Evaluating the Structure of Chunk Hierarchies in a Naturalistic Educational Task Using Gaussian Mixture Models

Peter C-H. Cheng (p.c.h.cheng@sussex.ac.uk),
Yanze Lui (yl680@sussex.ac.uk), Grecia Garcia Garcia (G.Garcia-Garcia@sussex.ac.uk)
Department of Informatics, University of Sussex, Brighton, BN1 9QJ, UK

Gabrielle Cayton-Hodges (gabrielle@readytolaunchec.com)
Ready to Launch, Educational Consulting, Princeton, New Jersey, USA

Abstract

Our knowledge of a topic such as mathematics is reliant upon the hierarchies of chunks we build in our memories. The time course of knowledge-based tasks, such as the transcription of algebraic formulas, provides rich signals that reflect the structure of the chunk hierarchies being processed. By building Gaussian Mixture Models, this paper provides evidence that decomposing the overall distribution of pauses between actions in a sequential task can give meaningful characterizations of the structure of the chunk hierarchy. We also examine whether individual competence in mathematics can be measured using a metric derived from the models.

Keywords: Chunking, sequential behaviour, Gaussian Mixture Models, transcription task, mathematics competence

Introduction

In human cognition the time course of behaviors in sequential tasks is substantially determined by the structure of chunks that the actor mentally processes (Chase & Simon, 1973; Egan & Schwartz, 1979; Moss, Gobet et al., 2001; Kotovsky, & Cagan, 2006; Cheng & van Genutchen, 2018). For example, participants can be given a stimulus with an explicit hierarchical structure and asked to reproduce it by memorizing it either as a whole or successively in parts. When the stimulus is then verbalized or written, the durations between actions, *pauses*, reflect the levels of the hierarchical structure (McLean & Gregg, 1967; Chase & Simon, 1973; Cheng & Rojas-Anaya, 2005, 2006). Clearly in general, it is useful for research in cognitive science to be able to assess the structure of chunks in the memory of participants as they perform tasks. Moreover, as chunk structures are acquired with learning, being able to assess the structures could serve as a foundation for assessments of individuals' knowledge or competence in a topic.

This paper considers whether *Gaussian Mixture Modelling* (GMM) (Reynolds, 2009) has potential as an approach to evaluate the structure of chunks in memory that is superior to pervious approaches.

Cheng and colleagues have previously investigated the possibility of assessing individuals' competence by asking them to perform transcriptions tasks. They called their overall approach *Competence Assessment by Chunk Hierarchy Evaluation with Transcription-tasks* (CACHET) (Colarusso, Cheng, Garcia Garcia, Stockdill, Raggi & Jamnik, 2023). CACHET has shown promise in domains including English as a second language (Zulfliki, 2013; Ismail & Cheng, 2021); mathematics (Cheng, 2014; 2015), and programming (Albehaijan & Cheng, 2019). Their more recent work has also moved beyond topics with linear notations to 2D charts and graphs (Colarusso et al., 2023). Drawing on chunking theory they identified various behaviors which may serve as the foundation for metrics of competence, including: (a) pauses between completing the production (copying) of one element and the start of the subsequent one; (b) the time to reproduce those parts of a stimulus that can be comfortably held in working memory immediately following a brief voluntary view of the stimulus; (c) the overall number of views of the stimulus needed to reproduce it in full.

In this paper we will focus on the pause-based chunk signals. Cheng & colleagues examined metrics that characterize the overall shape of the distribution of pauses and found that the *third quartile of pauses* yields strong correlations with independent measures of competence (Zulfliki, 2013; Cheng, 2014). This is because the distributions of pauses typically have a strong positive skew. The third quartile pause metric is not only sensitive to the relative difficult of stimuli items, but it reflects participants' competence well, with correlations against independent measures of competence up to 0.7 (Albehaijan, 2013; Cheng, 2014; Ismail, 2025). Correlations at this level are acceptable for high-stake assessments of educational competence (Educational Testing Service, 2010).

Although the third quartile of pauses has reasonable empirical support as a metric for competence, it has two related theoretical weaknesses. First, it is based on the rank of pauses, so it discards information about the actual value of the pauses. The value of two participants' third quartile of pauses could be identical even when one participant has a more skewed distribution above the third quartile. The second weakness is that the third quartile metric is a statistic of the overall shape of the pause distribution, but this ignores how the overall distribution of pauses arises from distinct sets of processes for components at different levels of the chunk hierarchy. Each of these sets of processes have their own sub-distribution of pauses. This implies that the third quartile measure may be rather insensitive to changes in the structure of chunk hierarchies, as it aggregates over all the sub-distributions. But not all the pauses are equally important, because the sub-distribution associated with highest level in the chunk

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

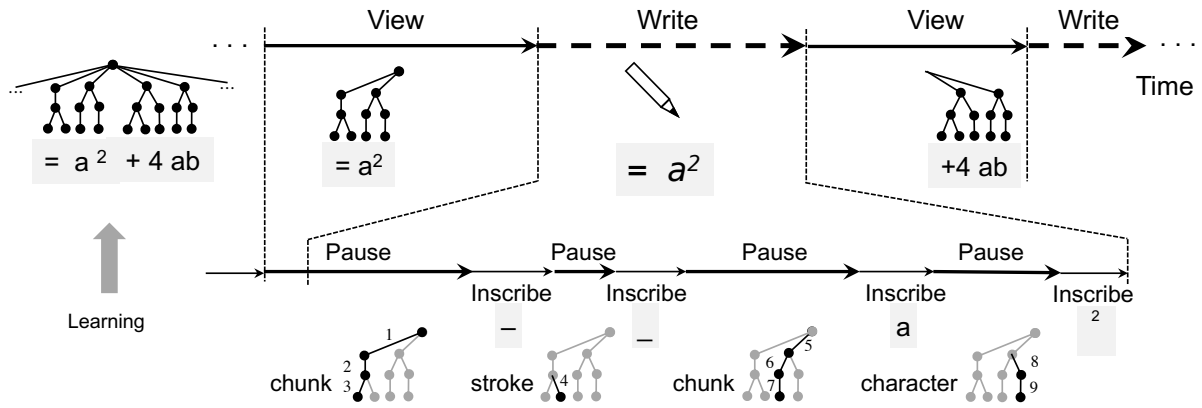

Figure 1. Processing of chunks in the transcription of (a part of) a mathematical formula.

hierarchy is the most affected during learning. Here, we investigate whether decomposing the overall pause distribution into separate Gaussian distributions, using *Gaussian Mixture Models* (GMMs), (a) may provide an effective approach to probing the overall structure of chunk hierarchies and (b) whether the pause distribution associated with the highest level in the chunk hierarchy has potential as a metric for the assessment of competence.

The next subsection considers how chunking theory justifies the modelling of the overall pause distribution as multiple sub-distributions. The subsection following that describes a machine learning method for extracting GMMs.

Chunking theory in transcription tasks

Why might a sequential task, such as transcribing a stimulus, be well characterized as a set of Gaussian distributions? Fig. 1 depicts the processes involved in the incremental copying of a mathematical equation. During the task, a participant repeats cycles of viewing parts of the stimulus, committing them to WM, and then writing them. Just a middle portion – $=a^2+4ab$ – of a larger stimulus is considered in this example. Suppose the participant is new to the topic of algebra, the tree on the left of Fig. 1 depicts a possible chunk hierarchy in their long-term memory, which they would have acquired through learning. It has four branches with small groups of characters, chunks: =, a^2, +4, and ab. (Technically speaking characters with multiple strokes are also chunks, or rather subchunks.) When the participant perceives the stimulus, they will retrieve these chunks. Now, assume that the participant encodes two chunks into WM, and that the process of producing the expression $=a^2$ will require nine processing steps, which are numbered in Figure 1: (1) the production of the first chunk is initiated; (2) the first (and only) character in the chunk is selected; (3) the first stroke of the character is chosen, its motor program is retrieved from memory and the pen moved to mark the stroke (top line of **=**); (4) the same is then done for the second stroke (bottom line of **=**); (5) as there are no more characters the production of the second chunk is initiated; (6) the first character in the chunk is selected (a); (7) the first (and only) stroke is prepared and made (***a***); (8) as there are no more strokes, the second character of the second chunk is chosen (a^2); (9) it is written (^{2}). Although this may seem to be a laborious sequence for just for the production of characters, it is consistent with the processing steps proposed by Cheng & van Genutchen's (2018) model of writing memorized sequential stimuli.

The chunk hierarchy is processed in a depth first manner. So on the completion of each stroke, control jumps back up to the lowest node which still has unprocessed sub-nodes. Thus, the amount of processing that occurs between the completion of a stroke and the start of the next – which determines the duration of the pause between strokes – will depend on the number of levels to be traversed in the hierarchy. Note also that the pause before the first stroke of the first letter of the first chunk includes the time to perceive the current set of chunks as the previous stroke would have been the last one from the last viewed set of chunks. As three levels of processing are present – *stroke*, *character* and *chunk* – there will be three distinct pause distributions. The size of their means increases in order. The lengths of the arrows in Fig.1 are indicative of the durations.

Importantly, with learning the small chunks will coalesce larger chunks, such as $=a^2+4ab$, and will become more familiar. This will reduce the mean duration of the chunk level pauses and also reduce the number of the chunk pauses, relative to the character and stoke pauses. Thus, participants with different levels of competence in algebra will have different overall shapes of the distribution of pauses, and in particular the positive skew of the distribution should lessen with more expertise. In the extreme, a highly competent individual might encode a whole equation as one chunk.

Cheng & Rojas-Anaya (2005, 2006, 2008) conducted experiments in which participants were taught sequences of numbers, letters, or words with an explicit grouping structure so the hierarchical chunk levels of each pause possessed by participants could be presumed. The stroke level pauses were ≈90 ms, the character level pauses were ≈260 ms, and the chunk level pauses were ≈440 ms. In Cheng & van Genutchen's (2010) study, adults memorised and wrote sentences with five explicit levels, which had the follow durations of pauses: strokes = 90 ms; letter = 273 ms; word = 374; phrase = 567 ms; sentences = 1134 ms (Cheng & van Genutchen, 2018). All this shows that we can measure the properties of a participant's levels of pauses, if we already know

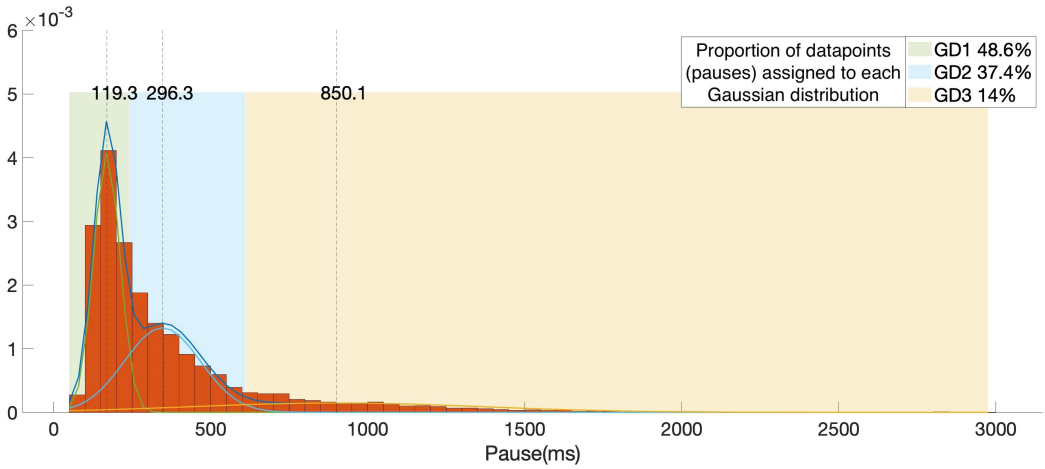

Figure 2. A histogram of data from a transcription task with three overlaid Gaussian distributions. The x-axis shows pause duration in ms, and the y-axis shows probability density (note the bar widths are 100 ms).

their chunk structure. However, in the context of competence assessment, the makeup of an individual's is what is to be determined. In this paper, we ask whether distributions found using GMMs are theoretically plausible and reflect individuals' levels of competence. We will use a transcription task similar to the previous CACHET studies outlined above, so three levels of pauses are expected, which will be called *GD1*, *GD2* and *GD3*. GD1 will often be associated with strokes, and GD2 with characters, but not necessarily so, because a stroke or a character's level will depend on the individual's familiarity of the stimulus. With an unfamiliar multi-stroke character each stroke might be treated individually as if it was like a character, or with a very familiar sequence of characters the sequence might be treated like it is one large character.

Gaussian Mixture Models (GMMs)

GMM is machine learning technique that attempts to find the best fit of a given number of Gaussian distributions (GDs) to a set of data, by exploring the space of means and standard deviations (SD) of the GDs. The Expectation-Maximization (EM) algorithm is commonly used to iteratively optimise these parameters (Dempster et al., 1977, McLachlan & Basford, 1988). It starts by randomly assigning parameters to the GDs, then estimates the probabilities of each datapoint belonging to each GD based on the current parameter values. Then using these probabilities, the means are pulled toward the densest region, the SDs shrink or spread out, and the weights grows for GDs with more datapoints. We used MATLAB's implementation of EM with the default parameters and a maximum iteration limit of 10,000.

The minimum recommended number of datapoints for reasonable GMMs follows the one-in-ten rule in statistics (Raudys & Jain, 1991, Peduzzi et al., 1996). Thus, it is possible to produce models for single participants, but as pause data in transcription tasks is inevitably noisy there is risk of overfitting the data.

Fig. 2 presents a sample of pause data from the experiment described below. The histogram shows the positive skew that is typical for transcription tasks. The overlaid curves are three GDs found using GMM and the overall (dark blue) curve is their sum. The means of the curves are marked by the dashed lines. The sum of data bars in the colored area for each GD give the proportion of datapoints that are most likely to belong to each GD.

Research questions and predictions

Can GMMs produce theoretically sensible Gaussian distributions (GDs) for sequential behaviors, such as a transcription task? This overall question can be broken down into three main questions relating to aspects of the task:

1. **Levels of chunking**. Do the modelled GDs reflect the hierarchical levels of chunking, with parameters of the GDs aligning with the theoretical expectations?
   a) The **means of pauses** of each GDs should match the predicted pauses for each level in the chunk hierarchy. The three GDs should be distinct, with means of increasing length. From the findings reviewed above, the difference between the mean pauses of each pair of consecutive levels should be 200 ms, approximately. Differences in the means are considered, rather than absolute values, because task and population factors are expected to influence pause durations.
   b) **Coefficient of variations** (CoVs) of pauses for the GD3 should be larger than that GD1 and GD2, because the greater number and variety of cognitive processes needed to execute chunk level actions will produce more relative variability than the lower-level actions (i.e., Fig. 1 steps 1 to 3, versus steps 8 to 9, or step 4). CoVs are considered, rather than SDs, to compensate for the typical growth in variance with larger means.
   c) The **proportion of pause datapoints** that the model assigns to GDs should reflect the expected decrease with altitude in the chunk hierarchy. The number of GD1 datapoints depends, in part, on the characters in the stimuli and how each participant writes characters. Table 1 lists the stimuli used here. The numbers of one stroke and two stroke characters are comparable.
2. **Competence**. The parameters of the GDs should reflect differences and similarities in participants' levels of competence, given that their chunk structures will vary.

Table 1. The stimuli with their complexity ranking. (Tasks numbered "1" were practice items and not shown.)

Task	Stimuli (Form 1)	Stimuli (Form 2)	Elements	Types of operators	Complexity rank
2.1	$x + y = y + x$	$x + y = y + x$	7	1	1
2.2	$(x + y) + z = x + (y + z)$	$x(y + z) = xy + xz$	15/12	2/3	=2
3.1	$(a + 2)(a + 3) = a^2 + 5a + 6$	$(a + 1)(a + 4) = a^2 + 5a + 4$	18	4	4
4.1	$y = ax^2 + bx + c$	$y = ax^2 + bx + c$	10	3	=2
4.2	$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$	$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$	16	6	5

Higher competence participants will have chunks containing more elements than those with lower competence. Thus, three predictions are made:

a) At the GD3 level, it is expected that the **mean pauses** of higher competent participants should be lower than that of less competent participants.
b) At the GD3 level, the **proportion of pauses** should be less for the high versus the low competent participants.
c) At the GD1 and GD2 levels, we would expect little change in the **pauses** with competence, as participants will be near ceiling in character writing performance.

3. **Stimuli complexity**. The parameters of the GDs should reflect the relative mathematical complexity of the stimuli, as participants will have less well-developed chunk structures for more complex stimuli, because they are harder to learn, and will have been introduced later than simpler material. Hence, simpler stimuli should include larger chunks hierarchies, so it is predicted that:
   a) The **means** of GD3 **pauses** will increase with stimuli complexity.
   b) At the GD3 level, the **proportion of pause data points** will increase with stimuli complexity.
   c) There will be little difference on **pauses** across stimuli complexity for GD1 and GD2, because participants' competence at making strokes and writing characters is expected to be independent of stimuli complexity.

The greater the number of positive responses to these questions the stronger the support for an affirmative response to the overall research question.

Experiment and Data

The pause data was obtained from the third part of a larger study (Cayton-Hodges & Fife, 2018). In that part of the study the transcription task was performed on tablet computers (iPads) with the stimulus always visible on screen. The tablet logged students' strokes (coordinates and timestamps of a finger or stylus on the tablet screen), from which pause durations between actions were computed. The stimuli are shown in Table 1. They were presented individually in each trial. There were different parallel sets (forms) for counterbalancing across the other parts of the larger study. The rank order of stimuli complexity is simply taken as the number of different types of mathematical operator in each equation.

A total of 248 students from a single school in the USA participated. Teachers were told that to be eligible to participate, the students should have seen the quadratic formula in class. The experiment was run during classes in introductory algebra towards the end of the school year.

One hundred and seventeen of the 232 students whose data was analyzed were female (50.4%) and 115 were male (49.6%). 90.1% were in grade 9, 7.8% in grade 10, 0.9% in grade 11, and 1.3% in grade 12. (For the whole school, SAT average for mathematics proficiency was 610 and SAT average for reading proficiency was 590). Of these only the 225 students for whom the teachers provided independent measures of mathematics competence are included. The school test scores range between 1 and 6 (A to F).

The Pearson correlation for the third quartile of pauses of each participant aggregated over all the stimuli against their school test scores was r=0.23 (N=255, p<.01).

The pause data was computed from the raw pen stroke data, where a pause is the time from the lifting the stylus or finger from the last stroke to the next contact on the tablet at the start of the following stroke. Initially a conventional approach to the treatment of outliers was adopted, with the longest 2% of pauses removed. However, the models produced included many GDs with means greater than 3 seconds, with small SDs, and that included few data points. Those GDs are likely to represent behaviors unrelated to the process of transcription (e.g., regression of the pen when it reaches the right side of the response window, or momentary distractions). Such long means are three times the expect values of the chunk level pauses, so not theoretically relevant. Thus, a threshold of 3 seconds was set for the removal of outliers. The number of data points included for modelling and percentage of datapoints removed are shown to the right of Table 2. In nearly all cases less than about 2% of the data were outliers, but oddly stimulus 4.1 (the quadratic polynomial) had a much higher proportion of outliers.

Results

Table 2 shows the data from the GMM modelling. The pairs of rows for each stimulus are in order of their complexity. Data is aggregated in the pairs of rows by high and lower competence participants. The means, SDs, CoVs and data proportions for each of the three GDs are in the four groups of columns. The blue data bars visualize the relative values in each group. The final the two columns give the number of datapoints and the percentage of outlier excluded.

Table 2. The mean, SD, and CoV for the pauses, and the proportions of data, for each GD, aggregated over participants for each level of competence and for each stimulus. The last pair of rows pools the data across all the stimuli. The in-cell data bars visualize the cell values with full scale values of: Mean=1,200ms; SD=1,200ms; CoV=1.0; Proportion=100%.

Stimulus (complexity)	Competence	Mean (ms)			S.D. (-)			CoV (-)			Proportion			Data points	Outliers
		GD1	GD2	GD3	GD1	GD2	GD3	GD1	GD2	GD3	GD1	GD2	GD3		
2.1 (1)	high	168	326	744	40.3	106	355	0.239	0.326	0.478	49.7%	38.9%	11.5%	1408	1.54%
	low	171	347	874	43.9	124	518	0.257	0.356	0.592	52.9%	35.4%	11.7%	1200	1.88%
2.2 (=2)	high	169	329	896	37.2	109	481	0.220	0.330	0.536	41.3%	41.6%	17.1%	2379	1.61%
	low	163	347	957	38.1	123	501	0.234	0.356	0.524	38.1%	43.5%	18.3%	2039	1.73%
4.1 (=2)	high	156	324	887	36.9	115	503	0.237	0.354	0.566	47.3%	41.9%	10.8%	1708	7.43%
	low	161	325	908	42.9	117	496	0.266	0.359	0.547	48.6%	36.8%	14.6%	1480	8.19%
3.1 (4)	high	176	355	922	48.4	123	458	0.275	0.345	0.496	46.4%	39.1%	14.5%	2750	1.19%
	low	167	352	952	44.6	125	501	0.267	0.355	0.527	41.3%	40.0%	18.7%	2372	1.90%
4.2 (5)	high	179	421	1006	46.1	159	471	0.258	0.378	0.468	34.8%	42.5%	22.8%	2308	1.79%
	low	185	446	1119	54.4	166	526	0.294	0.372	0.470	37.2%	38.8%	24.0%	1909	2.20%
All	high	169	346	900	41.8	120	463	0.247	0.348	0.514	42.8%	40.1%	17.1%	10553	2.52%
	low	170	366	990	45.4	135	514	0.268	0.368	0.519	43.1%	38.7%	18.2%	9000	3.02%

Levels of chunking (Question 1)

Do these GDs accurately predict (a) the means of pauses, (b) CoVs, and (c) the proportion of datapoints across the three levels of the GDs?

(Q1a) Graphs like Fig. 1 were produced for all the stimuli and inspected. Like Fig. 1 they all have distinct clearly separated means; there were no cases of overlapping GDs with similar means but one small SD and one large (e.g., to model a leptokurtotic sub distribution). Consider Table 2. The difference between the values of the means of pauses of GD2 and GD1 is approximately 180 ms, which is comparable to the expected value of 200 ms. The difference between GD3 and GD2 is 570 ms, is substantially greater than expected.

(Q1b) The **CoVs of pauses** for GD3 are greater than those of GD1 and GD2, by ≈2 times and 1.5 times, respectively, which is consistent with the prediction. The smaller CoVs of pauses for GD1 compared to GD2 is unsurprising.

(Q1c) The **proportion of datapoints** for each GD is as predicted. The proportion of GD1 and GD2 are approximately equal as anticipated, and approximately 2.5 times as frequent as that of GD3, which is in line with expectations.

Participant competence (Question 2)

The upper and lower rows in each pair of rows in Table 2 present data for the high and low competent participants.

(Q2a) The **mean pauses for GD3** show a consistent pattern in line with the prediction. For all stimuli, the mean pause is greater for the low compared to the high competent participants. As there are five stimuli, by the binomial theorem, assuming that performance on each stimulus is independent and there is an equal chance of either group having a higher mean, the chances of this pattern occurring by random is $½^5 = 1/32$ (i.e., $p<.05$). The differences range from 20 ms up to 131 ms across the stimuli, with a mean of 90 ms, which is meaningful in terms of cognitive operation times (Newell, 1990).

(Q2b) The **proportions of data points for GD3** show a consistent pattern in line with the prediction. For all stimuli, the proportion is greater for the low compared to the high competent participants. Again, the chances of this pattern occurring by random is 1/32 (i.e., $p<.05$). However, the range of the differences is less striking, in contrast to that for the mean pauses, with values from 0.3% to 4.2%.

(Q2c) As expected the differences between high and low competence participant for **means of pauses of GD1 and GD2** are small relative to the magnitude of the mean pauses for GD3 on each stimulus.

Stimuli Complexity (Question 3)

The pairs of rows in Table 2 are arranged in order of mathematical complexity given in Table 1. Note that stimuli 2.2 and 4.1 have equal complexity rankings.

(Q3a) A clear trend of increasing means of the pauses with increasing stimuli complexity is apparent in Table 2. For low competence participants the difference between the most and least complex stimuli (from 4.2 to 2.1) is 245 ms, and for the high competence participants it is 262 ms.

(Q3b) There is some indication of a trend of increasing proportion of GD3 datapoints in Table 2. For low competence participants the difference between the most and least complex stimuli is 12.3%, and for the high competence participants it is 11.3%, which corresponds to an approximate doubling of the number of datapoints with complexity.

(Q3c) There are no strong trends for variation of mean pauses or proportion of datapoints for the GD1 and GD2 with increasing stimulus complexity.

Individual participant competence (Q2 revisited)

To investigate the potential of GMMs at a finer grain of analysis, models with three GDs were computed for each participant individually, pooling their data over all the stimuli. The mean number of datapoints per participant is 122 (SD = 16.6), which meets the criteria for conducting GMM (as noted above). Table 3 presents data for means (and SDs) over participants (N=225) across the three GDs for the following: mean of pauses; SDs of pauses; CoVs of pauses; proportions of data.

The means of the participants' mean pauses for GD1, GD2, and GD3 are generally larger than for the data aggregated over stimuli (Table 2), with the values closer to those of the most complex stimuli (4.2). The CoV for GD1 and GD2 are comparable to those aggregated over stimuli, but for GD3 the

Table 3. Individual participant data.

Parameter	GD1	GD2	GD3
Mean of mean of pauses	186	416	1102
SD of mean of pauses	49.5	139	429
Mean of SDs of pauses	44.7	136	395
SD of SDs of pauses	24.3	63.8	183
Mean of CoVs	0.230	0.320	0.382
SD of CoVs	0.07	0.08	0.16
Mean of proportions	46.9%	40.2%	13.0%
SD of proportions	16.1%	13.4%	8.3%

mean CoV over the participants is notably smaller than the CoVs for each stimulus. The proportion of datapoints appears, generally, to be increased for GD1 and reduced for GD3, and approximately comparable for GD2. In summary, for individuals, in contrast to the data pooled for a stimulus, the modelling produces GD3's whose means tend to be larger and more closely fitting, as shown by the smaller proportion of data points and the smaller CoV.

What impact does this have on the relation between the mean of participants' GD3 and their competence? A Pearson correlation of the participants' GD3 mean pauses and their school test scores gives r=0.0012 (N=225, p>.05). The corresponding values for GD1 and GD2 are r=-0.073 (p>.05) and r=0.0014 (>.05). So, at the level of individuals, the mean of GD3 is far worse as an indicator of competence than the third quartile measure of competence (see above).

Discussion

Do Gaussian Mixture Models of pause data for a sequential task produce theoretically meaningful sub-distributions? For the task of transcribing mathematical equations there is affirmative converging evidence in support of the idea. GMM finds sub-distributions that are representative of the three expected chunk hierarchy levels. The form of three GDs in terms of the mean, CoV and proportions of data were as predicted. The impact of competence on the mean and proportion of data for the GD3 chunk level, but its absence at GD1 ("stroke") and GD2 ("character") levels, was as expected. The effect of stimulus complexity on the GD3 chunk level, but its absence on the GD1 and GD2 levels, was as also expected.

Contrary to previous findings on chunking transcription tasks, the durations of the pauses at all three levels were longer than expected. One possible explanation is simply differences in the populations (children vs. adults) or the medium of the task (finger/stylus on a touch screen vs. pen on paper (on a graphics tablet). More interestingly, another explanation is that the GDs are more accurately assigning datapoints to each level of the chunk hierarchy. In the previous experiments (Cheng & Rojas-Anaya, 2005, 2006, 2008; Cheng & van Genutchen, 2010) it was assumed that the stimuli elements were chunked in accordance to their predefined levels. However, those participants may have encoded some elements at a lower level than expected. For instance, high frequency bi-grams may have been treated as single compound characters: e.g., the *h* in "th" may have been processed at the stroke rather than the letter level. The same may apply at the word/chunk level with high frequency word pairs; e.g., "in the", "I am". Such processing would artificially depress the pause durations. All this implies that fitting GMMs to pause data may more accurately identify to which level a pause belongs than an a priori classification. This possibility is speculative, so requires further investigation, but it does highlight a potential advantage of the GMM approach.

One motivation to test GMMs was to find a metric of competence that overcomes the theoretical limitations of the third quartile of pauses. Although differences were found between low and high groups of participants, fitting GDs to individuals' data to extract their individual mean of pauses for the chunk level GD3 did not correlate with competence. This contrasts with the r=0.23 correlation that was obtained on the same data with the third quartile of pauses. There are two possible explanations for this null result. First, WM capacity varies by individual, and participants may have chosen to load their WM to varying degrees, which are both possible confounds. The explanation in Fig. 1 presumed two chunks would be perceived and processed on each cycle. In complex sequential tasks, WM loading is less than Miller's (1956) 7±2 chunks in idealized experimental tasks, with approximately four being typical (Cowan, 2001). The number of chunks may drop as low as two for particularly demanding tasks (Gobet & Clarkson, 2004). Albehajian & Cheng's (2013) experiment showed that normalizing performance with measure of each individual's WM capacity made little improvement to the correlation with competence, which implies that differences in WM capacity may not be the full explanation here.

The second explanation is that the GD3 for individual participants overfitted the data. The mean number of data points per participant was 122 and the proportion of data for GD3 was 13%, so each GD3 contains 16 datapoints. The low mean GD3 CoV (Table 3) compared to that for the main data (Table 2), and the smaller proportion of data for GD3, both add weight to the overfitting explanation. An implication is that for GMM to be useful at the individual level more data will be required, which means more or larger stimuli, which in turn degrades the overall practically potential of the chunk-based approach to competence measurement.

In conclusion, there is some support for the validity and utility of using GMMs to find meaningful sub-distributions of pause data from the sequential task of transcription. However, the accuracy of the modelling of distributions in the tail of the skewed distribution requires careful assessment of whether there is sufficient data to avoid overfitting.

Acknowledgments

We gratefully acknowledge the Educational Testing Service for supporting the collection of the experimental data and in particular to James Fife for his role in that. PC, YL and GGG would like to thank members of the Representational Systems lab for their contributions to the development of this work.

References

Albehaijan, N., (2013). *Applying temporal chunk signals analysis to measure programming competence by the transcription of Java program code.* Unpublished PhD thesis, Department of Informatics, University of Sussex.

Albehaijan, N., & Cheng, P. C.-H. (2019). Measuring programming competence by assessing chunk structures in a code transcription task. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 76-82). Austin, TX: Cognitive Science Society.

Cayton-Hodges, G.A. & Fife, J. (2018). Entering Equations: Comparison of handwriting recognition and equation editors. In T.E. Hodges, G. J. Roy, & A. M. Tyminski, (Eds.), *Proceedings of the 40th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 1163-1170). Greenville, SC: University of South Carolina & Clemson University.

Chase, W., & Simon, H. A. (1973). The mind eye's in chess. In W. Chase (Ed.), *Visual information processing*. New York, N.Y.: Academic press.

Cheng, P. C.-H. (2014). Copying equations to assess mathematical competence: An evaluation of pause measures using graphical protocol analysis. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 319-324). Austin, TX: Cognitive Science Society.

Cheng, P. C.-H. (2015). Analyzing chunk pauses to measure mathematical competence: Copying equations using 'centre-click' interaction. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, C. D. T. Matlock, Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 345-350). Austin, TX: Cognitive Science Society.

Cheng, P. C. H., & Rojas-Anaya, H. (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 the Twenty Seventh Annual Conference of the Cognitive Science Society* (pp. 424-429). Mahwah, NJ: Lawrence Erlbaum.

Cheng, P. C. H., & Rojas-Anaya, H. (2006). A temporal signal reveals chunk structure in the writing of word phrases. In R. Sun & N. Miyake (Eds.), *Proceedings of the Twenty Eighth Annual Conference of the Cognitive Science Society* (pp. 160-165). Mahwah, NJ: Cognitive Science Society.

Cheng, P. C. H., & Rojas-Anaya, H. (2008). A Graphical Chunk Production Model: Evaluation Using Graphical Protocol Analysis with Artificial Sentences. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society* (pp. 1972-1977). Austin, TX: Cognitive Science Society.

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. doi:10.1111/cogs.12606

Colarusso, F., Cheng, P. C.-H., Garcia Garcia, G., Stockdill, A., Raggi, D., & Jamnik, M. (2023). A novel interaction for competence assessment using micro-behaviors: Extending CACHET to graphs and charts. In A. Schmidt & K. Väänänen (Eds.), *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-14): Association of Computing Machinery.

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.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society: series B (methodological), 39(1), 1-22.

Educational Testing Service. (2010). *Linking TOEFL iBT Scores to IELTS Scores – A Research Report*. Princeton, NJ: Education Testing Service

Egan, D. E., & Schwartz, B. J. (1979). Chunking in recall of symbolic drawings. *Memory and Cognition, 7*(2), 149-158.

Gobet, F., & Clarkson, G. (2004). Chunks in expert memory: Evidence for the magical number four … or is it two? *Memory, 12*(6), 732-747. doi: 10.1080/09658210344000530

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.

Ismail, H. (2025). Competence Assessment by Stimulus Matching (CASM): A Novel Approach to Language Assessment by Chunk Transcription. Unpublished PhD thesis, Department of Informatics, University of Sussex.

Ismail, H., & Cheng, P. C.-H. (2021). Competence assessment by stimulus matching: An application of GOMS to assess chunks in memory. In J. Vandekerckhove & T. Stewart (Eds.), *19th International Conference on Cognitive Modeling* (pp. 1-7): Society for Mathematical Psychology.

McLachlan, G. J., & Basford, K. E. (1988). Mixture models: Inference and applications to clustering, Marcel Dekker. Inc. New York, 10-18.

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

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.

Moss, J., Kotovsky, K., & Cagan, J. (2006). The role of functionality in the mental representations of engineering students: some differences in the early stages of expertise. *Cognitive Science, 30*, 65-93.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. Journal of clinical epidemiology, 49(12), 1373-1379.

Raudys, Š., & Jain, A. K. (1991). Small sample size problems in designing artificial neural networks. In Machine Intelligence and Pattern Recognition (Vol. 11, pp. 33-50). North-Holland.

Reynolds, D. A. (2009). Gaussian mixture models. Encyclopedia of biometrics, 741(659-663).

Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, 461-464.

Zulfliki, P. A. M. (2013). *Applying pause analysis to explore cognitive processes in the copying of sentences by second language users.* Unpublished PhD thesis, Department of Informatics, University of Sussex.