Measuring Mathematic Formula Writing Competence: An Application of Graphical Protocol Analysis

Peter C-H. Cheng (p.c.h.cheng@sussex.ac.uk)
Hector Rojas-Anaya (h.rojas@sussex.ac.uk)
Representation and Cognition Research Group
Department of Informatics, University of Sussex
Brighton, BN1 9QH, UK

Abstract

Graphical protocol analysis (GPA) is a novel method for studying chunk-based cognitive performance using semi-automated analysis of a temporal chunk signal in writing protocols. This study applies GPA to the writing of mathematical equations by participants with different levels of expertise. Multiple levels of competence can be distinguished on a single individual and single task basis.
Keywords: chunks, graphical protocol analysis, novice-expertise difference, mathematical formulas.

Introduction

Chunking is a key theoretical concept in Cognitive Science. It is well understood that temporal patterns in behaviour may reveal the structure of chunks in memory (e.g., McLean & Gregg, 1967; Egan & Schwartz, 1979; Chase & Simon,). In particular, when a series of actions are executed the duration of the pause before a given action is typically taken to be indicative of the amount of processing required to produce the output, with longer pauses indicating boundaries between different chunks in memory. Such patterns of such pauses can be a rich and valuable source of evidence to address many issues in cognitive science. However, the use of this temporal signal is hampered by the laboriousness of extracting and analysing pause data and the theoretical uncertainties of interpreting such data with respect to complex task contexts.

The study reported here is a further step in an ongoing research programme that is attempting to make the extraction and interpretation of such temporal chunking signals more practical and reliable. The programme is developing *Graphical Protocol Analysis* (GPA) as a method to identify the structure of chunks in an individual's memory by analysing the processes of writing and drawing. The potential benefits of GPA include: the use of modern, economical, simple to use graphics tablet technology; raw data that is rich (hi-frequency), accurate and precise; automatic initial extraction, analysis and coding of digital behaviour protocols by computer (although current tools are research prototypes); the capture and analysis of continuous extended behaviour sequences encompassing multiple chunks; the use of relatively naturalistic tasks even in an experimental context.

Our previous work on GPA has demonstrated the existence of a strong and robust temporal signal that reveals the structure of chunks in memory (Cheng, McFadzean & Copeland, 2001; Cheng & Rojas-Anaya, 2005, 2006). In those studies participants memorised simple geometrical patterns, sequences of numbers or word phrases. The stimuli were created with predetermined structure and the specific stimuli learning procedures ensured the participants possessed chunks with those structures. A pause duration for a given graphical element is the time between the lifting of the pen at the end of the previous element and the placing of the pen to begin the given element. During the graphical production the pattern of pauses reflects the induced chunk structure. This temporal chunk signal is apparent in data for individual participants doing a single task/trial. Meaningful patterns can be found without the need to aggregate data over trials and participants. The strength and robustness of the temporal chunk signal suggests that it has the potential to be the basis for methods to probe the structure of chunks when they are not known *a priori*.

Three different ranges of pause durations can be distinguished: *long pauses* typically indicate the *inter-chunk* recall and preparation to write a new chunk (L2); *medium length* pauses corresponding to the *intra-chunk* production of sub-chunks that are symbols within a chunk (L1); *short pauses* occur with strokes within a particular symbol (L0), such as the second line of a '=' sign. In the previous study with the drawing of simple geometric objects it was found that L1≈410 ms and L2≈620 ms (Cheng, McFadzean & Copeland, 2001). For the writing of simple number sequences L1≈280 ms and L2≈440 ms (Cheng & Rojas-Anaya, 2005). For familiar and unfamiliar word phrases L1≈270 ms and L2≈400 ms (Cheng & Rojas-Anaya, 2006). The similarity between the pairs of times for the written tasks is noteworthy. One possible explanation for the longer pauses with the drawing tasks is the mode of graphical production used, i.e. drawing versus writing. Another explanation is a task difference, as the drawing of geometric objects was cued by names of the object whereas the production of the written sequences was through direct recall. Hence, the greater duration, particularly for L1, may be due to the extra step of retrieving the geometric pattern into working memory. The implication is that specific differences in the information processing steps needed for particular forms of graphical production may results in diagnostically useful differences in the inter- and intra-chunk pause durations, which could be identified by GPA.

The overall goal of the present study was to demonstrate and extend the utility of GPA. It had three related aims. First, it investigated the copying of meaningful mathematical formulas rather than the production of arbi-

Table 1. Target formulas

Group	Name	Formulas	Task order
Standard formulas	Quadratic solution	# $x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$	1
	Voltage divider	# $v_2 = \frac{v_s R_2}{R_1 + R_2}$	2
	Quadratic	# $(a+b)^2 = a^2 + 2ab + b^2$	5
	Cubic-1	# $(a-b)^3 = a^3 - 3a^2b + 3ab^2 - b^3$	7
	Cubic-2	# $(x-3)^3 = x^3 - 9x^2 + 27x - 27$	8
Bernoulli	Bernoulli's Eq. 1	# $P_1/\rho g + v_1^2/2g + z_1 = P_2/\rho g + v_2^2/2g + z_2$	3
	Bernoulli's Eq. 2	# $P_1/\rho + v_1^2/2 + z_1 g = P_2/\rho + v_2^2/2 + z_2 g$	4
Errors	Jumbled quadratic	# $a(b+^2)2()b = (+b+)$	6
	Cubic error 1	# $(a-b)^3 = a^2 - 5a^2b + 3a^4b^5 + 7b^3$	7
	Cubic error 2	# $(x-3)^3 = x^6 - 2x^2 + 27x^3 - 2$	9
Cubic expansion	Cubic exp. 1/3	# $(x-3)^3 = (x^2 - 6x + 9)(x-3)$	10
	Cubic exp. 2/3	$= x^3 - 6x^2 + 9x - 3x^2 + 18x - 27$	11
	Cubic exp. 3/3	$= x^3 - 9x^2 + 27x - 27$	12
Cubic expansion error	Cubic exp. error 1/3	# $(x-3)^3 = (x^7 - 6x + 1)(3x - 3)$	13
	Cubic exp. error 2/3	$= x^7 + 6x^2 + 9x^3 + 3x^2 + 2x + 27$	14
	Cubic exp. error 3/3	$= x^7 + 6x^3 + 5x + 27$	15

trary patterns or number/word sequences. Is the temporal chunk signal strong and robust in this context? Copying formulas presents a new challenge for GPA. They have greater spatial and typographical complexity. Participants with different levels of experience may chunk the formulas in different ways and may even adopt quite different task strategies. Copying not only involves graphical production but also the initial reading and encoding of the stimuli. Thus, patterns that correspond to a temporal chunk signal could be masked by these perturbing factors.

Second, the study tested whether the temporal chunk signal can be used to distinguish individuals with different levels of expertise or competency in writing such formulas. Can such differences be reliably shown at the level of single participants?

The third aim of the study was to propose and evaluate two different measures of individual competency/expertise based on the temporal chunk signal. Such measures are needed for GPA to be able to assess participant performance in different tasks when the structure of chunks is not known *a priori*, as would normally be the case and in contrast to our previous demonstration studies. The measures exploit the finding that inter-chunk pauses are longer than intra-chunk pauses and take into account individual differences in drawing and writing behaviours that are apparent as variable L1 and L2 durations across individuals. Hence, both measures use an estimate of L1 for each participant as a *baseline* to individually calibrate the measures. The baseline was obtained from the analysis of each participant writing their own name on the presumption that it is one of the most highly practiced things they write (see below). The first of the two measures, *Long Pause Count* (LPC) is simply the number of pauses, which are greater than some *long-pause-threshold*, in proportion to the number of between symbol pauses (i.e., not including L0 pauses). The long-pause-threshold is equal to some multiple of the baseline, where the multiplier of the baseline was investigated as part of the study. The second measure is the *Long Pause Duration* (LPD), which is the mean for all symbols of the ratio of the difference between the pause duration and baseline to the baseline itself.

Fig. 1. Quadratic Solution written by P3

On first sight LPD may seem the best measure because it takes into account the actual magnitude of the pauses, whereas LPC just considers the number of long pauses. However, given the range of possible perturbing factors, that may adversely affect the chunk signal, as mentioned above, LPD may be noisy and less reliable than LPC. Hence, it was valuable to consider both measures.

It was predicted that the more competent participants will use fewer large chunks to complete the tasks, so they will have smaller numbers of long pauses and hence the LPC score will decrease with increasing experience. It was also predicted that the LPD score will also follow a similar pattern, because the mean duration of pauses will also increase with greater numbers of longer pauses.

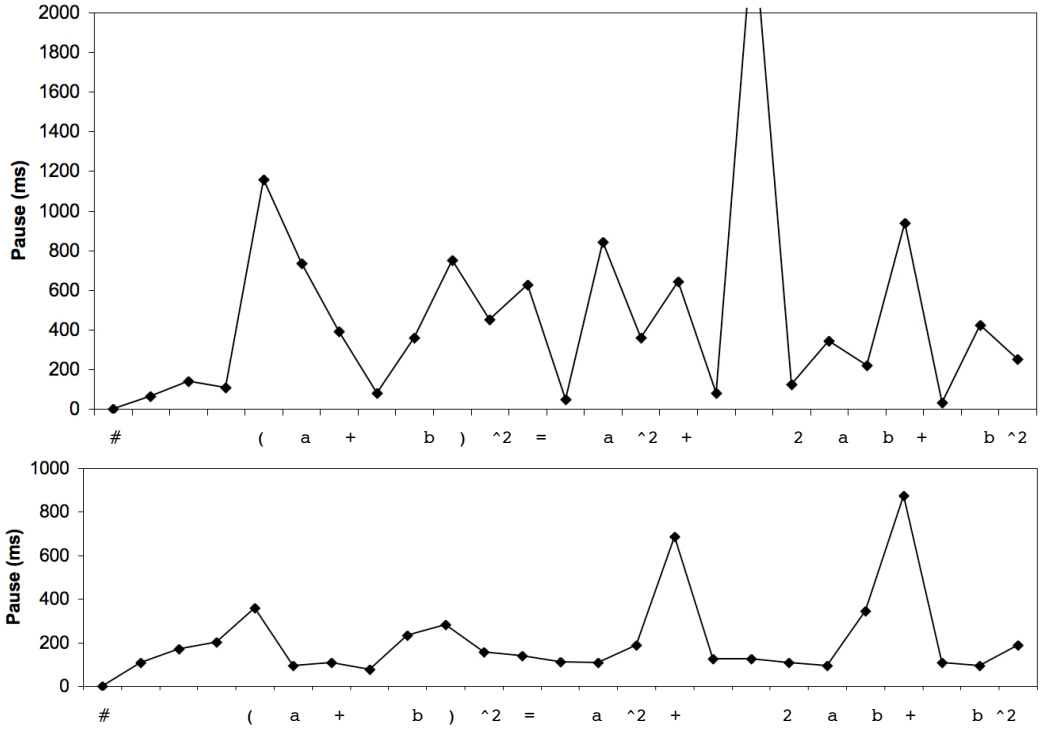

Fig. 2. Temporal protocol graphs for copying the quadratic solution: (a) P1, top; (b) P4, bottom

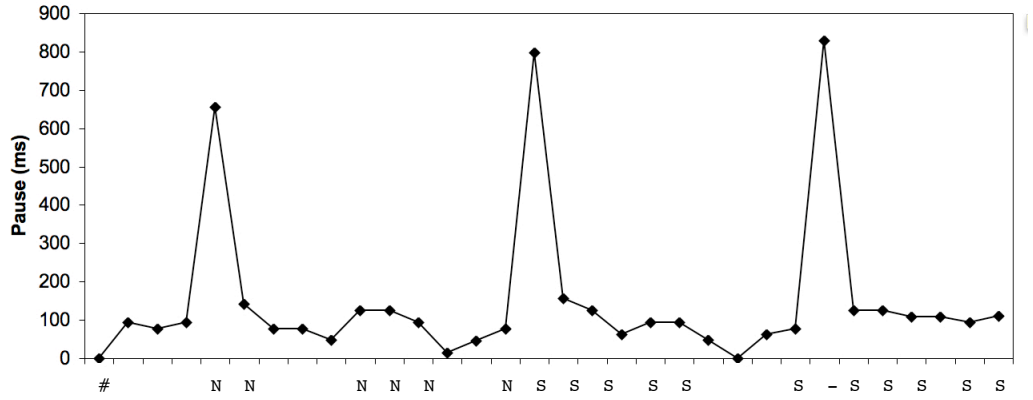

Fig. 3. Sample temporal protocol graphs for P2 writing their name

Method

Four volunteer participants, at the University of Sussex, were chosen to represent four distinct levels of experience in writing mathematical formulas. They were: a porter (P1) with high school level education; two postgraduate students, one in anthropology (P2) and one in electrical engineering (P3); an experienced lecturer in electrical engineering (P4).

At the beginning of the experiment the participants wrote their first name and family name. Table 1 shows the 16 targets formulas, which are arranged into five groups. They were presented to the participants and after brief familiarization (≈2 minutes) they were asked to copy them into blank boxes underneath each equation. A hash (#) was written at the beginning of each equation to ensure that the writing process was well underway before the first element. The order of copying is given in the 'task order' column.

A standard graphics tablet (Wacom Intuos$^{2®}$) and specially designed drawing/writing analysis software, TRACE (Cheng & Rojas-Anaya, 2004), were used to record the writing actions, to extract the pen positions and times, and to analyse the duration of pauses between drawn elements. Fig. 1 shows a snapshot from the TRACE graphical recording and analysis program, for the writing of the Quadratic solution. The small circles superimposed on each written symbol indicate the beginning and end of the production of elements in those symbols as the pen touches or leaves the paper. The lines between the elements indicate transitions where

the pen is off the paper. Note the two pairs of dots on the '4' as it was written in two parts.

Results

Examples of protocol graphs

Fig. 2 shows graphs of the pause durations for the writing of the quadratic solution, by the P1 and P4 who have the least and most formula writing competence. The pattern of pauses suggest that P4 writes this formula as small number of chunks but that P1 is treating it as a large number of elements. Inspection of the graphs for all of the tasks and participant reveals similar patterns. This suggests that the temporal signal is manifest in the copying of formulas and different levels of competence may be distinguished.

Name and L1 base-line

Fig. 3 shows the pauses for each mark made by P3 as he wrote his name. The letters on the x-axis correspond to each letter of P3's given name (N) and hyphenated surname (S). Distinct peaks are apparent for the first letter of each part of the name. The pauses for the other letters are much shorter and approximately equal. Pauses for stokes within a letter are even shorter. All participants show remarkably similar patterns of pauses for writing their names with a clear single chunk for each part of their name. This is in marked contrast to the clear differences between participants in overall patterns in the graphs for the writing of formulas, (e.g., cf. Figs. 2a and b).

Hence, the durations of the pauses for each letter within a name, which are intra-chunk L1 pauses, will be used as the baseline for each participant. The median L1 pauses are P1=109, P2=109, P3=94 and P4=148 ms.

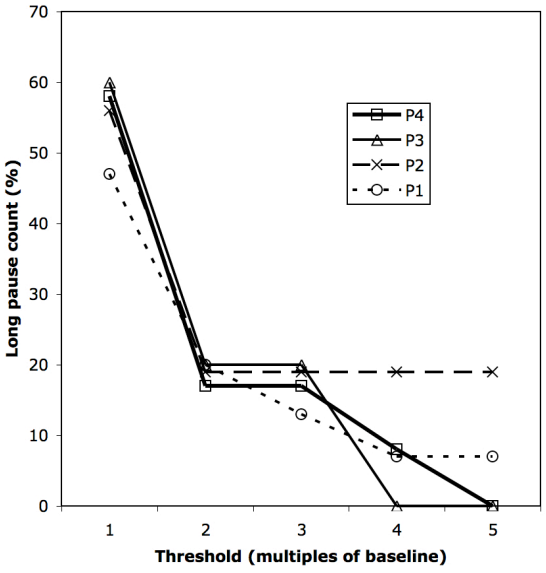

Fig. 4. Thresholds for LPC for writing of names

LPC threshold

The LPC measure counts of the number of long pauses as a proportion of the total number of pauses for symbols. A pause is taken to be long if its magnitude above the individuals' baseline is greater than some threshold. Consider possible thresholds that are multiples of the baseline. Fig. 4 shows values of LPC for the participants writing their names using different integer multiples of the baseline for the threshold. Given that participants are highly experienced at writing their own name, presumably at ceiling, a multiple of the baseline above three is not suitable, as the LPC values are varied. Some multiple of baseline below two are also not appropriate, because LPC values are too high, suggesting that about two in three pauses are long, even though each part of the participants' names have about 6 letters. A threshold of three times the baseline, rather than the two, is chosen as this will be more conservative in its classifying of pauses as long.

Comparison of LPC and LPD measures

Figs. 5 and 6 show the LPD and LPC scores for each participant on each copying task. The figures also include the respective values for the name writing. The data is presented in order of groups from Table 1. Inspecting the graphs reveals that the overall order of increasing LPD and LPC scores is P4<P3<P2<P1, consistent with the prediction that they should decrease with greater competency. As the data corresponds to single participants on single tasks, estimates of the significance of the pattern can be obtained by using the Binomial distribution as a model. The most severe case of a participant's score being out of order is P1 on the LPC measure, Fig 6, in which there are nine cases out of 16 where the value is not the greatest. For a given task the probability that a score for a participant is in the expected position may be taken to be p=0.25 (e.g., P1 is the highest). Considering all the copying tasks, the probability that nine or more out of 16 cases has P1 in the expected rank position is P=.007. Given that there are three other participants each with better matches of scores to expect rank order, it is unlikely that the overall pattern is due to chance.

Comparing the patterns of participants LPC and LPD scores across the tasks, the LPD measure more consistently ranks the participants in order of expected competency. The mean LPD for participants P1-P4 overall 16 tasks are 4.70, 3.18, 1.97 and 0.54, respectively. Similarly, the values for just the five standard formulae are 4.24, 3.01, 1.63 and 0.69, respectively. The LPD scores do not only distinguish the relative order participants satisfactorily, but the magnitudes of their scores are quite distinct.

Adding one to the LPD values and multiplying the results by the respective P1-P4 participants' baselines gives median pause durations, which are 632, 404, 228 and 223 ms, respectively.

Performance on each group of tasks

The consideration of the groups of tasks (Table 1) will focus solely on LPD scores, Fig 5. The first set of five *standard*

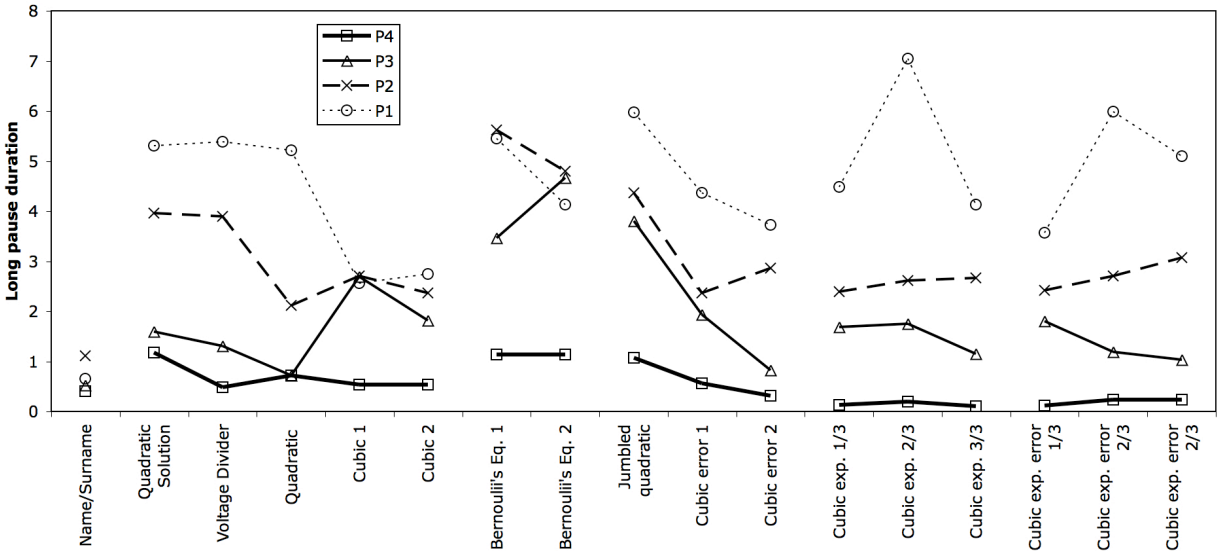

Fig. 5. LPD values for the different tasks and participants.

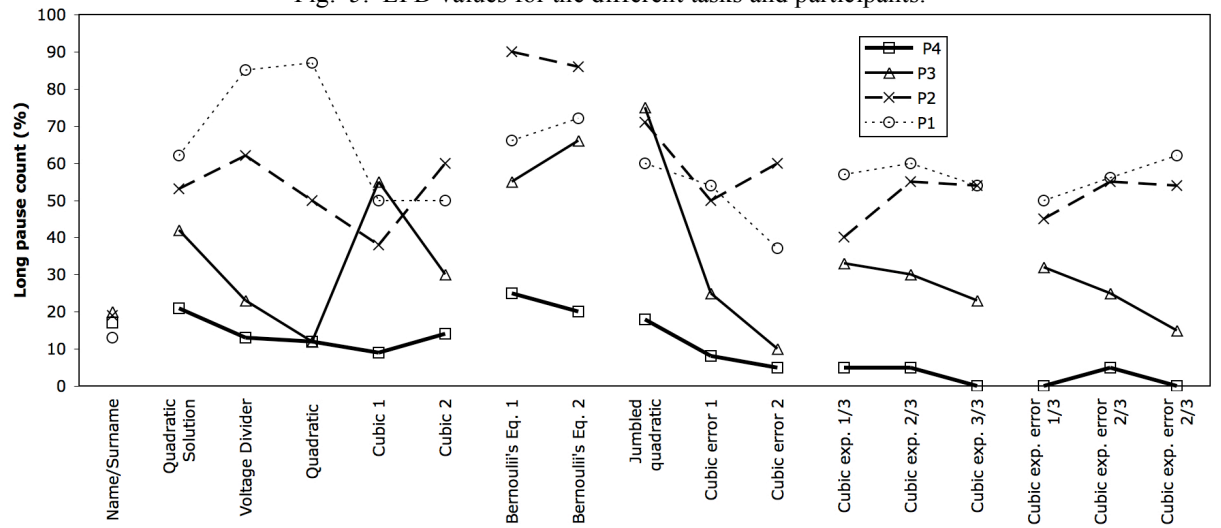

Fig. 6. LPC values for the various tasks and participant

formulas includes those with which P3 and P4 would expect to be familiar and P2 possibly familiar. The ranking of the participants is as predicted with exceptions in the quadratic and cubic-1 tasks. This pattern of orderings is unlikely to be due to chance. Using the Binomial distribution, as above, the probability that one score out of five tasks for a participant is out of place is P=.03, which applies to P1, P2 and P4. Similarly, for two scores, with P3, it is P=.09. Hence, an overall estimate of the probability, taking the product of all the probabilities is P<.001.

The Bernoulli group has two versions of the multi-term equations. None of the participants reported that they were familiar with them. There is little to distinguish P1-P3, but P4's scores are substantially lower and comparable to his scores in the standard formulas group, but just a little higher. A tentative interpretation is that P4's formula writing competency is robust and transfers well to novel items.

The Errors group of tasks contains three formulae that are incorrect versions of the last three tasks in the standard formulae group. The structure of the two Cubic-error equations is the same as the standard form of cubic formulas except that the signs and/or values of the coefficients are incorrect. The Jumbled quadratic is different in that symbols and operators have been mixed up systematically (e.g., '(' replaces each occurrence of 'a'). For all three tasks the order of the participants' corresponds to the predicted competency. Again, assuming the Binomial model the probability of the ranking occurring by chance is P=.016, given that there are three cases and the rank is as expected in each one. For each participant LPD scores for the Jumbled-quadratic are longer than their LPDs for the standard quadratic, which is a significant difference according to a one-tailed paired t-test: P=.04. There is no consistent pattern for the comparison of the Cubics (1 and 2) and their error versions, with t-

tests for cubic-1 versus cubic-error-1 P=.39 and cubic-2 versus cubic-error-2 P=.45. The difference between LPD scores for the quadratic and the mean of the Cubics is not significant (t-test, P=.43), but the difference between the Jumbled-quadratic and the mean of the Cubic-error tasks is (t-test, P=.01). The 'mixing' of the symbols in the Jumbled-quadratic significantly degrades the quality of copying and has a more deleterious effect than superficial errors in the value and sign of the coefficients.

The Cubic-expansion group contains three lines of working in which a cubic binomial is expanded. The corresponding Cubic-expansion-error-group includes formulas that have superficial errors. The overall ordering of participants is consistent with the predicted ranking in terms of competence for every one of the six lines in the two groups. Depending on whether the lines are taken as sets of three cases or one set of six cases, the probability of the pattern being due to chance are P=.016 or P=.0002, respectively. Comparing the two sets, the pattern of LPDs are clearly very similar, despite the errors in the second group.

Overall, the analysis in terms of the groups of tasks reveals that the ranks of participants are consistent with the predicted levels of participants' competence. Various interesting patterns of differences with and between the groups have been found.

Discussion

The overall goal of this study was to show that it is feasible to use Graphical Protocol Analysis (GPA) as a method to investigate chunk-based cognitive processing. This has been demonstrated for the task of copying mathematic equations. The temporal chunk signal appears to be strong and robust despite the range of factors that could be expected to substantially degrade the signal, such as the spatial and typographical complexity of the stimuli, the additional processes involved in copying compared to direct production from memory, and the potential use of different task strategies by different participants.

Setting baselines for individual calibration of the measures by exploiting the individuals writing their names appears to be a satisfactory approach. The values of L1 pauses obtained from participants writing their own names were approximately half the magnitude of L1 pauses in the previous number sequences and word phrase experiments. This is consistent with the greater level of familiarly of the names and may represent a lower bound for L1. Although they cannot be directly compared, it is noted that the median duration of pauses for P1 and P2 were comparable to the typical L2 value from the previous studies, whereas the values for P3 and P4 were comparable to the L1 values. This suggests some interesting relations connecting writing from memory and copying across different level of competence, which may be explored in future work.

The LPD and LPC measures both successfully identified the appropriate rank ordering of the level of competence in formula writing for the four participants. The LPD was better able to differentiate the competence levels of participants. This implies that factors such as the additional processes involved in copying may be entrained in the processing of chunks rather than acting independently. This is an issue for further investigation and better theoretical justification. There seems more to the chunk signal in copying than the merely effect of less competent writers using more chunks. However, LPD does now provide the required empirical leverage to enable the investigation of the subprocesses that are involved in graphical production. Performance on different groups of formulas is a case in point. There was little difference between individual performance on correct formulas and ones containing simple errors, such as sign and coefficient value changes. This suggests that the participants were not checking that the formulas were correct during mere copying. This includes the expansion of the cubic expansions. In contrast the Jumbled-quadratic shows a marked rise in LPD scores by all participants, which indicates that copying performance degrades when the overall structure of the equations are not in recognisable canonical forms.

It is noteworthy that the different level of competence were successfully distinguished on the basis of single participants, without the need to aggregate data over multiple individuals at the same level, or to aggregate data over several repetitions of each task. This implies that when the temporal chunk signal is encapsulated in a measure like LPD, the strength and robustness of the signal is maintained.

References

Chase, W., & Simon, H.(1973). Perception in chess. *Cognitive Psychology* **4**,55-81.

Cheng, P. C.-H., McFadzean, J., & Copeland, L. (2001). Drawing out the temporal structure of induced perceptual chunks. In *Proceedings of the Twenty Third Annual Conference of the Cognitive Science Society* (pp. 200-205). Mahwah, New Jersey: Lawrence Erbaum.

Cheng, P. C. H., & Rojas-Anaya, H. (2003). Writing out a temporal signal of chunks: patterns of pauses reflect the induced structure of written number sequences. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 424-429). Mahwah, NJ: Lawrence Erlbaum.

Cheng, P.C-H., & Rojas-Anaya, H. (2004). TRACE user guide (Unpublished Representational Systems Laboratory report).

Cheng, P. C.-H., & Rojas-Anaya, H. (2006). A temporal signal reveals chunk structure in the writing of word phrases. In *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum.

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

McLean, R., & Gregg, L. (1967). *Effects of Induced chunking on Temporal Aspects of Serial Recitation*. Journal of Experimental Psychology, 74(4), 455-459.