Making Sense of Time-Series Data: How Language Can Help Identify Long-Term Trends

Jordan Harold (jordan.harold@uea.ac.uk)
School of Psychology and Tyndall Centre for Climate Change Research,
University of East Anglia, Norwich, UK

Kenny R. Coventry (k.coventry@uea.ac.uk)
School of Psychology, University of East Anglia, Norwich, UK

Irene Lorenzoni (i.lorenzoni@uea.ac.uk)
School of Environmental Sciences and Tyndall Centre for Climate Change Research,
University of East Anglia, Norwich, UK

Thomas F. Shipley (tshipley@temple.edu)
Department of Psychology, Temple University, Philadelphia, USA

Abstract

Real-world time-series data can show substantial short-term variability as well as underlying long-term trends. Verbal descriptions from a pilot study, in which participants interpreted a real-world line graph about climate change, revealed that trend interpretation might be problematic (Experiment 1). The effect of providing a graph interpretation strategy, via a linguistic warning, on the encoding of long-term trends was then tested using eye tracking (Experiment 2). The linguistic warning was found to direct visual attention to task-relevant information thus enabling more detailed internal representations of the data to be formed. Language may therefore be an effective tool to support users in making appropriate spatial inferences about data.

Keywords: graph comprehension; language; visual attention

Line graphs can be a powerful communication tool to visually demonstrate important relationships in time-series data. They are ubiquitous in everyday life and graph interpretation is considered an important skill for a scientifically literate society (Glazer, 2011). Many types of real-world data exhibit substantial short-term variability as well as long-term trends, e.g. global mean surface temperature records (IPCC, 2013), share prices (Schwert, 2011), and incidence of certain diseases (e.g. Subak, 2003). In visualizations of such data, can users efficiently and accurately identify underlying long-term trends? If not, how might users be supported in doing so?

Comprehension of graphs involves an interaction between bottom-up sensory processes and top-down cognitive constraints, and is thought to involve two key cyclical processes (Carpenter & Shah, 1998; Freedman & Shah, 2002). First, users construct an internal representation of the display by encoding perceptual features of the graph, guided by prior knowledge. Then knowledge is applied to integrate the representation into a coherent mental model. If relevant information is represented directly in the graph and can be easily linked with existing knowledge, this integration phase is comparatively effortless. However, if information is not explicitly represented in the graph and/or the user lacks the required knowledge to form an accurate model, or cannot easily access the required knowledge, then comprehension is likely to require much more effort.

For example, a climate scientist will know to consider the long-term trend when interpreting temperature records and so may effortlessly transform and encode visual features from the data that support a representation of the long-term trend. In contrast, a climate science 'novice' may encode visual features that are explicitly represented in the graph, such as the amplitude of peaks or troughs, which may support an understanding of short-term fluctuations, but make inferences about the long-term trend rather effortful and less likely. Hence, graphs that organize and structure data, such that emergent visual properties explicitly reveal important relationships, e.g. based on Gestalt laws, may be particularly effective (Kosslyn, 1989; Zacks & Tversky, 1999), by reducing the cognitive effort that might otherwise be needed (Hegarty, 2011).

Although a line graph may be a single unit by the Gestalt law of connectedness (Ali & Peebles, 2013), a complex line may be decomposed into parts or 'chunks', based on local curvature extrema (Hoffman & Richards, 1984). Time-series datasets that show significant short-term variability may have numerous curvature extrema (e.g. trend reversals) creating multiple visual chunks. These chunks may serve as units on which inferential processes, required for interpretation, act (Freedman & Shah, 2002).

Trend reversals can increase study time, and also increase local content and decrease global content of verbal and written interpretations of line graphs (Carswell, Emery, & Lonon, 1993). In this study it was hypothesized that each set of continuous non-reversing data points constitutes a chunk of information in an individual's internal representation. Hence local curvature extrema may indicate boundaries in the perceptual grouping of connected lines thus creating numerous visual chunks for higher level cognitive processing. Interpreting long-term trends may therefore be difficult, because it requires integration of these visual

chunks, which may require effortful cognitive processes such as spatial transformations.

If this is the case, language might be a useful tool to support spatial cognition. Evidence suggests that attending to spatial language when encoding visual scenes can help construct representations that support spatial reasoning (Loewenstein & Gentner, 2005) and can influence memory of spatial scenes (Feist & Gentner, 2007). Furthermore, language can provide a user-goal during the study of a visual scene (i.e. a purpose for engaging with the scene), which may then activate relevant schema and guide visual-spatial attention (Brunyé & Taylor, 2009; Rothkopf, Ballard, & Hayhoe, 2007; Yarbus, 1967). Eye-tracking studies of relatively simple graphs indicate that visual attention appears to be driven by user-goals and graph knowledge (Carpenter & Shah, 1998; Peebles & Cheng, 2003) and hence using language to influence these top-down processes might help users to attend to and encode appropriate information in time-series line graphs.

The aim of Experiment 1 was to characterize difficulties, if any, in trend interpretation by asking participants to look at and then describe a real-world time-series graph that contained an underlying long-term trend as well as substantial short-term variability. Experiment 2 then asked whether a linguistic warning, providing an interpretation strategy, might improve encoding of long-term trends.

Experiment 1

To see if people correctly identify long-term trends from time-series graphs that also show significant short-term variability, verbal descriptions were collected from individuals exposed to a real-world graph showing such characteristics. The graph chosen (Figure 1) shows data for Northern Hemisphere spring snow cover extent between 1922-2012, published by the Intergovernmental Panel on Climate Change (IPCC, 2013). The IPCC is an international scientific body tasked with communicating policy-relevant scientific information to policy makers. The figure therefore has societal relevance. Furthermore, the data indicate a significant downward trend over the whole time-period, together with substantial inter-annual variability. The authors indicate that snow cover extent has decreased since the mid-20th century (IPCC, 2013), suggesting that this is an important communication goal.

Method

Participants Twelve undergraduate students (10 female, two male) from the University of East Anglia took part in the study in return for course credit or a nominal payment. Their average age was 21 years (range 19–29 years). None of the participants were studying environmental sciences.

Apparatus and Materials The stimulus was presented on a TFT LCD monitor (51cm x 29cm), set to 1280 x 720 pixels. Eprime Version 2.0 (Psychology Software Tools Inc., Sharpsburg, USA) was used to control stimulus presentation and record data. Verbal responses were captured via a headset microphone. The stimulus consisted of Figure SPM.3a from the IPCC Summary for Policy Makers (IPCC, 2013) (Figure 1).

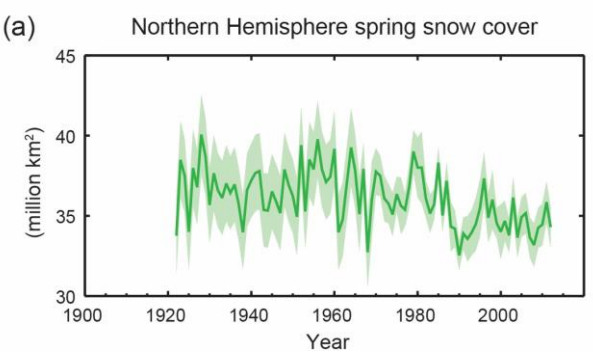

Figure 1: SPM.3a from Figure SPM.3: Multiple observed indicators of a changing global climate (IPCC, 2013).[1]

Procedure The figure was presented for 15 seconds – during this time, participants were asked to simply look at the figure. They then saw a 'Now describe' prompt and the same figure re-appeared on the screen, at which point participants were asked to describe what they thought it was trying to show. The figure remained on screen until the participant completed their verbal response, up to a maximum time limit of 45 seconds.

Coding Verbal descriptions were coded to assess the presence (1) or absence (0) of the following aspects: (a) the data represent changes in snow cover over time; (b) a general downward trend; (c) a downward trend between ~1960 and ~2012; (d) short-term variability/fluctuation.[2]

Results and Discussion

All twelve participants correctly identified that the data represented changes in snow cover over time, but only five participants (42%) described a downward trend over the whole data. One of these participants also described a downward trend between ~1960 and ~2012. Of the five participants who described either type of downward trend, one also described the short-term variability (20%), but of the seven participants who did not describe either downward trend, five described the short-term variability (71%) (p=.01, Fisher's Exact Test). These pilot data suggest that when presenting graphs that contain an underlying long-term trend and substantial short-term variability, spontaneous interpretation of the long-term trend may be far from guaranteed.

[1] Multiple observed indicators of a changing global climate: (a) Extent of Northern Hemisphere March-April (spring) average snow cover. All time-series (coloured lines indicating different data sets) show annual values, and where assessed, uncertainties are indicated by coloured shading.

[2] Inter-rater reliability across all aspects and all coding: κ = 1.000, p<.001.

Experiment 2

The pilot data from Experiment 1 indicate that the long-term trend may not be readily interpreted in graphs that also show substantial short-term variability. The aim of Experiment 2 was therefore to test whether a linguistic warning that provides a strategy for interpreting long-term trends (by ignoring task-irrelevant features) would improve encoding of the long-term trend; and if so, whether this is driven by changes in visual attention (measured using eye tracking). In addition, Experiment 2 investigated whether reducing, or removing intermediary x-axis tick marks and labels might have a beneficial effect on the encoding of long-term trends, as their presence might cue people to read-off data values or focus on short-term (inter-tick/-label) trends.

Method

Design To test spatial representations of the long-term trend (i.e. gradient) and short-term variability (i.e. amplitude), a forced choice task was employed in which participants were shown a graph to study and then asked to make a 'same' or 'different' judgment on a following test graph. The test graph was either identical to the study graph (same); had the same peaks and troughs as the study graph but with a different gradient (gradient different); had the same gradient as the study graph but with exaggerated peaks and troughs (amplitude different); or was completely different to the study graph (completely different). The number of x-axis ticks, either 2, 5 or 9, was varied across each type of test graph (see Figure 2 for examples).

To test the effect of a linguistic warning on cognition of the graph, participants were randomly allocated to either receive a warning asking them to ignore extreme values in order to consider the long-term trend (warning), or to receive no such warning (no warning). The experiment was therefore a 4 (trial type) x 3 (x-ticks) x 2 (warning) design, with trial type and x-ticks as within participant variables and warning as a between participant variable.

Participants Forty undergraduate students (29 female, 11 male) from the University of East Anglia took part in the study in return for course credit or a nominal payment. Their average age was 21 years (range 18-30 years).

Apparatus A Tobii TX300 Eye Tracker (Tobii Technology AB, Danderyd, Sweden) with integrated TFT LCD monitor (51cm x 29cm) set to 1280 x 720 pixels was used for stimulus presentation and collection of eye gaze data at 300Hz. Eprime Version 2.0 (Psychology Software Tools Inc., Sharpsburg, USA) was used to control stimulus presentation and record data. Responses for same-different trials were given using the 'Z' and 'M' keyboard keys. Response key mappings were reversed and counterbalanced between warning conditions. Verbal responses were recorded via a headset microphone. Eye gaze data were analyzed using OGAMA Version 4.5 (A. Voßkühler, Freie Universität Berlin, Germany), using default parameters for fixation detection.

Linguistic Warning The linguistic warning was displayed in 28pt Calibri and read: "WARNING When looking at graphs, people are often misled by extreme data points – short-term fluctuations in the data can obscure the long-term trend. To avoid errors, it is useful to ignore extreme data points to correctly identify the long-term trend."

Graph Stimuli Twenty-four study time-series graphs were created (1126 x 510 pixels), each plotting 17 data points. Graphs showed an underlying positive, negative or flat long-term trend. Data points for each graph were created by sampling residuals at random from a normal distribution, which were then applied to a baseline positive, negative or flat linear trend graph. The x-axis was labelled 'Years' and the y-axis was labelled either as 'Medication use (doses)', 'Infections (patients)', 'Temperature (^{o}C)', 'Rainfall (mm)', 'Income (GBP £)', or 'Expenditure (USD $)'. The x-axis covered a range of 16 years, with the starting year always between 1900 and 1994. A caption was created for each graph that simply read '[variable] over time.'

A positive, negative and flat trend study graph was allocated to each trial type. A test graph was then created for each study graph. Test graphs for the same condition were identical to their corresponding study graph. Test graphs for the gradient different condition were created by a transformation of the study graph that resulted in a visual rotation of the graph line by ±2 degrees. Test graphs for the amplitude different condition were created by multiplying the residuals of the study graph by a factor of 1.4. Three new graphs were created to serve as test graph pairings for the completely different trials. For each study and test graph pairing, three variants were created, each showing 2, 5 and 9 x-ticks (Figure 2). The remaining study graphs were allocated to true-false and describe filler trials, which also included variations for each level of x-ticks.

Areas of Interest (AOI) AOIs were defined for each study graph by first determining a circle around each data point with a maximum diameter that would avoid overlapping adjacent data points (58 pixels). A parallelogram with height 58 pixels, width 1002 pixels (2.0 x 34.5 degrees of visual angle), was then fitted over the line of best fit of the graph data, determined by linear least squares regression. This formed the line of best fit AOI (6.3% of screen area). A convex hull was then determined around the outer edges of these shapes, which formed the whole data AOI (mean 22.1% of screen area). An extreme data AOI was defined as the area of the whole data AOI that sat outside of the line of best fit AOI (mean 15.8% of screen area) (Figure 3).

Procedure Participants were informed that the study was investigating how people understand line graphs and they then received instructions on screen before a practice block of trials. The eye tracker was calibrated and then participants in the warning condition received the warning on screen and were instructed to read it before starting the first of three blocks of trials. Participants in the no

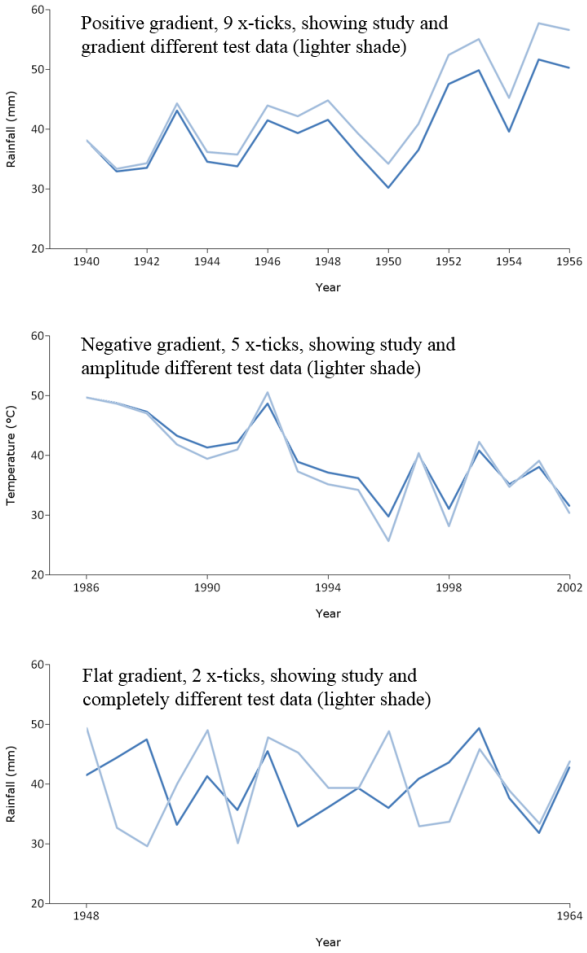

Figure 2: Three examples of the study and test graphs in Experiment 2.

warning condition simply started the first block of trials after eye tracker calibration. Each trial consisted of a study phase (Figure 4) during which participants were asked to look at and study the caption and the graph. The caption was presented prior to the graph to help control time spent reading the caption. The study phase was followed by one of three task cues (Figure 4). For same-different trials, participants had to make a same-different judgment about a test caption and then about a test graph in comparison to the study caption and study graph. Participants were instructed to give a response as quickly as possible when the caption/graph appeared.

Each block consisted of 12 same-different trials (three of each of the different trial types), presented in random order. Three true-false trials and three describe trials were included in each block to encourage participants to study the graphs in a naturalistic way and to ensure depth of encoding. Each x-tick variation of a given graph was presented in a different block. Blocks of trials were counterbalanced across participants and the eye tracker was re-calibrated at the start of each block. At the end of the third block, participants in the warning condition were asked what they remembered about the warning. The study lasted approximately 1 hour.

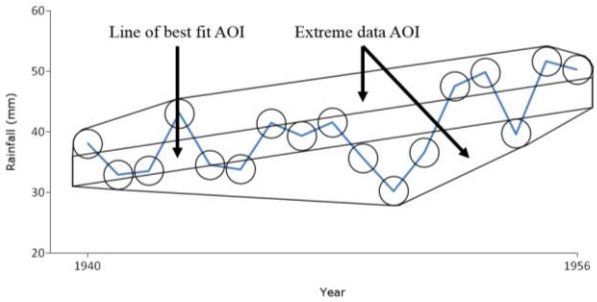

Figure 3: Line of best fit AOI and extreme data AOI for one of the 24 study graphs in Experiment 2.

Results and Discussion

Only same-different trials in which a correct response was given to the test caption and a response was given to the test graph were included in the analyses (i.e. trials in which participants correctly remembered the caption and then went on to make a judgement about the graph). Six participants were removed from further analyses: one participant who subsequently reported monocular vision impairment; one participant whose accuracy on completely different trials was 11% (lower than three *SD* from mean accuracy); and four participants in the warning condition who could not remember any detail about the warning when asked at the end of the study (and so may not have encoded it).

Task Performance Sensitivity to detect differences between the graphs of same-different trials was measured using *d'* in order to assess response accuracy with the effects of response bias removed. Participants' *d'* scores were analyzed with a 3 (trial type) x 3 (x-ticks) x 2 (warning) mixed ANOVA. There was a main effect of trial type, $F(2,64)=59.603$, $p<.001$, partial $\eta^2=.651$. Bonferroni post-hoc tests indicated a significant difference between amplitude different trials and completely different trials ($p<.001$), and gradient different trials and completely different trials ($p<.001$), indicating that participants had a greater ability to detect differences between study and test graphs when the test graph was completely different, than when only the amplitude or gradient was different.

There was no main effect of x-ticks, $F(2,64)=0.504$, $p=.606$; and no main effect of warning, $F(1,32)<0.001$, $p=.994$. However there was a significant interaction between trial type and warning, $F(2,64)=3.459$, $p=.037$, partial $\eta^2=.098$ (Figure 5). Post-hoc examination indicated that participants in the no warning condition performed significantly worse on gradient different trials ($M = 0.251$, 95% CI ±0.222) than amplitude different trials ($M = 0.667$, 95% CI ±0.274) ($p=.008$), whereas those in the warning condition performed about equally on gradient different trials ($M=0.504$, 95% CI ±0.293) and amplitude different trials ($M=0.479$, 95% CI ±0.349). There was no significant x-ticks x warning interaction, $F(2,64)=3.041$, $p=.055$; and no three-way interaction, $F(4,128)=1.162$, $p=.331$, indicating that the number of intermediary x-ticks did not influence sensitivity to detect changes in the long-term trend.

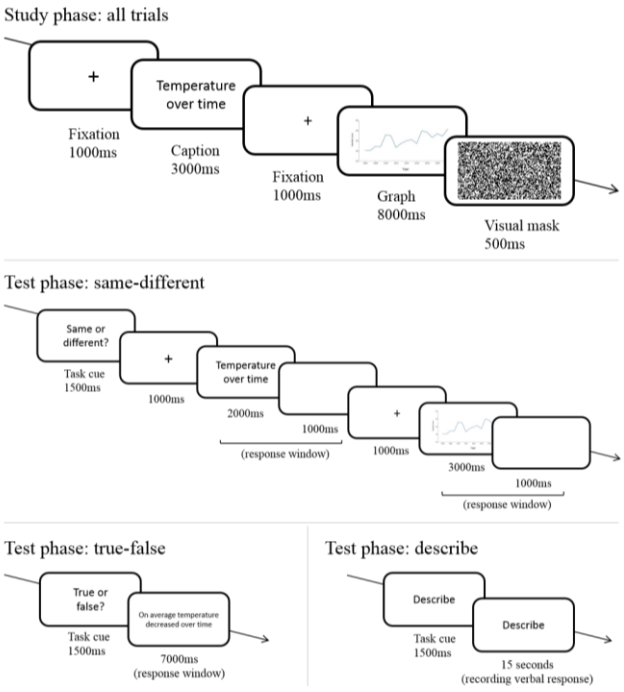

Figure 4: Presentation of same-different and filler trials.

Using language to provide task-relevant knowledge improved sensitivity to detect differences in task-relevant information (i.e. the long-term trend) relative to other information (i.e. amplitude). Furthermore, this did not appear to come at the expense of an impaired sensitivity to detect differences in the other information.

To investigate if the effect of the warning on gradient performance deteriorated over time, *d'* values were recalculated by collapsing data across x-ticks (as there was no significant x-ticks main effect or interaction), and then splitting out the data by block. A 2 (warning) x 3 (trial type) x 3 (block) mixed ANOVA was then performed. Results were consistent with the first mixed ANOVA, and there was no three way interaction between trial type, warning and block, $F(2.903,92.895)=0.189$, $p=.898$ (with Greenhouse-Geisser correction), indicating that there was no evidence to suggest that the trial type x warning interaction was modulated by the duration between the warning and the block of trials. This suggests that the warning was encoded into long-term memory and applied throughout the study. These results indicate that the warning had a lasting effect on participants' judgements, suggesting that in the absence of explicit user-goals, using language to impart graph knowledge may direct subsequent interpretation of the data.

Visual Attention To investigate if the improved discriminability of the gradient found in the warning condition might be driven by differences in visual attention during encoding, fixation durations for the AOIs of the study graphs were calculated. Fixations were calculated for same-different trials in which a correct response was given to the caption and a response was given to the test graph, all true-false trials in which a response was given, and all verbal trials. Trials for four participants were excluded from further analysis as they had poor eye tracking calibrations. Individual trials were excluded if >15% of eye tracking samples were missing, or if there was a continuous period >700ms of data missing (10.7% of trials). As there was no main effect or interaction of x-ticks in the *d'* data, fixation data were collapsed across x-ticks.

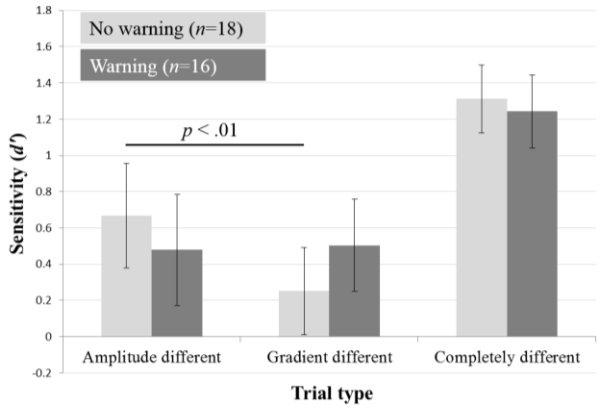

Figure 5: Average sensitivity (*d'*) for each trial type and warning group, with 95% confidence intervals.

At study, participants in the warning condition spent significantly longer fixating within the line of best fit area than participants who did not receive the warning, $t(19.802)=2.119$, $p=.024$ (one-tailed, equal variances not assumed) (Table 1). Conversely, there was no significant difference in total fixation duration of the extreme data area between the two groups, $t(25.137)=-0.352$, $p=.728$ (two-tailed, equal variances not assumed), nor a significant difference in total fixation duration in the whole data area, $t(28)=1.288$, $p=.208$ (two-tailed, equal variances assumed). Taken together, the task performance and visual attention results suggest that using language to provide graph knowledge can direct visual attention to task-relevant information during encoding, which then enables the creation of a more detailed internal representation of the graphed data (rather than merely an alternative representation) and can influence subsequent interpretation.

Table 1: Mean (*M*) and standard deviations (*SD*) of fixation duration in ms during study for each AOI.

Area of interest	**No warning (*n*=16)**		**Warning (*n*=14)**	
	M	***SD***	***M***	***SD***
Line of best fit	1426	(432)	1919	(772)
Extreme data	1587	(586)	1525	(356)
Whole data	3013	(884)	3444	(952)

General Discussion

The research presented here supports and builds on existing theoretical research on display comprehension and has important implications for communicators of time-series data. Pilot data from Experiment 1 found that interpretations

of a real-world time-series line graph that contained a high degree of short-term variability (and therefore many trend reversals) did not elicit correct descriptions of the long-term trend in more than half of the participants. This is consistent with the hypothesis that trend reversals provide salient visual cues that break down connected lines into separate visual chunks, which may then be difficult to integrate into a representation of the long-term trend. Experiment 2 found that in the absence of an explicit user-goal or an interpretation strategy, users created better representations of the short-term variability than the long-term trend. However, when provided with an interpretation strategy via a linguistic warning, participants encoded both the long-term trend and short-term variability equally well.

In contrast to previous research investigating changes to the layout and format of a display in order to make task-relevant patterns explicitly represented (e.g. Shah, Mayer, & Hegarty, 1999), the research presented here highlights top-down cognitive processes on the identification and interpretation of data patterns. Language may be an effective way of providing graph knowledge, which can then be drawn on to direct visual attention to relevant visual features and support appropriate spatial inferences.

This may be especially pertinent when communicating complex data sets that contain several communication goals. For example, climate scientists may wish to communicate the long-term trends of indicators of a changing climate, as well as enabling individuals to understand that short-term variability in these indicators exists. Language may provide a useful tool to direct users to consider aspects that require complex inferential processes (such as the long-term trend) in addition to the salient patterns in the display. Given the need for individuals to interpret graphs to make informed decisions and play an active role in society, there is a need to extend our theoretical understanding of display comprehension, and to apply and test out theoretical insights in real-world communication problems. The research presented here supports both of these aims.

Acknowledgments

JH was supported by a PhD studentship from the School of Psychology, University of East Anglia and a travel grant from the Spatial Intelligence & Learning Centre (SILC), Temple University. (SBE-1041707 from the National Science Foundation).

References

Ali, N., & Peebles, D. (2013). The effect of gestalt laws of perceptual organization on the comprehension of three-variable bar and line graphs. *Human Factors*, *55*(1), 183-203.

Brunyé, T. T., & Taylor, H. A. (2009). When goals constrain: Eye movements and memory for goal-oriented map study. *Applied Cognitive Psychology*, *23*(6), 772-787.

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, *4*(2), 75-100.

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.

Feist, M. I., & Gentner, D. (2007). Spatial language influences memory for spatial scenes. *Memory & Cognition*, *35*(2), 283-296.

Freedman, E. G., & Shah, P. (2002). Toward a model of knowledge-based graph comprehension. In M. Hegarty, B. Meyer, & N. H. Narayanan (Eds.), *LNAI 2317: Diagrammatic Representation and Inference*. Berlin, Germany: Springer.

Glazer, N. (2011). Challenges with graph interpretation: A review of the literature. *Studies in Science Education*, *47*(2), 183-210.

Hegarty, M. (2011). The cognitive science of visual-spatial displays: Implications for design. *Topics in Cognitive Science*, *3*(3), 446-474.

Hoffman, D. D., & Richards, W. A. (1984). Parts of recognition. *Cognition*, *18*(1), 65-96.

IPCC (2013). Summary for Policymakers. In T.F. Stocker, D. Qin, G.-K. Plattner, *et al*. (Eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK and New York, USA: Cambridge University Press.

Kosslyn, S. M. (1989). Understanding charts and graphs. *Applied Cognitive Psychology*, *3*(3), 185-226.

Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive Psychology*, *50*(4), 315-353.

Peebles, D., & Cheng, P. (2003). Modeling the effect of task and graphical representation on response latency in a graph reading task. *Human Factors*, *45*(1), 28-46.

Rothkopf, C. A., Ballard, D. H., & Hayhoe M. M. (2007). Task and context determine where you look. *Journal of Vision*, *7*(14), 16, 1-20.

Schwert, G. W. (2011). Stock volatility during the recent financial crisis. *European Financial Management*, *17*(5), 789-805.

Shah, P., Mayer, R. E., & Hegarty, M. (1999). Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph comprehension. *Journal of Educational Psychology*, *91*(4), 690-702.

Subak, S. (2003). Effects of climate on variability in Lyme disease incidence in the northeastern United States. *American Journal of Epidemiology*, *157*(6), 531-538.

Yarbus, A. L. (1967). *Eye movements and vision*. New York, USA: Plenum Press.

Zacks, J., & Tversky, B. (1999). Bars and lines: A study of graphic communication. *Memory & Cognition*, *27*(6), 1073-1079.