Cheng, P. C.-H., & Albehaijan, N. (2020). Some determinants of chunk size in sequential behavior: individual differences in the transcription of alphanumeric strings. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), Proceedings of the 42nd Annual Conference of the Cognitive Science Society (pp. 1164-1170). Austin, TX: Cognitive Science Society. https://cognitivesciencesociety.org/cogsci20/papers/0214/0214.pdf

Some Determinants of Chunk Size in Sequential Behavior: Individual Differences in the Transcription of Alphanumeric Strings

Peter C-H. Cheng[1] **(p.c.h.cheng@sussex.ac.uk)**
Noorah Albehaijan[1,2] **(N.albehaijan@sussex.ac.uk, naalbehaijan@iau.edu.sa)**
[1]Department of Informatics, University of Sussex,
Brighton, BN1 9QJ, UK
[1,2]Department of Computer Science, Imam Abdulrahman Bin Faisal University,
Jubail, P.O. Box 12020, Saudi Arabia

Abstract

Studies have shown that temporal chunk measures in transcription tasks can be used to assess competence in various domains. However, in other tasks, chunking strategies and thus performance differences can be highly variable across participants. If such individual differences are also large in transcription tasks this would undermine the use of chunk-based competence measures. Using four stimuli with fixed spatial structures this experiment demonstrates that there is good consistency in chunking strategy across 52 participants in two types of transcription tasks. The experiment spans 16,000 data points.

Keywords: chunking; individual differences; sequential behavior; transcription; competence measurement.

Introduction

Chunking is a foundational concept of cognitive science with a long history (Bryan & Harter, 1897; Miller, 1956). It underpins many theoretical explanations of cognitive phenomena, such as: perception (Miller, 1956); working memory (Cowan, 2002); learning (e.g., Anderson, 2000; Gobet et al., 2001); expert–novice differences (Chi, Glaser & Farr, 1988); control of sequential behavior (Rosenbaum, Kenny & Derr, 1983; Cheng & van Genutchen, 2018); stages of processing, including perception and encoding versus motor program preparation and physical output (Verwey, Shea & Wright, 2015). A chunk is a group of mental elements that have strong associations with other elements within the group and weak associations to elements in other groups. How big is a chunk? Miller (1956) suggests 7±2 components, Cowan (2001) four, Gobet & Clarkson (2004) three or even two. The difference in these estimates is explained by the specific definition of chunk adopted and the nature of the target task. Miller (1958) focused on the perceptual span of short-term memory (STM), Cowan (2001) on tasks that fully engage working memory (WM), and Gobet & Clarkson (2004) specifically on chunks as perceptual schemas, *templates*, in the domain of chess.

As learning involves the acquisition of a rich network of chunks (Gobet, at al., 2001), it should theoretically be feasible to assess an individual's comprehension of a target knowledge domain by measuring properties of that individual's network of chunks. In other words, we might be able to evaluate a person's knowledge or competence by determining the size of their chunks, notwithstanding the previous comments about chunk size.

Competence assessment using chunk measurement

Cheng and colleagues (Albehaijan & Cheng, 2019; Cheng, 2014, 2015; Cheng & Rojas-Anaya, 2006; Zulkifli, 2013) have developed assessments of comprehension with *temporal chunk measures* using simple transcription tasks in order to assess the structure of chunks in learners' memories. They chose transcription tasks, rather than, say, recall or problem-solving activities, because test takers will generate nearly the same overall content, which has valuable benefits in terms of the coding of sequences of responses and ease of comparison across test takers (Cheng, 2014). Two methods have been developed using *temporal chunk signals* that operate at different time scales.

In the *pauses in writing* method (Cheng, 2014; Cheng & Rojas-Anaya, 2006; Zulkifli, 2013) chunking behavior was measured using the duration between the writing (scribing) of individual characters, with the stimulus constantly on display. For instance, in non-cursive writing, a pause is the time between the placement of the pen on the paper at the start of a target stroke and the earlier time of lifting the pen from the paper at the end of the preceding stroke. Pauses occur between characters and sometimes within characters (e.g., the '-' in "t"). Pause-based measures have a characteristic timescale of the order of 100 ms. In transcription tasks, pauses at the beginning of chunks are substantially longer than pauses within chunks – at least twofold is typical. Further, Cheng & Rojas-Anaya (2008) found three distinct levels of pauses in the writing of artificial sentences. Van Genutchen & Cheng (2010) found four levels in the writing of sentences from memory at the start of sentences, phrases, words and letters. As a measure of competence, the third quartile of pauses, *Q3*, is a particularly useful general measure across task types and domains (Cheng, 2014; Cheng, 2015; Zulkifli, 2013).

The second method is *cluster-based* (Albehaijan & Cheng, 2019). In this method participants perform a deliberate action of pressing a key in order to view the stimulus, which is masked during periods of writing when the key is released. The small set of characters that participants chose to process together on each view is called a *cluster*. Three chunk measures were based on such clusters: (i) the *number* of components per cluster; (ii) the *view duration* taken by the

participant to select and encode a cluster from the stimulus into memory; (iii) the *writing duration* to reproduce a cluster from memory (following each view of the stimulus). The cluster-based method has a characteristic timescale of the order of 1 second. Note that a cluster is not necessarily a chunk, because a cluster may include more than one chunk.

In the present experiment, both methods are investigated. The pauses in writing method is addressed in the *Constant Display* (CD) condition of the experiment. The cluster-based method is addressed in the *Deliberate View* (DV) condition.

Empirical evaluations of both methods have spanned the domains of mathematics (Cheng, 2014, 2015), English as a second language (Zulfliki, 2013), and programming (Albehaijan & Cheng, 2019). Strong correlations between the temporal chunk measures and independent measures of domain competence were obtained with values commonly in the range 0.6-0.7. Hence, the approach appears to have some potential for competence measurement.

Other sub-second temporal measures of action features in continuous writing and drawing tasks have been developed in order to assess competence (Ovaite et al., 2018; Stahovich & Lin, 2016). Those approaches used machine learning techniques to find predictive combinations of features, but require the logging of large amounts of data across many sessions. In contrast, the methods of Cheng and colleagues require relatively small amounts of data, because the measures are theoretically derived from chunking theory.

Although the assessment of domain knowledge using temporal chunking signals appears to promise some utility, theoretical concerns may be raised about using measures of chunking in transcription tasks. One concern is whether performance manifest in simple transcription tasks can realistically represent the full range of component skills in which competence might be expressed. Another concern is the potential presence of substantial individual differences in sequential tasks – this is the focus of this paper.

Do individual differences impact chunk measures?

Newell's (1973) injunction about the inappropriateness of aggregating psychological data over different task strategies applies to the assessment of competence using temporal chunk measures. Differences in the strategies individuals use to encode or to produce chunks in a task is a potential source of variability that may confound the accuracy of chunk measures by introducing a factor that is unconnected to the test taker's chunk structure. For instance, John (1996) noted the existence of strategic differences in transcription typing. More severely, Gray and Boehm-Davis (2000) found that differences of just tens of milliseconds may produce different task strategies. Yet more dramatically, Cheng & van Genutchen (2018) discovered that substantial strategic differences occur between individuals in their writing of the same sentence from memory. Participants memorized sentences with a fixed five-level hierarchical structure, which they then wrote from memory in a non-cursive fashion. Cheng & van Genutchen (2018) analyzed pauses between successive strokes, letters, words, phrases and sentences. They found 23

S1. **9 g 2 b 6 d 7 f 5 w 1 c . . .**
S2. **z8 3b n2 7a y5 3e m9 . . .**
S3. **2b3 n5w 6t2 a7n 9g4 . . .**
S4. **2d6v y7b4 3e9k 2t6r . . .**

Figure 1. Sample stimuli (initial 12 to 16 characters)

different strategies were exhibited by their 32 participants, which they explained in terms of the alternative scheduling patterns that were used to interleave the retrieval of chunks from memory and the initiation and execution of motor actions. The different strategies impact the duration of pauses at the start of the writing of new chunks by as much as 500 ms, which is noteworthy as this is in the absence of any effects due to stimulus perception and encoding processes, as stimuli were written from memory.

The presence of such strategic differences might substantially hamper the reliability of temporal chunk measures by being sources of variability that is quite unrelated to performance differences due to the chunk structures, which the chunk measures are intended to record. So, this paper examines whether temporal chunk measures of competence may be susceptible to such individual differences that would undermine the viability of stimulus transcription approaches to competence measurement.

The approach we adopt follows McLean & Gregg's (1967) and Cheng & Rojas-Anaya's (2005, 2008) paused-based chunking studies. They induced chunk structures into participants' memory in order to achieve consistent hierarchical chunk structures across participants. In an analogous fashion, participants in the present experiment will transcribe arbitrary stimuli with a fixed structure determined by their spatial layout. Fig. 1 shows truncated examples of the stimuli. The characters alternate between a single letter and a single number, which were selected at random. Stimuli have the following structure: S1 – uniform list; S2 – sets of 2 characters; S3 – sets of 3 characters; S4 – sets of 4 characters. S3 is 39 characters long and the rest 40.

Each character in a stimulus is an *element* of this domain as participants are familiar with roman letters and Hindu-Arabic numerals. Being randomly generated, no particular sequence of characters is likely to be meaningful to participants, so it is presumed that participants will process characters in small groups that are bound temporarily in working memory. Such groups of characters are called *clusters* rather than *chunks*, for two reasons. First, the connection between the characters is one of *binding* in WM rather than *association* in long term memory. Second, in a single view of the stimulus a participant may encode sets of characters as one chunk or as a hierarchical structure with two (or more) groups containing two (or more) characters (Rosenbaum, Kenny & Derr, 1983), which leaves open whether the whole group or each sub-group should be properly designated as chunks.

How can the transcription of these stimuli inform whether individual differences may degrade temporal chunk signal measures of competence? This is the logic. (a) Encoding: assume that participants use the sets in the stimuli to encode the clusters of characters they use during transcription. (b)

Production: given clusters of the same size, assume participants are consistent in their output of those clusters, then the size of groups of characters produced will be consistent and the patterns of pauses associated with clusters of that size should also be consistent. If (a) and (b) both hold, then this provides good evidence that effects of individual differences are not substantial. However, if participants are not consistent in their size of clusters they encode or not consistent in their patterns of pauses, then it is impossible to claim that individual differences are not having an adverse impact, because either (a) or (b) may be responsible. If (b) is responsible then individual differences have a negative impact. If (a) is responsible the experiment cannot provide evidence for the absence of any negative impact of individual differences. Either way the claim that individual differences are not an issue cannot be dismissed. With S1 it is likely that participants will vary in the size of clusters, so encoding condition (a) may not hold. As we will see, both conditions (a) and (b) appear to hold for S2, S3 and S4.

Method

The experiment was conducted as part of a larger study about programming competence using chunk measures.

Design. The experiment was a within participants design, with two transcriptions modes: *Constant Display* (CD) and *Deliberate View* (DV). The CD mode implements the paused-based method and the DV mode the cluster-based method. Stimuli S1 to S4 (Fig. 1) were used.

Participants. The 52 participants were students and members of staff in the Department of Informatics at the University of Sussex. Their age ranged from 18 to 59 years (mean=22.3), and 35 were male, 16 females and 1 unspecified. They received £8 for participating.

Materials. The experiment was conducted using a standard graphics tablet (Wacom – Intuous3). It was connected to a PC running a logging program written in our lab that provides millisecond accuracy for the capture of pen strokes. Participants wrote with an inking pen on a response sheet on the tablet. The response sheet was printed with a grid of 17 lines; each consisting of 42 spaces for the writing of separate characters. The sheet was designed for non-cursive writing in order to provide character level pause data. Participants adapted to this style of writing quickly and like Cheng (2014) and Albehaijan & Cheng (2019). It does not appear to have adversely affected other aspects of their performance.

Procedure. Half of the participants started in the CD mode and the other half started with DV mode. In both CD and DV modes, the stimuli were presented to the participants in the same order, from S1 through to S4.

Participants held the pen in their preferred hand. They were trained to: start writing at the beginning of each line; start writing as soon as the stimulus is revealed; copy the stimuli as quickly and as accurately as possible, but without spaces; continue writing without correcting if they made a mistake; start each trial with writing a hash (#) on the line above the response field; in the DV mode, to hold down the special key to reveal the stimulus and write only when the stimulus key is released. The participants easily complied with these requirements and were fluent after a few practice trials. The purpose of the initial # is to ensure a pause data point is available for the first character. Cheng (2014) and Albehaijan & Cheng (2019) reported that imposing similar trial requirements did not affect the reliability of the results.

From the data logs in the DV transcription mode we calculated the number of characters per cluster. For both the DV and the CD modes the pause before the making of each pen stroke was computed and locations of strokes were used to label strokes as the first or subsequent stroke in a character.

Table 1. Mean size of clusters for each stimulus

Stimulus	S1	S2	S3	S4	Mean
(1) Mean size	4.06	4.30	4.13	4.01	4.13
SD	0.699	1.099	0.978	0.559	0.833
(2) Mean individual consistency	0.600	0.707	0.683	0.713	0.676
SD	0.184	0.189	0.212	0.210	0.199
(3) Mean consistency mode=stim	0.669	0.740	0.776	0.768	0.738
SD	0.067	0.229	0.007	0.187	0.123

Results

The DV and CD modes are analyzed in turn. All between participants comparisons use 2-tail t tests and are reported as four-tuples in the form [n, t, p, d]. In most cases n=52, but data was lost in two trials due to a technical problem.

Cluster size in DV mode – cluster-based method

The cluster sizes (number of characters between stimuli views) and mean cluster sizes are given in row 1 of Table 1. There is little difference in size among the stimuli, with a value close to four, which matches Cowan's (2001) chunk size.

However, this uniformity is an artefact produced by aggregation across participants. The modal cluster size for each participant was identified and the frequency of these cluster sizes are plotted in Fig. 2 for each stimulus. The mode is used because clusters are composed of discrete numbers of elements and to provide an unambiguous comparison to the character set sizes of the stimuli. (In the few cases with more than one modal cluster size, the contribution of each cluster size to the mean frequency was weighted in proportion to the number of modes.) Participants on S4 are highly consistent with nearly all having the most common modal cluster size of 4. The majority of participants (≥30) in S1, S2 and S3 have the same most frequent modal cluster size of 4, 4 and 3, respectively, so are reasonably consistent. Of the modal cluster sizes that are not the most frequent in S2 and S3, it is interesting that they are double or half the size of the most common mode for that stimulus. In contrast, S1's next most frequent clusters sizes are 3 and 5, which suggests that cluster size in S1 may reflect participants' WM capacity, with a typical size of 4 elements per cluster (c.f., Cowan, 2001) and a normal distribution of variability around that.

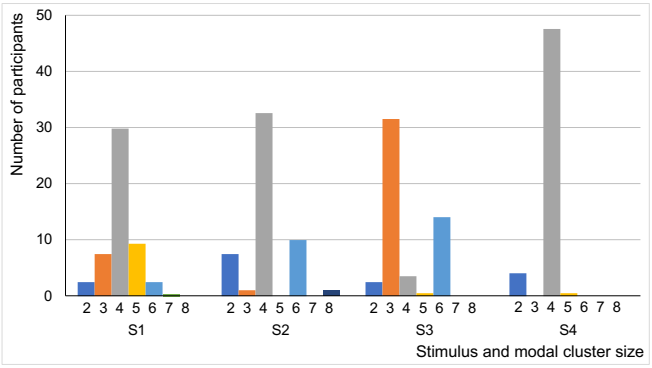

Figure 2. Frequency of modal cluster size across stimuli.

At a lower level, the *individual consistency* is the proportion of a participant's clusters whose size equals their personal modal size (i.e., number of clusters of modal size divided by total number of clusters, which is unity when all a participant's clusters are the same size). Table 1 row (2) gives the mean individual consistency across participants. Overall, participants use their modal cluster size about two thirds of the time, but this is clearly lower for S1 compared to the rest with medium, or near medium, effect sizes (in t tests S1 vs S2 = [52, 3.70, 0.0005, 0.554], S1 vs S3 = [52, 2.68, 0.006, 0.411], S1 vs S4 = [52, 4.11, 0.001, 0.630]).

Mean *consistency mode=stim*, Table 1 row 3, is the individual consistency for just those participants whose personal modal cluster size equals the most common cluster size of the stimulus (i.e., S1-4, S2-4, S3-3, S4–4; see Fig. 2). Overall, these participants use their modal cluster size about three quarters of the time, but this declines to two thirds of the time for S1. So, participants have a moderate level of intra-task consistency, especially for S2, S3 and S4.

Drilling down further, Fig. 3 shows the distributions of cluster sizes used by each participant across the stimuli. Clusters with low frequency may be due to omissions of characters in writing, so the few isolated instances of clusters size 3 in S4 may be disregarded, for instance. As expected, Figs. 3.S1-S4 shows modal cluster sizes that are consistent with the values in Fig. 2. Close inspection of the figures reveals an interesting pattern among stimuli in terms of the numbers of clusters of each size. S4 is dominated by clusters of 4, with small "tails" of size 2, 3 and 5 clusters. S3 is dominated by

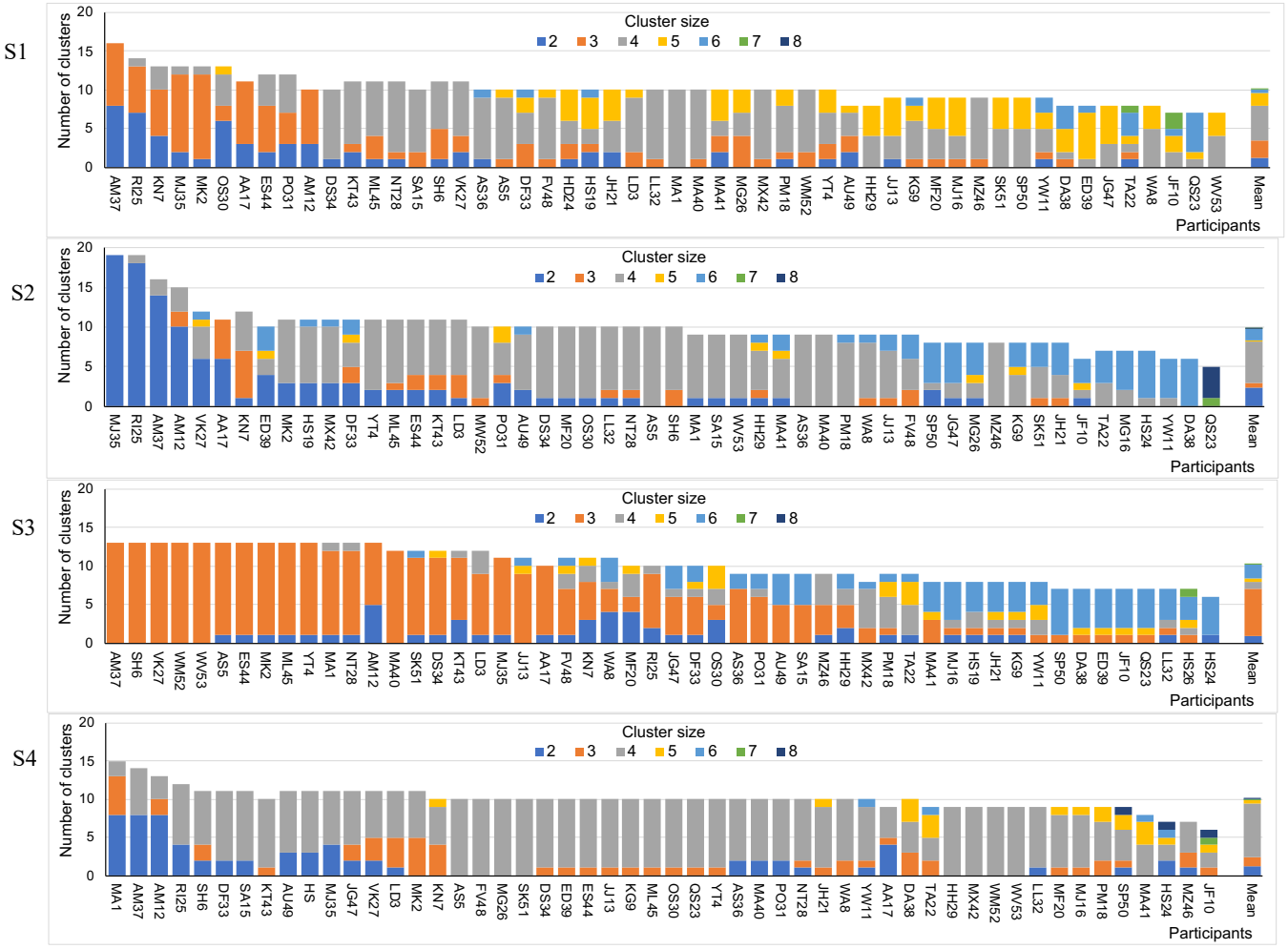

Figure 3. Cluster patterns across stimuli for each participant. Participants are ordered by their total number of clusters.
(The bar heights do not decline monotonically because the few instances of cluster sizes of 1 and ≥8 are not shown.)

Figure 4. Stacked graphs for stimuli's z-score pauses for successive characters.
Whole stimulus (left) and positions within groups of size 2 to 5 (right). Horizontal line through each curve is z=0 for that stimuli and distance between lines equals 3 z units. Thick lines – means; thin lines – SDs.

size 3 clusters, but less so than the size 4 clusters in S4. S3 also has tails of size 2, 3 and especially of size 6, which are clearly more substantial than the tails of S4. In S2 clusters of size 4 are most frequent, but less consistent than the most dominant clusters in S3 and S4. Further, S2 has prominent tails of cluster size 2 and 6. This pattern of declining dominance of the most frequent cluster size and the growing size of tails is most apparent in S1, which has clear tails for cluster sizes of 2, 3 and 5. In sum, the consistency is low for S1 and grows to a maximum for S4.

Pause measures in CD mode – pauses in writing

Pauses are the primary measure in the CD mode. First quartile, Q1, and third quartile, Q3, have been useful in previous studies of transcription performance (e.g., Cheng, 2014, 2015). Q1 represents the typical pause before a character (within a cluster) and is relatively constant across stimuli. Q3 represents the duration of long pauses that likely occur at the beginning of clusters. Table 2 shows mean Q1 and mean Q3 pauses of all strokes. Note the greater than twofold difference between Q1 and Q3 values, which are large effect sizes (Table 2, Q3 vs Q1). Q3 decreases from S1 to S4, monotonically, and the successive differences are clear and large for S1 to S2, and S2 to S3, but less so for S3 to S4 (Table 2, Q3: S*k* to S*k-1*). The longer pauses suggests that participants do more processing in stimuli with the small sets of characters.

Two approaches are taken to analyze the consistency of cluster sizes in the CD mode. Because we are interested in strategic differences rather than simple differences in latencies, both approaches use mean z-scores of pauses (computed using the mean and SD of each participant on each stimulus).

Table 2. Q1 and Q3 pauses across stimuli

Measure	S1	S2	S3	S4
Q1	302	289	261	279
SD	109	100	97	101
Q3	1035	955	823	774
SD	328	349	293	256
Q3 vs Q1 [n, t, p, d]	[51, 16.1, $<10^{-6}$, 1.66]	[52, 13.7, $<10^{-6}$,1.52]	[51, 14.8, $<10^{-6}$, 1.58]	[51, 16.3, $<10^{-6}$, 1.57]
Q3: S*k* to S*k-1*	–	[51, 3.51, 0.001, 0.24]	[51, 3.35, 0.002, 0.41]	[51, 2.51, 0.054, 0.165]

The first approach finds the duration of pauses of the first stroke made in each cell of the response sheet, see Fig. 4. S1 is at the bottom of the stack of graphs and the long sequences are pause z-scores for each of the 39 or 40 characters in a stimulus. The four short sequences on the right are groupings with different numbers of successive characters (e.g., position '1' in the size 4 group is the mean z-score of characters 1, 5, 9, 13, etc. in the response sequence). In effect, they examine whether different cluster sizes have unique patterns of pauses.

S4 has a distinct pattern of size 4 with a long pause followed by three short pauses, with the middle of these three being longer than the other two – a *long-short-medium-short* pattern. The absolute value of the long pauses are three times that of the rest. Comparing the groups of different size, it is clear that size 4 is most appropriate for S4. The pause of the first character in size 2 group appears to be a simple aggerate of the first and third pauses in size 4 group. Neither the size 3 nor size 5 groups show a position-related pattern of pauses.

Similarly, S3 also has a consistent pattern with a group size of 3 – a *long-short-medium* pattern. None of the other group sizes have a position-related pattern of pauses.

A *long-short* pattern is clearly apparent for S2, but it is less consistent than those of S4 and S3: pauses in positions 7, 35 and 39 are lower than usual, but are still longer than their following pauses. The group size of 4 also has a distinct pattern in S2 formed of a pair of *long-short* sequences: pauses at position 1 is shorter than in S4, the pauses in position 3 are longer than in S4.

Clearly, by inspection, the patterns to the left of Fig. 4, within and between S4, S3 and S2 are most unlikely to be due to chance. Nevertheless, let us apply the binomial test to the sequences. The chance of the first pause in each cluster being the longest are 1/4, 1/3, and 1/2, the number of clusters are 10, 13 and 20, for S4, S3 and S2, respectively. As the first value in each and every cluster is the longest, the probability of these patterns occurring by chance are all $p=1*10^{-6}$, $p=0.7*10^{-6}$ and $p=1*10^{-6}$, respectively.

S1's graph in Fig. 4 clearly shows it is the least consistent of all the stimuli. Group sizes of 2, 3 and 4 have distinct, but weak, patterns suggesting that some participants may have dominant clusters sizes of 2, 3 or 4. Note the very elevated pause of the first character – an absolute pause of 2.2 seconds.

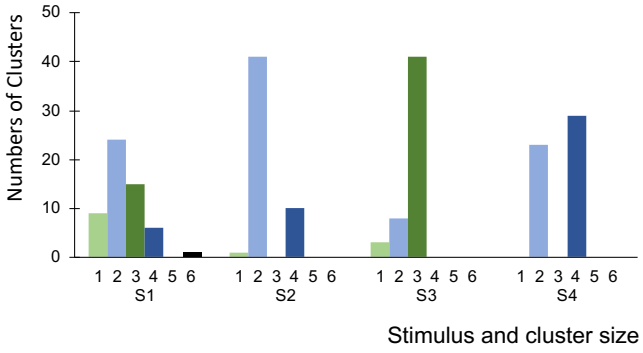

Figure 5. Number clusters with 1 to 6 characters across all participants.

This may reflect the time participants needed to initially decide upon a chunking strategy for S1.

A concern with the first analysis is that it assumes participants reliably produced clusters with a constant size. If a participant switches cluster size or accidently misses a character this will offset the position of the following characters and so scramble the analysis of the rest of the sequence. The second analysis method addresses this concern by working in the opposite direction to determine cluster sizes by counting the number of characters between long pauses. The overall approach is to find a pause threshold that, ideally, produces one modal cluster size for each participant. Five steps are applied iteratively. (1) A pause z-score threshold is set to identify long pauses. (2) Using that threshold clusters are found and their sizes noted. (3) For each participant, large *cluster-groups* with five or more clusters, each containing the same number of characters, are counted. Typically there is one such group. (4) A total of the large cluster-groups is tallied across participants. (5) Iteratively, the threshold in step (1) is adjusted so that a total number of large cluster-groups in step (3) equals the number of participants (i.e., 52). Ideally, each participant would have one large cluster-group (of ≥5 clusters), but some may have none or more than one, for a given pause threshold. Fortunately, the outcome is close to this ideal with just 1 to 4 participants, per stimulus, having none or more than one large cluster-group.

The z-score pause thresholds for S1 to S4 are 0.16, 0.13, 0.33 and 0.38, respectively. The higher thresholds for S3 and S4 suggests that the chunking behaviour in S3 and S4 is more regular than in S1 and S2. Fig. 5 shows the frequency of clusters by size for all participants for these thresholds. S2 and S3 are consistent with dominant cluster sizes of 2 and 3, respectively. S4's size 4 is in the majority, but size 2 is also substantial. S1 is the least consistent with a spread of sizes from one to six characters. Different combinations of large cluster-group frequency (other than 5) and pause thresholds, within reason, produce similar distributions.

Overall, the two analyses for the CD transcription mode show participants are relatively consistent with each other in S2, S3 and S4, but relatively inconsistent in S1.

Discussion

This experiment examined whether there are substantial individual differences in chunking strategy in the process of transcription. The consistency is high for S2, S3, and especially S4, across both the deliberate view (DV) and constant display (CD) modes. This suggests that individual differences in chunking strategy may not substantially impact the cluster-based or the paused-based measures of chunking behaviors developed by Albehaijan & Cheng (2019), Cheng (2014, 2015), Cheng & Rojas-Anaya (2006) and Zulkifli (2013). In terms of the logic of the experiment, it is likely that both (a) the encoding of the spatial structure of the stimuli and (b) the written production of clusters are consistent across participants. It is implausible that there could be large inconsistencies in (a) that just happen to cancel out inconsistencies in (b), because they are different processes. So, the experiment gives additional reassurance that chunk measures of competence are reliable measure of domain competence.

It should be noted that the task – transcription – and the experiment's test environment are similar to that used in our previous work on competence methods, so in these regards the present findings can reasonably be considered to support the validity of those methods. However, the stimuli are artificial meaningless strings, which raises the question of whether the observed consistency is due to some unaccounted for feature of the stimuli design. Such concerns are allayed by the systematic difference in participants consistency on S3 and S4 compared to S1, and to a lesser extent S2, which differ only in the size of the sets. Participants behaviour can sensibly be attributed to chunking processes related to the overt structure of the stimuli.

Further, the findings for S1 show that individual differences can occur in transcription tasks. S1 does not provide a spatial structure, so participants must pick their own cluster size, which leads to substantial individual differences in both DV and CD modes. S1 has the longest Q3 pauses (Table 2) and starts with an elevated pause (Figure 4), which may reflect the time for such a strategic decision about chunk size, that are seemingly absent in S2, S3 and S4.

The dominant cluster size in the DV mode is 4 characters for S1, S2 and S4 and size 3 for S3 (Figs. 2 & 3). In the CD mode clusters of size 2 features strongly in S1, S2 and S4, and size 3 in S3 (Figs. 4 & 5), and also size 4 in S4. Cluster sizes of four or less are within Cowan's (2001) WM chunk size estimate. The smaller cluster sizes in the CD mode seems to indicate that participants are choosing not to load their WM fully, as they could easily reinspect the stimuli with quick eye movements. The performance on S4, in particular, supports this view given the high proportion of size 2 chunks (Fig. 5) compared to the DV mode. The size 2 chunks could be generated by splitting each set of 4 characters in two. In contrast, the DV mode might have encouraged the filling of WM in order to limit the number of relatively laborious actions needed to reveal the stimulus.

The results between and within the stimuli suggest that one cluster is one chunk in transcription. However, this is not the full picture. The pause z-scores of size 4 clusters for S4, and

size 3 clusters for S3, in Fig. 4 suggest that these cluster may consist of two sub-chunks, because of the elevated pause of the third character in both cases. The elevated pauses suggests that some additional processing is required in order to produce the second pair of characters in S4 clusters, or the third character in S3 clusters; processes such as the retrieval or reactivation of the second sub-chunk. Rosenbaum, Kenny & Derr (1983) explain long-short-medium-(short) pause patterns in terms of hierarchical encoding and processing of chunks.

Finally, we note two points. First, it would be interesting to extend the experiment to stimuli with sets of five and six characters to investigate whether participants can use larger clusters and how they decompose them into sub-chunks. Second, the assessments of cluster sizes presented here suggest the possibility of new types of temporal chunk measures, which we plan to investigate in future studies.

Acknowledgments

Our thanks go to all the members of the Representational Systems Lab for their support of this work and comments on the paper. We would also like to thank the three anonymous reviewers for their useful recommendations.

References

Albehaijan, N., & Cheng, P. C.-H. (2019). Measuring programming competence by assessing chunk structures in a code transcription task. In *Proc. 41st Ann. Conf. of the Cognitive Science Society* (pp. 76-82).

Anderson, J. R. (2000). *Learning and memory : an integrated approach* (2nd ed.. ed.). New York, N.Y.: Wiley.

Bryan, W. L., & Harter, N. (1897). Studies in the physiology and psychology of the telegraphic language. *Psychological Review, 4*(1), 27-53.

Cheng, P. C.-H. (2015). Analyzing chunk pauses to measure mathematical competence: Copying equations using 'centre-click' interaction. In *Proc. 37th Ann. Conf. Cognitive Science Society* (pp. 345-350).

Cheng, P. C.-H. (2014). Copying equations to assess mathematical competence: An evaluation of pause measures using graphical protocol analysis. In *Proc. 36th Ann. Conf. Cognitive Science Society* (pp. 319-324).

Cheng, P. C. H., & Rojas-Anaya, H. (2005). Writing out a temporal signal of chunks. In *Proc. 27th Ann. Conf. of the Cognitive Science Society* (pp. 424-429).

Cheng, P. C. H., & Rojas-Anaya, H. (2007). Measuring Mathematical Formula Writing Competence. In *Proc.of the 29th Ann. Conf. of the Cognitive Science Society* (pp. 869-874).

Cheng, P. C. H., & Rojas-Anaya, H. (2008). A Graphical Chunk Production Model. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proc. of the 30th Ann. Conf. of the Cognitive Science Society* (pp. 1972-1977).

Cheng, P. C.-H., & van Genuchten, E. (2018). Combinations of simple mechanisms explain diverse strategies in the free-hand writing of memorised sentences. *Cognitive Science*, **42**, 1070–1109.

Chi, M. T. H., Glaser, R., & Farr, M. J. (Eds.). (1988). *The Nature of Expertise*. Hillsdale, N.J.: L. Erlbaum Assoc.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Science, 24*(1), 87-114.

Gray, W. D., & Boehm-Davis, D. A. (2000). Milliseconds matter: An introduction to microstrategies and to their use in describing and predicting interactive behavior. *Journal of Experimental Psychology: Applied, 6*(4), 322-335.

Gobet, F., & Clarkson, G. (2004). Chunks in expert memory: Evidence for the magical number four … or is it two? *Memory, 12*(6), 732-747.

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C.-H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Science, 5*(6), 1236-1243.

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

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for information processing. *Psychological Review, 63*, 81-97.

Oviatt, S., Hang, K., Zhou, J., Yu, K., & Chen, F. (2018). Dynamic handwriting signal features predict domain expertise. *ACM Transactions on Interactive Intelligent Systems, 8*(1), 1-21.

McLean, R. S., & Gregg, L. W. (1967). Effects of induced chunking on temporal aspects of serial recitation. *Journal of Experimental Psychology, 74*, 455-459.

Newell, A. (1973). You can't play 20 questions with nature and win. In W. G. Chase (Ed.), *Visual Information Processing* (pp. 283-308). New York, N.Y.: Academic Press.

Rosenbaum, D. A., Kenny, S. B., & Derr, M. A. (1983). Hierarchical control of rapid movement sequences. *Journal of Experimental Psychology: Human Perception and Performance, 9*(1), 86–102.

Stahovich, T. F., & Lin, H. (2016). Enabling data mining of handwritten coursework. *Computers & Graphics, 57*, 31-45.

van Genuchten, E., & Cheng, P. C.-H. (2010). Temporal chunk signal reflecting five hierarchical levels in writing sentences. In *Proc. 32nd Ann. Conf. of the Cognitive Science Society* (pp. 1922-1927).

Verwey, W. B., Shea, C. H., & Wright, D. L. (2015). A cognitive framework for explaining serial processing and sequence execution strategies. *Psychonomic Bulletin & Review, 22*(1), 54-77.

Zulkifli, M. (2013). *Applying Pause Analysis to Explore Cognitive Processes in the Copying of Sentences by Second Language Users.* (PhD), University of Sussex, Brighton, UK.