Memory & Cognition
https://doi.org/10.3758/s13421-020-01101-w

Zooming in on visual narrative comprehension

Tom Foulsham[1] · Neil Cohn[2]

Accepted: 16 September 2020

Abstract
The comprehension of visual narratives requires paying attention to certain elements and integrating them across a sequence of images. To study this process, we developed a new approach that modified comic strips according to where observers looked while viewing each sequence. Across three self-paced experiments, we presented sequences of six panels that were sometimes automatically "zoomed-in" or re-framed in order to highlight parts of the image that had been fixated by another group of observers. Fixation zoom panels were rated as easier to understand and produced viewing times more similar to the original comic than panels modified to contain non-fixated or incongruous regions. When a single panel depicting the start of an action was cropped to show only the most fixated region, viewing times were similar to the original narrative despite the reduced information. Modifying such panels also had an impact on the viewing time on subsequent panels, both when zoomed in and when regions were highlighted through an "inset" panel. These findings demonstrate that fixations in a visual narrative are guided to informative elements, and that these elements influence both the current panel and the processing of the sequence.

Keywords Visual language · Visual narrative · Comics · Attention

Introduction

Visual narratives, such as those in comics, carry meaning via a sequence of images. The way in which observers "read" such sequences has become a topic of interest for a range of cognitive scientists (Cohn, 2020; Foulsham, Wybrow & Cohn, 2016; Loschky, Magliano, Larson, & Smith, 2020). This interest has focused both on the constraints guiding narrative sequencing (Cohn, 2020) and on the way that context interacts with one's perception of an image (Foulsham, Wybrow, & Cohn, 2016; Loschky et al., 2020; Hutson, Magliano, & Loschky, 2018). Rather than being perceived in isolation, each panel of a comic strip is designed to be appreciated in the context of the surrounding events. Visual narratives can be comprehended quickly (Laubrock, Hohenstein, & Kümmerer, 2018), even when they do not contain words or formal language (Hagmann & Cohn, 2016; Inui & Miyamoto, 1981). However, there is little experimental work investigating what people attend to within a visual narrative, or how this may affect the speed or ease of comprehension. In the present study, we manipulate the framing of panels in a comic strip and examine the effects on viewing time (VT). It could be that removing any visual content will make the process of visual narrative comprehension more difficult – because additional inference will be required. However, here we frame panels to selectively highlight the details that are fixated by human eye movements (Foulsham et al., 2016). This provides a test of whether key information is sufficient for viewing a sequence normally, and whether observers attend to this information during natural comprehension.

Comic artists often frame images in order to focus attention (see Fig. 1). For example, some sequences start off with a "wide-angle" illustration of several characters, before zooming in to show a particular facial expression or action. The artist is therefore guiding the viewer towards items from a larger scene. In films, which often begin life as a drawn storyboard, the filmmaker may also direct attention by using different shot distances (see Smith, Levin, & Cutting, 2012). Interestingly, close-up shots in Hollywood cinema have become more common over time (Cutting, Brunick, & Candan, 2012), and there is a similar trend for more "zoom" panels showing a single character in American comics (Cohn, Taylor, & Pederson, 2017b). The implication is that artists

✉ Tom Foulsham
foulsham@essex.ac.uk

[1] Department of Psychology, Wivenhoe Park, Colchester, Essex CO4 3SQ, UK

[2] Department of Communication and Cognition, Tilburg University, Tilburg, The Netherlands

Published online: 06 October 2020

Fig. 1 Visual narratives can show a full scene (e.g., top row), but they can also focus viewers' attention by "zooming in" on key regions to tell a story (e.g., bottom row)

are increasingly directing viewers towards the key content of a scene. Here, we test whether this key content can be determined from fixations alone, and whether narrative comprehension is disrupted by presenting zoomed information without the rest of the panel.

In comics, content may be presented "zoomed-in," or it may be highlighted in other ways (e.g., as an "inset" panel inside a bigger panel). Since panel framing can act as a window on particular information, some work has posited that panels can simulate attention across a scene (Cohn, 2013). To what extent does the framing reflect actual overt attention? When viewing an image, we move our eyes to locations that are semantically meaningful or pertinent to the task (Henderson & Hayes, 2017; Mackworth & Morandi, 1967; Yarbus, 1967). Although some locations in images may attract fixations because of their bottom-up features (i.e., because they are bright or stand out from their background; Itti & Koch, 2000), an individual's scanning is also somewhat idiosyncratic (Foulsham & Kingstone, 2013) and varies depending on the task (Mills et al., 2011; Yarbus, 1967) and the observer's knowledge (Underwood, Foulsham, & Humphrey, 2009).

While there are relatively few studies examining attention in images that are part of a sequence, recent research has begun to investigate comics and graphic novels using eye-tracking methods. For example, Laubrock, Hohenstein, and Kümmerer (2018) report an eye movement corpus from 100 participants freely reading six graphic novels. In their analysis, viewers attended to key details in a panel, such as the main character, often making only a few fixations on the image. The first fixation on a panel indicated that viewers are able to preselect the most important material, by previewing it in peripheral vision. This raises the question of whether comic readers might be able to follow a visual narrative using only the key information in the panel, or whether they also need the surrounding context, even though this is fixated less often.

In Foulsham et al. (2016), we presented visual narratives that could either appear in the original order or in a randomized sequence that made less sense. Panels shown in the scrambled context were processed more slowly and attracted more and longer fixations. There was also evidence that people looked at different parts of an image according to the order. For example, they fixated the consequences of an action when, in the correctly ordered sequence, they had seen the build-up to this action. A conceptually similar result was observed in Foulsham and Kingstone (2017), in which photographs were presented from the perspective of a person walking down the street. When these scenes were presented in a randomized order, different locations were inspected (in comparison to when the scenes were presented in the logical order). The "snapshots" of the real world generated expectations about what would come next (further down the street), which could only guide attention in the coherently ordered condition.

The structure of visual narratives provides expectations about what is coming next, speeding comprehension. Previous work has shown that manipulating specific cues in panels, such as characters' posture, can change the processing of subsequent images in a sequence (Cohn & Paczynski, 2013; Cohn & Maher, 2015). With eye tracking, work has shown that omission of characters from a comic led to longer VTs and lower comprehension (Tseng, Laubrock & Pflaeging, 2018), and that cues that inform inferences attract more attention when prior information has been omitted (Hutson et al., 2018).

The present study

In the present study we framed panels within a comic based on where viewers actually look. Specifically, we developed a method to automatically construct "zoom" and "inset" panels based on fixations made by people while freely viewing a

comic strip (see Fig. 2 for a summary). In brief, we took the fixations made when viewing a particular panel and defined a density map, representing how likely each region was to be fixated. By thresholding this map, we could automatically define modified panels that preserved only the most (or least) fixated content. This process was completely automatic and unsupervised. We did not alter, crop, or screen panels on the basis of esthetic or semantic requirements. Here, we test whether these modified "fixation zooms" affect the viewing of different parts of the narrative. We do so by comparing self-paced VTs and ratings of difficulty.

In photographs, regions that are frequently fixated are rated as more meaningful (Henderson & Hayes, 2017; Mackworth & Morandi, 1967) and remembered more accurately (Foulsham & Kingstone, 2013) than other regions. To our knowledge, only one study has experimentally linked fixations to meaning in a sequence of images. Hutson et al. (2018) asked participants to click on the parts of a picture book that were most important for making an inference. They found that these regions were more likely to be fixated when a bridging inference was necessary (because previous pages had been omitted). If fixations throughout a sequence are efficient at picking out the key information, then a visual narrative should be understandable even when only fixated regions are displayed. However, an alternative view is that other details from the panel are necessary for comprehension, even when not fixated. For example, peripheral details could be important for the "gist" or general layout (Larson & Loschky, 2009). Since artists already intentionally frame panels in a particular way, distilling an image based on fixations might remove critical context. Indeed, in previous self-paced viewing studies, the removal of information from a sequence leads to increased VTs and EEG signatures of increased processing difficulty (Cohn & Maher, 2015; Cohn et al., 2017a; Tseng et al., 2018). In a related self-paced task, the "dwell-time paradigm," Hard et al. (2011) showed that VTs in a slideshow are sensitive to event segmentation. Kosie and Baldwin (2019) reported elevated VTs when slides were removed from a sequence, although viewers still dwelled longer on images that indicated an event boundary. Although VT in this paradigm has been interpreted as reflecting attention to particular slides, attention within a slide has not been examined.

In sum, it is not currently clear from the existing literature, which is mostly on individual images, whether non-fixated information can be "edited out" without making the sequences harder to comprehend. The fixation zoom method described here – where we examine whether fixated regions are necessary and sufficient for normal paced viewing – is complementary to other approaches where fixations are correlated with image features or semantic ratings (Henderson & Hayes, 2017; Itti & Koch, 2000). To the extent that fixations select specific content, we would of course expect the fixation zoom panels to be different from other parts of the image (e.g., they might be more likely to contain objects or faces). However, our stimulus-generation method was data driven, and did not

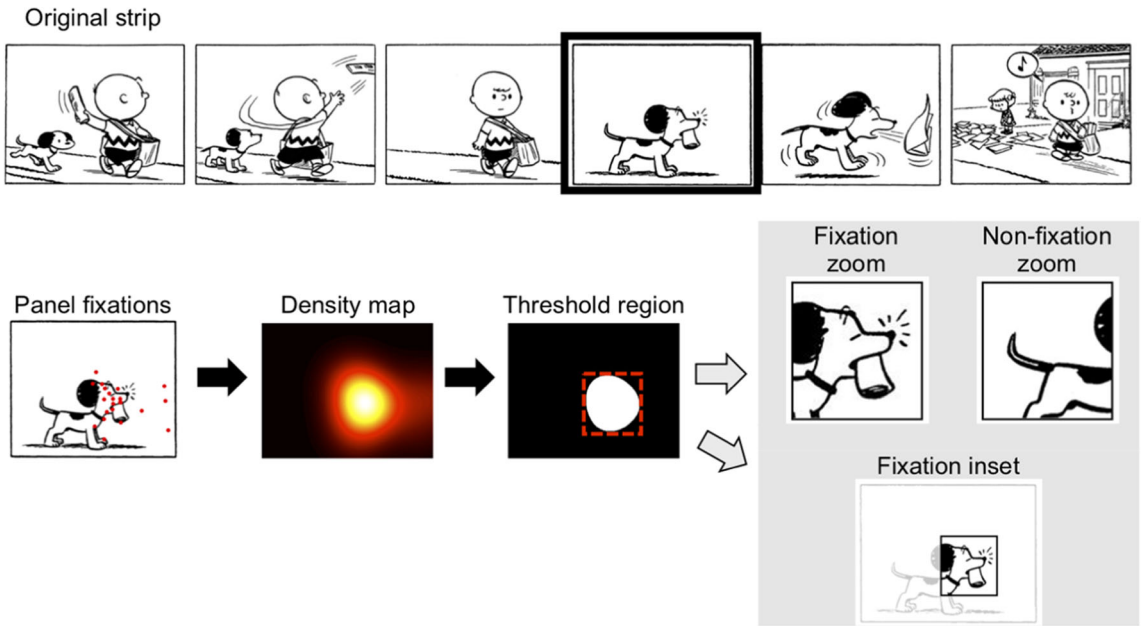

Fig. 2 Stimulus generation for Experiments 1–3. Original comic strips consisted of six panels (top row). The locations that were fixated in each panel were transformed into a density map that was thresholded to select panel content. This content was then used to automatically define modified panels (bottom right), which were cropped around fixations or control non-fixation regions. For more details, see the *Method* section. Peanuts artwork is © Peanuts Worldwide LLC

require assumptions about what these differences might be. We also investigated how changes to framing on one panel might affect processing of subsequent panels. If the VT on a panel mostly reflects the current content, then the framing of the previous panel should have little effect. If key information is selected by fixations, then fixation zoom panels might also be better at "priming" the interpretation of the subsequent panel. This would therefore be good evidence that information is built up over a sequence and that attention affects this accumulation.

Experiment 1: Zooming in on fixations

Our first experiment modified all the panels in a narrative. We compared viewing of automatically "zoomed-in" panels that were generated from fixations with control patches and the original images. This resulted in sequences of panels that were fixation zooms, non-fixation zooms, or a full-view of the original images.

Method

Participants

In all of the studies reported here, we aimed for a sample size of at least 50, which gives greater than 95% power for detecting a within-subjects effect on VT of the magnitude reported in Foulsham et al. (2016; for a t-test where $dz > 0.6$). Fifty-five participants from the University of Essex took part in Experiment 1 online in return for course credit. There were 39 females and the mean age was 21.0 years. As a measure of experience with comics, we also calculated participants' Visual Language Fluency Index (VLFI), derived from the self-reported frequency with which they read comics now and while growing up (Cohn et al., 2012a). Participants showed a range of VLFI scores (2–26) with a mean of 9.0, considered a low average.

Stimulus generation

As described above, our stimuli were created based on a previous eye-tracking experiment (Foulsham et al., 2016). In that experiment, participants viewed comic strips while we tracked their eye movements. We chose 24 strips, each of which had six panels depicting a wordless narrative about Snoopy and other characters from Charles Schulz's *Peanuts* series (see Fig. 2 for an example). All images were grayscale, with the same size and resolution.

To define our zoomed-in panels, we used all the fixations made by 14 observers in the "original" condition of Experiment 2 from Foulsham et al. (2016). In this condition, participants saw all the panels on a single screen before pressing a key to advance to the next sequence. There were an average of approximately six fixations per panel per participant (excluding the first fixation on the strip, which was constrained to be in the centre). We then developed an automatic workflow to crop panels based on these fixations (see Fig. 2). Image processing was accomplished in MATLAB and consisted of the following steps. First, the fixations from all observers were used to define a continuous density map (visualized as a heatmap; e.g., Wooding 2002). Maps were made by adding a two-dimensional Gaussian at the location of each fixation, producing a landscape of peaks and troughs corresponding to the number of fixations at each point. In this implementation we used a Gaussian kernel of approximately 2° in width, which allows for eye-tracker error and the visibility of information around the fovea. We did not take into account fixation duration in this implementation, although this would be straightforward.

Next, we used a threshold to select regions with the highest fixation density. We selected those pixels that were in the top 10% of the density map. We then used the MATLAB function "bwlabel," which uses connected component labeling to extract blobs from a binary image. In many cases this resulted in a single region reflecting the peak of the distribution, but in panels where there were multiple regions, we chose the one with the largest area. The closest-fitting rectangular bounding box was defined around this key region, and cropped to form the zoomed panel. With the addition of the bounding box, the selected regions encompassed 13.1% of the original panel's area, on average (SD = 1.5%). For displaying, zoomed panels were enlarged to the height of the original panel and a black border matching the original was added.

The workflow described above was automatic and data-driven. We did not manually label or pre-select any of the zoomed panels. It remained an open question, therefore, whether the fixation zoom panels would be at all meaningful.

In order to compare fixated content with other regions, we used a non-fixation control condition. Non-fixation zooms were created in the same way, but so as to select content that was *not* highly fixated. Choosing a non-fixated region using the minima of the fixation distribution proved difficult because it was a very different shape and size from the peaks. Instead, we defined a region that was the same shape and size as the fixation-zoom region but with minimal overlap. The number of overlapping pixels at all possible locations was calculated, and the selected control region was chosen randomly from those locations with the least overlap. The result was a "non-fixation zoom" panel that was the same size and shape as the fixation zoom but did not contain details that had been frequently fixated (see Figs. 2 and 3 for examples). To confirm that these panels were different in terms of the attention that they received, we calculated that the average percentage of fixations on the fixation zoom regions was 70%, while it was only 4% on the non-fixation zoom regions (using the data from Foulsham et al., 2016).

Experimental stimuli and design

The type of panel was a within-subjects factor with three conditions: full panel, in which original panels were presented; fixation zoom, in which all six panels were replaced with the fixation zoom; and non-fixation zoom, in which all six panels were replaced with the non-fixation zoom (see Fig. 3). Each participant saw all 24 strips, divided equally between the three conditions. Across participants, we counterbalanced stimuli and conditions into three sets using a Latin square design, such that each participant saw each strip once, but across the experiment all strips appeared in all conditions.

Procedure

The experiment was programmed and controlled using jspsych (de Leeuw, 2015) and Qualtrics, and conducted within each participant's internet browser. Jspsych uses JavaScript to present stimuli and record responses, with good timing performance (de Leeuw & Motz, 2016).

Participants viewed each sequence in a self-paced viewing task. They were asked to try to understand the narrative, one panel at a time, pressing a key on the keyboard when they were ready to move to the next panel. Each panel was presented in the centre of a white background at a size of 222 pixels vertically. After each strip, participants rated their comprehension using a Likert-type scale ("How easy was it to understand this comic strip?"), labeled from 1 ("very difficult") to 7 ("very easy"). After pressing a number on the keyboard, the next sequence began. All 24 strips were displayed in a random order, with the three conditions randomly interleaved. After viewing all the strips, participants answered the VLFI questions about comic-reading experience.

Results

Data processing

Data and analysis code are available online at https://osf.io/qf5ev/.

The raw data consisted of VTs for each panel and self-ratings of ease of comprehension for each strip. VT was recorded by the browser as the time from panel display to the participant's advancing key press. As is common with response time distributions, VT was positively skewed and we were also concerned about potential heterogeneity from the online data collection. The fifth and 95th percentiles across all panel viewing times were 288 ms and 3,623 ms, respectively. On the basis of this distribution, we chose to exclude outliers with VTs lower than 200 ms (which would likely be anticipatory errors) or higher than 5,000 ms (unusually delayed responses). This was 4.5% of all panel VTs. As conservative measures, we excluded the whole strip where any of the six component panels were outliers, and we excluded four participants who had less than 50% of their VT data remaining after this step.

Means are reported alongside within-subjects 95% confidence intervals (CIs; calculated according to the Cousineau-Morey method; Morey, 2008). For statistical analysis, VTs were transformed using the reciprocal 1/x to correct for skew. Data were analyzed with linear mixed effects (LME) models, allowing us to control for varying intercepts by participant and strip. Models were fit using the lme4 package (Bates et al., 2015) in R and used treatment coding to contrast the fixation zoom and non-fixation zoom conditions with the full panel strips. We report p-values based on likelihood ratio tests between different models. Predictors associated with a t-value of greater than 2 were considered statistically significant.

Panel viewing time

Figure 3 summarizes the panel VT data across the whole sequence. We plot the mean VT and 95% within-subjects CIs across participants.

On average, the fixation zoom condition led to longer VTs than the non-fixation zoom condition (means [95% CIs] per panel of 1,019 ms [983–1,054] and 857 ms [799–914], respectively), but shorter VTs than the original full panels (1,084 ms [1,038–1,129]). We fit an LME model, predicting transformed VT with random effects of participant and strip. A fixed effect of condition improved this model ($\chi^2(2) = 71.0$, $p < .001$). Compared to the full-panel condition, the non-fixation zoom panels were viewed more quickly ($\beta = 0.00033$, SE = 0.000039, t = 8.4). The fixation zoom condition was also viewed more quickly, though here the difference was less pronounced ($\beta = 0.000079$, SE = 0.000039, t = 2.0).

All conditions show longer VTs at the start of the strip (and slightly elevated VTs on the final panel). Adding a continuous effect of panel index improved the LME model ($\chi^2(1) = 45.9$, $p < .001$), and the interaction between panel index and condition improved it further ($\chi^2(2) = 9.1$, $p = .01$). This emerged because, although all the conditions showed a similar pattern over the sequence, the drop-off in the non-fixation zoom condition was more pronounced. In fact, the fixation zoom panels were viewed in the same way as the full panels at all points in the sequence with the exception of the first and last panel.

Comprehension ratings

There were clear differences between self-reported ease of comprehension of the three different conditions, as rated on the 7-point scale. Although the fixation zoom condition (M = 4.2; 95% CI [4.0–4.4]) was rated as more difficult to understand than the full-panel condition (M = 6.1; 95 CI% [5.9–6.3]), it was rated as much easier compared to the non-fixation control (M = 1.8; 95% CI [1.6–1.9]). An LME model

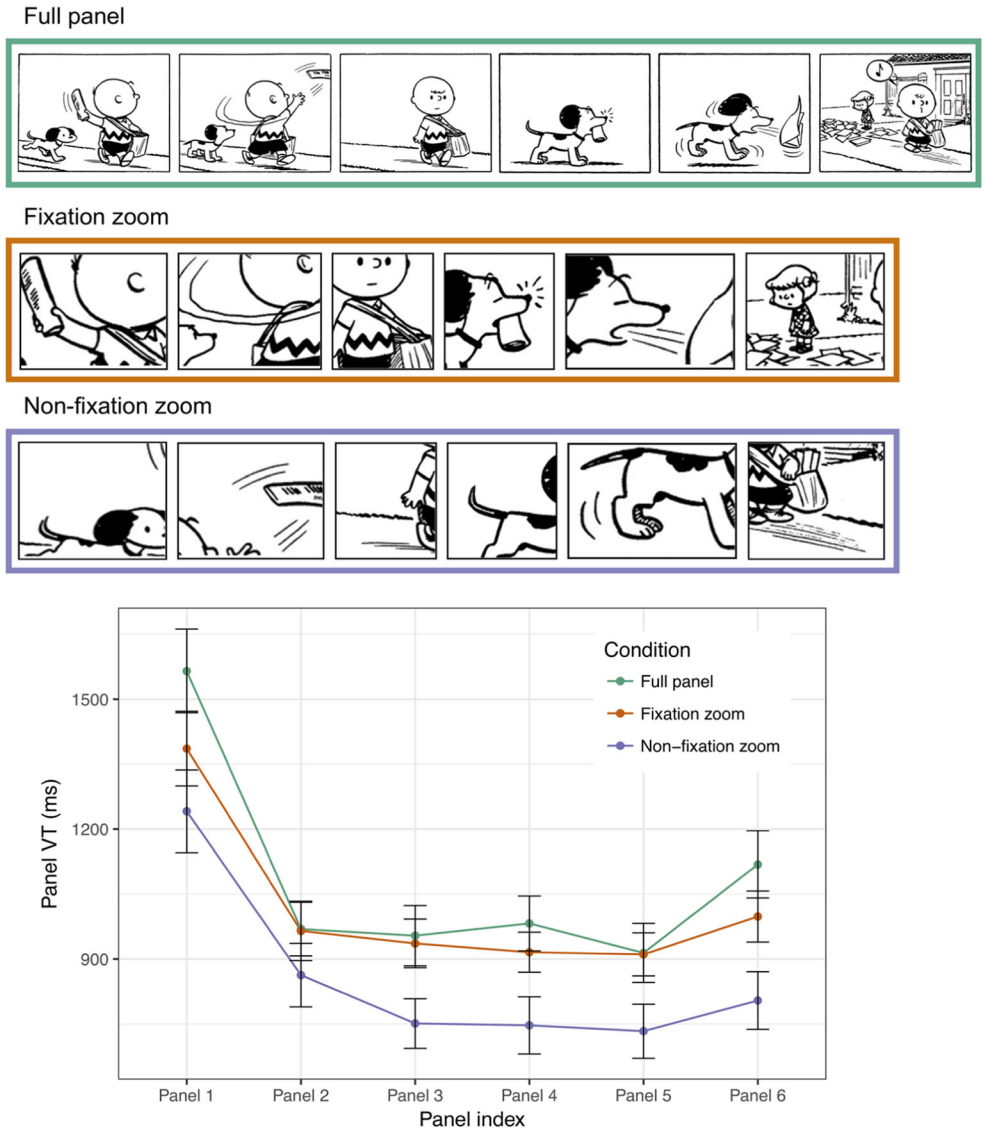

Fig. 3 Example stimuli from each condition (**top**) and viewing time (VT) results (**bottom**) from Experiment 1. VT is shown as a function of panel index (means, with 95% within-subjects confidence intervals)

predicting rating from the fixed effect of condition yielded a reliable model fit ($\chi^2(2) = 1{,}040.5$, $p < .001$). Both zoomed conditions were significantly different from the original full panel (fixation zoom: $\beta = -1.93$, SE = 0.099, $t = 19.7$; non-fixation zoom: $\beta = -4.25$, SE = 0.099, $t = 43.1$).

The condition associated with lower ease ratings (non-fixation zoom) also elicited the shortest VT. This is not what is normally expected in terms of fluency of comprehension. Our interpretation is that participants skipped quickly through the non-fixation zoom panels that may have contained blank space or uninformative details. When we entered comprehension rating as a continuous, fixed predictor of (transformed) strip VT there was only an interactive effect with condition ($\chi^2(3) = 33.9$, $p < .001$). In the full-panel condition, there was a small and non-significant negative relationship between VT and comprehension rating, such that strips that were rated as easier to understand were viewed more quickly ($t = 1.87$). In contrast, in the non-fixation zoom condition, ease was a significant *positive* predictor ($t = 3.47$). In this condition, strips that were rated as easier to understand were viewed for longer. As in our other measures, the fixation zoom condition was between these extremes, with a negligible and non-significant relationship in this condition ($t = 0.47$).

Individual differences in comic reading

Since we had information on participants' comic expertise (the VLFI questionnaire), we investigated whether this was correlated with performance. There was a negative correlation between VLFI score and mean VT, $r(49) = -.32$, $p = .02$. In contrast, there was a positive correlation with mean comprehension rating, $r(49) = .42$, $p = .002$. Participants with higher

expertise tended to view the panels more quickly and rated them easier to understand. We also repeated the previously reported LME models, including VLFI as a fixed effect, but the results were unchanged.

Discussion

This experiment examined visual narratives where each panel depicted a full scene, a fixation zoom, or a non-fixation zoom. The non-fixation zoom condition was rated as most difficult to understand. An informal inspection of the panels in this condition shows that they are more likely to contain background objects and empty space, and less likely to contain character's faces, and this is probably why they did not enable easy comprehension. Laubrock et al. (2018) found that observers tended to focus on faces within comics. The fixation zoom condition, in which panels were cropped to the most fixated features, produced much higher comprehension ratings (towards an "easy" estimate on our self-report measure). At least some of the time, therefore, the fixation zoom panels provided enough information for observers to rate them as being easy to understand. This is good evidence that fixations on a visual narrative are directed at regions that are, on average, more informative than non-fixated regions (see Mackworth & Morandi, 1967, and Foulsham & Kingstone, 2013, for similar logic in single photographs).

We had anticipated that fixation zooms would be viewed for less time because they "pre-selected" important information and were easier to understand. In fact, we found the opposite, with these panels viewed for more time, on average, than the non-fixation zooms. There are two related explanations for this pattern. First, panels that had less meaningful content were read quickly because there was less to process. Second, participants simply skipped panels that were not easily interpretable. Participants were likely not very motivated to try and interpret non-fixation zoom panels. When we investigated the change in VTs across the six-panel sequence, we found that the non-fixation zoom condition got faster over time, and more so than in the other conditions. This supports the idea that participants in that condition may have started skipping quickly through these panels when it became obvious they were less informative.

All three conditions showed a clear pattern over the sequence, with prolonged VTs on the first panel. This pattern has been observed in other studies (Cohn et al., 2012a; Foulsham et al., 2016), and likely reflects a process of "laying the foundation" for a new narrative (e.g., establishing the setting and characters). After the first panel, the fixation zoom and the full-panel condition were viewed at similar speeds. The last panel also showed evidence for slightly longer VTs, which may be a "wrap-up" effect as reported elsewhere (Cohn & Wittenberg, 2015; Foulsham et al., 2016) or an extra process associated with the subsequent rating task.

The full-panel condition was rated as easiest to understand and viewed for the longest time. The most fixated regions were helpful, but not sufficient to enable normal processing. This means that the background information contributed to understanding, either because there were multiple foci within the image or because peripheral information gave a useful context.

Experiment 2: Key panel fixation zoom

Restricting the visual content of the whole strip had a large effect. However, the manipulation would have been fairly obvious from the very first panel, which may have encouraged skipping behavior in the non-fixation zoom condition. In published visual narratives, restrictions in viewpoint are not typical across a whole sequence. It is more common for authors to toggle framing, by providing wide viewpoints in some panels, and focal viewpoints in others (Cohn, Taylor-Weiner, & Grossman, 2012b).

Each panel plays a narrative role within a sequence, and manipulating this content will have different effects depending on where they occur in this structure (Cohn, 2020). The rationale of Experiment 2 was to replicate the effects of fixation zooms using a more subtle manipulation on a single panel. We ask whether this manipulation will affect VTs and comprehension ratings, and whether these will "spill over" onto subsequent panels. We also use an extra control condition, inserting an incongruous, zoomed panel from another strip. This condition will reproduce the change in framing, and contain details that were fixated, but it should not provide useful information for the current strip.

Method

Participants

Sixty-one participants (33 female) took part in this online study, responding to adverts on social media or for payment on prolific.ac. As this was a non-student sample, there was a wider range of ages than in Experiment 1 (18–72 years; M = 35.8 years). VLFI scores showed a range of experience (2–43), with a slightly higher average than in Experiment 1 (M = 13.2).

Stimuli and design

The same comics (24 strips of six panels each) were used as in Experiment 1, and they were counterbalanced in the same way. In this experiment only a single panel in each strip was modified.

We chose the critical panel according to the narrative structure (Cohn, 2020). Our focus was on panels that begin a sequence of actions (e.g., Snoopy the dog beginning to sneeze; categorized as an "Initial") or manifest a climax (e.g., Snoopy

sneezing and dropping the newspaper; categorized as a "Peak"; see Figs. 2 and 3). We altered one Initial panel in each strip. These panels are key in that they "set up" subsequent actions, and each Initial has a clear relation with the next Peak, often sponsoring anticipations for what happens next (Cohn & Paczynski, 2013; Cohn et al., 2017a). We can therefore look for effects on both the manipulated panel and the subsequent peak. The ordinal position of the critical panel varied and was unpredictable, but was never the first or last panel.

The 24 strips were divided equally between four conditions (see Fig. 4). Full-panel, fixation zoom and non-fixation zoom conditions were as described in Experiment 1 (but only applied to the critical panel with all other panels being unchanged). The fourth condition was an incongruous zoom, where we swapped the original with a fixation zoom from a randomly chosen, *different* strip. This condition introduced a panel that had a change in viewpoint (because it was a zoom) and meaningful content (because it was based on fixated regions) but that would not fit the overall narrative. The incongruent zooms will also match the fixation zooms, on average, in terms of low-level features and complexity (since they are the same panels). As in Experiment 1, the four conditions were counterbalanced across participants so that each individual strip appeared with all four types of manipulation.

Procedure

The procedure and experimental program was exactly the same as in Experiment 1. Participants saw all strips in a random order, with conditions interleaved. We recorded VTs on each panel and ratings for each strip.

Results

Outliers were excluded using the same criteria as in Experiment 1, which removed 5.6% of panels. Eight participants were excluded because they did not finish or had fewer than 50% of trials remaining after outlier removal. The remaining sample consisted of 53 participants and approximately 1,500 trials per condition.

Overall viewing time

We began by examining VT for the whole strip (i.e., the sum across all six panels). Only one panel in each of the strips was different between conditions, and the other five panels were identical. Participants viewed strips in the full-panel condition for a mean of 9.6 s (95% CI [9.3–9.9]). This was similar in the fixation zoom condition (9.5 s [9.2–9.8]), but prolonged in the non-fixation zoom (10.1 s [9.8–10.4]) and incongruous (10.3 s [10.0–10.6]) conditions.

Modelling the fixed effect of condition on reciprocal-transformed VT led to an improved fit over and above a null model (with random effects of participant and strip; $\chi^2(3) = 30.1$, $p < .001$). Model estimates (see Table 1) confirmed that the non-fixation zoom and incongruous zoom conditions were viewed for longer than the full panels. The fixation zoom condition did not differ significantly from the original full panel.

By manipulating a single key panel we therefore succeeded in producing slower VTs when the content was incongruous or showed non-fixated items. Critically, there was no penalty when a zoomed in version was created based on the most fixated content. This suggests that the fixation zoom contained all the necessary information to understand the narrative.

Viewing time on key panels

All conditions showed a pattern of VTs across the sequence that was similar to Experiment 1, with the longest VTs on the first panel. Condition VTs were indistinguishable at the beginning of the sequence (as expected, since these panels were

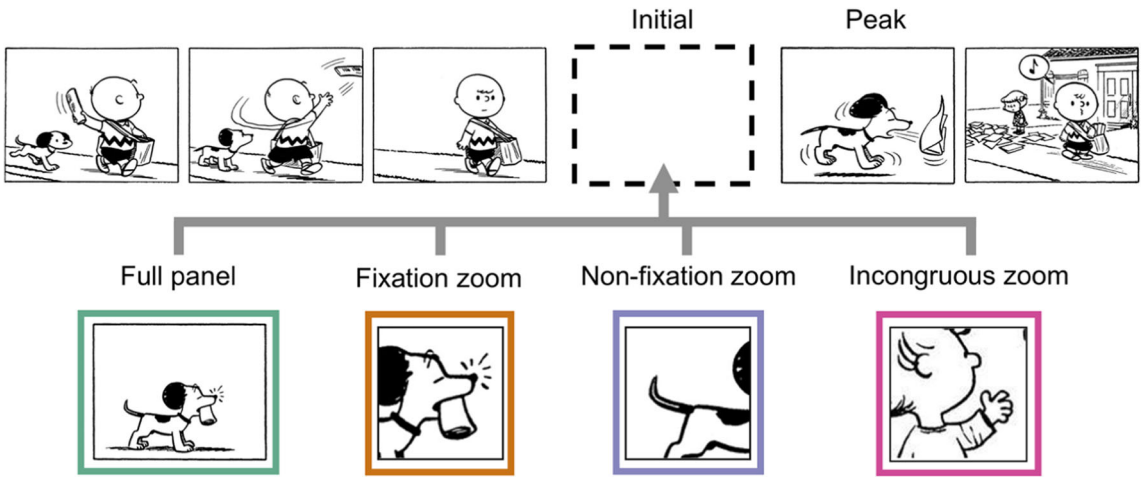

Fig. 4 Example stimuli from Experiment 2. The critical panel (here, panel 4) was either the original full panel or was replaced with one of three different zooms. This panel was an "Initial," setting up an action that climaxed in the following "Peak" panel

Table 1 Linear mixed effects (LME) model for predicting viewing times (VTs) from condition. The dependent variable is reciprocal-transformed VT, and hence negative estimates indicate an increase in untransformed VT. Levels are compared to the reference category, the full-panel condition

Predictor	β	*SE*	*t*
Intercept	0.00012	0.0000061	19.62*
Condition (Fixation zoom)	0.00000015	0.0000023	0.06
Condition (Incongruous zoom)	-0.000010	0.0000023	4.38*
Condition (Non-fixation zoom)	-0.0000074	0.0000023	3.24*

Asterisks indicate statistically significant predictors

identical) but diverged by the fourth panel (i.e., after the manipulated panel had been presented in most strips). We analyzed the VT on the key panel (i.e., the one that had been altered). This panel was chosen as an "Initial" panel to narratively set up the following "Peak" panel. If self-paced VT on a panel only reflects the processing of the current information, then we would expect results at the strip level to be driven by changes on the manipulated panel (N). Alternatively, changes to the VT of the Peak panel (N+1), which is identical in all conditions, would indicate the effect of prior information or integration within the narrative. Such differences are also important at the Peak because it has been posited to motivate the meaning of the whole strip (as its climax), while the preceding manipulated Initial functions to set up this information.

Figure 5 shows the VTs for both types of panel. Peaks took slightly longer to read than initial panels, on average. More importantly, condition had an effect on the time spent viewing both types of panels (both $\chi^2(3) > 38$, $p < .001$). Table 2 shows the LME model estimates, predicting transformed VT from condition, with random effects of participant and strip. Both panels showed the same pattern, with incongruous and non-fixation zoom panels being viewed longer than the full panel. In both cases, fixation zoom panels were not reliably different from the original full panel.

Comprehension ratings

Our subtler manipulation in this experiment led to smaller differences in the self-rated ease of comprehension. These differences corresponded to the VT data, such that the condition with the slowest VTs (incongruous zoom; M = 4.2; 95% CI [4.0–4.4]) was rated as less easy to understand than the non-fixation zoom condition (M = 4.6; 95% CI [4.4–4.8]), the fixation zoom condition (M = 5.2; 95% CI [5.1–5.4]) and the full-panel condition (M = 5.5; 95% CI [5.3–5.7]). Condition was a significant predictor of ease rating ($\chi^2(3) = 143.7$, $p < .001$). Fixation zooms resulted in lower ratings than the full-panel condition (β = -0.30, SE = 0.12, t = 2.5). However, incongruous (β = -1.3, SE = 0.12, t = -11.4) and non-fixation zoom (β = -0.89, SE = 0.12, t = 7.6) strips received much lower ratings than both the other conditions.

Individual differences in comic reading

Unlike in Experiment 1, in this experiment VLFI scores did not correlate significantly with either overall VT ($r(51) = .12$, $p = .38$) or comprehension ratings ($r(51) = .07$, $p = .61$). Including VLFI scores as a fixed effect in the LME models made no difference to the results.

Panel content and faces

The results so far suggest that fixation zoom panels are processed in a similar way to the full panels, and in this experiment they were read faster than the non-fixation zooms. Fixation zooms also primed the subsequent Peak panels to a greater extent than non-fixation zooms.

Our approach – automatically defining zooms based on separate participants' fixation distribution – does not require assumptions about what particular features comprise the key information. Nonetheless, we can examine the differences in content between zoom conditions and investigate how these differences affect VTs. Informal inspection reveals that the non-fixation zooms are less likely to contain full objects and characters than the fixation zooms (see Figs. 2, 3, and 5 for some examples). The non-fixation zooms are also more likely to contain background and empty space (although they did often include regions of detail or less important characters, as in Fig. 3).

We chose one candidate feature – faces – which are known to attract attention in many situations (Foulsham et al., 2010; Yarbus, 1967), including in visual narratives (Laubrock et al., 2018). First, we manually coded all the zoom panels according to whether they contained a face. Eighty-two percent of the fixation zoom panels contained a full or partial face, compared with only 13% of the non-fixation zoom panels (a large and statistically significant difference: $\chi^2(1, N = 288) = 136$, $p < .001$). This confirms that fixations are frequently made to regions containing faces.

Next, we estimated some supplementary models, including the dichotomous variable of face presence as a fixed predictor of VT. We compared the fixation zoom and the non-fixation zoom conditions, and we asked whether the difference between these conditions was moderated by face presence. In Experiment 1, adding face presence (and its interaction) to the fixed effect of condition improved the LME model ($\chi^2(2) = 21.5$, $p < .001$). Face presence was a significant predictor (β = 0.00020, SE = 0.000039, t = 5.0), but this was qualified by an interaction with condition (β = 0.00046, SE = 0.000060, t = 7.8). When both zoom panels contained a face, they were

Table 2 Linear mixed effects (LME) model for predicting viewing times (VTs) from condition. The dependent variable is reciprocal-transformed VT, and hence negative estimates indicate an increase in untransformed VT. Levels are compared to the full-panel condition

Dependent variable	Predictor	B	*SE*	*t*
Panel VT (Initial)	Intercept	0.00085	0.000045	18.84*
	Condition (Fixation zoom)	-0.000010	0.000024	0.44
	Condition (Incongruous zoom)	-0.00013	0.000023	5.40*
	Condition (Non-fixation zoom)	-0.000079	0.000023	3.41*
Panel VT (Peak)	Intercept	0.00078	0.000043	17.89*
	Condition (Fixation zoom)	-0.000034	0.000020	1.74
	Condition (Incongruous zoom)	-0.00012	0.000020	6.32*
	Condition (Non-fixation zoom)	-0.000097	0.000020	4.98*

Asterisks indicate statistically significant predictors

viewed for approximately the same amount of time ($\chi^2(1) < 1$). However, when the zoom panels did not contain a face, the fixation zoom panels were viewed for longer (the pattern observed in the original analysis; $\chi^2(1) = 93.6$, $p < .001$). This indicates that, while face presence makes a difference to VTs, even in the absence of faces the fixation zooms were processed differently.

In the data from Experiment 2, face presence contributed to the difference between fixation zoom and non-fixation zoom initial panels ($\chi^2(2) = 7.9$, $p = .019$). Zoom panels with faces were viewed for less time ($\beta = 0.00015$, SE = 0.000059, t = 2.6) and there was a non-significant interaction between condition and face presence ($\beta = 0.\ 000086$, SE = 0.000076, t = 1.1). As in Experiment 1, the difference between conditions remained when analysis was restricted to panels without a face. VTs on Peak panels were not affected significantly by the presence of a face in the preceeding zoom panel ($\chi^2(2)=2.9$, $p=.235$).

Discussion

This experiment showed that altering the framing, even in just a single panel, made a reliable difference to how comic strips were viewed and rated. Even though most of the content was exactly the same in the different conditions, there was a "cost" of approximately 500 ms when one of the panels was replaced

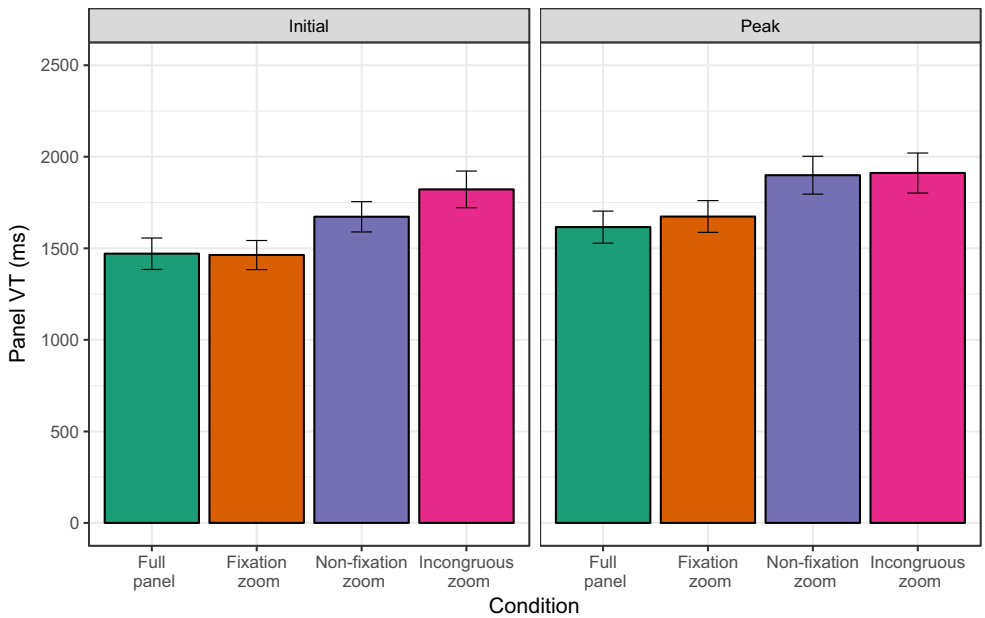

Fig. 5 Viewing times (VTs) for the manipulated panel (the Initial) and the following Peak in each condition. Bars show the mean with 95% within-subjects confidence intervals

with an incongruous panel or with non-fixated information. This is consistent with other costs shown to incongruities in visual narratives (Cohn & Wittenberg, 2015; West & Holcomb, 2002). Importantly, this cost was not observed with a fixation zoom, indicating that the fixated information on the Initial panel was sufficient. The relationship between self-reported ease and VT was more straightforward in this experiment and we did not observe the skipping behavior seen in Experiment 1 (presumably because most of the panels were not modified).

By looking at VTs we were able to examine both the direct effect on the critical panel and the influence on the subsequent panel. Because these panels formed a constituent (an initial and a peak, together illustrating an action subsequence), we had a good theoretical basis for further investigation. As expected, inserting an incongruous zoom from another strip led to elevated VTs on that panel, consistent with readers trying to resolve this out-of-place information (Cohn & Wittenberg, 2015; West & Holcomb, 2002). The incongruous zoom panels were equal to the fixation zoom panels in terms of complexity and framing, but they were not helpful for viewing the sequence quickly, and they were rated as the least easiest to understand. A non-fixation zoom also led to longer VTs, presumably because participants had to spend more time trying to understand the degraded information and integrating it with the sequence. The constrast between the fixation zoom condition and the control conditions demonstrates that VTs are not only sensitive to the amount of information in a panel, or its framing, but also to the information in context. Additional analysis of the zoom panels demonstrated that the fixation zooms were more likely to contain faces, but that this was not sufficient for explaining their advantage over non-fixation zooms.

Our manipulations also produced a "spill-over" effect on VTs for the next (unchanged) Peak panel. This would not happen if participants were inspecting each panel in isolation and it illustrates how visual narratives build over a sequence. Peak panels were viewed for more time when the preceding Initial panel was changed or degraded. When the Initial panel was intact (the full-panel condition) or focused on the key information (the fixation zoom condition), Peak panels could be processed more quickly. This is likely because the Initial had begun to set up the event that climaxes in the Peak (Cohn & Paczynski, 2013; Cohn et al., 2017a). Our results show that focal information as selected by fixations is sufficient for this.

Experiment 3: Key panel fixation inset

The VTs in Experiment 2 suggested that a fixation zoom provided enough information for normal comprehension. However, strips with a fixation zoom were still rated as slightly less easy to understand. This may be because background is useful for providing the context to a panel, even though it is not frequently fixated. "Zooming in" on particular content is only one of several ways in which framing in comics can draw attention to elements within a panel. An alternative is to highlight parts of a scene with an "inset" panel, which is a naturalistic convention of comics where content becomes emphasized with an outline and presented over the original background to form a "panel in a panel" (Fig. 6). We therefore carried out a third study, with two aims. First, we aimed to demonstrate that inset panels could also be automatically generated from fixations. Second, we sought to replicate the effects of fixated content from Experiment 2, while also providing the wider context. A recent model suggests that one way in which narrative understanding can affect perception is by prompting participants to search for particular information (Loschky et al., 2020). By selecting specific parts of a panel, insets could thereby serve as an explicit cue, drawing attention and facilitating the visual search that a comprehender might make during viewing.

Method

Participants

Fifty-seven undergraduate participants (49 female) from the University of Essex took part in this online study in return for course credit. The mean age was 21.3 years. VLFI scores ranged from 1 to 29, with a mean score of 7.8, indicating a low level of experience with comics.

Stimuli and design

The stimuli and design were the same as in Experiment 2, but with a different manipulation on the critical panel (see Fig. 6). To produce the fixation-inset condition, rather than cropping around the most fixated region, we drew a rectangular box around this region in order to create an "inset" panel similar to those sometimes used in comics (Cohn, 2013; Postema, 2013). The bounding box was drawn in black, and the gray-level of the surrounding, non-selected pixels was increased in order to fade out the background, while still leaving it visible. A non-fixation-inset condition was created in the same way, but with the inset selecting non-fixated content (as described in Experiment 1).

The 24 strips were divided equally into three conditions: full panel, fixation inset, and non-fixation inset. Only the critical panel was manipulated. Conditions and strips were counterbalanced so that each participant saw all strips only once, but across the whole experiment each strip appeared in all conditions.

Procedure

The internet-browser-based procedure was the same as previously. All strips were presented in a random order and we recorded VTs via self-paced keyboard presses and ratings of comprehension difficulty.

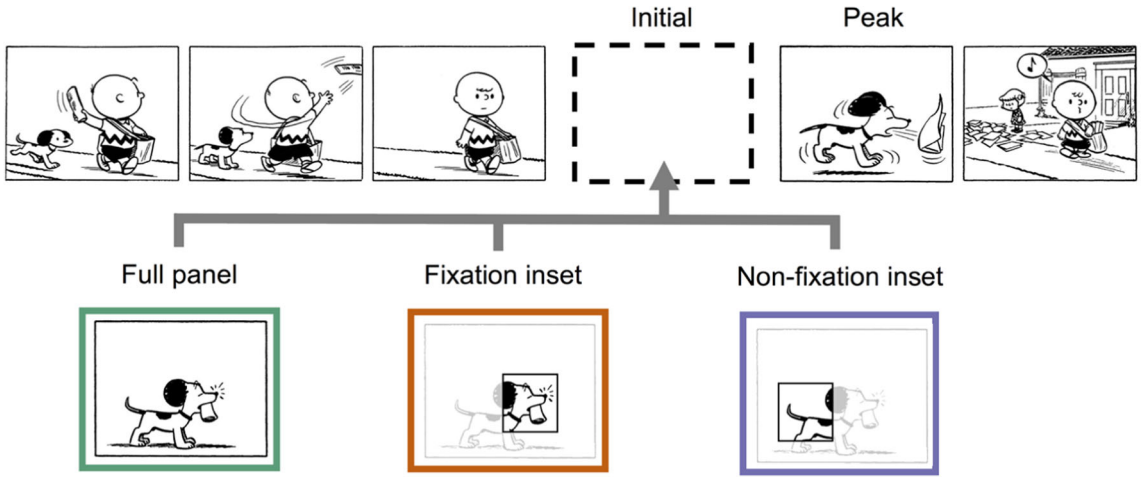

Fig. 6 Example stimuli from Experiment 3. The critical panel was either the original full panel or was replaced with an "inset" panel that highlighted fixated or non-fixated content

Results

As in previous studies, a small number (5.7%) of panels were excluded as outliers. Five participants were excluded with fewer than 50% of strips remaining after this step, leaving a sample of 52 participants.

Overall viewing time

Participants viewed strips in the full-panel condition for a mean of 6.5 s (95% CI [6.3–6.7]). VTs were longer in the fixation-inset (6.9 s [6.6–7.2]) and non-fixation-inset conditions (7.3 s [7.0–7.5]). LME analysis demonstrated a significant effect of condition on transformed VT ($\chi^2(2) = 21.9$, $p < .001$). Both inset conditions resulted in significantly slower self-paced reading (Fixation inset: $\beta = -0.000013$, SE = 0.0000045, t = 3.0; Non-fixation inset: $\beta = -0.000021$, SE = 0.0000045, t = 4.6).

Viewing time of key panels

The first panel in each sequence was viewed for longer, and, as in Experiment 2, conditions diverged after the manipulated panel. We examined the responses separately for the modified Initial panel and the subsequent Peak (see Fig. 7).

Inset panels had a disruptive effect, especially in Initial panels. In these panels, condition was a significant predictor of transformed VT ($\chi^2(2) = 105.7$, $p < .001$). The fixation-inset condition was associated with significantly longer viewing times than the full-panel condition ($\beta = -0.00028$, SE = 0.000041, t = 6.8). The longest VTs were seen in the non-fixation-inset condition, which was also significantly different from the full-panel condition ($\beta = -0.00043$, SE = 0.000042, t = 10.4).

In Peak panels, condition was also a reliable predictor ($\chi^2(2) = 35.0$, $p < .001$), with the same pattern of VTs. Preceding fixation inset panels led to a subsequent Peak to be viewed for longer than those following a full panel ($\beta = -0.00012$, SE = 0.000035, t = 3.3), as were those following panels with non-fixation inset panels ($\beta = -0.00021$, SE = 0.000036, t = 6.0). Follow-up LMEs confirmed that in both Peak and Initial panels the fixation-inset condition was viewed for less time than the non-fixation-inset condition.

Comprehension ratings

Condition was also a significant predictor of the comprehension ratings given to each strip ($\chi^2(2) = 14.3$, $p < .001$). However, the differences in this experiment were smaller than in Experiments 1 and 2, and all of the conditions received a mean rating greater than 5, which was not the case in the previous studies. This is likely because, even in the non-fixation-inset condition, the full-panel content was still visible. The non-fixation-inset condition (M = 5.4; 95% CI [5.2–5.6]) was associated with significantly lower ratings than the full-panel condition (M = 5.7; 95% CI [5.5–5.9]; $\beta = -0.338$, SE = 0.091, t = 3.7). Ratings in the fixation-inset condition (M = 5.6; 95% CI [5.4–5.7]) were not significantly different from the unmodified strips ($\beta = -0.094$, SE = 0.091, t = 1.0).

Individual differences in comic reading

Correlations between VLFI scores and performance measures were weak and did not reach statistical significance (overall VT: r(50) = .19, p = .16; comprehension ratings: r(50) = .21, p = .14). As in previous experiments, including VLFI scores as a fixed effect in the LME models made no difference to the results.

Discussion

The results of this experiment showed again that manipulating a single panel was enough to change comic viewing. Inset panels had an effect, even though the whole panel was still available to be inspected. Directing people to look at a certain place was

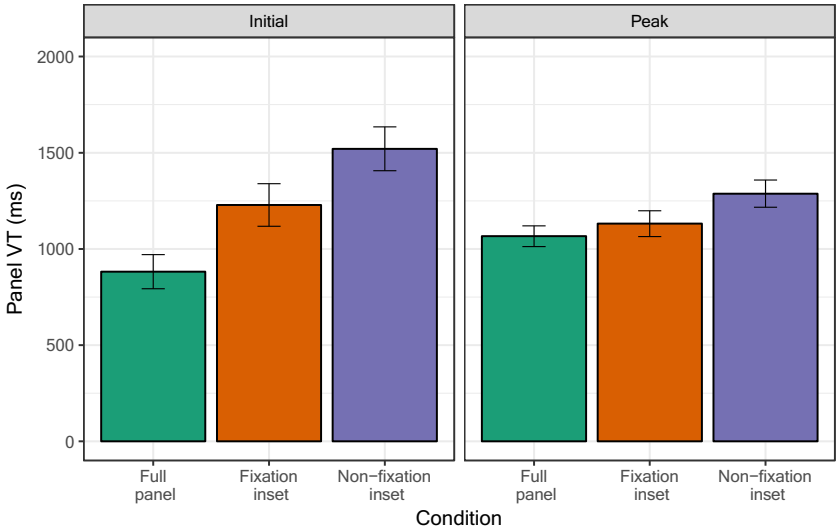

Fig. 7 Panel viewing times (VTs) in Experiment 3, for the manipulated initial and the subsequent peak panel

disruptive, especially when the inset was based on non-fixated regions. In some ways, the insets increase the complexity of the image –adding a sub-panel—which may account for the longer VTs in these conditions. Nevertheless, fixation inset panels were more helpful than highlighting unfixated regions. As in Experiment 2, the time spent viewing a subsequent Peak panel was affected by the content of the preceding Initial, demonstrating how the cues of one image can affect the processing of subsequent information (Cohn & Paczynski, 2013; Foulsham et al., 2016).

General discussion

These experiments tested whether panels can be framed in a way that mimics visual attention, by focusing on key elements and editing out less important information. These manipulations were based on the prediction that overt attention – as measured by fixation positions – will select important content.

Results across the three experiments were consistent. Modified panels that were based on content fixated by a different set of observers were rated as easier to understand than panels from a non-fixated region or from a different comic strip. Fixated information as zoom or inset panels was viewed in a more similar manner to the original un-modified strips. This is a striking result considering that the zoom panels contained only around 13% of the pixels from the originals. They were also automatically defined in a data-driven fashion, instead of being drawn and preselected by an artist.

Fixations select informative content within a visual narrative

We modified panels using a novel approach that automatically selected the peak of the fixation distribution. To our knowledge, this is the first time that fixation-driven content has been tested in this way, and particularly within the context of visual narratives and multiple images in a sequence.

Comprehension of the subset of information that is selected most often by eye movements provides a strong test of the degree to which fixations are allocated according to meaning. This question remains a topic of interest for researchers in visual attention and scene perception (Henderson & Hayes, 2017; Foulsham, 2015). At one extreme, it is possible that fixations might be allocated randomly, or regardless of the particular image. Given that visual narratives require rapid switching between panels, often with visual and textual elements, it is not unfeasible that a default strategy of responding to all panels the same way might be observed. Instead, our results suggest that fixations are clustered on areas of information, and that viewers are able to understand the narrative even when only the most-fixated parts are presented. This insight is similar to previous work investigating isolated natural scenes. Mackworth and Morandi (1967) divided photographs into 64 small regions and asked participants to rate their "informativeness." Participants tended to fixate the regions that were rated as being more informative, and they neglected other parts. More recently, subjective ratings have been used to create a "meaning map" for predicting fixations (Henderson & Hayes, 2017). One implication of the current study is that fixated regions presented in isolation are mostly sufficient for viewing these visual narratives at a normal pace. This insight is not possible from studies that correlate fixations and ratings, since they do not control where people look or the availability of information in peripheral vision.

It is notable that the advantage for fixated content differed across the experiments. In Experiments 2 and 3, fixation-based panels were viewed more quickly and rated as easier to understand, which is consistent with the interpretation of reaction times in self-paced reading. Our assumption in these

cases is that strips took longer to process when the important, fixated content was missing (non-fixation zoom panels). In theory, fixation zoom panels might even be viewed more quickly than the original drawn images, since they pre-select the key information and thus do not require participants to search around (Hutson et al., 2018; Loschky et al., 2020). However, that did not seem to occur here. In Experiment 3, fixation inset panels also identified the important content, although this may have been offset by the increased complexity of adding a border and making the background more difficult to see. In Experiment 1, where all six panels were modified, we found a different pattern of VTs and difficulty ratings, such that the conditions rated most difficult (the non-fixation zooms) were actually viewed for the shortest amount of time. This seems to be because participants moved quickly through the less-meaningful material. It also illustrates that while focused content might be sufficient sometimes, on other panels a wider composition, with background and context, is more helpful. Further research with the current technique could systematically test the role of background context at different points in a sequence.

Differences in features between zoom conditions

The generation of panels from fixations resulted in the editing out of pixels from the original comics. The self-paced viewing data provides one way of characterizing what was selected (the fixation zooms) and what was not (the non-fixation zooms). These data demonstrate that fixated regions are viewed in a more similar way to the full panel (and rated as easier to understand). In Experiment 2, fixated regions were also better at priming VTs on the following panel. These effects cannot be explained by the size or shape of the panels, since the non-fixation zooms had the same dimensions. Moreover, the incongruous zoom condition showed a completely different pattern of VTs in Experiment 2, despite being equivalent to the fixation zooms in terms of complexity and novelty.

What was the key information selected by the fixation zooms? Further research could examine this in detail by describing the differences between zoom conditions, perhaps using computer-vision methods. One clear difference was that the fixation zooms were more likely to contain characters' faces than the non-fixation zooms. Zoom panels containing faces were viewed for less time in Experiment 2. Thus one of the ways in which fixation zooms focused on key information was by selecting regions with faces, which tend to drive the narrative due to character's actions and facial expressions. However, the effects on VTs were not entirely explained by faces, since fixation zooms were viewed for less time even when not containing a face. The presence of a face in the zoom did not modulate the effect on the subsequent Peak panel. Thus fixations are also selecting other information-rich regions such as non-face objects.

Comprehension of panels across a sequence

Unlike looking at a single image, visual narratives must be understood by combining elements across a sequence. We observed that the first image in each sequence was inspected for a longer duration than subsequent panels, a pattern that has been reported previously in both visual narratives (Cohn & Wittenberg, 2015; Foulsham et al., 2016) and verbal narratives (Haberlandt, 1980). A likely explanation is "laying the foundation" (Cohn, 2019; Gernsbacher, 1990), an initial acquisition of information (setting and characters) that can be used to interpret subsequent information more quickly. It is interesting to note that we observed this pattern even in the non-fixation zoom condition, where all panels were replaced with uninformative information. Thus participants were still trying to make sense of what they saw, and they did so in a consistent fashion across the panels in the sequence.

Only a few studies have examined how specific "morphological cues" affect the comprehension of the narrative sequencing (Cohn et al., 2017a; Hutson et al., 2018; Tseng et al., 2018). In Experiments 2 and 3, we used our data-driven method to change the focus of one key panel. This key panel was an Initial, preparatory image followed by a Peak that depicted the climax. Previous work has often manipulated the Peak panels, thereby informing about how alteration of the primary information of a sequence changed its processing (Cohn & Wittenberg, 2015; Hutson et al., 2018). Here, manipulation of the Intial panel allowed us to see how anticipatory information might change the processing of subsequent climactic information (Cohn & Paczynski, 2013; Cohn et al., 2017a).

Our results showed clearly that viewing was disrupted both on the modified panel and on the subsequent Peak (which was identical in all conditions). When the Initial panel was incongruous, interpretation of the Peak suffered. This was also the case when the Initial highlighted non-fixated information. Importantly, the fixation zoom and fixation-inset conditions did not prolong viewing to the same extent (and in Experiment 2, the fixation zoom condition was viewed at the same speed as the original strips). The content selected by fixations was sufficient to prime comprehension of the second, Peak panel. The difference in results between Experiment 2 (where background content was removed in the fixation zooms) and Experiment 3 (where the fixated information was highlighted but background was still available) is also informative. It was easier to process only the focal content than a panel with focal and background content highlighted. This could be because peripheral information was fixated, even though it was not highlighted, and because it was harder to see since it was partially grayed out.

Challenges and limitations

As described above, there were many differences between the fixation and non-fixation zooms, and some of these "low-level" differences will have affected the VT (e.g., the complexity of the panel). Nonetheless, the pattern of VTs across experiments – with full panels and fixation zooms showing prolonged VTs in Experiment 1, and speeded VTs in Experiment 2—are most consistent with fixated regions being more meaningful in the narrative context. This raises the question of how fixations are guided to these key locations within a full panel. This guidance could be through recognition of scene elements in peripheral vision, or through low level cues such as salience, and artists may well manipulate these cues when composing images. Future research manipulating these factors separately could investigate guidance within panels in more detail.

We note that our measures are only indirect indicators of comprehension. As in studies of self-paced text reading, VT may reflect a number of processes beyond simple comprehension, and processing may also proceed differently in a naturalistic context (where all items in a sequence are available for "looking back" to). We also only have self-report measures of comprehension difficulty, and future studies could test understanding more directly (e.g., by asking participants questions about the narrative). The current stimuli may also be useful for controlled presentation in experiments with EEG, where the neuropsychological correlates of narrative processing and integration can be examined on-line.

It would also be possible to control the elements that are fixated within an image in other ways, such as through the use of symbolic arrow cues or onset cues, or by asking participants to follow a fixation dot. These could have the advantage of leaving peripheral regions visible, although they might add load from the additional task of following the cue. Since visual narratives already use zoom and inset panels, our approach offers a manipulation that is both natural and flexible in terms of the features that are highlighted or obscured.

Conclusion

Using a novel procedure, we constrained the framing of a visual narrative based on where previous observers had fixated. The results confirmed that fixated regions are informative for understanding a comic strip. Focusing on content in one panel also had an effect on the processing of subsequent panels. The elements that we pay attention to therefore both support and are supported by the narrative context.

Open practices statement The data and analysis code for all experiments are available at https://osf.io/qf5ev/ and Experiment 3 was preregistered.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. doi:https://doi.org/10.18637/jss.v067.i01.

Cohn, N. (2013). *The visual language of comics: Introduction to the structure and cognition of sequential images*. London, UK: Bloomsbury.

Cohn, N. (2020). Your brain on comics: A cognitive model of visual narrative comprehension. Topics in Cognitive Science, 12 (1), 352-386. doi: https://doi.org/10.1111/tops.12421.

Cohn, N., & Maher, S. (2015). The notion of the motion: The neurocognition of motion lines in visual narratives. *Brain Research*, 1601, 73–84. https://doi.org/10.1016/j.brainres.2015.01.018.

Cohn, N., & Paczynski, M. (2013). Prediction, events, and the advantage of Agents: The processing of semantic roles in visual narrative. *Cognitive Psychology*, 67(3), 73–97. https://doi.org/10.1016/j.cogpsych.2013.07.002.

Cohn, N., Paczynski, M., Jackendoff, R., Holcomb, P. J., & Kuperberg, G. R. (2012a). (Pea)nuts and bolts of visual narrative: Structure and meaning in sequential image comprehension. *Cognitive Psychology*, 65(1), 1–38. https://doi.org/10.1016/j.cogpsych.2012.01.003.

Cohn, N., Paczynski, M., & Kutas, M. (2017a). Not so secret agents: Event-related potentials to semantic roles in visual event comprehension. *Brain and Cognition*, 119, 1–9. https://doi.org/10.1016/j.bandc.2017.09.001.

Cohn, N., Taylor, R., & Pederson, K. (2017b). A picture is worth more words over time: Multimodality and narrative structure across eight decades of American superhero comics. *Multimodal Communication*, *6*(1), 19-37.

Cohn, N., Taylor-Weiner, A., & Grossman, S. (2012b). Framing attention in Japanese and American comics: cross-cultural differences in attentional structure. *Frontiers in psychology*, *3*, 349.

Cohn, N., & Wittenberg, E. (2015). Action starring narratives and events: Structure and inference in visual narrative comprehension. *Journal of Cognitive Psychology*, 27(7), 812–828.

Cutting, J. E., Brunick, K. L., & Candan, A. (2012). Perceiving event dynamics and parsing Hollywood films. *Journal of experimental psychology: human perception and performance*, *38*(6), 1476.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior research methods*, *47*(1), 1-12.

de Leeuw, J. R., & Motz, B. A. (2016). Psychophysics in a Web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. *Behavior Research Methods*, *48*(1), 1-12.

Foulsham, T. (2015). Scene Perception. In Fawcett, Risko & Kingstone (Eds.). *The Handbook of Attention*. MIT Press.

Foulsham, T., Cheng, J. T., Tracy, J. L., Henrich, J., & Kingstone, A. (2010). Gaze allocation in a dynamic situation: Effects of social status and speaking. *Cognition*, *117*(3), 319-331.

Foulsham, T., & Kingstone, A. (2013). Fixation-dependent memory for natural scenes: An experimental test of scanpath theory. *Journal of Experimental Psychology: General*, *142*(1), 41.

Foulsham, T., & Kingstone, A. (2017). Are fixations in static natural scenes a useful predictor of attention in the real world?. *Canadian Journal of Experimental Psychology*, *71*(2), 172.

Foulsham, T., Wybrow, D., & Cohn, N. (2016). Reading without words: Eye movements in the comprehension of comic strips. *Applied Cognitive Psychology*, 30, 566–579. https://doi.org/10.1002/acp.3229.

Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Lawrence Earlbaum.

Haberlandt, K. (1980). Story grammar and reading time of story constituents. *Poetics*, 9(1–3), 99–118.

Hagmann, C. E., & Cohn, N. (2016). The pieces fit: Constituent structure and global coherence of visual narrative in RSVP. *Acta Psychologica*, *164*, 157-164.

Hard, B. M., Recchia, G., & Tversky, B. (2011). The shape of action. *Journal of Experimental Psychology: General*, *140*(4), 586.

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, *1*(10), 743.

Hutson, J. P., Magliano, J. P., & Loschky, L. C. (2018). Understanding Moment-to-Moment Processing of Visual Narratives. *Cognitive science*, *42*(8), 2999-3033.

Inui, T., & Miyamoto, K. (1981). The time needed to judge the order of a meaningful string of pictures. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5), 393.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10), 1489-1506.

Kosie, J. E., & Baldwin, D. (2019). Attentional profiles linked to event segmentation are robust to missing information. *Cognitive Research: Principles and Implications*, *4*(1), 1-18.

Larson, A. M., & Loschky, L. C. (2009). The contributions of central versus peripheral vision to scene gist recognition. *Journal of Vision*, *9*(10), 6-6.

Laubrock, J., Hohenstein, S., & Kümmerer, M. (2018). Attention to comics: Cognitive processing during the reading of graphic literature. In A. Dunst, J. Laubrock, & J. Wildfeuer (Eds.), *Empirical comics research: Digital, multimodal, and cognitive methods* (pp. 239–263). New York: Routledge.

Loschky, L. C., Magliano, J., Larson, A. M., & Smith, T. J. (2020). The Scene Perception & Event Comprehension Theory (SPECT) applied to visual narratives. *Topics in Cognitive Science*, 12(1), 311-351.

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception & Psychophysics*, *2*(11), 547-552.

Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., & Dodd, M. D. (2011). Examining the influence of task set on eye movements and fixations. *Journal of Vision*, *11*(8), 17-17.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *reason*, *4*(2), 61-64.

Postema, B. (2013). *Narrative Structure in Comics: Making Sense of Fragments*. Rochester, NY: RIT Press.

Smith, T. J., Levin, D., & Cutting, J. E. (2012). A window on reality: Perceiving edited moving images. *Current Directions in Psychological Science*, *21*(2), 107-113.

Tseng, C. I., Laubrock, J., & Pflaeging, J. (2018). Character Developments in Comics and Graphic Novels: A Systematic Analytical Scheme. In A. Dunst, J. Laubrock, & J. Wildfeuer (Eds.), *Empirical comics research: Digital, multimodal, and cognitive methods* (pp. 239–263). New York: Routledge.

Underwood, G., Foulsham, T., & Humphrey, K. (2009). Saliency and scan patterns in the inspection of real-world scenes: Eye movements during encoding and recognition. *Visual Cognition*, *17*(6-7), 812-834.

West, W. C., & Holcomb, P. (2002). Event-related potentials during discourse-level semantic integration of complex pictures. *Cognitive Brain Research*, 13, 363–375.

Wooding, D. S. (2002). Eye movements of large populations: II. Deriving regions of interest, coverage, and similarity using fixation maps. *Behavior Research Methods, Instruments, & Computers, 34*(4), 518-528.

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