Competence Assessment by Stimulus Matching: An Application of GOMS to Assess Chunks in Memory

Hadeel Ismail (hi71@sussex.ac.uk)
Department of Informatics, University of Sussex
Brighton, BN1 9QJ, UK

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

Abstract

It has been shown that in hand-written transcription tasks temporal micro-behavioral chunk signals hold promise as measures of competence in various domains (e.g., Cheng, 2014). But data capture under that an approach requires the use of graphics tablets which are relatively uncommon. In this paper we propose and explore an alternative method – Competence Assessment by Stimulus Matching (CASM). This new method uses simple mouse-driven interfaces to produce temporal chunk signals as measures of learner's ability. However, it is not obvious what features of CASM will produce effective competence measures and the design space of CASM tasks is large. Thus, this paper uses GOMS modelling in order to explore the design space to find factors that will maximize the discrimination of chunk measures of competence. Results of a pilot experiment show that CASM has potential in using chunk signals to measure competence in the domain of English language.

Keywords: chunking; GOMS; language competence; pause analysis; stimulus matching

Introduction

This paper concerns the assessment of learners' competence in knowledge rich domains, using the analysis of computer logs of micro-behaviors in task activities. Moss, Kotovsky, and Cagan (2006), in the domain of engineering, and Arslan, Keehner, Gong, Katz, & Yan (2020), in the domain of mathematics, used drag and drop tasks to examine the underlying cognitive processes in either replicating subject-related diagrams or solving mathematical problems, respectively. Another study analyzed pauses during text composition by means of key-stroke logging (Schilperoord, 2002). These methods were successful in extracting and associating behavioral signals with cognitive processes, by logging actions at a time scale of ≈10 seconds.

An alternative approach that holds some promise is to log and analyze micro-behaviors at a time scale of 1 second and less. Machine learning was used to analyze large amounts of data logged during freehand writing (Stahovich & Lin, 2016) and drawing (Oviatt, Hang, Zhou, Yu, & Chen, 2018) during problem-solving tasks. Their findings revealed significant correlations between pause durations and proficiency levels.

In contrast, Cheng and colleagues have used cognitive chunking theory to develop methods that require less data using short transcription tasks. According to Cowan (2001) and Miller (1956), "chunking" is a process by which perceived information are grouped and stored in working memory (WM), and since information is presented as units, people tend to group these units into "chunks" of meaningful information. The number of "chunks" stored is constrained by one's mental capacity, however Cowan (2001) also points out that the capacity is also affected by the amount of prior knowledge one holds in long term memory in the expert domain. So, in the experiments carried out by Cheng and colleagues, they examined differences in pause behavior of novices and experts whilst engaging in transcription tasks to probe chunk structures in memory. Cheng and Rojas-Anaya (2007) observed individuals copying mathematical equations freehand and could distinguish level of experience. However, their sample size was small and participants had large differences in their mathematical expertise. Extending the approach Cheng (2014) showed strong correlations between competence and the third quartile (Q3) pauses. Similarly, Zulkifli (2013) asked learners of English as a second language to transcribe English sentences freehand and found Q3 to be an effective measure of competence. Albehaijan and Cheng (2019) show the possibility of measuring programming competency using the same method. Overall, it seems that pause based measures in transcription tasks have some potential for assessing competence in various domains.

Despite the promise of freehand transcription, one limitation is the need for a graphics tablet, an uncommon IT equipment. Thus, it would be desirable to combine mouse driven tasks (Arslan et al., 2020; Moss et al., 2006) with the benefits of capturing micro-behaviors. Cheng (2015) used a mouse and a response grid on a screen to measure temporal chunk signals related to mathematical competency. Participants copied the stimuli by clicking on the matching symbols that appeared on the grid. Results showed that clicking to select symbols has potential as a means to measure mathematical competence.

In this paper we propose an alternative approach to the assessment of competence administered on a standard computer by means of mouse clicking: *Competence Assessment by Stimulus Matching* (CASM). A preliminary CASM task design has been created (Fig. 1), that takes into consideration the different factors that would encourage the

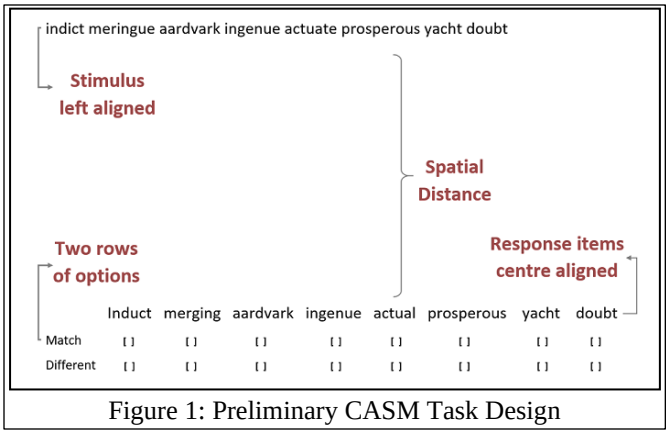

Figure 1: Preliminary CASM Task Design

use of chunking. The task is presented as a split screen with the stimulus at the top and the response area at the bottom. The response area includes words that either match or differ from the stimulus. Participants are expected to verify the match or mismatch and use the mouse to mark their responses as quickly and as accurately as possible. The time course of clicks in the check boxes will reflect certain aspects of the test takers language competence, so measures of competence may be devised for the task.

A key problem is how to design CASM tasks to produce behaviors that maximally differentiate high and low competence. Will the micro-behaviors of experts and novices differ substantially on the task in Fig.1, and so potentially provide effective temporal chunks measures of competence? This paper considers the possibilities, but the possible space of design is large. Some of the factors influencing this task include: the large spatial distance, the deliberate misalignment of words, the use of low frequency words and multi-syllabic words. So, how can we effectively yet efficiently explore the space? A task analysis approach is adopted, in particular a somewhat novel approach to the application of GOMS modeling is used to assess chunks in memory in order to further determine how the different design factors impact the task environment.

Task Design Space

The aim is to develop chunk-based Competence Assessment by Stimulus Matching (CASM) tasks that rely on mouse clicking, in contrast to Cheng and colleagues pen-on-paper transcription approach. The key issue is the design space, where many variables provide us with a plethora of design choices, from which we must choose those that impact the distribution of pauses that maximally differentiate experts from novices.

Screen Layout and Stimulus Positioning: The layout may encourage the use of chunking to provide experts with an advantage over novices. Firstly, the spatial distance between the stimulus and the response areas may be made deliberately large to impose a task load on individuals, who must shift their gaze vertically. In turn this may encourage them to chunk as much as possible. Cheng (2014, 2015) used distant positioning to improve the Q3 pause measures of competence. Secondly, the misalignment of the stimulus and the response is assumed to encourage experts to use chunks to save the effort of switch gaze, and place some difficulty on the novices who, because of their limited language knowledge, might take longer to locate the point where they last left as they shift their gaze.

Presentation Mode: In presenting the stimuli, one approach is to have it visible throughout the duration of the task; "constant display" (Cheng & Rojas-Anaya, 2007; Cheng, 2014; Zulkifli, 2013). The second is "voluntary view", where the appearance of the stimuli requires an action by the individual (Albehaijan and Cheng, 2019).

Stimulus and Response Composition: The general approach here is to play with effects of stimulus and response composition or decomposition. This applies at the whole stimulus (sentence), word (compound words) and part word (syllable) levels. If working at the word level, one option is to present stimuli words in a way that, if two were combined, they would make up a compound word which may differentially benefit the expert by increasing their chunk size by treating the two words as one unit rather than two for a novice (e.g., "counter measure"). We would expect the benefit to be reflected in the pauses in the task and hence in measures of competence.

Stimulus Content: Content manipulations include word frequency (high and low), word length, sentence structure (simple, complex, incorrect), semantic meaning, etc. Zulkifli (2013) shows that such manipulations can be applied in ways that benefit experts to use their knowledge which may be revealed in chunking measures.

Method

The steps taken to carry out the task analysis are: (1) Design a number of task variations. (2) Use GOMS to develop flow charts that predict the processes employed by experts and novices. (3) Calculate the durations for each process, to predict differences in pause distributions and lengths. (4) Run a pilot study to evaluate the modeling results.

GOMS, is a well-established systematic approach to cognitive task analysis that is usually applied during system design to test for usability aspects, choose between candidate designs and understand user behavior (Card, Moran & Newell, 1983). However, our motivation is not to understand user performance, per se, but rather to find designs that constrain their behavior so that micro-behavioral signals of competence are as robust as possible.

While the GOMS models are usually applied to understand how the external task environment affects the individual's behavior, we on the other hand apply the analysis in a way to understand the internal processing of chunks, leading to how that impacts the design of the task. So, within the framework of GOMS, in our approach, ***goals*** are related to the size of the chunk an individual can hold in memory. Not only this is affected by the layout of the interface *(externally)* but its largely constrained by their level of familiarity with the words presented *(internally)*. Among the ***operators*** of particular interest to us are those classified as cognitive operators. Those that deal with the

decomposition of a chunk are decisional processes concerned with whether certain elements makeup a chunk or not. Others are related to retrieving chunks from memory, comparing and verifying. The ***methods*** are the internal loop processing by which the sequence of operators to achieve a certain sub-goal. ***Selection rules*** are choices that test takers will make to choose between alternative methods based on the chunks they possess, which will be manifest as different micro-behaviors and that chunk measures will attempt to measure.

Allocating Time Durations

All *operators* are allocated specific time durations that were mostly extracted from past GOMS studies.

1. **Word/syllable recognition:** The time for recognizing a six-letter word, a syllable or a letter is 340ms (John & Newell, 1989).
2. **Cognitive operators:** Cognitive operators include those processes that involve holding a chunk in memory, decision making, verifying, and comparing. According to the literature, the average duration for mental processes is between 50 and 70ms (Gray & Boehm-Davis, 2000; Olson & Olson, 1990; John & Newell, 1989). The proposed tasks involve low-level cognitive processing, so 50ms is chosen.
3. **Chunk retrieval:** This process was allocated a duration of 50ms, following similar studies involving immediate copying (John, 1988, as cited in Olson & Olson, 1990).
4. **Mouse move:** A quick pilot experiment was conducted on the author and an additional participant. The average time for moving between response items was 500ms, and 700ms for moving from the top screen to the bottom. The second was used as the duration of the action to reveal stimuli in voluntary display tasks.
5. **Eye movement:** The time for a saccade is 30ms (Russo, 1978, cited in Card et al.,1983).

Analysis

Task Analysis: Flowcharts

Since the design space is large it is impossible to examine all combinations of variables here, so we focus on the design in Fig. 1 as an exemplar. The main features of the design are the layout, use of low frequency words, inclusion of disyllabic and trisyllabic words, and presenting the stimulus in constant display mode. The flowcharts in Fig. 2

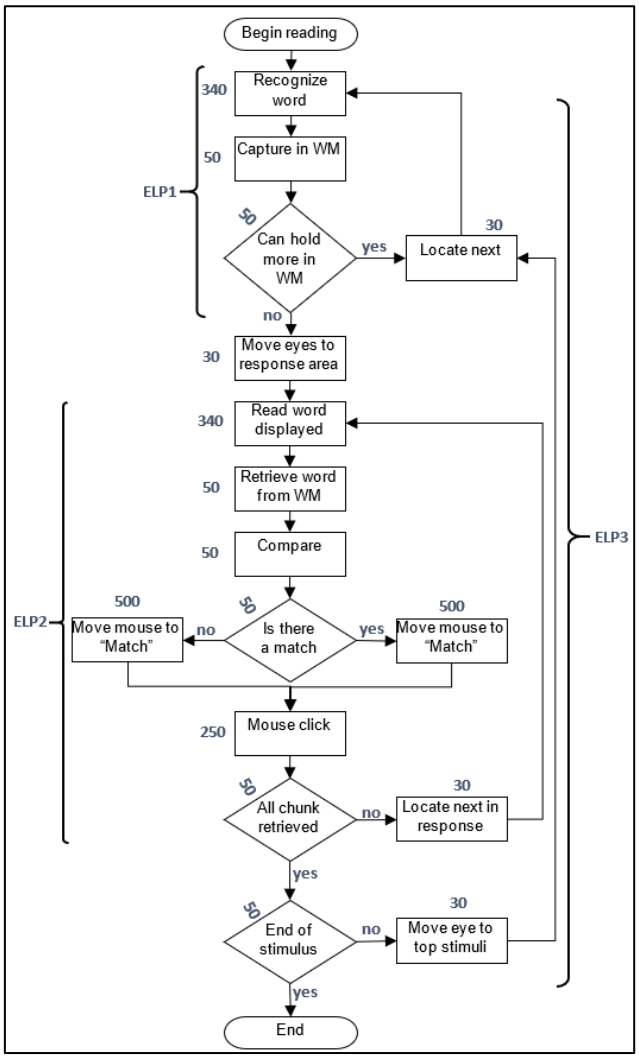

Figure 2: Expert Flowchart

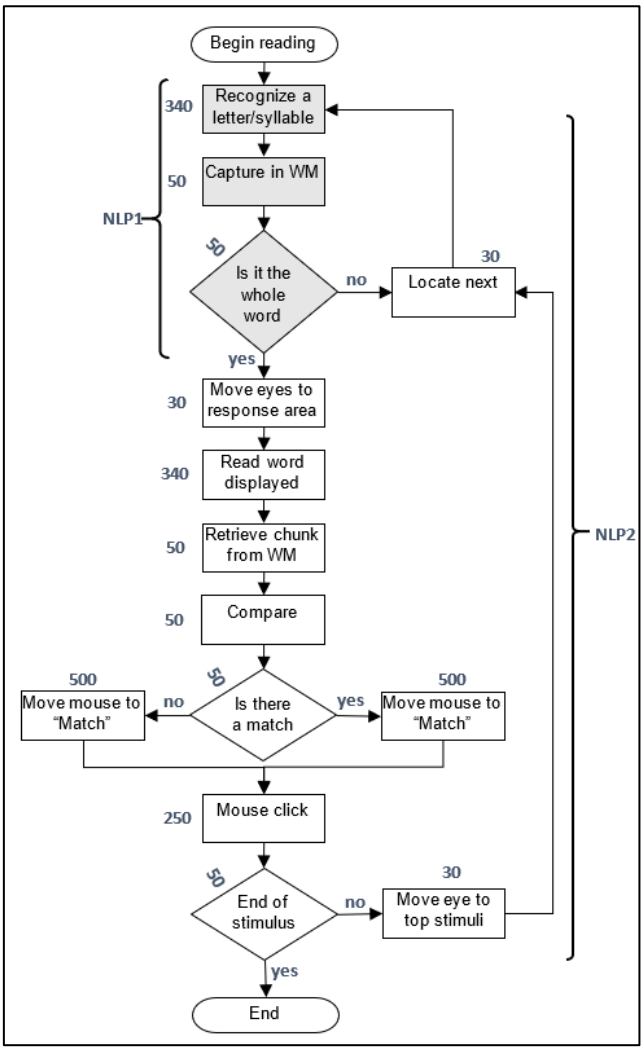

Figure3: Novice Flowchart

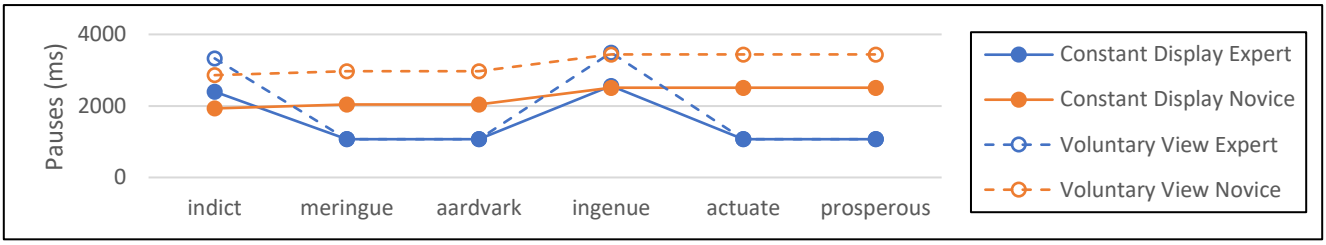

Figure 4: Predicted pattern of pauses for experts and novices in the constant display and voluntary view conditions

and Fig. 3 are high-level representations that conceptualize how an ideal expert and an ideal novice, in the English language, would approach the indicated task. For the purpose of this study, our definition of an expert is someone who possess a vocabulary that enables them to fluently read a piece of text while simultaneously processing its meaning. A novice, on the other hand, is someone with a small vocabulary size, and therefore their reading is much slower as they exert much of their cognitive effort in phonetically processing presented words.

In general, the processing of chunks suggested in both flowcharts act in nested loops. This is similar to Crump and Logan's (2010) inner-outer loop theory of typing, where the outer loop receives words from reading that are then individually passed to an inner loop that translates the word into letters for keystrokes. In our case, there are different loops that work together in a nested fashion for grouping bits of a chunk, decomposing them, transferring them individually to be compared, and then back again to process the next chunk.

Expert Flowchart, Fig. 2: For the sake of analysis the expert is assumed to chunk three words at a time, so they are predicted to have the following pattern of steps:

1. Begin by viewing stimulus, looping three times around ELP1 to create a chunk of three words. By the third loop, the WM is assumed to have reached its capacity and therefore a decision is made to end WM loading. Time elapsed to this point totals 1380ms ($3 \times (340 + 50 + 50) + (2 \times 30)$).
2. The eyes shift to the response area (time duration 30ms). With this movement, the second loop of processes (ELP2) is triggered, which includes reading the word displayed, selecting target word from WM, comparing the words, deciding and finally moving the mouse to click. Accordingly, the step duration is 990ms. The total time, from the start to the first mouse click, the initial pause, is 2400ms.
3. The clicking action of the first word takes 250ms.
4. The expert would then continue to loop through ELP2 to make their second and third response for the words "meringue" and "aardvark" respectively (Fig 4). Pauses for these two responses are both 1070ms each.
5. Once the first three-word chunk is complete, they loop up to the stimulus to gather the next chunk of three words (ELP3). The process of deciding to do this and looking up takes 160ms. This duration is the first part of the pause that precedes the first click in the next group of words.

This analysis is depicted on the solid blue line in Fig 4, which shows pause duration for successive words. The first point is the pause before "indict", comprised of steps 1 and 2. The second and third points are the result of step 4. The fourth point, the pause prior to "ingenue", is comprised of step 5 and 1 again. Hence, experts are expected to exhibit long pauses for grouping words into chunks, with short pauses between responses from within the chunks.

Novice Flowchart, Fig. 3: A novice is assumed to process unfamiliar words by breaking them into parts and then regrouping them to form a chunk. Therefore, for modeling purposes a novice would process a word by the number of syllables it contains. In Fig. 4, the first half of the words are disyllabic while the others are trisyllabic. Hence, a novice's steps for processing are assumed as follows:

1. Begin by looping through NLP1 twice taking 910ms ($2 \times (340 + 50 + 50) + 30$). They then move their eyes to the response area (30ms) to process the presented word and make a move to click (990ms). So, prior to making their first click their total pause would be 1930ms.
2. Next, they click to make a response (250ms).
3. Finally, they would loop up for the next word, NLP2, with the duration for deciding, gazing up and locating the next item is 110ms. This will be calculated as part of the pause that precedes the next response click. These pause durations are represented on the solid orange line in Fig 4. While the first point is comprised of process 1, the rest are composed of processes 1 and 3. The small rise in the duration of the final three points to 3440ms is the result of processing trisyllabic words, where the number of times they loop through NLP1 (in step 1) would increase to three. Accordingly, a novice is predicted to experience long pauses between all clicks, and slightly longer pauses when the number of syllables in a word increases. Overall, the predicted profiles of the expert and novice are substantially different.

Effects of Various Factors

Other factors and their potential effects were analyzed in the same manner. By changing the display of the stimuli from constant display to voluntary view, the stimulus is now concealed and may only be revealed by hovering the mouse over it in the top area. As a result, extra processes are added to the expert's and novice's models for the hover actions. This increases the lengths of long pauses, so further increases the difference in profiles between experts and novices in Fig. 4 for the voluntary view condition, with two of the expert's pauses increasing, whereas all the novice's pauses are higher. The first half of rows in Table 1 summarizes all of the separate pieces of analysis for the

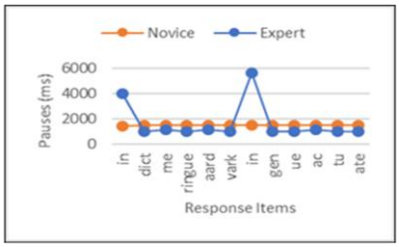

Figure 5: Pause pattern in matching parts of words with parts of words

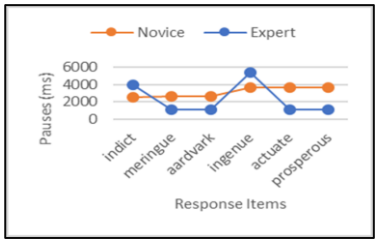

Figure 6: Pause pattern in matching parts of words with words

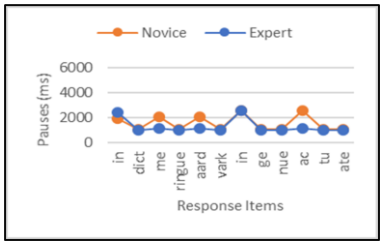

Figure 7: Pause pattern in matching words with parts of words

presentation factor, showing higher discrimination among individuals under the voluntary view mode. The median was chosen to represent the data, however in calculating the mean, a similar pattern of data existed; showing no difference in the overall results.

Models were created to analyze the effect of pairing different types of stimuli with responses, the range of data between the first row and last row of the first half of Table 1 summarizes these modelling results. In addition to matching words with words, we looked at the possibility of pairing parts of word in the stimuli with parts of words in the response (i.e., syllables with syllables). Such presentation alters the expert's model to include two additional loops, one at the start to group syllables into words, and one at the end to decompose the chunked words back to their syllables. This in turn affects the shape of their pause pattern (Fig. 5). A novice on the other hand, is predicted to treat each syllable as a separate chunk, processing each syllable in one large loop causing them to shift their gaze frequently between syllables. Accordingly, their pause pattern is a straight line (Fig. 5).

The other possibility is to pair parts of words in the stimuli with words in the response, for example matching the syllables "in" "dict" with the word "indict". As with the previous task, experts are expected to chunk syllables and form words in their WM and then matching them directly with whole words in the response. The graph, Fig. 6, for this model predicts that an expert's pause pattern would be similar to that found in Fig. 4, however with an increase in the long pauses, in particular, prior to chunking trisyllabic words. If novices were assumed to treat each syllable as a separate chunk, the model predicts that they would be shifting their gaze many times prior to clicking a response causing their overall pause durations to be higher than previously seen (Fig. 6). The difference in pause measures is the highest for this task design (Table 1).

Finally matching words in the stimuli with parts of words (opposed to the above task) was tested. The expert's pattern of pauses is similar to those found in Fig.5 however, with a decrease in the overall duration (Fig.7). On the other hand, a novice's pause pattern differs from those depicted in Figs. 4, 5 and 6 with long pauses prior to matching the first part of a word followed by shorter pauses for each subsequent part of that particular word (Fig. 7). The reason behind the change in pattern is due to the number of loops experienced by the novice. While their processing was always composed of either one or two loops, in this task a third loop appears at the bottom of the model for decomposing the chunk, and comparing parts. This design has the least effect on the pause measures (Table 1).

Evaluating Model Results

To test the model, a pilot study was conducted with two participants. The participants were picked and classified after assessing their vocabulary size using a standard vocabulary size test (Nation & Beglar, 2007), with the high competent (HC) individual scoring at the 16,800-word level and the less competent (LC) at the 8,100-word level.

Based on the predictions in pause measures, the pilot was developed to include four blocks of twelve trials under the conditions of matching word for word and part to word in both constant display and voluntary view. Although, the number of participants was limited, the amount of data was substantial; 48 pause measures were extracted from 384 mouse clicks per individual. The mean of median pauses was calculated for each block separately (Table 1). Overall, findings reveal that patterns between the model and observations are consistent, with the LC experiencing higher pause durations than the HC across all types of tasks. Specifically, out of the 48 trials, only two of the LC trials scored better, i.e., having shorter pauses. It is worth noting however that the value of those measures were small

Table 1: The effects of design variables on pause durations

Model vs. Pilot	Type of Display	S-R Composition	Median		
			Novice	Expert	Differ.
Modelling Results	Constant Display (CD)	Word to word	2275	1070	1205
		Part to word	3175	1120	2055
		Part to part	1520	1020	500
		Word to part	1070	1020	50
	Voluntary View (VV)	Word to word	3205	1070	2135
	Type of Display	S-R composition	Mean of Medians		
			Novice	Expert	Differ.
Pilot Experiment Results	Constant Display (CD)	Word to word	2269	1287	982
		Part to word	3856	2502	1354
	Voluntary View (VV)	Word to word	2116	942	1174
		Part to word	4235	1569	2666

(≈150msc), occurring on items that contained low frequency words and would not be expected to distinguish participants well. Furthermore, confirming our predictions, higher discriminations were observed under voluntary view, especially when combined with part to word tasks.

Discussion

The aim of the present study was twofold. Firstly, to introduce the method of Competence Assessment by Stimulus Matching. CASM attempts to combine the benefits of mouse driven tasks for assessing chunking behavior (c.f., Arslan et al., 2020; Moss et al., 2006) with the benefits of temporal chunk measures for micro-behavior analysis (c.f., Albehaijan & Cheng, 2019; Cheng, 2014; Zulkifli, 2013). In other words, CASM aims to obtain measures of competence based on rich chunk level data at a time scale of 1s with the convenience of standard IT interface devices. From the preliminary results it appears that CASM has potential to compete with the freehand transcription approach and also Cheng's (2015) method that used a mouse and a selection grid. The magnitudes of predicted differences of pauses between the expert and novice are comparable to the magnitudes observed in our pilot as well as the empirical evaluation of those previous approaches.

The second aim was to explore some of the large design space of CASM tasks by using GOMS models to examine the effects of different factors on the processes of chunks. A reason for using GOMS and not a sophisticated cognitive model such as ACT-R (Anderson, 1998), is that we were looking at an efficient method for finding effective designs without all of the detail and effort required to build a full cognitive model. The aim is not to explain in precise detail all of the cognitive steps associated with doing the task, therefore what we needed was an engineering tool and not a scientific one. The produced models provided us with useful guides for designing CASM tasks, as they represent general differences in the processes of an ideal expert and an ideal novice. In between these two models would exist various intermediate levels. Someone who is gradually learning the language may behave according to a mixture of the models. Their decomposition of words may vary depending on their level of familiarity with the words presented, so their looping structure would differ. Variations at the level of individual loop structures would not affect the overall results as these differences would be reflected on the expert's and novice's models, however the number of each type of loop that exist within a model determines the difference.

In using GOMS to analyze the tasks, it was possible to assess chunks in memory and predict pause behaviors. The modelling results show how different patterns of nested loops affect the shape of pause distributions. In the task of matching words with words (Fig 4), an expert's pattern included few long pauses separated by successive short pauses, while novices were shown to have long pauses between clicks. This is explained by how their language knowledge affects the process of chunking. Experts are expected to recognize words in a fluent manner, providing them with the advantage of loading into their memory as many words as possible (see ELP1 in Fig. 2), explaining the few long pauses. The short pauses however, are due to the transfer of words in memory from ELP1 to ELP2. Novices, on the other hand, spend time in processing a word, by breaking it apart into syllables and then regrouping them (see NLP1 in Fig. 3). This lengthy process is expected to load their WM, limiting their ability to hold one word in a chunk and causing frequent gaze shifting between responses. This indicates that behaviors are very much determined by the chunking structure of the participants.

In terms of the design space what task factors are predicted to mostly distinguish between different competence levels? First, the spatial distance between the stimulus and response plays a role in encouraging the use of chunks (Cheng, 2014). If they were close, then experts and novices might rely on quick gazes rather than chunking, causing both to exhibit similar patterns.

Second, for the presentation mode, the analysis showed no effect on the pattern of pauses but a greater difference between pause measures was identified under voluntary view (Table 1). Confirmed by the pilot study, this mode seems potentially more effective than constant display.

Third, with respect to stimulus and response composition, pairing syllables in the stimuli with words in the response seems to be the most effective option. According to GOMS, constructing the stimulus in this way makes it easier for novices to recognize a syllable and move to the response for comparison. However, the complexity of having multi-syllabic words in the response forces novices to shift their gaze as many times as required to have all parts of the word matched. Predictions were confirmed by the results of the pilot study showing longer pauses for novices in these types of tasks, making it seem most effective in exploiting the difference between experts and novices (Table 1).

Fourth, the difference between the model and pilot results are reasonably close, which drives us to conclude that there is potential for such approach. However, one explanation for the absolute difference between the model and each participant being relatively large may be due to variations in strategies within each participant. To control for that, task instructions are being tightened.

GOMS has helped in visualizing the kind of designs most suitable for developing CASM tasks that use temporal chunk measures to assess competency in natural language. We are planning on carrying out further empirical studies.

References

Anderson, J. R. (1998). The atomic components of thought. Mahwah, N.J.: Lawrence Erlbaum Associates.

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

Arslan, B., Jiang, Y., Keehner, M., Gong, T., Katz, I. R., & Yan, F. (2020). The Effect of Drag-and-Drop Item

Features on Test-Taker Performance and Response Strategies. *Educational Measurement: Issues and Practice, 39*(2), 96-106.

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction. Hillsdale*, NJ: Lawrence Erlbaum Associates, Inc

Cheng, P. C-H., & Rojas-Anaya, H. (2007). Measuring Mathematical Formula Writing Competence: An Application of Graphical Protocol Analysis. In *Proc. of the 29th Ann. Conf. of the Cog Sci Soc* (pp.869-874).

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

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

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behav Brain Sci, 24(1), 87-114. doi:10.1017/S0140525X01003922

Crump, M. J. C., & Logan, G. D. (2010). Hierarchical Control and Skilled Typing: Evidence for Word-Level Control over the Execution of Individual Keystrokes. *J Exp Psychol Learn Mem Cogn, 36*(6), 1369-1380.

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

John, B. E., & Newell, A. (1989). Cumulating the science of HCI: from s-R compatibility to transcription typing. In *Proc. of the SIGCHI Confrence on Human Factors in Computing Systems* (pp.109-114).

Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. Psychological review, 101(2), 343-352. doi:10.1037/0033-295X.101.2.343

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*(1), 65-93.

Nation, I.S.P. & Beglar, D. (2007) A vocabulary size test. The Language Teacher, 31(7), 9-13.

Olson, J. R., & Olson, G. M. (1990). The Growth of Cognitive Modeling in Human-Computer Interaction Since GOMS. *Human-comp. interaction*, 5(2-3), 221-265.

Oviatt, S., Hang, K., Zhou, J., Yu, K., & Chen, F. (2018). Dynamic Handwriting Signal Features Predict Domain Expertise. *ACM Trans. Interact. Intell. Syst., 8*(1), 1-21

Schilperoord, J. (2003). On the cognitive status of pauses in discourse production. In T. Olive & C. M. Levy (Eds.), *Contemporary tools and techniques for studying writing* (pp.61-88). Dordrecht: Kluwer.

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

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