Decoding Expertise: Exploring Cognitive Micro-Behavioural Measurements for Graph Comprehension

Fiorenzo Colarusso[1], Peter C-H. Cheng[1], Ronald Grau[1], Grecia Garcia Garcia[1], Daniel Raggi[2], Mateja Jamnik[2]

[1]Department of Informatics, University of Sussex, Brighton, UK
[2]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK

{f.colarusso, p.c.h.cheng, r.r.grau, gg44}@sussex.ac.uk
{daniel.raggi, mateja.jamnik}@cl.cam.ac.uk

Abstract

Transcription with Incremental Presentation of the Stimulus (TIPS) is a novel approach relying on micro-behaviours proposed by Colarusso and colleagues (2023) to study users' cognition with data visualizations. The study in this paper has two primary objectives: (a) investigate whether TIPS can measure an individual's competence with data visualizations; and (b) explore the potential enhancement of TIPS measures by normalizing them with the individual's performance on tests of visuo-spatial abilities and memory capacity. We test 30 participants with different expertise and cognitive skills. Results reveal that TIPS provides some promise for individual competence assessment, but only when normalized with the individual's performance on a test of rigid transformation of mental images. Other tests measuring visuospatial abilities or memory capacity did not produce effective normalizations.

Keywords: Competence Assessment, Individual Differences, Spatial Cognition, Learning Analytics, Visualization Competence, Graph Comprehension.

Introduction

Visualization literacy is distinguished from visualization competence. While visualization literacy entails the ability and skills to read, interpret and extract information from data visualizations (Lee et al., 2017), visualization competence refers to the cognitive processes and practices needed to adequately use and fluently transform visual display in the form of tables, diagrams and graphs (Gilbert, 2005; Gilbert et al., 2008). Four major components constitute visualization competence: constructing (i.e., generating a set of external representations to form graphs or diagrams), interpreting (i.e., understanding the meaning), transforming (i.e., creating new external visualizations to replace or complete the original), and critiquing (i.e., evaluating external representations) (Chang & Tzeng, 2018). This paper concerns graph comprehension, an important aspect of visualization competence. As defined by Fox (2023), graph comprehension is the process of deriving meaningful insights from graphs, a task deeply anchored in visuospatial processing that develops through a blend of guided learning and practice.

As confirmed by previous studies (Gilbert, 2005; Hinze et al., 2013; Nitz et al., 2014) users' prior knowledge plays a crucial role in visualization competence. Similarly, graph comprehension theories (Freedman & Shah, 2002; Hegarty, 2005; Kriz & Hegarty, 2007) underscore that the interaction between prior knowledge and graphical features is vital for effective chunking operations, enabling a proficient comprehension of the visual display. However, they did not investigate the role of spatial processing in graph comprehension (Ratwani et al., 2008).

Previous research has shown the importance of spatial transformation processes in graph comprehension and in predicting user performance with information visualization. This has been demonstrated through various methods including: verbal protocols (Trafton & Trickett, 2001; Trickett & Trafton, 2004; Trickett & Trafton, 2006) and spatial transformation tests such as the Mental Rotation and Paper Folding (Luo, 2019; Stewart et al., 2008; Tandon et al., 2023; VanderPlas & Hofmann, 2016; Vicente et al., 1987). Spatial transformations are cognitive operations executed on both external (i.e., visualization) and internal representation in order to improve the comprehension of graphical displays through manipulations (i.e., adding or removing features) and comparison of representations (Stewart et al., 2008; Tandon et al., 2023; Trafton et al., 2002; Trafton et al., 2005; Trickett & Trafton, 2006). These operations are essential in graph comprehension tasks as they facilitate the construction, interpretation, transformation, and critique of external representations.

There is a limited number of methods to assess graph comprehension, and these can be categorized into four groups: (a) self-assessment (Xi, 2016), (b) multiple choice questions (MCQ) (Cui et al., 2023; Feeney et al., 2000; Fox, 2019; Ge et al., 2023; Lee et al., 2017), (c) written descriptions (Carswell et al., 1993; Golparvar & Azizsahra, 2023) and (d) oral descriptions (Shah & Freedman, 2011; Xi, 2016). However, these methods can be time-consuming because many MCQs are needed or need manual grading by an instructor.

An alternative approach, *Transcription with Incremental Presentation of the Stimulus* (TIPS) (Colarusso et al., 2023) was recently introduced to measure graph comprehension. TIPS quantify the users' familiarity with a given visualization by recording and analysing keystrokes and pen-strokes data during cycles of viewing and drawing, providing a direct and quantitative approach for competence assessment.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Since copying requires the construction of a visual representation through spatial processing, as well as the hierarchical organization of motor chunks for execution (Sommers, 1989), TIPS could be a suitable tool to measure graph comprehension.

Visualization competence using micro-behaviours

TIPS is a member of the *Competence Assessment by Chunk Hierarchy Evaluation with Transcription-tasks* (CACHET) family of methods that were designed to assess competence of technical skills (Cheng, 2014; Colarusso et al., 2023). CACHET and TIPS can assess a users' competence by focusing on micro-behavioural signals, that is interactions occurring in the range between 100ms to 2s duration.

Previous CACHET research evaluated competence on *linear notations* (e.g., strings of code, mathematical equations, and formulas). Various designs were investigated, including: constantly available stimulus presentation (Cheng, 2014); voluntary stimulus presentation controlled by the user (Albehaijan & Cheng, 2019); or response verification with sequential multiple-choice items (Ismail & Cheng, 2021). The first attempt to apply CACHET to *graphical material* was conducted by Colarusso and colleagues (2021) revealing chunking operations in copying a line graph and a bar chart. However, a noisy micro-behavioural signal was reported due to the different drawing strategies used by the participants. These limitations have been addressed by introducing TIPS (Colarusso et al., 2023) and allowing a dynamic stimulus presentation under the user control to amplify the chunking signal while limiting individual differences in drawing strategies. The results highlighted the effectiveness of TIPS to study the users' familiarity with a given visualization.

The TIPS design (Figure 1) has three main phases: Familiarization, Visualization, and Copying. It starts with a brief familiarization of the stimulus (Figure 1, step [1]), to form an overall mental image. Afterwards, a blank mask is shown in step [2]. Next, the users' task is to display the amount of information that they are able to copy in one burst by pressing the View-Key (i.e., Right arrow key) consecutively (steps [3-5]). Subsequently, they choose to start copying (step [6]) the displayed information by pressing the Draw-Key (i.e., Down arrow key). Once they finish copying, they can view more of the stimulus (step [7]), repeating the process until the whole stimulus has been copied. Colarusso and colleagues (2023) demonstrated the strengths of this design by applying a Cognitive, Perceptual and Motor GOMS (CPM-GOMS) model (John & Kieras, 1996) to reveal the cognitive processes occurring during the pauses in the Visualization Phase (Figure 1, steps [3-5]). Shorter View Pauses (steps [3-4]) reflecting quicker recognition processes for familiar stimuli were found, while longer Draw Pauses (steps [5-6]) were required to integrate the displayed items before copying.

What is the theory underpinning the design of CACHET and TIPS? As people's expertise increases, they are able to process more information and organize it more efficiently in their memory. Hence, micro-behaviours recorded during the task can provide insights about (i) how much information a user is able to process, and (ii) the structure of the information stored in memory. Several theories from Cognitive Science support this approach. First, according to the working memory chunking theory (Cowan, 2001; Miller, 1956), the information is processed by the mind as *chunks,* clusters of information where intra-chunk elements have stronger semantic associations between them than elements belonging to different chunks. Furthermore, according to the long-term memory chunking theory (William G. Chase & Herbert A. Simon, 1973; Gobet et al., 2001; Johnson, 1970), learning increases the hierarchical chunks of information stored in memory. Hence, expertise in a particular domain can be explained by different hierarchical sub-chunk sizes where experts have bigger sub-chunks that enable to recognize meaningful patterns more quickly when compared to novices, whose smaller sub-chunks rely on perceptual features. Hence, the hierarchical organization of chunks can be studied using micro-behavioural signals in the form of pauses between actions (W. G. Chase & H. A. Simon, 1973; Cheng & van Genuchten, 2018; Cheng & Rojas-Anaya, 2007; Egan & Schwartz, 1979; Obaidellah & Cheng, 2015; Roller & Cheng, 2014; Thompson et al., 2017). Experts have bigger chunk sizes showing few long inter-chunk pauses and many short intra-chunk pauses. In contrast, due to the smaller amount of information recalled within each chunk, novices exhibit more and short intra-chunk pauses (W. G. Chase & H. A. Simon, 1973; Gobet & Simon, 1996) with many inter-chunk pauses longer than those of the experts (Cheng & van Genuchten, 2018). Previous evaluations of CACHET have produced correlations in the range of 0.6-0.7 between independent measure of competence and the micro-behavioural chunk-based measures in various domains such mathematics (Cheng, 2014, 2015; Cheng & Rojas-Anaya, 2007), programming (Albehaijan & Cheng, 2019), and English (Cheng & Zulkifli, 2009; Ismail & Cheng, 2021).

Those theories inspired the design of two categories of competence measures within CACHET and TIPS: (i) chunk size and (ii) temporal micro-behaviours (Figure 1D). A descriptive list of each measure is provided.

(1) *Viewing Episode Size*: the number of consecutive View-Key presses for each Visualization phase (Figure 1, steps [3-5]). It may reflect chunking as it represents the amount of information a user is able to display before starting the Copying Phase.

(2) *Viewing Time*: the duration of each Visualization Phase (Figure 1B). Its duration is influenced by the number of elements that are encoded.

(3) *Number of Drawing Episodes*: the total count of Copying phases for each stimulus (Figure 1, step [6]). It may reflect chunking being inversely proportional to the amount of information that is processed. A larger amount of chunk content suggests a smaller Number of Drawing Episodes to copy the stimulus.

(4) *Drawing Time*: the duration of the Copying phase (Figure 1, step [6]). This reflects chunking as a larger chunk content implies a longer drawing time.

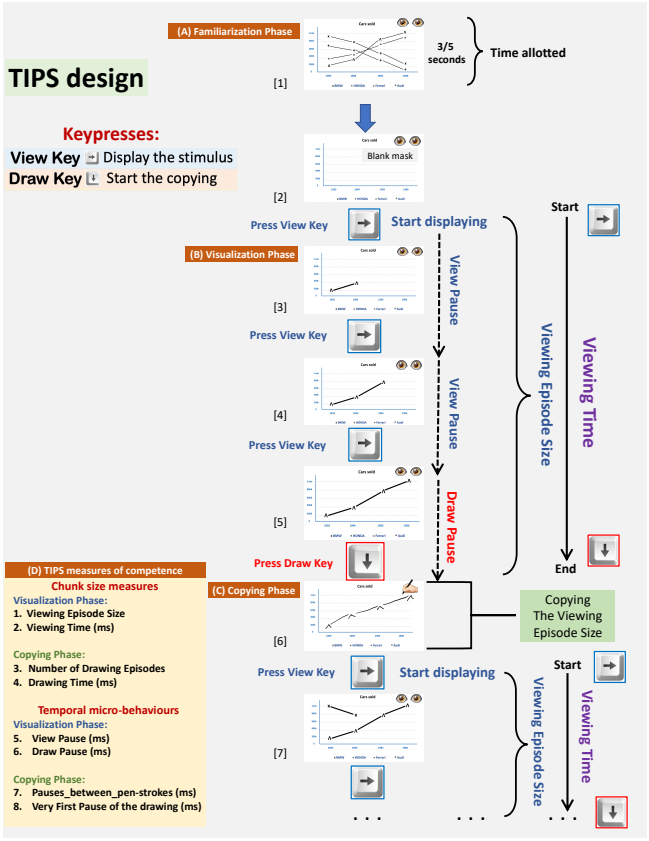

Figure 1: The TIPS method applied to a simple line graph. TIPS is an interactive task, based on cycles of stimulus viewing (Panel B) and copying (Panel C). The users choose how much of the stimulus to reveal incrementally in the Visualization Phase by pressing the View Key multiple times. Afterwards, they press the Draw Key to start copying the remembered elements. A list containing the competence measures recorded by TIPS is provided (Panel D).

(5) *View Pause*: the duration before pressing the next View-Key within the same Viewing Episode (Figure 1, steps [3-4]). A shorter View Pause indicates quicker recognition processes when the user is familiar with the displayed information before pressing the next View-Key.

(6) *Draw Pause*: the duration of time before starting the Copying phase by pressing the Draw Key (Figure 1, steps [5-6]). A longer Draw Pause reflects the integration processes to combine the to-be-copied chunk contents into a new representation.

(7) *Pauses_between_pen-strokes*: the duration of the pauses among all the pen-strokes performed to draw each stimulus. It may reflect the participants hierarchical organization of chunks.

(8) *Very first pause of the drawing*: the time between the Draw-key press and the very first pen-stroke of each copying phase after viewing the stimulus. This represents the cognitive effort required to recall the to-be-copied chunk contents.

Given this wide set of measures designed to decode different facets of expertise, the first set of research questions motivating the experiment is:

RQ1) *Can TIPS be used to measure graph comprehension? What are the TIPS measures that best correlate with users' experience in using graphs?*

Refining competence via Normalization

Alongside prior knowledge and expertise, visuo-spatial abilities also play an important role in graph comprehension tasks (Stewart et al., 2008; Tandon et al., 2023; Trickett & Trafton, 2006). Hence, they may mask the true relationship between the participants' expertise and TIPS measures, compromising our RQ1. Several psychometric tests measuring visuo-spatial abilities were included in this experiment with the aim to improve the accuracy of any correlations between our independent measure of competence and TIPS' dependent measures of competence. By normalizing, we aim to increase the strength of the graph comprehension signal by controlling the influence of potential masking variables that may affect the direct relationship between our variables of interest.

In previous CACHET experiments (Cheng, 2014, 2015), dependent measures of competence were normalized in order to improve the accuracy of temporal chunk signals during the copying, by removing the impact of individual differences due to underpinning skills. Specifically, the third quartile (Q3) of the pauses distribution during the copying of basic mathematics skill was subtracted from the Q3 of the pauses distribution derived by the copying of complex equations. This subtraction resulted in a new normalized value that better correlated with the independent measures of competence. Hence, there is a second research question to be investigated:

RQ2) *Does normalizing TIPS measures of competence with visuo-spatial abilities measures enhance the accuracy of the baseline graph comprehension signal?*

Experiment

Design

The experiment has a within-participants design with 4 stimuli featuring line graphs and bar charts of different complexity already used in previous TIPS pilot work (Colarusso et al., 2023). The line graph in Figure 1 is an example. The other three stimuli are: (i) a bar chat with an inverted U-shaped distribution; (ii) a line graph with two bell curves, two sigmoid functions and one irregular trend; (iii) a bar chart showing separate distributions such as bimodal, negative skewness, normal, and uniform. The self-assessment questionnaire already used by Colarusso et al. (2023) serves as the independent measure of competence. It features a 7-point scale comprising three sub-scales: (i) *experience using graph* (GE), (ii) *graph reading ability* (GRA) and the (iii) *typical reactions to graphs* (GR). TIPS measures will be the dependent variables (Figure 1D). Visuo-

spatial ability tests are used for the normalization to address potential masking variables.

Participants

The experiment was conducted with 30 university students at University of Sussex. All of them received a £15 Amazon voucher as compensation for their time. We recruited participants with a wide range of experience with graphs by targeting students from STEM and NON-STEM departments. Ethics clearance was obtained by the appropriate committee of the University. An a priori power analysis was conducted using G*Power version 3.1.9.6 (Faul et al., 2007) to determine the minimum sample size required to test the research questions. Results indicated the required sample size to achieve 80% power for detecting a large effect (i.e., 0.5), at a significance criterion of $\alpha = .05$, was $N = 29$ for a two-tailed correlation. Thus, the collected sample size of $N = 30$ is adequate to study the research questions.

Questionnaire and Cognitive Assessment

Similar to previous CACHET research (Albehaijan & Cheng, 2019), the experiment began with an online questionnaire on Qualtrics where the participants provided demographic information and self-assessed themselves on our independent measure of competence. Afterwards, they completed the Mental Rotation test (MRT) (Shepard & Metzler, 1971; Vandenberg & Kuse, 1978) and Paper Folding test (PFT) (Ekstrom, 1976), each comprising 20 items.

Next, the participants completed two additional visuo-spatial working memory tests in our lab using the Corsi Block Tapping Test (Corsi, 1972) and the Visual Pattern Test (Della Sala et al., 1999). Both tests were administered digitally on an iPad Pro. The Corsi Block Tapping Test was administered in its forward (CF) and backward (CB) version using Psytoolkit (Stoet, 2010, 2017). This spatial-working memory test records the participants' *spatial span* by requiring them to tap sequences of blocks of increasing complexity, from a 3-block series to a 9-block sequence. Visual working memory was measured on a different online platform (Castro-Alonso et al., 2018) using the Visual Pattern Test (VPT). VPT records the participants percentage of correctness in memorizing and recalling visual patterns of different complexity ranging from a 2x2 grid to 6x5 grid. These tests were also used for the normalization of TIPS measures.

Material and TIPS procedure

Once the cognitive assessment in the lab was completed, participants started the TIPS copying task on a laptop running InteractLog version 2.3[1] together with a Cintiq 16 Wacom tablet. Prior training with example stimuli was provided in order to familiarize participants with the TIPS procedure. Stimuli used for the copying task were randomized and interaction logs from each participant were collected. These interactions were used to compute the eight TIPS competence measures, as listed in Figure 1D. For each measure, the median value was recorded, except for the Number of Drawing Episodes, where the total was calculated.

Results

Graph Familiarity Questionnaire

The correlations between the three sub-scales were calculated to determine the most appropriate independent measure of competence. All correlations were strong and statistically significant (d.f.= 28, two-tailed, r_{crit}= 0.570, p<.001). Specifically, GE showed a correlation of r= 0.65 with GRA and r= 0.70 with GR. Additionally, a correlation of r= 0.69 was found between GRA and GR. Given these strong and significant correlations, and as chunking involves efficient organization and processing of information based on prior knowledge and experience, GE alone was chosen as the most appropriate sub-scale to be the independent measure of competence.

RQ1: Correlations among GE and TIPS measures

Our first research question (RQ1) examined the feasibility of using TIPS as a tool for measuring graph comprehension.

The initial baseline correlations between GE and TIPS measure of competence were investigated for each stimulus (Figure 2, top bar (brown) in each group). To facilitate the interpretation, all correlations are shown as absolute values (GE is negatively correlated with *Number of Drawing Episodes*, *View Pause*, *Pauses_between_pen_strokes*, and *Very First Pause of the drawing*). This is of no consequence for RQ1. Overall, weak baseline correlations were found between GE and TIPS measures of competence on each stimulus with only 05/32 correlations being significant. The *Number of Drawing Episodes* and the *Viewing Episode Size* were the only TIPS measures that were significant. Specifically, the *Number of Drawing Episodes* had significant correlations, on the simple bar chart r(28) =. 40, p < .05, complex bar chart r(28) =. 42, p<.05 and complex line graph r(28) = .43, p < .05; all two tailed. Significant correlations for the *Viewing Episode Size* were found on the simple bar chart r(28) = .46, p < .05, and complex line graph r(28) = .39, p<.05. Thus, the reported ratio of 05/32 significant baseline correlations showed that TIPS works poorly to reveal the participants' graph comprehension with the give visualizations since the only two measures with significant correlations are the *Number of Drawing Episodes* and *Viewing Episode Size,* and only on specific stimuli.

RQ2: Normalization with cognitive measures

RQ2 concerns the investigation of normalization methods, using visuo-spatial ability tests, to improve the baseline graph comprehension signal. Normalized values are adjusted values that eliminate the effect of potential masking variables in the data, which allows a more accurate interpretation. In our case, we want to make TIPS measures regardless of visuo-spatial

[1] https://github.com/rrgrau/InteractLog.

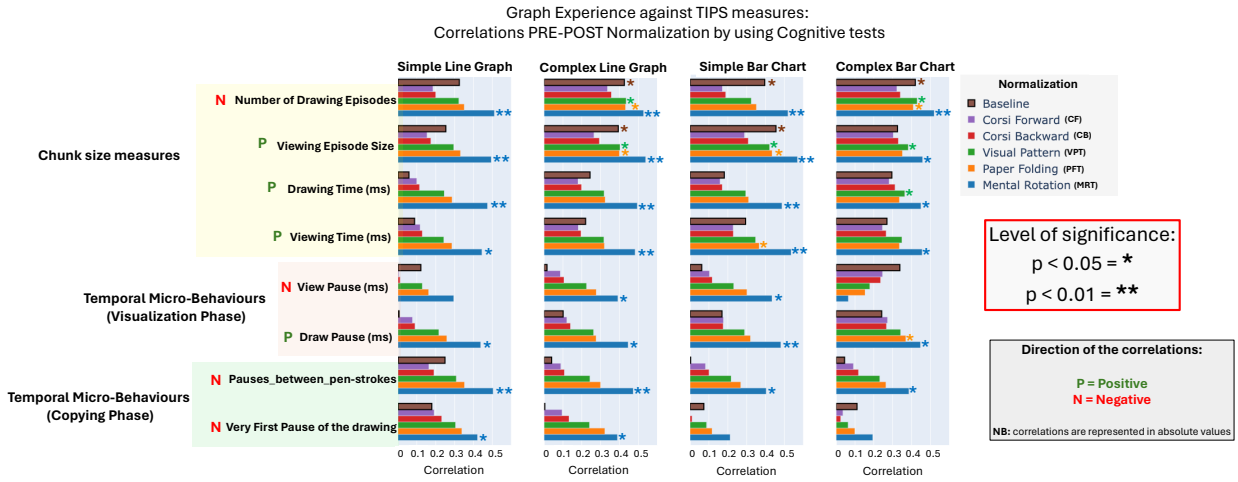

Figure 2: Correlations between Graph Experience (GE) and TIPS measures on each stimulus across different normalization approaches. The results are plotted using the absolute values of the correlations. The best improvement from the baseline correlations (brown bars) was obtained by normalizing with MRT (blue bars). All TIPS chunk size measures improved significantly across the stimuli by normalizing with MRT. However, some temporal micro-behaviours were not significant (i.e., *View Pause*: simple line graph, complex bar chart; *Very First Pause of the drawing*: simple and complex bar chart).

ability in order to isolate the correlations between GE and TIPS measures. This investigation is carried out in three steps examining: (i) individual differences between High and Low Competence participants on cognitive tests, (ii) the relationship between the GE and cognitive abilities, (iii) the normalization of TIPS measures by removing the influence of visuo-spatial abilities.

Differences between groups on cognitive tests. The dataset was divided applying the median split method based on GE, our independent variable. Participants with a score above the median were labelled as High Competence (HC) (N= 14) whereas those with a score below as Low Competence (LC) (N= 16). Normality assumptions were met for MRT, PFT and VPT ($p>0.05$) but not for CF ($p<0.01$) and CB ($p<0.001$). Hence, the non-parametric Mann-Whitney U test was used (instead of t-test) to compare the two groups. Superior performance of HC group was found on all the cognitive tests. Comparing the two groups on MRT, HC scored higher ($M = 8.36$, $SD = 3.56$) than LC ($M = 5.88$, $SD = 2.94$) showing a statistically significance difference ($t = -2.06$, $p < 0.05$, Cohen's $d = 0.76$). Differences on the border of significance were found for PFT ($t = -2.02$, $p = 0.05$, Cohen's $d = 0.73$) and VPT ($t = -2.04$, $p = 0.05$, Cohen's $d = 0.73$), with HC outperforming LC (PFT: $M = 12.86$, $SD = 3.21$ vs. $M = 10.31$, $SD = 3.68$; VPT: $M = 84.86$, $SD= 6.13$ vs. $M= 79.06$, $SD = 9.28$). No statistically significant difference was found between the two groups on CF (U-value = 87, $p = 0.29$, HC median = 6.5, LC median = 5.5) and CB (U-value = 88, $p = 0.30$, HC median = 5.5, LC median = 5.0).

In summary, the superior performance of HC on all the cognitive tests, while borderline or not significant in some cases (i.e., PFT, VPT, CF, CB), provided a first indication about the role of visuo-spatial abilities as masking variables.

Correlation between GE and cognitive tests. To continue the detection of potential masking variables, the correlations between GE and visuo-spatial ability tests were also investigated. A moderate and significant correlation was found between GE and MRT $r(28) = .49$, $p<.01$, indicating a meaningful relationship between graph experience and performance on MRT. However, weaker and non-significant correlations were reported between GE and the PFT $r(28) = .30$, $p>.05$, VPT $r(28) = .26$, $p>.05$, CF $r(28) = .09$, $p>.05$, and CB $r(28) =.11$, $p>.05$. Thus, considering both results provided by the comparisons and correlation analysis, the normalization of TIPS measures by using the cognitive measures is warranted.

Correlations post normalizations using cognitive tests. We now directly address RQ2 regarding the impact of normalizing TIPS measures of competence with visuo-spatial ability tests (i.e., MRT, PFT, VPT, CF, and CB) to enhance their initial baseline correlations with GE. All the variables were standardized using their *z-scores*. TIPS measures positively correlated with GE (i.e., *Viewing Episode Size, Drawing time, Viewing time, Draw Pause*), were normalized by summating them with each cognitive measure, while negatively correlated measures (i.e., *Number of Drawing Episodes*, *Pauses_between_pen strokes, View Pause, Very First Pause of the drawing*) were adjusted by subtraction.

Correlations pre and post normalization are presented in Figure 2. Alongside the proportion of 05/32 significant correlations between GE and the TIPS measures (brown bar on each stimulus), the correlations adjusted to account for the individual differences in visuo-spatial abilities are presented in different colors (Figure 2). The proportion of significant correlations is given for each cognitive measure. The best improvement from the baseline correlations was obtained

when normalizing with MRT (28/32) (blue bars). Minor improvements were reported for PFT (06/32) (orange bars) and VPT (06/32) (green bars), while no notable changes occurred for CF (0/32) (purple bars) and CB (0/32) (red bars).

A repeated measures ANOVA was applied to investigate whether there are significant differences in scores pre versus post normalization, as well as to examine variations between TIPS measures of competence too. A within-subject factor included baseline and post-normalization correlations values using MRT, PFT, VPT, CF, and CB. Additionally, a between-subject factor incorporated TIPS measures of competence. Assumption checks identified a violation of the sphericity assumption, addressed with Greenhouse-Geisser correction ($\varepsilon = 0.291$). The analysis revealed significant main effects of correlations pre-post normalization and TIPS measures ($p<.001$), without an interaction effect. Post hoc tests using Bonferroni correction indicated significant differences between baseline correlations and post-normalization correlations using MRT, PFT, and VPT ($p<.001$). Notably, no significant difference was found for CF ($p = 0.256$) and CB ($p = 1.000$). The most substantial improvement was observed for MRT with a mean difference from the baseline of -0.235. Smaller improvements were observed for PFT (-0.107) and VPT (-0.082), while no improvement was found for CF (0.033) and CB (0.017). These results emphasize the MRT normalization's effectiveness in refining the graph comprehension signal compared to other cognitive measures. Thus, when evaluating graph comprehension with TIPS, accounting for both individual differences in prior knowledge and spatial processing is essential for an accurate results interpretation.

DISCUSSION

This paper evaluated and extended TIPS (Colarusso et al., 2023), a computerized learning analytics tool for assessing graph comprehension. Two research questions were posed.

RQ1 aimed to determine if TIPS can measure graph comprehension and find out which of the TIPS measures would be most effective for this (Figure 1D). When applied to *linear notations,* the typical range of correlations in CACHET was 0.6-0.7. Hence, the findings for RQ1 yielded limited support for TIPS as graph comprehension measurement tool, with only 05/32 within a range of 0.4-0.5 being significant across the four data graphs stimuli.

As TIPS targets visualizations and given that spatial processing is important for graph comprehension (Stewart et al., 2008; Tandon et al., 2023; Trickett & Trafton, 2006) and copying tasks (Sommers, 1989), it is possible that individual differences in visuo-spatial processing may have influenced the chunking signal upon which the TIPS measures rely. RQ2 aimed to enhance the baseline graph comprehension signal by normalizing with visuo-spatial abilities tests. Normalization with MRT revealed a substantial improvement from the baseline with 28/32 significant correlations in a range of 0.4-0.5, showing that TIPS has some potential and somewhat supporting RQ1. Notably, within this proportion, all TIPS chunk size measures turned out significant on each stimulus. However, some temporal micro-behavioral measures during the visualization and copying phase did not yield significant results. Specifically, the *View Pause* was not significant with the simple line graph and complex bar chart while the *Very First Pause of the Drawing* did not turn significant on the simple and complex bar chart.

While both MRT and PFT assess spatial transformation abilities, normalization with MRT proved to be more effective due to its alignment with the TIPS design. MRT requires rigid transformations (i.e., distances between points in an object is preserved) that correspond to the spatial transformations required by TIPS in both internal and external representation every time the View Key is pressed. Although TIPS may seem to require non-rigid transformations as the PFT probes, the Draw Pause allows participants to apply rigid spatial transformation to integrate the displayed information into a new mental representation that is then used for the Copying phase. For instance, similar to the MRT, once the amount of chunk contents that is going to be used for the Copying Phase is displayed (Figure 1), participants may use the Draw Pause to compare and combine the items displayed into a unified mental representation. In line with previous research, (Chen, 2000; Toker et al., 2013; Velez et al., 2005; Vicente et al., 1987) VPT is less relevant to reveal performance with information visualization than spatial transformation tests. In addition, the Corsi Tests may have not worked due to their simplicity compared to MRT, PFT and the visuo-spatial processing required by TIPS. Moreover, to the best of our knowledge, this is the first application of the Corsi tests in visualization research.

The experiment has a number of limitations. First, this study only employed a small range of visualizations such as line graphs, bar charts and histograms. Further investigations on different visualizations such as stacked bar charts, box plots, chord or Sankey diagrams will be needed. Second, stimuli complexity has to be considered. The simplicity of some stimuli (e.g., line graph) could be the reason for the absence of any chunking signal in the baseline measures and weaker correlations than found in previous CACHET experiments.

Although successfully employed in previous CACHET research (Albehaijan & Cheng, 2019), another limitation concerns the self-assessment questionnaire used as an independent measure in this experiment. Unlike data visualization literacy tests that measure different skills in reading and extracting information from data (Cui et al., 2023; Lee et al., 2017), the items included in the adopted questionnaire are rather generic and do not aim to measure any specific skills involved in graph comprehension. Given the importance of chunking and micro-behaviours in the TIPS design, a tailored test assessing key aspects represented by the visualizations is needed. This includes pattern recognition and users' knowledge of the visualization being used in TIPS. For instance, recognizing a negative skew in a plot of a distribution and what a negative skewness represents, could provide a more accurate measure of familiarity with a given visualization.

Acknowledgments

This work was supported by the EPSRC grants EP/R030642/1, EP/T019603/1, EP/T019034/1 and EP/R030650/1.

References

Albehaijan, N., & Cheng, P. (2019). *Measuring Programming Competence by Assessing Chunk Structures in a Code Transcription Task.* Proceedings of the 41st Annual Conference of the Cognitive Science Society (pp. 76-82), Austin, TX: Cognitive Science society.

Carswell, C. M., Emery, C., & Lonon, A. M. (1993). Stimulus Complexity and Information Integration in the Spontaneous Interpretations of Line Graphs. *Applied Cognitive Psychology*, *7*(4), 341-357. https://doi.org/10.1002/acp.2350070407

Chang, H. Y., & Tzeng, S. F. (2018). Investigating Taiwanese Students' Visualization Competence of Matter at the Particulate Level. *International Journal of Science and Mathematics Education*, *16*(7), 1207-1226. https://doi.org/10.1007/s10763-017-9834-2

Chase, W. G., & Simon, H. A. (1973). The Mind's Eye in Chess. In W. G. Chase (Ed.), *Visual Information Processing* (pp. 215-281). Academic Press. https://doi.org/10.1016/b978-0-12-170150-5.50011-1

Chase, W. G., & Simon, H. A. (1973). Perception in Chess. *Cognitive Psychology*, *4*(1), 55-81. https://doi.org/10.1016/0010-0285(73)90004-2

Chen, C. (2000). Individual differences in a spatial-semantic virtual environment. *Journal of the American society for information science*, *51*(6), 529-542.

Cheng, P. (2014). *Copying equations to assess mathematical competence: An evaluation of pause measures using graphical protocol analysis.* Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 36, No 36).

Cheng, P. (2015). *Analyzing chunk pauses to measure mathematical competence: Copying equations using'centre-click'interaction.* in CogSci.,

Cheng, P., & Zulkifli, P. A. M. (2009). *Exploring GPA as a Tool in Measuring Pauses of Writer's Language Competency.* In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol.31, No. 31). ,

Cheng, P. C., & van Genuchten, E. (2018). Combinations of Simple Mechanisms Explain Diverse Strategies in the Freehand Writing of Memorized Sentences. *Cognitive Science*, *42*(4), 1070-1109. https://doi.org/10.1111/cogs.12606

Cheng, P. C.-H., & Rojas-Anaya, H. (2007). *Measuring Mathematical Formula Writing Competence: An Application of Graphical Protocol Analysis.* In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 29, No. 29).

Colarusso, F., Cheng, P. C.-H., Garcia Garcia, G., Raggi, D., & Jamnik, M. (2021). Observing Strategies of Drawing Data Representations. In *International Conference on Theory and Application of Diagrams (pp. 537-552).* Springer. https://doi.org/10.1007/978-3-030-86062-2_55

Colarusso, F., Cheng, P. C. H., Garcia Garcia, G., Stockdill, A., Raggi, D., & Jamnik, M. (2023). A novel interaction for competence assessment using micro-behaviors: Extending CACHET to graphs and charts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23) April 23-28, 2023,Hamburg, Germany.* (pp. 1-14. https://doi.org/10.1145/3544548.3581519). ACM, New York, NY, USA. https://doi.org/https://doi.org/10.1145/3544548.3581519

Corsi, P. M. (1972). Human memory and the medial temporal region of the brain.

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav Brain Sci*, *24*(1), 87-114; discussion 114-185. https://doi.org/10.1017/s0140525x01003922

Cui, Y., Ge, L. W., Ding, Y., Yang, F., Harrison, L., & Kay, M. (2023). Adaptive Assessment of Visualization Literacy. *IEEE Transactions on Visualization and Computer Graphics.* https://doi.org/10.1109/TVCG.2023.3327165

Della Sala, S., Gray, C., Baddeley, A., Allamano, N., & Wilson, L. (1999). Pattern span: a tool for unwelding visuo-spatial memory. *Neuropsychologia*, *37*(10), 1189-1199. https://doi.org/10.1016/s0028-3932(98)00159-6

Egan, D. E., & Schwartz, B. J. (1979). Chunking in recall of symbolic drawings. *Memory & Cognition*, *7*, 149-158. https://doi.org/10.3758/bf03197595

Ekstrom, R. B. (1976). *Kit of factor-referenced cognitive tests*. Educational Testing Service.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191. https://doi.org/10.3758/bf03193146

Feeney, A., Hola, A., Liversedge, S., Findlay, J., & Metcalf, R. (2000). *How People Extract Information from Graphs: Evidence from a Sentence-Graph Verification Paradigm*. https://doi.org/10.1007/3-540-44590-0_16

Fox, A. R. (2019). When Graph Comprehension Is An Insight Problem. Annual Meeting of the Cognitive Science Society,

Fox, A. R. (2023). Theories and Models in Graph Comprehension. *Visualization Psychology*, 39-64.

Freedman, E. G., & Shah, P. (2002). Toward a model of knowledge-based graph comprehension. In International conference on theory and application of diagrams (pp. 18-30). ,

Ge, L. W., Cui, Y., & Kay, M. (2023). *CALVI: Critical Thinking Assessment for Literacy in Visualizations.* Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems., Hamburg, Germany. https://doi.org/10.1145/3544548.3581406

Gilbert, J. K. (2005). Visualization: A Metacognitive Skill in Science and Science Education. In Visualization in science education (pp. 9-27).

Gilbert, J. K., Reiner, M., & Nakhleh, M. B. (2008). *Visualization : theory and practice in science education (Vol.3)*.

Gobet, F., Lane, P. C., Croker, S., Cheng, P. C., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in cognitive sciences*, *5*(6), 236-243. https://doi.org/10.1016/s1364-6613(00)01662-4

Gobet, F., & Simon, H. A. (1996). Templates in chess memory: a mechanism for recalling several boards. *Cognitive Psychology*, *31*(1), 1-40. https://doi.org/10.1006/cogp.1996.0011

Golparvar, S. E., & Azizsahra, M. (2023). The effect of graph complexity and planning on graph writing performance and descriptive strategies. In *Foreign Language Annals* (Vol. 56, pp. 117-143). https://doi.org/10.1111/flan.12676

Hegarty, M. (2005). Multimedia Learning About Physical Systems. In *The Cambridge handbook of multimedia learning.* (pp. 447-465). Cambridge University Press. https://doi.org/10.1017/CBO9780511816819.029

Hinze, S. R., Rapp, D. N., Williamson, V. M., Shultz, M. J., Deslongchamps, G., & Williamson, K. C. (2013). Beyond ball-and-stick: Students' processing of novel STEM visualizations. *Learning and Instruction*, *26*, 12-21. https://doi.org/10.1016/j.learninstruc.2012.12.002

Ismail, H., & Cheng, P. (2021). Competence Assessment by Stimulus Matching: An Application of GOMS to Assess Chunks in Memory. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 43, No. 43),

John, B. E., & Kieras, D. E. (1996). The GOMS family of user interface analysis techniques. *ACM Transactions on Computer-Human Interaction*, *3*(4), 320-351. https://doi.org/10.1145/235833.236054

Johnson, N. F. (1970). The Role of Chunking and Organization in The Process of Recall. *Psychology of Learning and Motivation*, *4*, 171-247. https://doi.org/10.1016/s0079-7421(08)60432-6

Kriz, S., & Hegarty, M. (2007). Top-down and bottom-up influences on learning from animations. *International Journal of Human-Computer Studies*, *65*(11), 911-930. https://doi.org/10.1016/j.ijhcs.2007.06.005

Lee, S., Kim, S. H., & Kwon, B. C. (2017). VLAT: Development of a Visualization Literacy Assessment Test. *IEEE Trans Vis Comput Graph*, *23*(1), 551-560. https://doi.org/10.1109/TVCG.2016.2598920

Luo, W. H. (2019). User choice of interactive data visualization format: The effects of cognitive style and spatial ability. *Decision Support Systems*, *122*, 113061. https://doi.org/https://doi.org/10.1016/j.dss.2019.05.001

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81.

Nitz, S., Ainsworth, S. E., Nerdel, C., & Prechtl, H. (2014). Do student perceptions of teaching predict the development of representational competence and biological knowledge? *Learning and Instruction*, *31*(1), 13-22.

Obaidellah, U. H., & Cheng, P. C. (2015). The role of chunking in drawing Rey complex figure. *Perceptual and Motor Skills.*, *120*(2), 535-555. https://doi.org/10.2466/24.PMS.120v17x6

Ratwani, R. M., Trafton, J. G., & Boehm-Davis, D. A. (2008). Thinking graphically: Connecting vision and cognition during graph comprehension. *J Exp Psychol Appl*, *14*(1), 36-49. https://doi.org/10.1037/1076-898X.14.1.36

Roller, R., & Cheng, P. (2014). *Observed strategies in the freehand drawing of complex hierarchical diagrams* Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 36, No. 36).

Shah, P., & Freedman, E. G. (2011). Bar and line graph comprehension: an interaction of top-down and bottom-up processes. *Topics in cognitive science.*, *3*(3), 560-578. https://doi.org/10.1111/j.1756-8765.2009.01066.x

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701-703. https://doi.org/10.1126/science.171.3972.701

Sommers, P. v. (1989). A system for drawing and drawing-related neuropsychology. *Cognitive Neuropsychology*, *6*(2), 117-164.

Stewart, B. M., Hunter, A. C., & Best, L. A. (2008). *The Relationship between Graph Comprehension and Spatial Imagery: Support for an Integrative Theory of Graph Cognition.* Diagrammatic Representation and Inference: 5th International Conference, Diagrams 2008, Herrsching, Germany, September 19-21, 2008. Proceedings 5 (pp. 415-418).

Stoet, G. (2010). PsyToolkit: a software package for programming psychological experiments using Linux. *Behavior research methods.*, *42*(4), 1096-1104. https://doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2017). PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments. *Teaching of Psychology*, *44*(1), 24-31. https://doi.org/10.1177/0098628316677643

Tandon, S., Abdul-Rahman, A., & Borgo, R. (2023, Jan). *Measuring Effects of Spatial Visualization and Domain on Visualization Task Performance: A Comparative Study* IEEE Transactions on Visualization & Computer Graphics,

Thompson, J. J., McColeman, C. M., Stepanova, E. R., & Blair, M. R. (2017). Using Video Game Telemetry Data to Research Motor Chunking, Action Latencies, and Complex Cognitive-Motor Skill Learning. *Topics in cognitive science 9*(2), 467-484. https://doi.org/10.1111/tops.12254

Toker, D., Conati, C., Steichen, B., & Carenini, G. (2013). Individual user characteristics and information visualization: connecting the dots through eye tracking. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 295-304).

Trafton, J. G., Marshall, S. P., Mintz, F., & Trickett, S. B. (2002). *Extracting Explicit and Implict Information from Complex Visualizations* In Diagrammatic Representation and Inference: Second International Conference, Diagrams 2002 Callaway Gardens, GA, USA, April 18–20, 2002 Proceedings 2 (pp. 206-220).

Trafton, J. G., & Trickett, S. B. (2001). *A New Model of Graph and Visualization Usage* In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 23, No. 23).

Trafton, J. G., Trickett, S. B., & Mintz, F. E. (2005). Connecting Internal and External Representations: Spatial Transformations of Scientific Visualizations. *Foundations of Science*, *10*(1), 89-106. https://doi.org/10.1007/s10699-005-3007-4

Trickett, S. B., & Trafton, J. G. (2004). Spatial transformations in graph comprehension. Diagrammatic Representation and Inference: Third International Conference, Diagrams 2004, Cambridge, UK, March 22-24, 2004. Proceedings 3 (pp. 372-375).

Trickett, S. B., & Trafton, J. G. (2006). *Toward a comprehensive model of graph comprehension: Making the case for spatial cognition.* In International Conference on Theory and Application of Diagrams (pp. 286-300). Berlin, Heidelberg: Springer Berlin Heidelberg.,

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills.*, *47*(2), 599-604. https://doi.org/10.2466/pms.1978.47.2.599

VanderPlas, S., & Hofmann, H. (2016). Spatial Reasoning and Data Displays. *IEEE Trans Vis Comput Graph*, *22*(1), 459-468. https://doi.org/10.1109/TVCG.2015.2469125

Velez, M. C., Silver, D., & Tremaine, M. (2005). Understanding visualization through spatial ability differences. IEEE Visualization 2005, Minneapolis, MN, USA.

Vicente, K. J., Hayes, B. C., & Williges, R. C. (1987). Assaying and isolating individual differences in searching a hierarchical file system. *Human Factors*, *29*(3), 349-359. https://doi.org/10.1177/001872088702900308

Xi, X. (2016). Do visual chunks and planning impact performance on the graph description task in the SPEAK exam? *Language Testing*, *22*(4), 463-508. https://doi.org/10.1191/0265532205lt305oa