>
<td>300</td>
<td>400</td>
</tr>
<tr>
<td>18</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>6</td>
<td>X</td>
<td>2</td>
</tr>
<tr>
<td>16, 19</td>
<td>300</td>
<td>400</td>
</tr>
<tr>
<td>10</td>
<td>600</td>
<td>600</td>
</tr>
<tr>
<td>20</td>
<td>X</td>
<td>4</td>
</tr>
<tr>
<td>12, 21</td>
<td>400</td>
<td>600</td>
</tr>
<tr>
<td>5</td>
<td>500</td>
<td>700</td>
</tr>
<tr>
<td>1.25</td>
<td>300</td>
<td>400</td>
</tr>
<tr>
<td>32</td>
<td>400</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>23</td>
<td>600</td>
<td>600</td>
</tr>
<tr>
<td>17, 27, 28</td>
<td>X</td>
<td>6</td>
</tr>
<tr>
<td>8, 9, 26, 29</td>
<td>X</td>
<td>4</td>
</tr>
<tr>
<td>14, 30</td>
<td>X</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>4</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>24</td>
<td>X</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>500</td>
<td>600</td>
</tr>
<tr>
<td>31</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Counts (N=32)</td>
<td>31</td>
<td>39</td>
</tr>
</tbody>
</table>

† Move pattern change that is not the smallest for all of this group’s models.

For each participant all the feasible models were initially enumerated. For example, participant 13 has pauses at 200, 200, 200 and 400ms across the chunk levels. There is one component that predicts pauses of 200ms at letter level, and similarly at the word and phrase levels. At the sentence level, however, three components predict a 400ms pause (i.e., S1C, S2B or S3A). Thus, participant 13 has three (1*1*1*3) feasible models. The number of feasible models for the participants are given in Table 1 (centre). The total number of feasible models is 132, which is an order of magnitude less than the 1944 potential models.

To determine the best model for each participant, the consistency principles (C16 and C17) were applied as filters to the feasible models. Consider the application of the cognitive consistency principle first. As noted above (C17), the operationalization of this principle focuses on changes to patterns of triggers across chunk levels. Specifically, pairs of components on successive chunk levels (e.g., word to phrase chunks) were examined and the number of elements in the second (phrase) component that had a different trigger than in the first (word) component were counted. For example,
consider the transition from word component ‘WW’ to phrase component ‘PP’ in Fig. 5. In WW the trigger for \( W_2 \) retrieval is a jump, but in PP the trigger of \( W_1 \) is the retrieval of \( P_2 \). For the other elements, the triggers do not change: triggers for jumps are retrievals of motor programmes; triggers for inscriptions are jumps; triggers for letter retrieval are the ends of inscriptions; triggers for motor program retrievals are letters. Thus, two trigger changes are needed for WW to PP. Repeating this for all the pairs of successive components in a model yields a count of trigger changes, and dividing this count by the number of transitions between components gives the mean number of trigger changes as a measure of the consistency of the model. (In some cases there are fewer than three transitions, because of the absent ‘X’ components.) The minimum mean count of trigger changes is 1, because every transition introduces an additional chunk which must have a trigger. A participant’s model with the smallest mean count of trigger changes is taken to be their best model. These models are listed in Table 1 along with their mean number of trigger changes. In just three cases the value is greater than the mean number of trigger changes across all 132 feasible models, which using a binomial test, assuming \( P=0.5 \) and \( n=22 \), is significant, \( p<.01 \). So, the subset of best models is meaningfully different to the set of feasible models.

The application of the production consistency principle (C16) is simpler and considers switches of move-patterns in the components between successive chunk levels. The fewer the switches the better. Ideally a model should have no changes and would have components labels that include identical digits (e.g., models M1, M11 and M18). For the best models – i.e., those with the lowest mean trigger changes – the right-most column of Table 1 gives the mean number of move-pattern changes. For the majority of the models the move-pattern is constant and the mean move-pattern change is zero. The other cases have values between 0.33 and 1. Across all 132 feasible models the mean proportion of move changes is 0.46 (SD=0.28). Of the 22 best models, just 5 involve a change of move-pattern, which according to a binomial test (\( P=0.5 \) and \( n=22 \)) is significant, \( p<.01 \). Further, in only one case does a feasible model have a lower move-pattern change than the best model (M9). In other words, the best models according to the cognitive consistency principle coincide with the best models under the production consistency principle, with just one exception.

**GPOMS models evaluation**

How well does the set of GPoMS models explain the observations? For ease of comparison, the grouping of participants in box plots Fig. 3A–C match the three groups defined by the letter level pauses in Table 1, and their order in each plot matches the order in Table 1.

**Obs-A.** The general overall increase in the duration of observed pauses with chunk level is consistent with the general increase in the duration of components over successive levels. The mean pause duration of the components are 300, 350, 400 and 450ms across the chunk levels from letter to sentence, respectively.

**Obs-B.** The growing range of the components’ pause values at each level potentially explains the increase in variability of the observed pauses with chunk level. For each move-pattern, the range of components’ pause values are zero, 100, 200 and 300ms, respectively, from letter to sentence levels.

**Obs-C.** At the letter level, the range of the component’s pause values covers all of the observed pause values, with just one exception (participant 2).

**Obs-D.** At the word level, the range of components’ pause values covers all of the observed pause values, with just two exceptions (participants 2 and 31).

**Obs-E.** At the phrase level, the range of components’ pause values covers all of the observed pause values, with just three exceptions (participants 2, 31, and 14 and 30, together).
Obs-F. At the sentence level, the range of model pause values covers 13 of the participants’ pauses. The predictions are poor for the models that involve components with longer pauses in general.

Obs-G. The large variety of individual differences is potentially explained by the variety of theoretically plausible combinations of components in the models, even when limited to the best component combinations.

Obs-H. Together, models M1 & M11, and M18, explain how participants (a) 13, (b) 1 & 25, and (c) 15 can have nearly constant pauses across the first three chunks levels. The constant pause values across chunk levels may be attributed to components with chunk-schedules that retrieve words and phrases in parallel with the previous inscription; components with labels ending in ‘A’, in Fig. 4. Thus, these retrieval times are hidden by the inscriptions rather than being added to the letter retrieval times. The ≈100ms difference in the pauses of groups (a), (b), and (c) may be explained by the three move-patterns and their 100ms differences.

Obs-I. Together, models M6, M16 and M10 explain the approximately linear increase of pauses for (d) participants 16 & 19 (across all chunk levels), (e) participants 8, 9, 26 & 29 (across the first three chunk levels), and (f) participant 5 (last three chunk levels), despite the difference in the pause magnitudes of the groups. The increase across chunk levels may be explained by chunk-schedules in which word and phrase (and sentence) retrievals follow the completion of the prior inscription; components with labels ending in ‘B’, ‘C’ and ‘D’ for the word, phrase and sentence levels, respectively, in Fig. 4. As these chunk-schedules trigger the retrieval of the letters in strict sequence, they incrementally add time with each successive chunk level. Again, the ≈100ms difference in groups (d), (e), and (f) may be explained by the three move-patterns and their 100ms differences.

The explanations for Obs-H and Obs-I represent the extremes of parallel and serial processing in the models. The remaining models are mixtures that fall between these extremes.

Obs-J. A partial potential explanation for why the pauses at the sentence level increase at a rate out of proportion to the lower levels is suggested by the absence of any of the first chunk-schedules at the sentence level: components S1A, S2A and S3A are absent from models in Table 1. Inspecting these components in Fig. 4D (top row), reveals that the time required to retrieve sentence, phrase and words in sequence equals the duration of an inscription. However, as the 300ms for inscriptions is merely a representative median value, in half of the cases, the inscription time will be too short for the retrieval of all three chunks, so the chunk-schedule pattern must change to S1B, S2B and S3B, or even S1C, S2C and S3C, if an inscription is particularly short. Thus, in general, pause durations at the sentence level may tend to be longer than expected, because there is often insufficient time for multiple retrievals in parallel with inscriptions. A case in point are participants 1 and 25 (model M11) who have nearly constant pauses in the 300ms band at lower chunk levels, but pauses in the 400ms band at the sentence level. All of the pauses of the model in Fig. 5 are 300ms, so it matches these participants at the lower levels, but inclusion of component S2A is unrealistic. The 400ms sentence pause for participants 1 and 25 could be modelled by any of the components S1C, S2B or S3A, but S2B is the best because no change of move-pattern is needed, as required by constraint C16.

Although the previous analysis explains why there are no short sentence level pauses, it does not explain why many sentence level pauses are greater than the longest pause for a sentence component (700ms); Xs in Table 1. GPoMS gives a potential explanation in relation to WM load. Consider again Fig. 5, but ignore the initiation of the second sentence. The count of the WM-lines peaks at six chunks, but is often 5 chunks, which is greater than the typical WM capacity for complex tasks (C14). Thus, items that entered WM least recently are likely to be lost from WM. In particular, the stimulus chunk is likely to be forgotten when processing the words of the last phrase of a sentence, simply due to fading activation. So, information about the next sentence may not be available and
therefore will require a deliberate cognitive action to retrieve that information, which will have time costs (Altmann & Trafton, 2002). It might add an extra 200ms to a sentence pause for operations, including recognising the need for the retrieval (50ms), deciding to take that action (50ms) and the retrieval itself (100ms). Obs F includes a cluster of participants with median sentence pauses around 900ms, which are compatible with this explanation. Further, if the stimulus chunk has decayed substantially, then a particular sentence may need to be retrieved by rehearsing the whole stimulus up to that target sentence. Such situations might explain the third cluster of participants whose sentence pauses group around 1500ms, Obs F. Further, participant 14 and 30’s unusually long phrase level pauses, Fig. 3B, might also be explained by a similar argument one level down at the sentence level.

Obs-K. Participant 2 is an outlier, with all pause durations exceeding the maximum predicted by the components. However, this participant’s median pause is 453ms which is just above the upper end of the band (450ms) for classification as a L3 letter component. So, Participant 2 and 31 might be considered as similar. Now, the previous argument may be applied down one further level, if it is assumed that some participants engage in phrase rehearsal in a similar fashion to sentence rehearsal. Such behaviour would be necessary for individuals’ with small WMs. Unfortunately, the WM capacity of the participants was not measured. Clearly, the explanations of Obs-J and Obs-K will require further empirical support, but GPoMS does at least provide a theoretically grounded basis for predicting excess sentence durations that extends but does not contradict the theoretical constraints, and which scales consistently over chunk levels.

In summary, GPoMS provides specific explanations of nine of the observations and potential explanations of the other two. Some observations range over all 32 participants, whilst others are specific to particular sub-groups or individuals. The framework of 22 models provides a coherent explanation of the observations that cover the quantitative differences among the groups and that systematically spans the four chunk levels.

Model discussion

Three questions were posed above for the GPoMS models to address, which are answered in turn.

(1) The first question was “What is the set of mechanisms (micro-strategies) that enables the interleaving of the cognitive and motor actions?” Two mechanisms have been identified to explain the interleaving of cognitive and motor operators that are consistent with the basic modelling constraints, C1-C15. The move-patterns control the movement of the pen between successive inscriptions of letters and the chunk-schedules control when chunks can be retrieved. The two types of mechanism are integrated by using each other’s operations as triggers. At each chunk level, the interactions among the two mechanisms generate a set of components for that level. A notable feature of the model is the complexity of the control involved, with the critical path switching between the cognitive operators and motor actions for the production of every letter in all variants of the GPoMS models. Another important feature is the just in time processing of chunks, which will be discussed below.

(2) The second question was “How can the interaction of the mechanisms accommodate the processing across the chunk levels given in Fig. 1, in a fashion consistent with the overall magnitudes and general patterns of measured pauses (Obs A-F)?” The overall range and general pattern of observed pauses is explained by the range of components at each chunk level. The set of components at each level is just sufficient to span the shortest to the longest pauses for most participants at the letter, word and phrase levels: in other words, the envelope of the predicted pauses matches the envelope of observed pauses. Although the components do not in themselves account for unusually long sentence pauses, these may be explained by the need to reactivate chunks from near the top of the chunk hierarchy (Fig. 1), which are lost from WM due to the high cognitive demands of the task.
The third question was “How can the interaction of the mechanisms explain the observed individual differences among the participants (Obs G-L)?” The explanation of the wide variety of participants’ individual pause patterns is challenging, because it must hold across participants with diverse behaviours, including those who exhibit short pauses (e.g., participant 13) and long pauses (e.g., participant 24), but also participants with distinct patterns, such as constant pauses (e.g., participant 13) or linearly increasing pauses (e.g., participant 8). However, GPoMS provides an explanation on three levels. First, theoretically diverse models are made possible through the combinatorics of the variants of mechanisms and the range of components, but the application of the production consistency principle (C16) and the cognitive consistency principle (C17) reduces this to a single best model. Second, among the best models, two main factors account for the large individual differences. (a) The move-patterns split the participants into the three main groups (the 3 box plots in Fig 3). (b) Within each of these groups, individual differences are produced by the varying degree to which cognitive processing occurs in parallel with motor processing, specifically the different chunk-schedules. Third, the WM demands of the models varies with the amount of parallelism, which may explain the adoption of different strategies by individuals with different WM capacities (see general discussion below).

The adequacy of GPoMS may be evaluated in terms parsimony and the quality of the fit to the data. At a fundamental level, GPoMS is parsimonious as it is based on two fundamental mechanisms and two key principles. Combinations of variants of the two mechanisms permit the GPoMS to model patterns of pauses at four chunk levels across diverse individual differences, but the principles identify a single ideal model for each participant. In relation of the fit of the models to the data, a particular concern is whether the multiple components at each chunk level provide so many degrees of freedom that the model for each participant constitutes an over-fitting of the data. The freedom available for matching an ideal model to participant data is relatively limited because the production consistency (C16) and cognitive consistency (C17) principles are strong constraints. C16 eliminates two thirds of the components, because the move-pattern should be consistent at all chunk levels and the specific move-pattern is determined by the pause band for letter chunks. For example, given a median letter pause in the 200ms band, the only components on the left hand side of Fig. 4 are available to a model. C17 limits the selection of components on each successive chunk level to components that require the fewest changes of triggers: Model-1, Fig. 5, is preferable to Model-2, Fig. 6. The extent to which the best models for the participants match this constraint is shown by their significantly smaller mean number of trigger changes compared to all plausible models (Table 1). The simultaneous satisfaction of both principles by each model means that they are not arbitrary and suggests that despite the large individual differences, the process of writing has a much underlying simplicity.

Further, the choice of the banding for the pauses in Table 1 provides another potential degree of freedom for modelling, however the choice of 100ms bins was theoretically motivated and small changes to the size of the bands would only impact a few participants at the boundaries between bands, so overall pattern of pauses and the overall distribution of the models would not be greatly affected.

The match of models to participants’ pauses is good at the letter, word and phrase levels but poor at the sentence level. The spread of sentence pause values is large, with just 13 participants fitting within the expected pause bands. GPoMS provides two natural explanations. The first explanation was considered above, in which participants with small WMs may execute extra retrievals to refresh sentence (and perhaps phrase) chunks in WM. Another explanation is that participant might not encode every word as single chunk but as a compound of two or more sub-chunks (e.g., “be-cause”). Such cases would introduce an additional level into the chunk hierarchy between the letter and word levels with a consequent increase in pause durations of phrase and sentence levels of 100ms. This
explanation requires the relaxation of constraints C1 and C18, as edits to the chunk hierarchy will be required for particular parts of some stimuli and for specific participants.

GPoMS does not explain individual differences through different settings of parameter values, rather it follows those models that explain performance differences in terms of alternative sequence of actions (e.g., John, 1996, section 3.5; Gray & Boehm-Davis, 2000, Vera et al., 2005). Such models examine how different action durations or relatively local changes in the order of actions impact the critical path and hence change the overall processing times. The approach to developing GPoMS differs in various ways. (1) Rather than making predictions about the overall duration for a task, GPoMS attempts to match multiple temporal measurements at different levels of the task, that is, the pauses at alternative chunk levels. (2) GPoMS attempts to systematically enumerate the space of possible components derived as variants of the two fundamental mechanisms. (3) The models were constructed for individual participants, or small groups, rather than attempting to build generalised models of data aggregated over a large number of participants in different experimental conditions. (4) One ideal model from among the feasible models for each participant was found by applying the two hypothesised principles to “triangulate” upon a single best model for each participant.

Overall discussion

How is writing controlled in this task? At the most general level, the GPoMS models are consistent with the serial depth first processing of chunks (Johnson, 1970). The control of the task may be characterised as the coordination of two processes, the retrieval of chunks from the hierarchy and the production of the succession of individual letters. This coordination has various key features. First, across all models, the critical path switches back and forth between cognitive operations and motors actions for each letter (Figs. 5 & 6) in a seemingly complex fashion. However, the process possesses an underlying simplicity in various ways. It revolves around the just-in-time scheduling of letter production with processing primarily organised around the retrieval and outputting one letter at a time. What differs between models is how the retrieval of higher level chunks is coordinated in relation to the letter processing. There is also simplicity in the invariance of the particular move-pattern mechanism and the consistency in the action triggering conditions at different chunk levels. The WM demands of all the models is predicted to be high, at the upper limits of the range of WM capacity reported for continuous tasks (Cowan, 2001).

A striking finding of the present experiment is the existence of substantial individual differences in performance, despite the imposition of many constraints to minimise recognised sources of individual differences. The task design attempted to induce a common chunk hierarchy for each stimulus in the memory of participants and to avoid potential differences in perceptual and stimulus encoding. The motor skill demands were made uniform through the use of a common response grid. At an elementary level, the multiple components available from the combination of the variants of the two basic mechanisms provides a large space of possible models. In terms of the best models for each participant, GPoMS explains the range of differences in terms of the three different move-patterns and the degree of parallelism between cognitive and motor actions. Individual differences have previously been attributed to varying degrees of parallelism, but with respect to overlapping perceptual and cognitive behaviours, for instance made possible through variations in task goal structure (Janssen, 2012). In contrast, the individual differences found here occur even when the task structure is the same (and among models with the same move-pattern) and further GPoMS posits a link between the degree of parallelism and WM demands.

The GPoMS model provides an interesting comparison to Crump and Logan’s (2010) two-loop model of transcription typing, in which an outer loop processes at the word level and passes words to the inner loop to be typed. The GPoMS models suggest that a characterization of writing as two
distinct loops may be specific to transcription typing, because some of GPoMS’s components for the processing of words, phrases and sentences involve the inscription of a letter from the current word at the same time as the retrieval of next word, phrase or even sentence, specifically components W1A-W3A, P1A-P3A and S1A-S3A in Fig. 4. Hence, the privileged role of words in the two loop model may be specific to the task of transcription, the activity of typing or both. The relatively long duration of motor actions in pen-based inscription of letters, as opposed to keypresses, provides an opportunity for the overlapping to occur in the processing of words.

The present findings reinforce Gray and Boehm-Davis’s (2000) claims that analysis at the milliseconds level is necessary for accounts of interactive behaviour, because micro-strategies have a causal impact on strategies at greater times scales. GPoMS goes further by demonstrating that even in a relatively simple task a myriad of different strategies may exist that produce large performance differences. The existence of diverse strategies in such a highly constrained task re-emphasises Newell’s (1973) caution about the illegitimacy of aggregating data over subgroups in the same experimental condition when participants are using different strategies, because those different strategies constitute an uncontrolled factor.

In some previous models, individual differences have been modelled by positing alternative sequences of operators at a local level within models (e.g., Gray & Boehm-Davis, 2000; John, 1996; Vera et al., 2005). GPoMS builds on such models but the overall approach differs. GPoMS assumes there are independent basic mechanisms – move-patterns and chunk-schedules – whose variants are combined into components. Then, from feasible models built from the components, ideal models were identified on the basis of two core principles – production consistency and cognitive consistency. This approach has some theoretical justification. Positing distinct mechanisms is reasonable, because writing naturally appears to involve types of activities that occur independently in other tasks. On the one hand, tools are often moved from location to location where some discrete action is performed, such as spreading butter or slicing food, without the cognitive processing of a complex hierarchy of chunks. On the other hand, certain tasks involve the processing chunk hierarchies without the use of a tool, such as silently reciting a poem or verbalizing learnt material (e.g., Buschke, 1976; McLean & Gregg, 1967). The production consistency principle asserts that an individual will be consistent in their use of a single pen movement strategy across chunk levels. If this was not the case, then two unrealistic consequences occur. (a) Deliberate decisions would have to be made to change strategy during the on-going process of writing, which would involve additional processes, such as the recognition of the need to change, creating an associated sub-goal, selecting a new strategy and its retrieval. However, the fine grained pause data and the GPoMS model shows there is insufficient spare time for such processes. (b) Alternatively, particular strategies may be specifically associated with each chunk level and automatically selected, which would not involve addition processes. However, this begs the question of why and how such associations would have been learnt in the first place.

Similar arguments may be made for the cognitive consistency principle, but they will apply to triggers rather than whole strategies. Further, the initial separate consideration of motor operations and cognitive operations gives clarity in their specification and the systematic enumeration of all pairs of mechanisms, which aids the systemic identification of feasible components and helps guard against some of the recognised sources of error that may occur in building task analytic models by hand (Vera, John, Matessa, Rimington & Freed, 2002). Our overall approach is reminiscent of Vera et al.’s (2002) composition of templates, but it differs in its initial emphasis on enumerating components.

The substantial individual differences invite two theoretical questions about the nature of optimal task performance that relate to approaches assuming the principle of bounded rationality, in which the cognitive system is presumed to seek a strategy from among feasible strategies by optimising some utility function (e.g., Howes, Lewis & Vera, 2009). An analyst may define the function in terms
of maximizing task performance score or minimizing task time, for instance. So, the first question is whether there is a plausible model better than the fastest GPoMS model? Such a model would challenge the constraints upon which GPoMS is based. The fastest plausible GPoMS model is comprised of components L1-W1A-P1A-S1A, which matches participant 13, up to the phrase level. Theoretically, the quickest imaginable model has a critical path comprised just of motor actions with no pauses between the jumps and inscriptions. The construction of this model reveals that it is not cognitively feasible for various reasons. The activation of the next letter chunk must occur whilst the current letter is being inscribed, that is abandoning constraint C7, which would likely result in inscriptions errors due to multiple active motor programs. The WM demands of the model are larger than the fastest plausible model, violating C14, and continuously high because chunks must be buffered well before they are needed. The retrieval of sentence chunks occurs two letters prior to the inscription of its first letter and involves a trigger that is rather arbitrary as it as an action from a different branch of the chunk hierarchy. The implausibility of this hypothetical model reinforces the apparent benefits of the just-in-time retrieval of letters before inscription.

The second optimality related question is why such a large range of individual differences exists among the participants given that they are adults who would have had ample opportunity over years to adapt their writing performance to some best overall strategy? Again, this may be considered from motor and cognitive perspectives. Each move-pattern provides a distinct set of triggers for retrievals and inscriptions, so there is a cost to switching between move-patterns; this idea underpins the production consistency principle. As someone begins to learn to write, any preference for one move-pattern could mean that it is practiced more than the others, which would eventually result it becoming preferred and automatic. Then cost of switching could be considerable and thus potential benefits of using other move-patterns, when more complex writing tasks are encountered, may not be sufficient to overcome the cost of switching. A bias towards some operations may develop because they have become more automated through practice, even though their initial selection may have been relatively arbitrary. So more optimal strategies may be blocked by the natural inertia of over-learnt sub-optimal strategies acquired earlier in task skill development.

![Fig. 7. GPoMS “Model-3” with only serial components L2-W2B-W2C-S2D.](image)

Another reason why individual may not switch to more efficient strategies is the differences in WM demands of alternative cognitive chunk-schedules, which are due to the varying degrees of parallelism of the chunk retrieval with inscriptions. Fig. 7 depicts Model-3 L2-W2B-W2C-S2D that
has no parallelism, which in inscriptions are completed before new retrievals are made. Thus, the WM demands are typically less than 5 chunks, whereas in Figs. 5 five or more chunks must be present in WM. As the effective WM for ongoing tasks is about 4 chunks, it is plausible that individuals with smaller WM capacities might select the more serial chunk-schedules as means to manage their WM load, despite the available performance efficiencies available from other schedules. Unfortunately, WM capacities of participants were not measured as part of the experiment, so this hypothesis cannot be tested here.

The above explanations of the persistence of the individual differences are speculative and will require further investigation, but nevertheless there is a wider implication for studies of human bounded rationality that seeks to identify tasks strategies on the basis of optimizing a utility function (Howes, et al., 2009; Gray, et al., 2006; Payne & Howes, 2013). For certain tasks, sophisticated utility functions may necessary. Utility function may need to address trade-offs between multiple competing factors, such as task completion speed, performance accuracy and WM demands. They may need to incorporate independent factors to address selection pressures that act separately on different mechanisms underpinning the overall task strategy, as witnessed by the different explanations for move-pattern versus chunk-schedule mechanisms selection. Further, optimality models may require utility functions, or other means, to encompass historical inertia for certain strategies arising from prior learning early in task acquisition.

Finally, this study has introduced two methodologically innovations. First, we have introduced a visualisation for the development of CPM-GOMS models, that is a refinement of Pert-style charts (cf., John et al, 2002; Patton & Gray, 2010). The construction of the diagrams in Figs. 5-7 constitutes an investigation of the nature of critical paths, as represented by aligned horizontally operators and actions. These diagrams have the benefit that the instantaneous WM can be read directly from the diagrams. The second innovation is the use of individual differences in a highly constrained task to explore sources of strategic differences. The modelling of the variation among individuals provided the main source of empirical leverage to identify plausible mechanisms in the model. This contrasts to approaches in which a prototypical model is built and model variants are used to explain data aggregated over participants in separate experimental conditions, with individual differences attributed to stochastic variability.

References


Cheng, Peter C. H., & Rojas-Anaya, Hector. (2005). Writing out a temporal signal of chunks: patterns of pauses reflect the induced structure of written number sequences. In B. G. Bara, L. Barsalou & M. Bucciarelli (Eds.), *Proceedings of


