A novel interaction for competence assessment using micro-behaviors: Extending CACHET to graphs and charts

Fiorenzo, Colarusso, Department of Informatics, University of Sussex, Brighton, UK, f.colarusso@sussex.ac.uk
Peter C-H., Cheng, Department of Informatics, University of Sussex, Brighton, UK, p.h.cheng@sussex.ac.uk
Grecia, Garcia Garcia, Department of Informatics, University of Sussex, Brighton, UK, gg44@sussex.ac.uk
Aaron, Stockdill, Department of Informatics, University of Sussex, Brighton, UK, a.a.stockdill@sussex.ac.uk
Dr495@cam.ac.uk
Daniel, Raggi, Department of Computer Science and Technology, University of Cambridge, dr495@cam.ac.uk
Mateja, Jamnik, Department of Computer Science and Technology, University of Cambridge, mateja.jamnik@cl.cam.ac.uk

ABSTRACT

Competence Assessment by Chunk Hierarchy Evaluation with Transcription-tasks (CACHET) was proposed by Cheng [14]. It analyses micro-behaviors captured during cycles of stimulus viewing and copying in order to probe chunk structures in memory. This study extends CACHET by applying it to the domain of graphs and charts. Since drawing strategies are diverse, a new interactive stimulus presentation method is introduced: Transcription with Incremental Presentation of the Stimulus (TIPS). TIPS aims to reduce strategy variations that mask the chunking signal by giving users manual element-by-element control over the display of the stimulus. The potential of TIPS, is shown by the analysis of six participants transcriptions of stimuli of different levels of familiarity and complexity that reveal clear signals of chunking. To understand how the chunk size and individual differences drive TIPS measurements, a CPM-GOMS model was constructed to formalize the cognitive process involved in stimulus comprehension and chunk creation.

CCS CONCEPTS

• Human-centered computing; • Human computer interaction (HCI); • HCI theory, concepts and models;

KEYWORDS

Competence assessment, Chunking, Task Analysis, GOMS, Individual Differences

ACM Reference Format:


Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s). CHI ’23, April 23–28, 2023, Hamburg, Germany © 2023 Copyright held by the owner/author(s). ACM ISBN 978-1-4503-9421-5/23/04. https://doi.org/10.1145/3544548.3581519

1 INTRODUCTION

HCI is in a propitious position to contribute to the development of novel learning analytic methods of computer-based assessment because it sits at the intersection of computer science, interaction technologies and cognitive science. As defined by Siemens [57], “Learning Analytics (LA) is the measurement, collection, analysis and reporting of data about learners and their contexts, for the purposes of understanding and optimizing learning and the environment in which it occurs”. The interconnection between HCI and LA is much stronger than one might imagine as an important part of LA research is measuring the individual performance with the learning material. Although keystrokes and mouse actions are still the primary sources of data, handwritten activities play an important role in the learning activities ranging from simple note-taking to the sketching of diagrams representing more complex topics, so significant potential exists for the capture and analysis of handwritten responses in learning analytics. Much research has been conducted on how to analyze data of students’ conventional interactions with software [2]. Theories from cognitive science may be used to refine such approaches by identifying behavioral features that may be particularly strong indicators of learning and to guide the design of interactions to maximize those signals [35].

Evidence supporting the feasibility of using micro-behaviors for competence assessment exists. For instance, Landy & Goldstein [37, 38] found that the distance between characters of handwritten mathematical equations reflects writers’ knowledge of the syntactic structure of those equations. With digital pens, graphics tablets and touch screens it is possible to capture large amounts of low-level user interactions down to the scribing of individual characters. Further, the growth of pen-centric systems [47], and the increasing presence of tablet computers with stylus or finger driven interfaces, further argue for the importance of research into this form of learning analytics. Such research is exemplified by the work of Oviatt et al. [47] and Stahovich & Lin [59] who collected data of pen movements at a millisecond time scale from students as they did problem solving tasks over multiple sessions. They used machine learning techniques to find signals indicative of learning by combining large numbers of micro-behavioral features and obtained good fits between measures based on micro-behavioral signals and independent measures of competence.

Oviatt and colleagues [47] investigated how to classify students as “experts” or “non-experts” using data captured with digital pens
as they worked in small groups on geometry and algebra problems of varying difficulty. From the raw position and time data (logged at a frequency of 37 Hz), stroke distance, stoke duration, pen pressure (force), and pen speed were derived as measures after various stages of filtering. In initial work, experts were clearly differentiated from non-experts by shorter strokes, shorter stroke durations and less pressure, with \( R^2 \) effective sizes up to .27. However, stroke speed did not separate the two categories. Some weak trends within each group were also noted, the most consistent being a decrease in stroke duration and an increase in stroke distance with increasing problem difficulty. In further analysis [47] the individual stokes were manually classified as diagrammatic, numbers, mathematical symbols, or linguistic (words). Then three standard machine learning methods (support vector machine, random forest, and naïve classifiers) were applied to different combinations of the feature sets to examine classification performance. Applying the ML methods to all the stroke measures, along with levels of problem difficulty and stoke types, accuracies in the range 67-71% were obtained. With further partitioning of the data by problem difficulty and stroke type, plus selections of particular stroke combinations, accuracies in the range 79-92% were obtained. However, the strength of the relation between students’ degree of expertise and some measure of competence based on the stroke measures was not reported. The coding of the type of stokes was important to Oviatt et al.’s approach but was completed with substantial manual effort, though the authors suggested the coding could be automated for performance software.

One step in Stahovich and Lin’s [59] approach is precisely to do such automatic coding of the type of strokes. Their dataset of nearly 30 million stokes was captured using digital pens from a total of over 900 students as they variously completed homework or solved examination problems in the domain of statics in engineering. Other steps performed by Stahovich and Lin’s coding tools include, in order, the identification of crossing out, diagram versus equation classification, equation grouping, character grouping, and character recognition. The tools produced a rich set of temporal, spatial, grouping, and crossing out features for use in assessments. Stahovich and Lin [59] showed that high level features are predictive of the correctness of solutions and of student competence, including the numbers of equation groups, alphabetic characters, units of measure, and pauses. Applying a machine learning technique (support vector machine) to these features, correlations of 0.67 with a teacher’s marks on the response to same exam problems were achieved. Of particular interest for considerations below are (a) the definition of the number of pauses as the number of the intercharacter durations that are longer than the median, and (b) the fact that a positive correlation of 0.41 was obtained between the number of pauses and the student exam scores, which indicates that long dwells (greater than the median) between characters is a sign of greater competence.

Although Oviatt et al.’s [47] and Stahovich and Lin’s [59] studies focus on handwriting data, their approach is similar to that of most educational learning analytics and datamining in the area of educational assessment, because they do not satisfy one of the key requirements of Pellegrino, Chudowsky, and Glaser’s [49] assessment triangle for the science and design of educational assessment. Specifically, the design of an assessment should, regardless of its purpose, draw upon a valid model of how students represent knowledge and develop competence in the subject domain.

Competence Assessment by Chunk Hierarchy Evaluation with Transcription-tasks, CACHET, is an interactive task for the assessment of student competence proposed by Cheng [14]. A chain of ideas underpin CACHET:

- First, learning about a target domain involves the accumulation of information in memory.
- During learning the amount of information in memory not only increases but gradually becomes ever more coherently structured in memory. The greater your learning the better the stored information is organized.
- Performing tasks in some domain invokes moment to moment mental processes to retrieve relevant information from memory and to manipulate the information in ways that meet the task goals.
- The quality of the structure of the information in memory affects the efficiency of that short timescale processing, which will be manifest in micro-behavioral signals observable in task performance.
- Those signals provide an opportunity to assess an individual’s knowledge or competence in a target domain by using the signals to probe the nature of structure of the information in memory and hence quantify how well the individual understands concepts in the domain.

In summary, it is theoretically feasible to design interactive computer-based methods to assess the competence of learners by analyzing patterns in their micro-behaviors as they do tasks in a target domain.

Studies on CACHET have investigated the use of various types of interactive tasks and interfaces, such as pens on graphics tablets and selections from a grid of symbols [11]. Variants of transcription tasks have been examined, such as: constantly available stimulus [14]; voluntary stimulus presentation under user control [1], or response verification with sequential multiple-choice items [28]. Evaluations of CACHET have shown good correlations between independent measures of competence and micro-behavioral chunk-based measures on a range of domains, including: algebra [11, 14], programming [1], and English [28].

This paper both challenges and extends the work of Cheng & colleagues on CACHET. In particular, previous incarnations of CACHET were tested on domains with linear notations, but here this paper extends CACHET to work in domains whose representations are graphical and two dimensional. The extension requires the introduction of a new technique which we call Transcription with Incremental Presentation of the Stimulus, TIPS. To motivate and contextualize TIPS, three questions, at increasing levels of specificity, are posed.

1.1 Overarching question: Why can micro-behaviors be used to assess competence, and how?

To answer the question of why micro-behavior approaches to competence assessment is even feasible we need a theoretical understanding of how the storage of information in human memory is structured. Cognitive science provides appropriate theories. One is
the working memory chunking theory [17, 42] that claims that the mind processes information as *chunks*, small clusters, where intra-chunk concepts are strongly associated but sparsely connected with inter-chunk concepts. Another is the long term memory chunking theory [10, 23, 33, 58] which asserts that knowledge of a domain is stored as hierarchies of chunks and during learning the contents of chunks grows. Experts’ superior abilities are explained by their possession of large hierarchical sub-chunks that enables them to recognize large meaningful patterns in stimuli, whereas novices must rely on simple intrinsic patterns (e.g., visual Gestalts) or rehearsed elements to hold information in working memory. These differences in memory structures produce a distinct temporal chunking signal in the form of variable duration pauses between actions. Pauses can distinguish experts from novices as they do tasks, such as reproducing the position of chess pieces from a briefly viewed board [9] or writing memorized sentences [15]. Experts have long bursts of activity that generate many elements with a long inter-chunk pause at the start and many short intra-chunk pauses in between. In contrast, novices exhibit a lower frequency of short intra-chunk pauses and reproduce short sequences of only a few elements [9, 22] and the duration of long inter-chunk may also be greater [15]. Thus, there is a robust temporal chunking signal that can be exploited.

A critical issue for the use of micro-behavioral chunking signals is the problem of differences in the specific strategies that people use to do a task. Such individual strategic differences will mask the chunking signal by introducing long duration pauses associated with reasoning that do not directly reflect the structure of chunks in memory. In CACHET tests pauses are negatively correlated with competence, but in Stahovich and Lin’s [59] method that used problem solving tasks the correlations were positive, which shows quite different processes are at work. This is why CACHET uses transcription tasks, because they dramatically reduce the strategic variability among individuals, while providing a task environment that depends on the processing of chunks. One reason that CACHET assessments require far shorter sampling periods and far less data than Ovitt et al.’s [47] and Stahovich & Lin’s [59] machine learning approaches is because the transcription tasks are much less noisy in terms of task strategy variability.

The purpose of this paper is to introduce a technique for learning analytics and competence assessment with spatial stimuli and 2D data visualizations extending the CACHET areas of application.

### 1.2 Intermediate question: Can CACHET be used in the domain of graphs and charts?

So far, CACHET has been evaluated on domains that use linear notations, with some success [1, 11, 14, 28]. Can micro-behavioral assessment of competence, like CACHET, be used to assess individuals’ level of experience with graphs and charts? Consider first what is known about understanding graphs and charts, and also existing approaches to assessing competence with them.

Graph comprehension has been defined as a set of skills which permit the effective reading and interpreting of graphs [53]. Graph comprehension can be studied through the lens of cognitive models which explain well the processing structures of graphs and charts. Kintsch’s [34] Constructional-Integration (CI) model of text comprehension was reconsidered and applied by Freedman and Shah [21] to explain graph comprehension. According to this model, the comprehension takes place in two phases: (i) a construction phase and (ii) a comprehension phase. Moreover, three pools of processing units were included in the model: visual features, domain knowledge, and interpretation proposition. During the construction phase, prior knowledge and goals interact to facilitate the chunking of the elements in the graph and an automatic activation of perceptual features guides the processing of the data depicted by the graph. During the integration phase, the available knowledge is combined to form a coherent mental representation, so when the information depicted by the graph can be linked to the prior knowledge the comprehension is effortless, otherwise when more inferences are needed the process becomes effortful. Likewise, the Model of Display Comprehension [26] explains how people construct a mental representation from a visualization, emphasizing the interaction between bottom-up information (design features) and top-down processes (prior knowledge). Thus, familiarity and background knowledge may influence the way attention is directed to the external display and how information is perceived, interpreted, and modelled internally [36].

One of the aims of this paper is to include in the CACHET catalogue of available approaches to competence assessment a method for assessment with diagrammatic representations and data visualizations. Common approaches to measure familiarity with data visualizations include: (A) multiple choice instruments and questionnaires to measure (i) prior experience using graphs, (ii) graph reading abilities, (iii) typical graph reactions [63]; (B) tests of skills such as (i) data reading, (ii) data comparisons and (iii) inference generation from data [20]. Moreover, other common methods to assess graph comprehension and graph interpretation abilities are written descriptions [8] and oral descriptions [54]. These studies demonstrated how the graph complexity affects the study time before writing description [8] and the contribution of the prior knowledge in generating main inferences and a better understanding from a graph [54]. Another approach is the card sorting task employed by Cox and Grawenmayer [18] to assess how people organize their knowledge of external representations (ERs). They found higher semantic distinctions and accurate naming of ERs with fewer categories in ER card-sorting tasks produced by high competence participants reflecting their better mental representations of ER knowledge since they perceived the semantic commonality between visually different ERs.

Cognitive science has studied chunking with spatial and diagrammatic representations. The classic work on chunking in expertise established chunking phenomena in the perception of chess boards [9], so findings about chunking are applicable to 2D representations. Studies with electronic circuit diagrams [19] show the basic theoretical claims applied to symbolic representations. Obaidullah & Cheng [46] show that chunking of geometrical diagrams is hierarchical and that the duration of pauses in tracing, transcription or drawing from memory reflects chunk structure. Roller & Cheng’s [52] study provides evidence that participants with different chunk structures exhibit different distinct drawing strategies and that the duration of pauses reflects processing levels within chunk hierarchies.
Thus, it seems feasible that CACHET may be applicable to competence assessment of spatial representations. Specifically, this paper presents a foundational study of CACHET to establish that chunking behaviour is manifest in a particular transcription technique for graphs and charts, which we propose in the next subsection.

1.3 Specific question: what form should interactive tasks for CACHET with graphs and charts take? TIPS.

In a previous experiment conducted by Colarusso and colleagues [16] the typical approach of CACHET was used in which the whole stimulus was accessed by the participants at will during transcription. Although some evidence of chunking was apparent, it was also clear that there were substantial strategic differences in the way that the participants approached the task of copying the stimuli. Those differences appeared to be interfering substantially with the chunk signal. Such strategic differences do not occur with linear notations as participants without exception reproduce them in the order they are written (left-right, top-bottom). With the diagrammatic representations the participants did not only choose different starting places but also drew sub-components of chunks in different orders.

To address this limitation with the Colarusso and colleagues [16] task, we have designed a novel interactive task that attempts to limit the undesirable individual differences in drawing strategies. In contrast to previous stimulus presentation interaction tasks, in which the whole stimulus was available each time in the user viewed it, the new technique requires the user to incrementally reveal the elements of the stimulus. Transcription with Incremental Presentation of the Stimulus, TIPS, is summarized in Figure 1. The new technique has been designed around two phases: a Familiarization Phase and a Visualization Phase. The targets for transcription are the data curves in the graph, with the axes system given as a background framework. The TIPS task starts with a short stimulus presentation after which a blank mask appears. The purpose of the Familiarization Phase is to allow the user to form an overall mental image of the stimulus. The presentation duration is set depending on the stimulus complexity (e.g., 3 or 5 seconds). The main part of TIPS concerns the Visualization Phase where the user has alternating view and draw phases. To enter the view phase a specific View-Key is pressed (right arrow key on the board), which shows the first element that is to be transcribed. Successive presses of the View-Key at the discretion of the user reveal further elements. Then the user starts the copying phase pressing the Draw-Key (down arrow key), which masks the display whilst they draw. The instructions to the users include: (i) reproducing faithfully the stimuli and the meaning represented by the shapes trying not to omit anything during the copying; (ii) copying the stimulus sequentially starting from the information displayed by the very first View-Key to the last; (iii) during each view phase display only as many elements as they believe they can comfortably copy during the following episode of drawing.

TIPS provides several measures that are relevant to chunk-based competence assessment. (a) The total viewing episode size corresponding to the number of viewing actions per view episode (number of consecutive View-Key presses before pressing the Draw-Key), which is the amount of information the user feels comfortable copying. (b) time of drawing episodes (duration between a Draw-Key press and the next View-Key press). (c) The viewing time, which is the cumulative duration of all View-Pauses and Draw-Pauses in an episode. The View-Pause corresponds to the latency before pressing the next View-Key within the same viewing episode while the Draw-Pause is the latency before starting the copying by pressing the Draw-Key.

1.4 Hypotheses, experiment, and cognitive task analysis

This paper presents a preliminary study of the efficacy of TIPS that focuses on whether a robust chunking signal is manifest. The study includes an experiment and cognitive tasks analysis modelling. To frame the investigation, four hypotheses are proposed built on the claims that since expertise increases the content of each chunk and WM capacity is essentially constant, the TIPS measures relating to chunk size should be sensitive to chunking effects. To manipulate chunk size, stimuli that will have different degrees of familiarity to the participants are used. We specifically sample individuals with similar levels of competence in order to minimize individual differences in their expertise. There are four hypotheses for the experiments.

H1) Viewing episode size: the amount of information displayed per viewing episode will be larger for familiar stimuli.

H2) Drawing episode time: the duration of drawing episodes will be longer for familiar stimuli.

H3) Viewing episode time: the durations of view episodes will be longer for familiar stimuli.

H4) Pause duration: View Pauses will be shorter for familiar stimuli than Draw Pauses.

All four hypotheses expect that their measure of competence will be positively correlated with the familiarity of the stimuli, because more familiar stimuli will be encoded with larger chunks. The viewing time per episode, the amount of information displayed per viewing episode, and the duration and drawing episodes may obviously seem to be closely related: the more elements in a viewing episode will require more time to review and more time to draw. However, they are not redundant, because different cognitive processes are involved, perception for viewing and production for drawing.

The purpose of the cognitive task analysis modelling is to provide a deeper understanding of the cognitive processes involved in TIPS, which will aid the interpretation of the experiment and inform refinements to TIPS. The modelling is presented after the experiment, which follows in the next section.

2 EXPERIMENT

2.1 Participants

Six right-hand participants (PhD student = 4; Faculty Staff = 2) from a University Department of Computing were recruited. Before joining in the lab, they completed an online survey platform the Xi’s [63] graph familiarity questionnaire redesigned with a Likert 7-points scale (the original version has a 6-points scale). As required, all the participants shared approximately the same level
A novel interaction for competence assessment using micro-behaviors: Extending CACHET to graphs and charts

2.2 Materials

For this experiment we designed six stimuli of different complexity (Figure 2) to reveal the contribution of the prior knowledge in chunking operations. Specifically, we introduced two representations that differed dramatically in familiarity, a drawing of a house and the Rey Figure, an unusual geometric figure used as a neuropsychological assessment to evaluate different cognitive functions such as visuospatial abilities, and visuo-spatial working memory. Four main stimuli of different complexity were also designed. For the simple main stimuli we used a line graph representing two upward and downward trends and a bar chart representing a normal distribution, whereas the complex main stimuli increased the total number of trends, data points and bars. The complex line graph represented two bell curves, two sigmoid functions and one irregular trend while the complex bar chart included several distributions such as bimodal, negative skewness, normal and uniform.

A laptop computer running InteractLog, a logging program written in our lab, combined with a Wacom Cintiq 16 tablet and a keyboard was used. InteractLog allows us to compute several potential measures of competence such as: (1) Viewing episode size, (2) Viewing episode time, (3) Drawing episode time. Moreover, it recorded the latency between successive displayed parts (View Pause) and before starting the copying (Draw Pause).

2.3 Procedure

Before starting the task, our participants received training through 3 practice stimuli to get familiar with the interactive procedure to display and draw the stimulus (Figure 1). Once the task started, the house and then Rey Figure were submitted first to each participant whereas the data visualizations stimuli were randomized. Each stimulus differed in the amount of information displayed by pressing the View-Key. For the House, the amount of information displayed by one View-Key was one simple element at a time starting from the outline going through the sub-patterns (e.g. outline → roof → door → single window). Likewise, the order of presentation for the parts constituting the Rey Figure, was designed following the Obaidellah & Cheng [46] results where the participants tendency during the drawing was to complete the outline of the figure first, then the internal elements and finally the outer elements. As concern the data visualizations stimuli, one bar at time was displayed for the bar charts by pressing the View-Key while one dot and one connecting line were displayed together for
the line graphs as shown in Figure 1. TIPS measures of chunking were computed from the logs of each participant on each stimulus recording, including: (a) the median of the Viewing episode size (View-Key presses between pairs of Draw-Key presses), (b) the median of drawing episode time (duration between a Draw-Key press and the next View-Key press); and, (c) the median of viewing episode time for each participant on each stimulus. Moreover, to better understand the user interactions with TIPS, the median for View Pause and Draw Pause was also recorded for each participant on each stimulus.

2.4 Results

Our first hypothesis (H1) concerning the Viewing episode size is addressed in Figure 3A where the differences about the total amount of information displayed per viewing episode across stimuli has been shown. How the prior knowledge affects the creation of bigger chunks is particularly evident in Figure 3A where the median of viewing episode sizes are all similar (3 chunks) except the Rey Figure (1.5 chunks). Since the unfamiliar Rey Figure was clearly transcribed using the smallest number of chunks, nobody had familiarity with it and therefore it is the stimulus with the smallest chunks displayed. For the familiar stimuli, the chunk size is in line with the Cowan’s [17] typical chunk size of approximately 4 elements for complex tasks.

The chunking measures drawing episode time (Figure 3B) and viewing episode time (Figure 3C) revealed how familiar the participants were with the given stimuli addressing our second (H2) and third (H3) hypotheses. The results plotted in Figure 3B and 3C, are consistent with the claim that the level of familiarity with the given representations affects the chunking. Indeed, comparing the median values of the drawing episode time and viewing episode time they are greater for the house and the data visualizations stimuli and lower with the Rey Figure. Specifically, the median values for drawing episode time (Figure 3B) show a bigger chunk size for the house and data visualizations reflected by a longer drawing time, within boundaries of 10s to 20s, compared to the Rey Figure whose median drawing episode time was less than 10s. The viewing episode time also revealed the chunking behavior between the stimuli. Thus, the longer viewing episode time for the House and the data visualization stimuli (Figure 3C), reflects the larger amount of information displayed before starting the drawing episode by pressing the Draw-Key (Figure 3A). Likewise, the smaller chunk size for the Rey Figure is also reflected by the smallest visualization time compared to the other stimuli. Except one participant who was able to reveal and copy the whole stimulus following the linear copying strategy (this happened for the House, Rey Figure and Simple Line Graph), all the other participants interacted with TIPS as expected chunking the stimulus during the Visualization phase and following linear copying strategies.

For the pause analysis relating to our fourth hypothesis (H4) we compare the performance of our 6 participants on each stimulus in relation to the measures of View Pause and Draw Pause. From the participant’s distribution on each stimulus, the median value was compared. The proportion of participants whose Draw Pause was longer than the View Pause on each stimulus are: House: 3/6; Rey Figure: 2/6; Simple Line Graph: 6/6; Simple Bar Chart: 4/6; Complex Line Graph: 5/6; Complex Bar Chart: 6/6. Moreover, to have a better overview about the difference between the Draw Pause and View Pause over the stimuli, the participants medians for each pause type were plotted (Figure 4). Looking at the medians in the graph, the View Pause duration results are shorter than the Draw Pause.
A novel interaction for competence assessment using micro-behaviors: Extending CACHET to graphs and charts

Figure 3: Differences for the chunk size over the stimuli are evident through the number of Viewing Episode Size (A) the Drawing Time (B) and (C) Viewing Time. In figure (A) and (B) outliers beyond the range of the axis are noted above the box plots.

except for the Rey Figure where the Draw Pause median is longer compared to the View Pause. In order to run a statistical analysis, a total of 231 Draw Pauses and 618 View Pauses coming from all the participants’ interactions has been collected and analyzed using Python and all the statistical tests run through the Pingouin library [62]. Before running a three-way ANOVA, the Shapiro-Wilk test was used to check the assumption of normality for the dependent variable (Pause Duration) and each independent variable (Stimulus, Pause Type and Participant). Since the data wasn’t normally distributed, a log transformation was applied and the Shapiro-Wilk test performed again. The test results show that each level of the independent variables presents mainly a normal distribution with some exceptions. Listed here the proportion of normal distributions for each independent variable: Participant 5/6; Pause type 1/2; Stimulus 4/6.

Once confirmed the normal distribution for our data, a three-way ANOVA (Table 1) was run using the transformed data to compare the main effect of Participants, Stimuli and Pause Type (independent variables) and their interaction effects on the Pause Duration (dependent variable). Participants included six levels (from P1 to P6), Stimuli six levels (the stimuli used for this experiment) and Pause type consisted of two levels (Draw Pause and View Pause). All the three main effects (Stimulus, Participants and Pause Type) and the two-way interactions (Stimulus * Participants; Stimulus * Pause Type; Pause Type * Participants) were statistically significant at the .05 significance level. However, the interaction effect for the three factors was not significant (Table 1).

2.5 Discussion Copying Task

TIPS addresses the limitations found in a previous work conducted by Colarusso and colleagues [16] where the chunking behavior in copying data visualizations was evident, through the number of views on the stimulus and the length of copying time [1], but affected by the diversity of drawing strategies employed by participants. Since complex diagrams do not follow a linear format, such as mathematical formulas [13] and strings of code [1], their copying can be carried out using different approaches [52]. We overcome this weakness by making the display procedure interactive and instructing our participants to copy following the stimulus presentation (i.e., copying the display in sequence). This design makes the copying of the stimulus more linear and suitable to reveal chunking operations both during the visualization phase and copying phase. An advantage provided by TIPS is the recording of the amount of information constituting a chunk before the copying production, which allows us to measure the amount of information an individual can process in WM for drawing. Hence, the introduction of stimuli of different familiarity, enables the comparison of chunking across stimuli, revealing how the Rey Figure (i.e., the most unfamiliar stimulus) was copied by producing smaller chunks than the house and the data visualizations and how this behavior is particularly
Figure 4: Distribution for the View Pause and Draw Pause in general (top figure) and over the stimuli (bottom figure) aggregating the participants medians for each pause. For all the stimuli, the median for the View Pause is shorter than the Draw Pause except for the Rey Figure.

Table 1: Three-way ANOVA table

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sums of Squares</th>
<th>df</th>
<th>Mean Sq.</th>
<th>F</th>
<th>p-value</th>
<th>Partial Eta Squared (η²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stimulus</td>
<td>7.50</td>
<td>5</td>
<td>1.50</td>
<td>3.38</td>
<td>0.005</td>
<td>0.021</td>
</tr>
<tr>
<td>Pause type</td>
<td>13.11</td>
<td>1</td>
<td>13.11</td>
<td>29.55</td>
<td>0.000</td>
<td>0.037</td>
</tr>
<tr>
<td>Participant</td>
<td>38.65</td>
<td>5</td>
<td>7.73</td>
<td>17.43</td>
<td>0.000</td>
<td>0.101</td>
</tr>
<tr>
<td>Stimulus * Pause type</td>
<td>7.44</td>
<td>5</td>
<td>1.49</td>
<td>3.36</td>
<td>0.005</td>
<td>0.021</td>
</tr>
<tr>
<td>Stimulus * Participant</td>
<td>30.59</td>
<td>25</td>
<td>1.22</td>
<td>2.76</td>
<td>0.000</td>
<td>0.082</td>
</tr>
<tr>
<td>Pause type * Participant</td>
<td>10.19</td>
<td>5</td>
<td>2.04</td>
<td>4.59</td>
<td>0.000</td>
<td>0.029</td>
</tr>
<tr>
<td>Stimulus * Pause type * Participant</td>
<td>14.16</td>
<td>25</td>
<td>0.57</td>
<td>1.27</td>
<td>0.165</td>
<td>0.039</td>
</tr>
<tr>
<td>Residual</td>
<td>344.65</td>
<td>777</td>
<td>0.44</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

evident to through all the measure of chunking introduced in this task.

Chunking operations can be also studied through the lens of temporal measures which reveal chunk boundaries [9, 12, 13, 15, 16, 19, 46, 51, 58]. Indeed, when performing a sequence of motor acts (i.e., motor chunk), the first action of the sequence tends to be slower compared to the subsequence intra-chunk motor actions due to the extra cognitive resources to retrieve and plan the motor execution of the whole chunk. Applying the principle of temporal measures on the visualization phase of this task we obtained a reversed effect of the chunk boundaries evident through the differences in the View Pause and Draw Pause latency. Indeed, since the stimulus was displayed part by part through the View Key (this happened after the short familiarization of 3/5 seconds at the beginning of the task), the View Pause was shorter due to recognition processes associated with the displayed sub-parts. In contrast, the Draw Pause latency was longer because the participants had to refine their mental representation creating a chunk they were able to draw from the viewing episode size. However, this was not true for the Rey Figure where the View Pause latency was longer, reflecting effortful recognition and comprehension resulting in smaller chunks displayed before copying (Figure 3C). Moreover, from the ANOVA results (Table 1), the statistically significant main effects and the 2-way interaction effects revealed that both individual factors, such as (i) individual differences between our participants, (ii) level of familiarity with each stimulus and (iii) the different nature of View Pause (recognize and comprehend the displayed part/sub-chunk) and Draw Pause (integrating the information displayed and creating a chunk before copying), and their interactions explained the differences in the pause durations.

All these results together are in line with Kris and Hegarty [36] which claimed how the interaction between prior knowledge and bottom-up features presented by the stimulus affects the perceptual processing and how chunks are made. Therefore, recognition processes should be evident through shorter View Pause, as happened for the House and the target stimuli whereas the chunk boundaries
can be marked using the Draw Pause. However, when there is no prior knowledge to support the perceptual processing, as the case of the Rey Figure, the View Pause increases considerably, becoming longer than Draw Pause.

Although TIPS provided good results to study the chunking behavior, an evident limitation for this study is surely the small sample size. Hence, further investigations are needed to test TIPS on a bigger sample and with groups of different familiarity with the learning material to study individual differences in chunking operations.

3 COGNITIVE TASK ANALYSIS

This part of the paper introduces an implemented version of Cognitive, Perceptual, and Motor GOMS (CPM-GOMS) model for task analysis [30] to understand and explain the participants behaviors reflected through the View-Pauses and Draw-Pauses during the Visualization Phase (Figure 1). The TIPS design to display the stimulus was internally guided by the participants who had the total freedom to (i) see each displayed part for as long as they wanted and (ii) set up the amount of information they were able to process. We defined the View Pause as the duration of the latency between two consecutive viewing episodes by pressing the View Key while the Draw Pause is the time frame used by a participant to create a chunk and decide to start the copying phase after pressing the View Key several times in a row. As pointed out by Kriz and Hegarty [36], the comprehension outcome strongly depends upon the interaction between the prior knowledge and the bottom-up features depicted in the representation so that the higher familiarity with the given representation should be reflected through better information processing and the creation of bigger chunks. Therefore, the second part of this paper aims to model the cognition behind the creation of chunks (Figure 3A) using loops of cognitive processes which explain and predict the View Pause and Draw Pause duration. Moreover, introducing this implemented CPM-GOMS we aim to complete and extend the results obtained in the first part of this work addressing our four main hypotheses.

GOMS is a family of modeling techniques used to predict, model and analyze the user performance with interactive systems [31]. Despite existing GOMS models of different complexity and granularity (e.g., KLM, CMN-GOMS, NGOMSL, CPM-GOMS), all the variants share the same hierarchical decomposition structure based on Goals, Operators, Methods, and Selection Rules (GOMS) [6]. Given a task, Goals are simply what the user is trying to accomplish, and they are usually divided into sub-goals. Operators are atomic elements with a fixed execution time representing the motor actions or cognitive processes required to accomplish a specific task. Methods are procedures describing how a specific task is accomplished by the user. However, when more than one method is available to achieve a given goal, Selection Rules represented in the form of conditional statement (if-else) are required.

As claimed by Hochstein [27], although CPM-GOMS can model a task in great detail, it presents several limitations: (i) predictions are only valid for expert user ignoring users who are learning and need to comprehend different aspects of the system or interface, (ii) only goal-directed tasks can be modelled neglecting the problem-solving nature of some tasks and (iii) individual differences between users are not considered. Hence, the challenge is to implement the CPM-GOMS approach making it suitable to predict and model problem-solving situations introducing reasoning operators in the form of decision boxes and loops of behaviors which better represent the cognitive steps to (i) recognize and comprehend the displayed parts from each stimulus by pressing the View Key (View Pause) and create chunks of information before making the decision of copying (Draw Pause).

3.1 Understanding GOMS and CPM-GOMS

We built our cognitive model using a GOMS approach for task analysis. GOMS models are a core part of HCI methods, and they can be classified as predictive, descriptive, prescriptive [27]. They are predictive because they predict the user performance to complete the task with millisecond accuracy estimating the time taken by each operator involved in the model. GOMS models are descriptive because through goals, subgoals, methods and selection rules they clearly represent the way a user is interacting with the system. Moreover, they are prescriptive because they can serve as a guide for developing training programs and help systems since they can be used to teach new users how to achieve the goal.

The Cognitive, Perceptual, Motor GOMS (CPM-GOMS) is the most appropriate GOMS technique to model our task because (1) it employs the Model Human Processor (MHP) [6], a framework which represents the cognitive, perceptual, and motor operators needed to accomplish the task, and (2) it can deal with the execution of these operators in parallel. In the MHP three separate but related components work together to efficiently achieve the task: (i) perceptual processor is responsible to convert the perceived information into a form that can be processed by the cognitive system; the cognitive processor works jointly with the motor system scheduling and planning actions using contents from working memory (WM) and long-term memory (LTM); and the motor processor is responsible to translate thoughts into actions.

3.2 CPM-GOMS modelling

To better understand and match the operators described in this section with the corresponding positions in the model (Figure 5), we will insert next to each operator the corresponding reference in the figure using parentheses [].

We designed our model starting from the end of the short familiarization time with the whole stimulus (3 or 5 seconds based on the complexity) when the blank mask was on [], after that the participants had the press the first View Key [7] to display the first part of the stimulus (Figure 1). The designed model has one main GOAL, one Sub-Goal and two Sub-Sub-Goals associated with the loops generated by the reasoning operators (Figure 5). Since the visualization phase is strongly related to the copying phase, the main GOAL is copying the stimulus by setting the appropriate chunk size for each copying phase by pressing the View Key multiple consecutive times and then the Draw Key to start the copying. The Sub-Goal is using the View Key to display the next part of the stimulus (sub-chunk) after comprehending the previous one. Moreover, two Sub-Sub-Goals of equal importance are allocated to (i) comprehend the visualized part before displaying the next
and (ii) refine the mental representation before starting the copying phase.

Although we introduced some non-standard CPM-GOMS operators to better represent the cognitive processes involved in the task (encoding, construction, comprehend, updating and reasoning operators), most of them and their corresponding values come from the literature [6, 24, 31, 32, 50]. After displaying each part of the viewing episode size by pressing the View Key [7], an eye-movement [10] of 30 ms is needed to perceive the displayed information through a perceptual operator [11] of 290 ms [30]. Once perceived the information, a non-standard CPM-GOMS encoding operator [12] of 100 ms has been introduced to convert the perceived information into symbolic form capable of being processed by the WM [3, 6]. This operator has been designed following the Tulving and Thompson [61] Encoding Specificity Principle, according to "specific encoding operations performed on what is perceived determine what is stored, and what is stored determines what retrieval cues are effective in providing access to what is stored". After the encoding, since GOMS models are not used to model comprehension tasks and spatial coordinates are prominent elements in diagrammatic representations, two consecutive cognitive operators of 50 ms have been introduced to verify information [13] and location [14]. Since graph schemata (i.e., viewers prior knowledge about the graphs) mediate the transition of visual features into a mental representation and they are responsible to facilitate chunking operations [21, 49], a non-standard cognitive Construction Operator [15] of 50 ms has been introduced to represent the graph schemata activation. Afterwards, a cognitive Comprehend Operator [16] of 100 ms is needed to finally construct the internal representation of the viewing episode size [3]. Moreover, due to the complexity and the interactive nature of the visualization task, it likely involves executive functions (EFs) [43, 44, 48]. EFs make possible formulating goals and carrying out plans to effectively accomplish them [41], they differ from cognitive functions (CFs) as they explain how and whether a person goes about doing something, rather than what they differ from cognitive functions (CFs) as they explain how and whether a person goes about doing something, rather than what the user perceives. The Updating Operator [17] of 100 ms [16] reflects the Updating EF presented by Miyake and colleagues [43, 44, 48] which is required to actively monitoring the incoming information and WM representation, replacing old not relevant information with newer more relevant information [45]. Hence, this operator is particularly useful to revise the WM representation every time a new part of the stimulus is displayed and the decision of starting to copy is made.

Since some decision-making actions are needed to achieve the final goal, we designed selection rules using loops in the form of decision boxes representing reasoning operators (i.e., choices by the participant whose outcome directs the flowchart to start to copy or to loop the behavior repeating the flow of operators in each loop). In the designed CPM-GOMS model, three reasoning operators [18, 19, 20] have been created to represent the decision processes occurring during the task. The first: “Did I understand what I displayed?” [18] occurs once each part of the stimulus is displayed and processed; its negative outcome loops the participant to process the displayed part again going through the Comprehension Loop reflecting the Sub→Sub→Goal "comprehend the stimulus". Otherwise, when the comprehension of the visualized part is successful, the participant is engaged with the second reasoning operator: “Can I visualize more?” [19]. If more information can be displayed and processed, the View Loop must be run in order to prepare the motor action for the next View Key press. Otherwise, when the limit of WM capacity is reached and no more information can be displayed, the participant deals with the third reasoning operator: “Can I copy what I have displayed?” [20]. If the amount of information displayed can be integrated into a coherent mental representation the final GOAL of copying can be accomplished, else if the created mental representation lacks some information, the participants must go through the Refresh WM Loop to accomplish the Sub→Sub→Goal refining the stimulus mental representation.

3.3 Modelling results

3.3.1 CPM-GOMS model results. The introduced loops in our model may be helpful to model the different duration of View Pause and Draw Pause and explain the cognitive steps in the creation of chunks. We decided to model the median values (Figure 4) of each pause on the stimuli in order to see how well the model predicts the time before the keypress actions. The comparison between the data produced by the model and the median values has been plotted and compared using the mean absolute percentage error (MAPE) (Figure 6). Following the Lewis scale [40] for the MAPE interpretation, the results from our model revealed a good forecasting in predicting the View Pause duration with a MAPE of 10.42 % while a highly accurate forecasting was found for the Draw Pause with a MAPE of 9.34 %. Moreover, the model considers how many loops are required before pressing each View Key and Draw Key. The generated values from our model predicted 2 loops in Comprehension Loop before pressing the next View Key for the House and all the data visualization stimuli except the Simple Bar Chart and the Rey Figure which required 3 and 4 loops respectively. Otherwise, before starting the copying phase, 3 loops in the Refresh WM Loop were required for the House, Rey Figure, Complex Line Graph and Simple Bar Chart, while 4 loops for the Simple Line Graph and Complex Bar Chart were needed. Overall, for the house and the data graphs the Draw Pause required more loops than the View Pause, while an opposite effect was found for the Rey Figure which require a higher number of loops for the View Pause. The only exception is for the Simple Bar Chart where the same number of loops where required both for the View Pause and the Draw Pause.

3.3.2 Modelling Discussion. The introduced GOMS model aimed to overcome several limitations of general CPM-GOMS models. First, since we employed stimuli of different complexity and familiarity, we demonstrated how our model is able to predict the variability of the users’ interactions across the stimuli although we collected only expert users. Second, the problem-solving nature of tasks is usually neglected by CPM-GOMS, hence introducing reasoning operators and loops we made CPM-GOMS suitable to deal with common reasoning operations involved in a complex task (i.e., comprehension tasks) without neglecting the modelling of task’s goals. Third, since CPM-GOMS doesn’t consider individual differences, we modelled the performance on different stimuli comparing the performance between them.

Furthermore, the recorded pauses and the CPM-GOMS model can be used to reveal internally and externally guided sequencing [4]. Externally guided sequencing is contingent on externally
Figure 5: Representation of our CPM-GOMS model for graph comprehension. Decision Boxes and loops have been used to reflect the GOALS hierarchy. The Comprehension Loop, (from the Attend Information operator to the first decision box) is responsible for the comprehension and the creation of WM representation. The View Loop serves to display the next part of the stimulus whereas the Refresh WM loop is required to refine the WM representation before starting the copying. Numbers on the right of each operator have been introduced as reference.

specified visual stimuli whereas the internally guided sequencing depends on the ability to program a sequence of future actions requiring a greater demand for memory and planning. Therefore, recognition and comprehension operations should be externally guided by bottom-up features presented by the stimulus reflecting automatic chunking processes [23] while the creation of chunks and mental representation should be internally guided and more demanding in terms of cognitive resources since it is goal driven for deliberate chunking [23]. On a high-level analysis, the difference between internally and externally guided sequencing should be evident just comparing the latency of View Pause and Draw Pause. Thus, using CPM-GOMS we can also justify the extra cognitive resources required to press the Draw Key since three reasoning operators and more loops to create the to be copied chunk are needed.

The observed differences in terms of loops between the View Pause and the Draw Pause over the stimuli can also provide additional support to the hypotheses addressed in the first part of this work. Indeed, the small number of loops concerning the View Pause for the house and the data visualizations reflect the faster recognition and comprehension processes associated with each part displayed within the viewing episode size. Thus, these results can be used to complete the explanation of our three main hypotheses where familiar stimuli have larger viewing episodes size, longer viewing time and drawing episode time supported by faster recognition and comprehension processes.

Moreover, according to the Lewis scale [40], our MAPE values indicates a reliable forecasting level for our model in predicting individual differences in comprehending diagrammatic representations and revealing chunk boundaries.

4 DISCUSSION

The paper proposes TIPS, a novel method for competence assessment of diagrammatic representations which aims to overcome the limitations and weaknesses found in previous work [16]. TIPS aims to be an innovative and reliable method for competence assessment of diagrammatic representations which goes beyond questionnaires, verbal descriptions, written descriptions or simple task which only measures competence indirectly [8, 18, 54, 63]. Indeed, TIPS implements and extends the available measures to study chunking operations, introducing in addition to drawing episode time and number of viewing episodes [1], additional measures such as the viewing episode time and the viewing episode size. Moreover, its design might have potential to assess competence and to study chunking even in stimuli which don’t have a linear format and can
Figure 6: Comparisons between the pauses predicted by our model using loops and the median value for each stimulus (Figure 6). Smaller number of loops for the View Pause was required by the house and data graphs than the Ray Figure. For the Draw Pause, three loops are needed by the House, Rey Figure, Complex Line Graph, Simple Bar Chart, while four for the Simple Line Graph and Simple Bar Chart. Overall, the View Pause required a smaller number of loops than the Draw Pause, except for the Rey Figure.

be copied following different strategies. The TIPS design has been tested through an interactive drawing task positively addressing four main hypotheses which aimed to explain the chunking behavior during the task through: (H1) the viewing episode size, (H2) drawing episode time, (H3) viewing episode time and (H4) different duration of View Pause and Draw Pause.

Another important feature implemented in our approach has been the recording of pauses during the visualization phase. Temporal measures are usually employed during the recalling phase to inspect chunk structures after participants have created a mental representation. To the best of our knowledge, there are no studies which use temporal measures to map the ongoing comprehension and reveal the creation of chunk structures during the stimulus visualization. Therefore, the View Pause reveals comprehension and recognition processes associated with the displayed sub-chunk while the Draw Pause latency is helpful to mark the creation of the chunks and chunk boundaries.

Thus, TIPS provides an extended catalogue of measures with which to record and analyze micro-behaviors to reveal the potential level of competence with the given representation. Furthermore, since diagrammatic representations, charts and graphs have a 2D nature, TIPS amplifies the chunking signals by standardizing the drawing strategies and eliminating the noise produced by alternative copying strategies. The positive results obtained addressing our four main hypotheses contribute to validate the new measures of competence making TIPS suitable to study how familiar users are with representations. Hence, given its strengths, TIPS can be potentially applied in educational settings to assess students’ familiarity with specific topics represented through visual materials such as data visualizations making it a tool for learning analytics. The learning analytics and HCI community may benefit from applying TIPS in competence assessment because it combines and analyzes in one task two key sources of data in learning analytics: keystrokes and pen-strokes. Since chunks underpin our knowledge of concepts, TIPS’s design reveals how familiar a particular user is, recording several measures of chunk size during the Visualization phase (keystrokes data) and the Copying phase (pen-strokes data). Moreover, others potential indicators of the competence may be the View Pause and Draw Pause as they reflect the latency to recognize
each new displayed part and to create chunks in memory before copying.

However, some of the limitations from the TIPS approach don’t make it easy to use. First, participants need training to get familiar with the visualization and drawing procedure. Second, the high length of time to complete the task.

The CPM-GOMS model has been designed to (i) overcome several limitations in the CPM-GOMS modelling [27], (ii) explain and addressed hypotheses using TIPS. To do that, reasoning operators and loops have been introduced to deal with problem-solving and decision-making processes required by the task. Moreover, the obtained MAPE values quantified the model’s reliability in predict and modelling the behavior behind the collected pauses.

ACKNOWLEDGMENTS

This work was supported by the EPSRC grants R/030642/1, EP/T019603/1, EP/T019034/1 and EP/R030650/1

HISTORY DATES

Received September 2022; accepted January 2023

REFERENCES


