Copying equations to assess mathematical competence: An evaluation of pause measures using graphical protocol analysis.

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

Abstract

Can mathematical competence be measured by analyzing the patterns of pauses between written elements in the freehand copying of mathematical equations? Twenty participants of varying levels of mathematical competence copied sets of equations and sequences of numbers on a graphics tablet. The third quartile of pauses is an effective measure, because it reflects the greater number of chunks and the longer time spent per chunk by novices as they processed the equations. To compensate for individual differences in speeds of elementary operations and skill in writing basic mathematical symbols, variants on the measure were devised and tested.

Keywords: Chunks; pause analysis; freehand copying; mathematical competence; Graphical Protocol Analysis

Introduction

In Cognitive Science it has been well established for decades that the duration of pauses in overt behaviours often reflects the amount of mental processing required for ongoing sub-tasks. Such pauses are found in behaviours such as verbalizations, writing, drawing and interaction with computer interfaces (e.g., Miller, 1956; Egan & Schwartz, 1979; Newell, 1990; Cowan, 2001; Cheng, McFadzean & Copeland, 2001). The analysis of pauses provides a technique to study cognitive phenomenon by identifying how cognitive demands vary during and between tasks. Further, it may be feasible to use patterns of pauses to identify the specific organization of chunks that an individual has in working memory, including when those structures are multilayer hierarchies (van Genuchten & Cheng, 2010). All this is possible as pauses reflect the amount of processing that is required to produce an action, which is less for an element within a chunk than the first element of a new chunk. For example, writing the first letter of a new word requires some processing of the word itself, whereas writing subsequent letters just requires the processing of the letters alone. The same is true for higher levels in the hierarchy of a stimulus; e.g., beginning to write the first word of a new sentence requires some processing of sentence level information that is unnecessary for subsequent words in the sentence (van Genuchten & Cheng, 2010).

This facility to directly access the amount of cognitive processing required for certain tasks and, potentially, the specific organization of information in memory raises the interesting possibility that the analysis of pauses may be used to quickly and efficiently measure levels of competence, or expertise, in knowledge-based tasks. We have used *graphical protocol analysis*, GPA, as a term for methods developed to exploit the temporal signal in the distribution of pause durations for the measurement of cognitive abilities and for investigations of the organization of information structures in memory. The focus in this paper is on pauses in simple copying tasks, in which a *pause* is the time between the pen touching the paper to begin a new stroke minus the time the pen was lifted from the paper at then end of the previous stroke.

Cheng & Rojas-Anaya (2007) used graphical protocol analysis in a study with four people of quite different levels of experience in mathematics, from elementary to expert, whilst they simply copied mathematical formulae freehand. To assess their competence a measure was devised – *Long Pause Duration* (LPD). To calculate LPD participants first write their names (given and family) several times and a baseline value consisting of the mean of the all pauses between letters, within each part of their name, is calculated. LPD equals the mean of a participant's pauses on a target stimulus minus that baseline, all divided by the baseline. Thus, LPD attempts to normalize pause values both respect to absolute and relative individual differences, by subtracting and dividing by the baseline, respectively. It was found that LPD correctly rank ordered the four participants on single test items, which is noteworthy as the participants were being reliably differentiated without any aggregation over trials or with repeated testing. However, there are some limitations to this result. First, given the large differences between the participants' abilities, it is an open question whether the overall approach can be used to differentiate people with finer grained differences in mathematical competence, which is necessary if GPA is to be used as a practical assessment technique. Second, LPD is just one of several imaginable measures based on pauses, so others should be investigated particularly as LPD's dual use of the same baseline for two forms of normalization lacks a good theoretical justification.

Zulfilki's (2013) PhD thesis studied various GPA techniques and measures for the assessment of English language competence of students for whom it was not their first language. It was found that students with different levels of language competence could be differentiated in the simple task of copying sentences. More reliable differentiation occurred when the students were not permitted any preparation time than when they briefly previewed the sentences. Explorations of a number of measures of competence revealed that the third quartile of the pauses appear to correlate strongly with independent language competence scores.

A theoretical justification can be made for a measure based on the third quartile of pauses by considering the differences in the shape of the distributions of pauses for experts versus novices. Experts possess chunks containing more elements and so the overall hierarchical organization

N1 # 193 826 745 931 584 726 247 315 896 915 438 589 462 731 762
N2 # 56789 12345 34567 54321 98765 76543 56789 23456 45678
N3 # 267 5193 833 4706 478 9621 582 6479 589 4627 314 5121
N4 # $12=2\times6=3\times4$ $12=3+3+3+3$ $3+4+3+4=14$ $2+3+2+3+2+3=15$
E1 # $(a+b)+c=a+(b+c)$ $d+e=e+d$ $g(h+i)=gh+gi$
E2 # $\sin^2 r+\cos^2 r=1$ $\tan x=\sin x/\cos x$ $\sin 2u=2\sin u\cos u$
E3 # $(x+1)^2=x^2+2x+1$ $(x-a)(x+a)=x^2-a^2$ $(x-y)^3=x^3-3x^2y+3xy^2-y^3$
E4 # $x_1,x_2=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$

Fig 1. Number (N) and Equation (E) stimuli.

of the chunks in memory will be less complex. In contrast novices encode fewer elements per chunk with the consequence that their hierarchical organization will likely have a greater span at each level and also will be deeper overall. Thus, in tasks involving the processing of structures of chunks, novices in contrast to experts will exhibit a greater number of transitions across chunk boundaries, with their associated longer pauses, but also their navigation of a larger tree of chunks will tend to mean those transitions will be more demanding, so the durations of the pauses will be longer than for experts. Therefore, the overall distribution of pauses will be positively skewed, so a measure that recognizes this characteristic seems promising. Further, as the number of elements per chunk in naturalistic tasks is often about four (Cowan, 2001), this suggests that the third quartile of pauses will be appropriate; rather than, say, the median or the 95th percentile, which are likely to include too many within chunk pauses, or too few between chunk pauses, respectively. See van Genuchten, & Cheng (2010) for a fuller discussion of the relation between pause duration and chunk structure.

The aim here is to extend the precision, rigour and scope of the findings of Cheng & Rojas-Anaya (2007). The experiment will test whether it is feasible to differentiate participants at a finer grained level of mathematical competence than the relatively gross levels of expertise in the previous experiment. In addition to comparing measures based on the third quartile of pauses to the LPD measure, two normalization methods to refine the measures will also be examined that attempt to take into account individual differences. (a) As a measure of individuals' baseline speed of elementary processing of written symbols, the first quartile of pauses will be used. This will be subtracted from third-quartile measure for each individual, which in effect is an attempt to align the main body of participants' distributions of pauses so that positive tails can be more precisely compared. The difference between the third and first quartile is of course the interquartile range (IQR). (b) In the copying of mathematical formulae there are basic things that will be familiar to everyone, because they are fundamental pieces of mathematical knowledge. One's understanding of numbers and simple numerical equations is one such *subordinate maths skill*. The experiment will attempt to factor out, normalize for, such experience by measuring copying performance at this level and subtracting a measure for this subordinate skill from each individual's performance on the more sophisticated target equation items.

Thus, the questions addressed by the experiment are:

1) Are measures of competence based on the 3rd quartile of pauses appropriate for a mathematical domain? The third quartile of pauses will be called $Pause_{Q3}$ for short.

2) Does the interquartile range of pauses, $Pause_{IQR}$, provide a means to compensate for individual differences in processing speed at the level of elementary operations?

3) Can the accuracy of the pause measures be improved by normalizing them with respect to baseline measures of subordinate maths skills, in order to focus upon aspects of performance more closely associated with higher maths competence? Here the target skill is the copying of complex equations, so the copying of the simple number sequences and sums will be used as the subordinate skill to establish a baseline measure for normalization. In other words, the measure from these stimuli will be subtracted from measures of the equation stimuli.

4) Are these measures an improvement on the LPD measures used in the previous experiment?

5) This is a supplementary exploratory question: can the pause measures be used to determine the relative difficulty of different test items between and also within test item types? This will be investigated by aggregating item scores over participants and attempting to relate them to particular characteristics of the test items.

Method

The experiment was run as a course work assignment of the *Psychological methods for system evaluation* module of the HCI MSc at the University of Sussex, in 2013. Four students on the module received course credit for collecting and collating the data.

Participants

Twenty participants were recruited by deliberately sampling participants with various level of maths competence from among the friends, family and colleagues of the four student experimenters. All were adults without impairments. There were approximately equal numbers of female and males.

Materials

Tests of mathematical competence were devised to measure three aspects of participants' competence. The first aspect concerned their prior experience of maths which involved nine graded questions about their level of education in maths from exams taken at age 16 through to degree level, and questions about whether participants used mathematics for their work and leisure pursuits. The second aspect involved answering eight graded multiple-choice questions on mathematical problems (e.g., 'The perimeter of a square is 20 units. Find its area.' 'What is the number of solutions to a quadratic equation?'). The third aspect involved the participants rating their confidence on their answers to each of the questions in aspect 2, using a seven point Likert scale. The tests were completed online using a commercial questionnaire delivery web site.

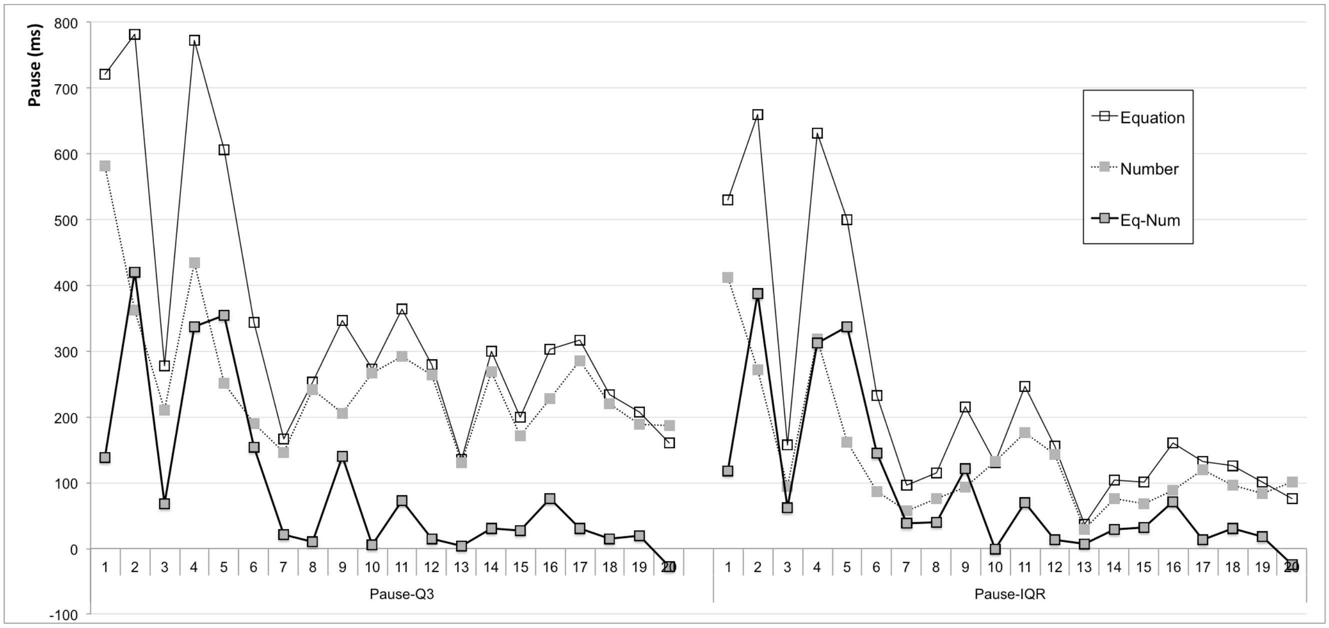

Fig 2. $Pause_{Q3}$ and $Pause_{IQR}$ measures for participants who are rank ordered by competence score. Data for equation and number stimuli are shown along with the normalized measure (Eq-Num) of equation pauses minus the number pauses.

There were eleven copying test items. Three practice stimuli consisted of words familiar to the participants: specifically, their own given and family name (twice) and 'University of Sussex' (twice) and 'Falmer Brighton' (twice). The rest of the experimental stimuli consisted of eight items divided into two sets, four *number* items and four *equation* items as shown in Fig 1. The first set of numbers consisted of triplets of arbitrary digits. The second was sets of five ascending or descending digits. The third has sets of alternating three and four arbitrary digits. The final set included four simple arithmetic equations. The equation stimuli were algebraic equations drawn from various maths topics that would be familiar to schools students who studied maths beyond the age of 16 in the UK.

The tasks were performed on a Wacom Graphics Tablet (Intuos3) connected to personal computers. The copying was done onto plain white A4 paper in landscape orientations using a Wacom inking pen. A Java program, *SMouseLog*, specially written in our lab was used to capture the pen movements, including the position and time of each touch and lift of the pen from the paper. The temporal accuracy of the logging was better than 1 ms.

Procedure

Each participant completed the three components of the mathematical competence questionnaire.

After familiarization with the experimental setup, the participants first completed the practice items in order. Sequences of alternating number and equation items where devised, such that equal numbers of each type of item occurred first, and items occurred in most positions through out the sequences. Sequences were allocated to participants in a pseudo-random fashion. Given the design of the experiment and the nature of GPA data, none of the observations or findings depends on the specifics of the organization of the test sequences or the manner of the assignments.

For each trial a response sheet was taped to the tablet. When the participant was ready a card with the test item was turned over and was placed 5 cm above the tablet and so was ≈10 cm above the top edge of the paper. Participants were told to start at the left and middle of the sheet and to copy the stimuli as accurately and quickly as possible, and if they made a mistake to carry on without going back and correcting it. They were trained to start by writing the hash (#) and to continue straight on, in order to ensure the pause for the first symbol of the stimuli was legitimate.

Results

For a general competence score, it was decided to use an equally weighted sum of normalized values of the experience, problems and confidence scores, because each reflects an aspect of mathematical competence and because participants' score were spread across the whole range of each of the three tests. Pearson's *r* correlations between this overall competence score and the individual component scores are .90 for the experience, .81 for the problems and .83 for confidence, which supports the decision to use an equally weighted sum as each score makes a meaningful contribution to the overall competence score. The overall competence scores ranged from 8 to 99% in a uniform distribution.

Pause durations for each pen stroke by each participant on all the test items were calculated from the computer logs. The median Pause_{Q3} and Pause_{IQR} values were calculated for each test per participant. Means of these medians were calculated by aggregating over the practice, number and equation test items. Fig 2 shows the means of the medians across the participants, who are rank ordered by their competence scores. The graph to the left shows Pause_{Q3} values

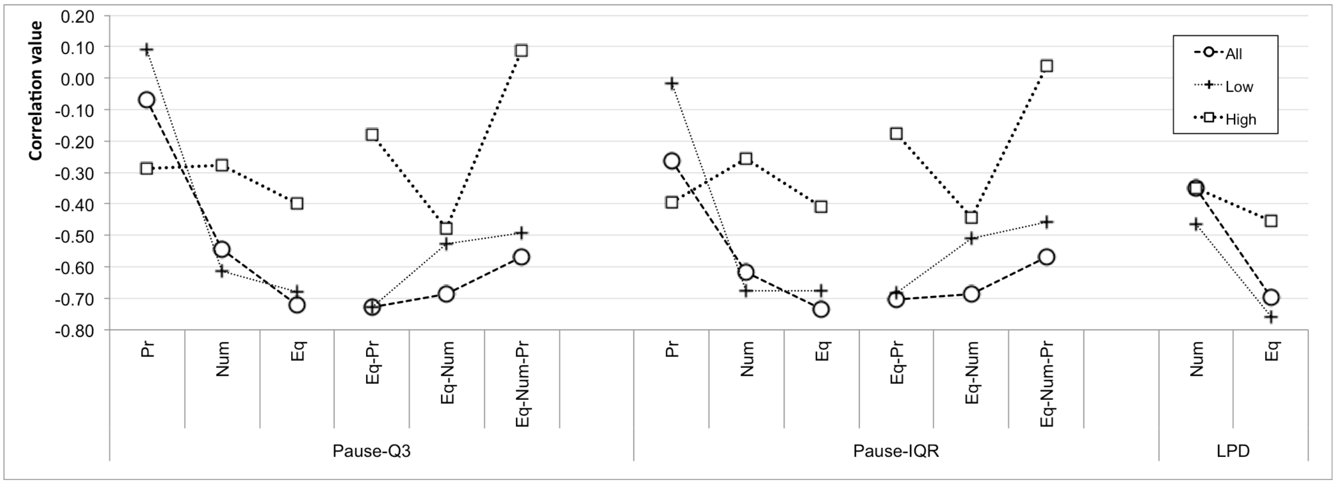

Fig 3. Correlations of $Pause_{Q3}$, $Pause_{IQR}$ and LPD measures against competence scores for levels of participants, across stimuli types and normalization methods. Stimuli: Pr – practice; Num – number; Eq – equation. Normalization: Eq-Pr – equation measure minus practice measure; Eq-Num – equation measure minus number measure; Eq-Num-Pr – equation minus both.

for equation, numbers and the normalized measure that is difference between the two (Eq-Num). Pause$_{Q3}$ for the equation items ranges from nearly 800 ms for the least competent participant down to less than 200 ms for the most competent. For all but the most competent participants, Pause$_{Q3}$ values for number items are less than the equation values. The heavier line lower in the graph (Eq-Num) is the equation Pause$_{Q3}$ normalized by subtracting the number Pause$_{Q3}$ from it. This has a range of over 400 ms. The graph to the right of Fig 2 shows the same information but for Pause$_{IQR}$. The overall pattern of the data is similar except that Pause$_{IQR}$ values for both equation and number items are typically 125 ms lower than Pause$_{Q3}$. For the normalized Eq-Num measures the mean of the differences between Pause$_{IQR}$ and Pause$_{IQR}$ is a mere 5 ms.

Graphs of the Pause$_{Q3}$ and Pause$_{IQR}$ values for the practice items show no overall patterns with respect to competence.

Clearly for both Pause$_{Q3}$ and Pause$_{IQR}$ across equation items, number items and the normalized measure (Eq-Num) there is a strong relation to participant competence scores. To examine these Pearson's *r* correlations were calculated between participant competence scores and Pause$_{Q3}$ and Pause$_{IQR}$ for (a) practice, (b) number and (c) equation items and the (d) Eq-Num normalization. Further, correlations were also calculated for (e) normalization with respect to practice items, *Eq-Pr*, and (f) normalization for both number and practice items, *Eq-Num-Pr*. These are shown in Fig 2, with round data points for all participants together. Note that the scale runs from -0.8 to 0.2. For one tail tests with N=20 (df=18) significant correlations at p<.05 and P<.01 have critical values of 0.378 and 0.516, respectively. The correlations for the practice items are not significant, as expected. However, for all the other measures they are significant in respect to Pause$_{IQR}$ and Pause$_{IQR}$.

To further investigate the nature of the correlations a binary split was performed on the competence scores to create two equal groups of participants, one *low* and one *high* competence group. The correlations for these groups are shown in Fig 2 as the cross and square data points, respectively. For one tail tests with N=10 (df=8) significant correlation at p<.05 have a critical value of 0.549. For both Pause$_{Q3}$ and Pause$_{IQR}$ the overall pattern of the data is similar. For the low competence group the number items, equation items and Eq-Pr normalization are significant for both Pause$_{Q3}$ and Pause$_{IQR}$. For the high competence participants none of the correlations are significant, but there are large interesting differences among the values. The strongest correlations are for the Eq-Num normalization, which are greater than for equations alone. However, when normalizing using the practice values (Eq-Pr and Eq-Num-Pr) the correlations drop substantially.

To the far right of Fig 3 LPD values are shown for both number and equation items. For the baseline of the LPD values the first quartile of each practice item was calculated and the mean across them found. The LPD value was calculated for each participant by subtracting their baseline from their median pause durations and dividing that by the baseline; as per the definition given above. For the participants taken as a whole the LPD correlation is significant for the equation items but not for the number items. For the subgroups only the low competence group on the equation items had a significant degree of correlation.

To explore the different types of test items in more detail Fig 4 shows mean Pause$_{Q3}$ values for each of the test items, aggregated over all participants and also separately for the low and high competence groups. For the practice items all three cases are indistinguishable. Considering all the participants there is a clear overall trend of increasing Pause$_{Q3}$ values for practice, numbers and equations. The breakdown into low and high competence reveals that it is the substantial increase from practice to number, and from number to equations, of the low competence group that is responsible for the overall trend. The mean across the four number items for the low and high competence groups are 289 and 244 ms, respectively, which is a marginally significant difference at P=.08 (t = 1.83, one-tail, df=18). The mean

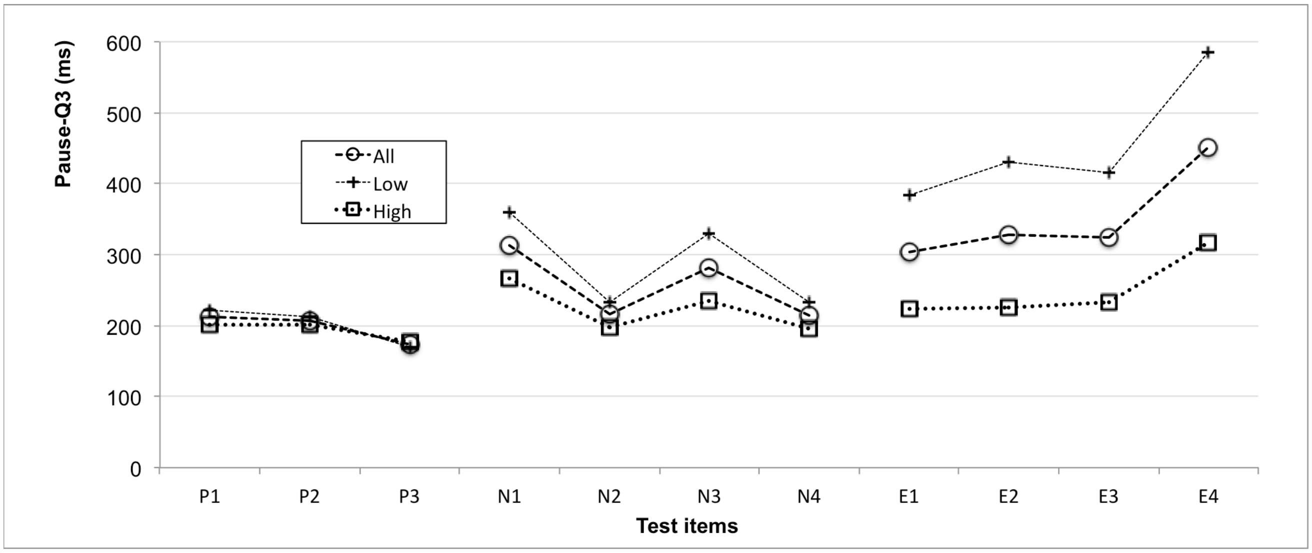

Fig 4. Pause_{Q3} values for each test item compared across the different competence levels of participants.

across the four equation items for the low and high competence groups are 454 and 250 ms, respectively, which is a significant difference at P<.01 (t=2.91, one-tail, df=18). Close inspection of the graph for the low competence group shows that none of the Pause_{Q3} values for the three types of items overlap. Perhaps the only noteworthy divergences of the high competence group are for N1 and E4, which are higher than the other values in the same test set (significantly so by one-tail matched t tests (with df=9) compared against the mean of the other three test items in the set: N1 vs. N2-4, p=.002, t=4.23; E4 vs. E1-4, p=.004, t=3.78).

Discussion

Overall, the experiment has demonstrated that it is feasible to use Graphical Protocol Analysis with simple freehand copying tasks as means to measure competence in an information rich domain. As such, the findings contribute to the evidence from previous studies (e.g., Cheng & Rojas-Anaya, 2007; van Genuchten & Cheng, 2010) that there is a rich temporal signal that reflects the level of cognitive demand that participants face when reading and reproducing meaningful stimuli, which is a function of their level of competence in the target domain. Low competence participants require more processing time in the copying tasks than those with high competence.

The first question posed in the Introduction concerns whether measures based on the third quartile of pauses can be used to assess mathematical competence: the answer is affirmative. For the participants taken as a whole, strong correlations exist between the competence scores and the particular measures of interest, specifically Pause_{Q3} and Pause_{IQR}, for equations considered alone and also for the Eq-Num normalized measure (Fig 3). Further, the correlations for both Pause_{Q3} and Pause_{IQR} were also relatively strong for the number items but less so than for the equation items. Although it might be expected that all participants would have equal facility at copying a simple sequences of numbers and basic sums, there appears to be some difference between the groups (Fig 4). For the equation items that difference is more substantial. It is this large difference between the low competence group's pause measures that is the major contributor to the pattern found for the whole group of all participants (Fig 4).

In Cheng & Rojas-Anaya (2007) it was shown that participants with large differences in levels of mathematical competence, from minimal experience to expert, could be distinguished using the LPD pause-based measure. The findings here provide stronger evidence of the value of GPA in simple copying tasks, because a larger number of participants were examined with finer differences in their levels of competence. Further, the strong correlations of the low competence group, who were spread uniformly across the lower half of the range of competence scores, demonstrates that the approach can be effective with a relatively narrow range of competence. Thus, this approach may have some potential utility as a tool for educational assessment.

A stronger claim could be made about the potential of GPA and copying had similar strong correlations been found for the high competence group, but they were not. Why were the correlations not as strong? One explanation is that test items were not sufficiently challenging to differentiate among the high competence participants. Evidence for this can be found in two places. First, there is a relatively flat distribution of Pause_{Q3} and Pause_{IQR} for Eq-Num values for participants 11 to 20 in Fig 2, which suggests that these participants may be at ceiling in their performance. Second, as seen in Fig 4, there is little difference between the equation items and either the number or practice items for the high competence group, which may mean that the equation stimuli are as familiar to them and as easy to process as the other sets of items. An important implication follows from this: measures of competence based on the analysis of pauses may be sensitive to the difficulty of stimuli relative to individuals. This suggests that tests might be designed which incorporate a set of items spanning a range of difficulty and that measures could be devised to exploit the differences in the levels of stimuli difficulty.

The second question is whether Pause_{IQR} is a better measure than $\text{Pause}_{Q3,}$ because subtracting first quartile pauses may provide a means to compensate for individual differences occurring at the level of elementary operations. Comparison of the two measures in Fig 3 shows little difference in their correlations with the competence scores. One explanation for this lack of difference is that the magnitude of first quartile pauses is small compared to Pause_{Q3}, on average a third of the size, so differences between participants first quartiles values will be small compared to the size of Pause_{IQR} and so will make little impact and may even act as a source of noise in the data

The third question is whether subtracting Pause_{Q3} values for number items from equations items may serve as a means to factor out, normalize for, generic subordinate skills related to the writing of numbers and basic mathematical symbols and so provide a focus on more sophisticated aspects of equation copying. Fig 3 shows that Eq-Num Pause_{Q3} and Pause_{IQR} correlations are no greater than the values for equation items alone, when all participants are considered together. In the case of the lower competence participants the correlations actually decline moderately, whereas with the high competence participants the Eq-Num measure increases the strength of the correlation, but only slightly. Thus, there appears to be little general value in attempting to normalize the measures using equations with the measures for numbers, because for the more competent group the improvement is small and for the low competent group it appears to be making the signal nosier.

The potential to degrade the pause measures, and hence a lesson to learn about this danger in general, can be seen by examining values of the Eq-Pr and Eq-Num-Pr measures in Fig 3, which presents an interested pattern of data. The subtraction of practice values from equation items alone or from the Eq-Num measure degrades the high competence group's correlations substantially, but has relatively little effect on the low competence group. This is probably due to the similarity in the magnitude of the high competent group's equation, number and practice item values, so when they are subtracted what is left is largely noise due to the natural variability in those measures.

The fourth question is how the present third quartile pause measures compare to the LPD measure proposed in Cheng & Rojas-Anaya (2007). Fig 3 shows that LPP is broadly similar to the present third quartile pause measures. As LPD is more complex and its theoretical justification less rigorous, it appears appropriate to favour Pause_{Q3} or Pause_{IQR} in general use.

The final question is whether by aggregating over participants it is possible to assess the relative difficulty of test items, which may be useful for the purpose of designing evaluations by combining items of different level of sophistication. As seen in Fig 4, the copying of the most complex of the equation items, E4, is distinct from the other items, with a significantly larger value of Pause_{Q3}. However, E4 has more 2D structure than the rest, which might lead to elevated pauses merely because of the spatial distribution of symbols, which is in contrast to the linear arrangement of the other items. Further, the simplest of the number items, N1, has Pause_{Q3} values larger than the rest of the number items. This may be because each of the three digits (see Fig 1) is processed as individual chunks and so the relative density of long pauses associated with new chunks is greater than the other stimuli where the number of elements may be nearer the typical value of four per chunk. The implication of all this is that care must be taken in the analysis of pauses, because there are multiple factors that may contribute to the shape of the distribution of pauses, including the relative frequency of chunk boundaries and also aspects of stimuli structure that elevate pause duration but that are not specifically related to mathematical difficulty. Nevertheless it has been shown the Graphical Protocol Analysis can be used to measure mathematical competence in tasks involving the simple copying of equations.

Acknowledgments

My thanks go to Ewen Cluley, Rebekah Egli, Hayden Saughter and Jo Rusecka for their thoroughness in the running the trails and their conscientiousness in the collation of the data.

References

Cheng, P. C-H., McFadzean, J., & Copeland, L. (2001). Drawing out the temporal structure of induced perceptual chunks. In J. D. Moore & K. Stenning (Eds.), *Proc. of the 23rd Ann. Conf. of the Coge Science Society* (pp. 200-205). Mahwah, New Jersey: Lawrence Erlbaum.

Cheng, P. C-H., & Rojas-Anaya, H. (2007). Measuring Mathematical Formula Writing Competence: An Application of Graphical Protocol Analysis. In D. S. McNamara & J. G. Trafton (Eds.), *Proc. of the 29th Ann. Conf. of the Cognitive Science Society* (pp. 869-874). Austin, TX: Cognitive Science Society.

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.

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

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.

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

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

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.