Attention, Perception, & Psychophysics (2018) 80:1011–1022
https://doi.org/10.3758/s13414-018-1484-0

Spatial legend compatibility within versus between graphs in multiple graph comprehension

Eva Riechelmann[1] · Lynn Huestegge[1]

Published online: 1 February 2018

Abstract

Previous research has shown that spatial compatibility between the data region and the legend of a graph is beneficial for comprehension. However, in multiple graphs, data–legend compatibility can come at the cost of spatial between-graph legend incompatibility. Here we aimed at determining which type of compatibility is most important for performance: global (legend–legend) compatibility between graphs, or local (data–legend) compatibility within graphs. Additionally, a baseline condition (incompatible) was included. Participants chose one out of several line graphs from a multiple panel as the answer to a data-related question. Compatibility type and the number of graphs per panel were varied. Whereas Experiment 1 involved simple graphs with only two lines/legend entries within each graph, Experiment 2 explored more complex graphs. The results indicated that compatibility speeds up comprehension, at least when a certain threshold of graph complexity is exceeded. Furthermore, we found evidence for an advantage of local over global data–legend compatibility under specific conditions. Taken together, the results further support the idea that compatibility principles strongly determine the ease of integration processes in graph comprehension and should thus be considered in multiple-panel design.

Keywords Spatial compatibility · Graph comprehension · Multiple panel · Display design · Visual complexity

Introduction

The omnipresence of graphs

Graphs have become omnipresent across a wide range of contexts in our everyday life (Purchase, 2014; Shah, Freedman, & Vekiri, 2005; Zacks, Levy, Tversky, & Schiano, 2002). Especially in the scientific world, they are a key ingredient for disseminating statements in a compact and yet powerful way. If well designed, graphs are a convenient way to communicate data, with many advantages over textual presentation (Larkin & Simon, 1987). These advantages include portraying complex data and relationships in an easy and understandable way, reducing reading time by presenting key findings in a readily visible manner, and reducing the overall word count (Franzblau & Chung, 2012). Given the omnipresence and potential effectiveness of graph usage, it is unfortunate that some graphs do not live up to their potential due to poor construction.

Early graph design guidelines often relied on common sense, positing plausible principles without strong empirical evidence (Bertin, 1983; Schmid & Schmid, 1979; Tufte, 1983). Over the years, however, graph comprehension theories have been backed up by empirical data (e.g., Carpenter & Shah, 1998; Cleveland & McGill, 1984; Pinker, 1990), and subsequent research has empirically addressed specific aspects of graph design. This resulted in various empirically informed design guidelines (Franzblau & Chung, 2012; Hollands & Spence, 1998; Kosslyn, 1994, 2006; Kumar & Benbasat, 2004; Shah & Carpenter, 1995; Shah & Hoeffner, 2002; Wickens, Hollands, Banbury, & Parasuraman, 2013), including guidelines for specific scientific disciplines (see, e.g., American Psychological Association, 2010, for psychological research). Note that these guidelines have mainly focused on the design of single graphs.

One example of a powerful display design principle is the well-known *proximity compatibility principle* (PCP) (Wickens & Carswell, 1995), which states that similarity (perceptual proximity) of graph elements fosters integration processes (processing proximity). The concept of similarity can

The original version of this article was revised: In the original article the authors' names were mistakenly included in the article title.

✉ Eva Riechelmann
eva.riechelmann@uni-wuerzburg.de

[1] Department of Psychology, Würzburg University, Röntgenring 11, 97070 Würzburg, Germany

refer to perceptual attributes (e.g., color and texture) as well as to absolute spatial position (Gillan, Wickens, Hollands, & Carswell, 1998; Wickens et al., 2013). The *principle of compatibility* is a concept related to the PCP. It is a common psychological principle with a long research tradition (e.g., Proctor & Vu, 2006), which is often referred to when designing effective graphs (Kosslyn, 2006) in terms of its benefits for performance—reflected, for example, in decreased error rates and/or response times (e.g., Hommel & Prinz, 1997; Huestegge & Philipp, 2011). Compatibility, as we refer to it in the present study, describes the degree to which the components of different elements in a display (i.e., stimulus–stimulus [S–S] compatibility) are spatially interrelated (Fitts & Simon, 1952; Proctor & Vu, 2006).

Huestegge and Philipp (2011) addressed S–S compatibility in graph comprehension by investigating the influence of spatially compatible versus incompatible data–legend relations. A facilitation of graph comprehension in terms of faster response times (RTs) and higher accuracy for spatially compatible (vs. incompatible) data–legend relations was revealed when judging the correspondence of a graph with a previously displayed statement. For example, participants were faster when reading a graph in which the upper data line was black and the upper legend entry also referred to this black line. Furthermore, it has been shown that the compatibility effect scales up with increasing complexity, in terms of both data pattern complexity (i.e., stimuli depicting interactions instead of main effects) and visual graph complexity (i.e., high vs. low amounts of data depicted within a graph). With this finding, the authors extended the original claim of the PCP, since they demonstrated that display proximity can also refer to *relative* spatial proximity, in terms of the relative position of elements in the legend and data regions. This finding is informative for theories of graph comprehension.

Theories of graph comprehension

Among others (e.g., Cleveland & McGill, 1984; Kosslyn, 1989; Lohse, 1993; Pinker, 1990; Simkin & Hastie, 1987), Carpenter and Shah (1998) introduced an influential, empirically informed model of graph comprehension. They proposed a multicycle, three-stage processing model. Every cycle starts with a *pattern recognition* phase, devoted to the encoding of visual patterns by forming visual chunks. An *interpretation* phase involves retrieving and constructing qualitative and quantitative meaning from the chunk (e.g., associating an ascending line with increase), and finally, an *integration* phase relates these meanings to the semantic referents inferred from legend, labels, and titles. Thus, to facilitate the process of information integration, the comprehensibility of the legend (and/or label and title) should be maximized. The model's assumption of a multicycle process was empirically supported by corresponding eye fixation patterns (i.e., frequent gaze transitions between elements of the graph). The importance of information integration processes for graph comprehension has been further supported by other studies (e.g., Ratwani, Trafton, & Boehm-Davis, 2008). Huestegge and Philipp (2011) addressed integration processes with special interest regarding the integration of elements of the data region and the legend by manipulating the spatial compatibility between these elements (spatially compatible vs. incompatible data–legend relations). Corresponding eyetracking data showed a decrease of gaze transitions between the data region and the legend in data–legend-compatible conditions, suggesting that data–legend compatibility facilitates integration processes in graph comprehension.

However, there is a substantial lack of empirically backed knowledge regarding the design of *multiple* panels. Multiple panels are widely used in all fields of science and refer to the combined presentation of several graphs showing (closely) related, yet different, sets of data (Wickens et al., 2013). Many research guidelines lack any recommendations for the design of multiple panels (e.g., American Psychological Association, 2010), or mention this issue only vaguely (e.g., Coghill & Garson, 2006). Given the obvious advantages and widespread use of multiple panels (Kosslyn, 2006), this is highly surprising. Hence, in the present study we aimed to focus on one specific open issue, namely spatial legend compatibility, that we consider relevant to maximizing the efficiency of information integration processes in multiple-panel graphs.

The present study

The central starting point of the present study was the consideration of two different, but equally plausible, design options for multiple-panel line graphs, with respect to compatibility within elements in the data and legend regions and between legends. First, optimizing each individual graph of the multiple panels (along the lines of Huestegge & Philipp, 2011), and thus applying the principle of data–legend compatibility (i.e., within-graph compatibility), would lead to *local* optimality, whereas global between-graph legend incompatibility might occur—given sufficient variability in the data presented (see Fig. 1a). Second, it is possible to achieve *global* (i.e., between-graph) legend compatibility, but at the cost of potential local data–legend incompatibility in several graphs of the multiple panels (see Fig. 1b). Where Kosslyn (2006) argued in favor of the concept of within-graph compatibility, Andre and Wickens (1992) considered global compatibility (albeit in the context of human–machine interface design) as an important design feature. The goal of the present study was to put the competing compatibility principles to an empirical test by comparing the effects of both kinds of compatibility (pitted against an incompatible baseline condition) on graph comprehension processes. Specifically, spatial compatibility was manipulated by

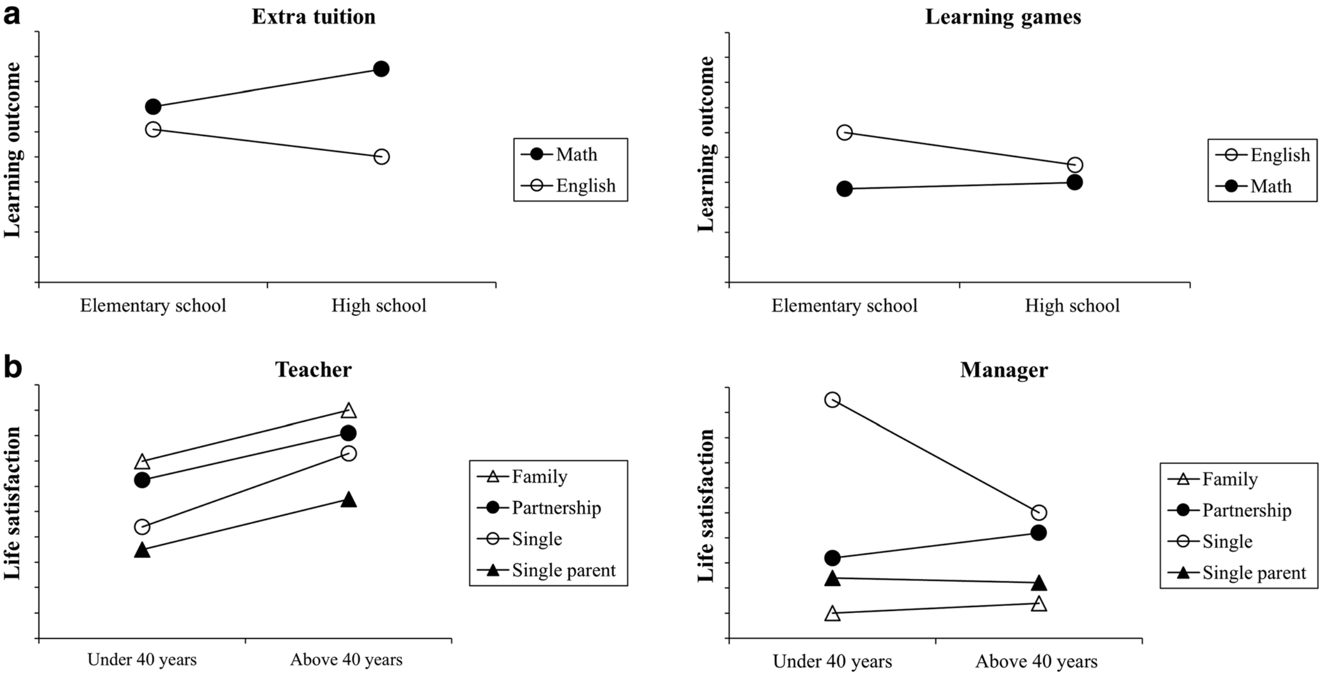

Fig. 1 Examples of the multiple panels used in (**a**) Experiment 1 (example for the local, within-graph-compatible condition) and (**b**) Experiment 2 (example for the global, between-graph-compatible condition). The original graphs were presented in the German language

changing the order of the legend entries relative to the data region.

We also addressed the issue of panel size (number of graphs in a panel: two vs. six) in the present experiments, because several studies have shown that spatial compatibility effects scale up with visual complexity (Huestegge & Philipp, 2011; Ratwani et al., 2008). These previous findings indicated that at a certain threshold of complexity, the likelihood of observing adverse effects of incompatibility is increased (Carpenter & Shah, 1998; Ratwani et al., 2008), probably due to the limits of working memory capacity.

Furthermore, the presentation order of the compatibility conditions (blocked vs. random) was varied for two reasons. First, it appears reasonable to assume that compatibility effects could become more pronounced when trials are presented in blocks as compared to a random presentation order, since in blocked conditions participants might learn to take advantage of the particular type of compatibility over the course of a block. Second, the blocked-design condition allows for a separate analysis of *initial* performance (i.e., performance in the first block of a particular compatibility condition, without having experienced any other compatibility conditions) versus total performance averaged across the experiment. We considered such an analysis of initial performance relevant because it best represents the rather spontaneous encounter with a single (nonchanging) type of graph design in everyday life. In contrast, we anticipated that the repeated processing of graphs with varying types of legend arrangements (in the random sequence, or across all blocks in the blocked sequence) might yield a special processing strategy in order to cope with all types of legend arrangements encountered throughout the experiment, eventually yielding potentially diluted effects of compatibility.

On the basis of the aforementioned research indicating that graph complexity plays a major role regarding the presence/size of spatial compatibility effects in graph comprehension, we additionally ran Experiment 2, in which the complexity *within each graph* was increased by using line graphs with four (vs. two in Exp. 1) lines/legend entries per graph. Thus, we examined visual complexity in our study in two ways: namely, regarding panel complexity (within each experiment) and regarding graph complexity (across experiments).

Taken together, we predicted that both between- and within-graph compatibility would facilitate graph processing, and that such compatibility effects should scale up with visual complexity. Thus, we expected to find more substantial evidence for compatibility effects in Experiment 2 than in Experiment 1, and a stronger compatibility effect for larger panels (containing six graphs) than for smaller panels (containing two graphs). Regarding the two types of compatibility, we reasoned that especially for larger panels, in which working memory limits (Baddeley, 1983; Baddeley & Hitch, 1974) should strongly constrain any integrated or parallel processing of multiple graphs due to the greater number of graphs to be processed, within-graph compatibility should yield better performance than between-graph legend compatibility. In within-graph-compatible arrangements, we assumed that graph readers would automatically generate expectations regarding the (spatial) data region layout when they encoded the legend, and that meeting these expectations might promote integration processes (Huestegge & Philipp, 2011).

Experiment 1

In Experiment 1, we examined legend compatibility in *simple* multiple panels—that is, panels in which each graph consisted of only two lines/legend entries. Every graph within the multiple panels consisted of two parts: the data region and the legend, which was placed at the center right of the data region. To manipulate compatibility, we varied the order of the legend entries. In within-graph-compatible multiple panels, the order of the graph lines corresponded to the order of the legend entries in each graph of the multiple panels. Between-graph compatibility was obtained by maintaining a constant legend order for every graph of the multiple panels. There was no spatial match—neither between data region and legend nor between several legends—in incompatible multiple panels.

Method

Participants Twenty-five university students (20 women, five men; age range: 20–29 years; M = 23.25, SE = 0.51) participated in the experiment and received credit points. One participant was excluded due to low accuracy (see the Results and discussion section for details). All of the remaining 24 participants reported normal or corrected-to-normal vision and had basic prior experience with statistics (e.g., due to statistics classes, work as a research assistant, or in the context of writing an empirical thesis). They gave informed consent. A power analysis using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), based on the very large observed effect sizes (regarding compatibility effects on RTs in line graphs) in the study of Huestegge and Philipp (2011), indicated that a sample size of four participants was sufficient to observe a spatial compatibility effect (power = .95, α = .05). Nevertheless, we opted for a larger sample size of n = 12 for each group, since we could not be sure that the compatibility effects in the present study could be expected to be as large as those in Huestegge and Philipp's study.

Stimuli Each trial consisted of the simultaneous presentation of several graphs (generated with Microsoft Excel), together forming multiple panels (consisting of either two or six graphs of equal size; see Fig. 2). There was a horizontal distance of 3.1° of visual angle between two graphs and a vertical distance of 0.9° between each of the three graphs in the six-graph panel.

The size of each graph amounted to 9.2° × 8.1° of visual angle (width × height). All graphs were black-and-white line graphs consisting of two uncrossed black lines each. The graphs depicted main effects and/or interactions, with both types of effects being represented within each of the multiple panels. The data point markers were black or white circles. Each legend (1.8°–2.5° × 1.5° of visual angle, depending on legend's content) was placed to the right of the data region (5.5°–6.3° × 5.7°, depending on the legend's size) and contained two entries, each consisting of a data marker (black/white circle) and a verbal label. The spatial separation between legend and data amounted to 0.3°, and the title and the data region were separated by 0.6° (see Fig. 1a).

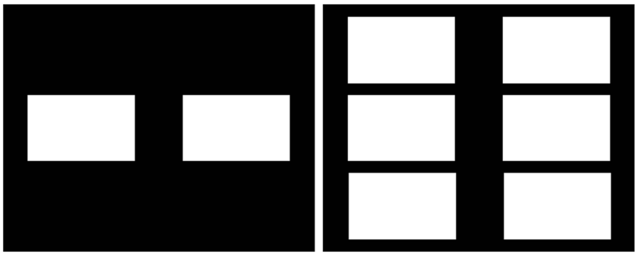

Fig. 2 Spatial arrangement of graphs in multiple panels consisting of two (left) or six (right) graphs

To increase generalizability, we generated graphs (identical in design) covering three different topics (fictional dependent variables). These variables were represented on the y-axis: namely *screen viewing time*, *life satisfaction*, and *learning outcome*. These measures were plotted as a function of three independent variables. First, the x-axis referred to a dichotomous variable and was more or less related to age, thereby following the recommendation of Wickens et al. (2013) to place a (quasi-)quantitative variable on the x-axis of line graphs. The two lines in the graph represented the second independent variable, defined through the legend. The third variable was also categorical and was presented above the data region, thus also representing a graph title (see Fig. 1a).

We designed two basic multiple-panel figures per each of the three topics: a two-graph multiple-panel figure, and a six-graph multiple-panel figure. These six basic multiple panels served as templates, and each multiple panel was designed with both spatially compatible (within-graph- and between-graph-compatible) and incompatible legends, resulting in 18 multiple panels. In between-graph-compatible multiple panels, the order of the black and white data point markers in the legend (e.g., black markers in upper position/white markers in lower position) was counterbalanced across trials.

Note that due to the restriction of depicting only two lines (i.e., graphs with two legend entries), it was not possible to create between-graph-compatible panels that did not involve some individual graphs with compatible data–legend arrangements (here, half of the graphs per panel). For example, when using a constant "black above white marker" legend, half of the graphs in the panel also contained "black above white line" data (otherwise, all data regions would have been arranged in the same manner, which would be unlikely in real data sets as well as an uninteresting case for the present research question).

Thus, global (between-graph) compatibility here was characterized by the absence of *consistent* local (within-graph) compatibility, not by the absence of *any* local compatibility.

Each trial consisted of the simultaneous presentation of a question (white font on black background), extending over seven to ten text lines (10° horizontally), and the multiple panels (see Fig. 1a for an example). For every question, there was a single correct answer, which corresponded to (the title of) one specific graph of the multiple panels (e.g., "For what kind of learning support is the learning outcome for the subject mathematics higher than that for the subject English?"—correct answer: "extra tuition" for two-graph multiple panels; "For what kind of learning support in high school is the learning outcome for the subject English most superior to the learning outcome for the subject mathematics?"—correct answer: "tutoring in small groups" for six-graph multiple panels). Half of the questions (for both two-graph and six-graph multiple panels) were designed to ask for main effects, and the other half to ask for simple main effects (considering the two temporal categories of the *x*-axis each as reference points). The questions always relied on vertical spatial terms (e.g., higher/lower) to indicate the task. We generated four questions for each of the three topics (screen viewing time, life satisfaction, and learning outcome) and for each kind of multiple panel (two-graph and six-graph multiple panels), resulting in 24 questions in total. Each of the 24 questions was combined with the corresponding graph (regarding the topic and number of graphs within the multiple panels) in its three different compatibility condition versions, resulting in 72 experimental trials in total.

Apparatus, task, and procedure The text and figures were presented centered on a 19-in. TFT screen (1,280 × 1,024 pixels) at a viewing distance of approximately 57 cm. A standard keyboard and a computer mouse were available as input devices for the participants. The experiment was run using the PsychoPy presentation software (Peirce, 2007).

Before the single-session experiment (about 30 min) started, participants read a visual instruction (white font on a black background) and underwent four practice trials to familiarize themselves with the task. Each trial started with a white fixation cross (0.5° × 0.5°) on a black screen, presented for 1 s and placed on the left side of the screen (see Fig. 3). After a 2-s black screen interval, the question and the multiple panel were presented simultaneously, with the question located at the position of the prior fixation cross and the multiple panel on the right side of the screen. With the onset of question and multiple panel, the mouse cursor appeared at the center of the multiple panel to ensure equal starting positions for each trial. Participants were asked to indicate (with a left mouse click) as quickly and accurately as possible the specific graph representing the correct answer. There was no implemented time limit for the answer. Each trial contained only one correct option to answer the question. After each click, performance feedback was provided for 1 s. The assignment of the correct answer to a position within the multiple panel was equally distributed across trials. For half of the participants, the trials were presented in a fully randomized order. For the other half, compatibility type (within-graph compatibility, between-graph compatibility, incompatible) was manipulated block-wise (six blocks altogether), with the block order being fully counterbalanced across participants.

Design and data analysis Compatibility (incompatibility vs. between-graph compatibility vs. within-graph compatibility, within-subjects variable), number of graphs (two vs. six graphs within the multiple panels, within-subjects variable), and context (random vs. blocked presentation of compatibility groups, between-subjects variable) served as independent variables. Corresponding three-way mixed analyses of variance (ANOVAs; α = .05 throughout) were conducted to analyze performance (RTs and error rates). Additionally, we assessed initial performance on the first block of trials in the blocked-presentation group (see the introduction for details) with two-way ANOVAs, treating compatibility as a between-subjects variable. In the case of sphericity violations, Greenhouse–Geisser corrections were applied.

Results and discussion

Outliers were calculated on the basis of correct trials and defined as trials with exceedingly long RTs (three *SD*s above the mean per participant, corresponding to 20 trials in total across all participants). These outliers, together with practice trials, were omitted from further analysis. Additionally, participants with very high error rates (three *SD*s above the mean, amounting to a cutoff value of 17.41%) or who lacked above-chance performance in at least one cell of the design were excluded (corresponding to one participant). In both the initial performance analysis and the global performance analysis, participants on average performed significantly better (in terms of faster RTs and fewer errors) when the multiple panels consisted of two (vs. six) graphs, $F(1, 9) = 43.27$, $p < .001$, $\eta_p^2 = .83$, and $F(1, 9) = 36.00$, $p < .001$, $\eta_p^2 = .80$, respectively, for RTs and errors in initials performance analysis, and $F(1, 22) = 276.56$, $p < .001$, $\eta_p^2 = .93$, and $F(1, 22) = 101.74$, $p < .001$, $\eta_p^2 = .82$, respectively, for RTs and errors in global performance analysis.

Initial performance analysis The two-way ANOVA with number of graphs as a within-subjects variable and compatibility as a group variable revealed a significant effect of compatibility neither on RTs, $F(2, 9) = 2.60$, $p = .128$, $\eta_p^2 = .37$, nor on error rates, $F < 1$. For both RTs and error rates, the number of graphs did not significantly interact with compatibility, both *F*s < 1.

Global performance analysis RTs were submitted to a three-way ANOVA with compatibility and number of graphs as

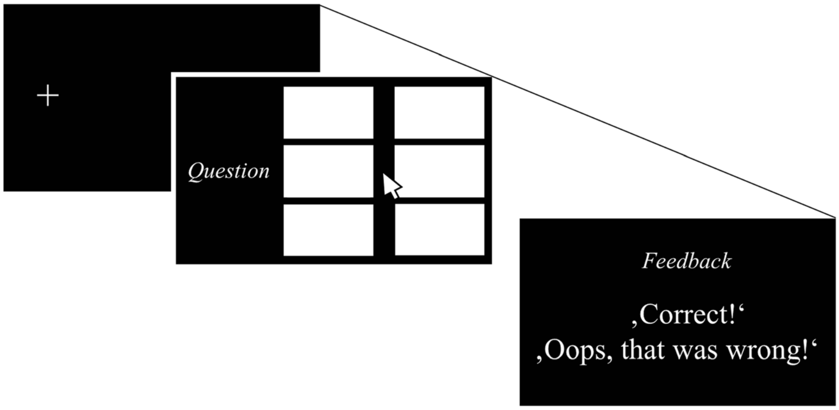

Fig. 3 Schematic representation of a trial: After the presentation of a black screen with a white fixation cross, a question and a multiple panel appeared simultaneously on the screen. After the participant's response (a mouse click on the corresponding graph representing the answer), performance feedback was provided. Per each compatibility condition (incompatible, between-graph compatible, within-graph compatible), 24 trials were presented, resulting in 72 trials altogether. Note that the mouse cursor is enlarged in this figure for the sake of visibility. In the actual experiment, the cursor was of default size, and there was no overlap between the cursor and any of the graphs

within-subjects variables, and context as group variable. We observed no significant main effect of compatibility, $F(2, 44) = 1.29$, $p = .285$, $\eta_p^2 = .06$, and no significant main effect of context, $F < 1$ (see Fig. 4a). Furthermore, none of the two-way interactions were significant, all Fs < 1. However, the three-way interaction was significant, $F(2, 44) = 3.49$, $p = .039$, $\eta_p^2 = .14$, indicating that compatibility had a different effect in the blocked (vs. the random) context, especially in the six-graph condition. However, when we conducted pairwise post-hoc t tests between the compatibility conditions, none of the comparisons approached significance, all ps > .10.

The mean overall error rate amounted to 9.20% (SE = 0.58). The main effect of compatibility on error rates was not significant, $F < 1$ (see Fig. 4b). None of the interactions revealed significant effects—neither the interaction of compatibility and context nor that of number of graphs and context (both Fs < 1), nor the three-way interaction, $F(2, 44) = 1.18$, $p = .316$, $\eta_p^2 = .07$. Only the interaction of compatibility and number of graphs was marginally significant, $F(2, 44) = 2.97$, $p = .062$, $\eta_p^2 = .09$. We decided to follow up on this marginal interaction by computing pairwise comparisons between the compatibility conditions. In the two-graph condition, the within-graph-compatible condition differed significantly from the incompatible condition, $p = .009$, whereas none of the remaining comparisons (including those in the six-graph condition) approached significance, all ps > .10. These results suggest that (specifically within-graph) compatibility tended to reduce error rates in the two-graph condition, but there clearly was no such tendency in the six-graph condition.

In sum, Experiment 1 revealed the expected result that graph comprehension is quicker and more accurate when the number of graphs depicted in the multiple panels is low. However, there was only sparse evidence (in terms of a fewer errors for within-graph-compatible designs in the two-graph condition) for a beneficial effect of spatial legend compatibility.

Probably this lack of a clear performance advantage for compatible relative to incompatible graphs in Experiment 1 can be attributed to the fact that the individual graphs within each panel were all rather simple, consisting of only two lines and legend entries. A reason for not finding performance differences between the two compatibility conditions may be that in the between-graph-compatible condition, half of the individual graphs per panel were still data–legend compatible (due to the restriction of depicting only two lines; see the Method section), which may have reduced the actual difference in design between the two compatibility options.

The explanation that the lack of clear effects could have been due to the lack of individual graph complexity corresponds to previous findings showing strongly attenuated (or absent) data–legend compatibility effects in single-graph panels with simple (two-line) graphs, whereas much stronger effects emerged for more complex graphs consisting of more than two lines (Huestegge & Philipp, 2011). To explicitly test this explanation, we conducted Experiment 2, which involved more complex graphs consisting of four (instead of two) lines per graph.

Experiment 2

In Experiment 2, we focused on visually complex graphs (e.g., graphs depicting more data) by raising the number of lines and legend entries for each graph from two to four. On the basis of previous research suggesting that legend compatibility effects scale up with graph complexity, we reasoned that we should observe clearer compatibility effects in Experiment 2 than in Experiment 1, as well as stronger compatibility effects for large (vs. small) panels.

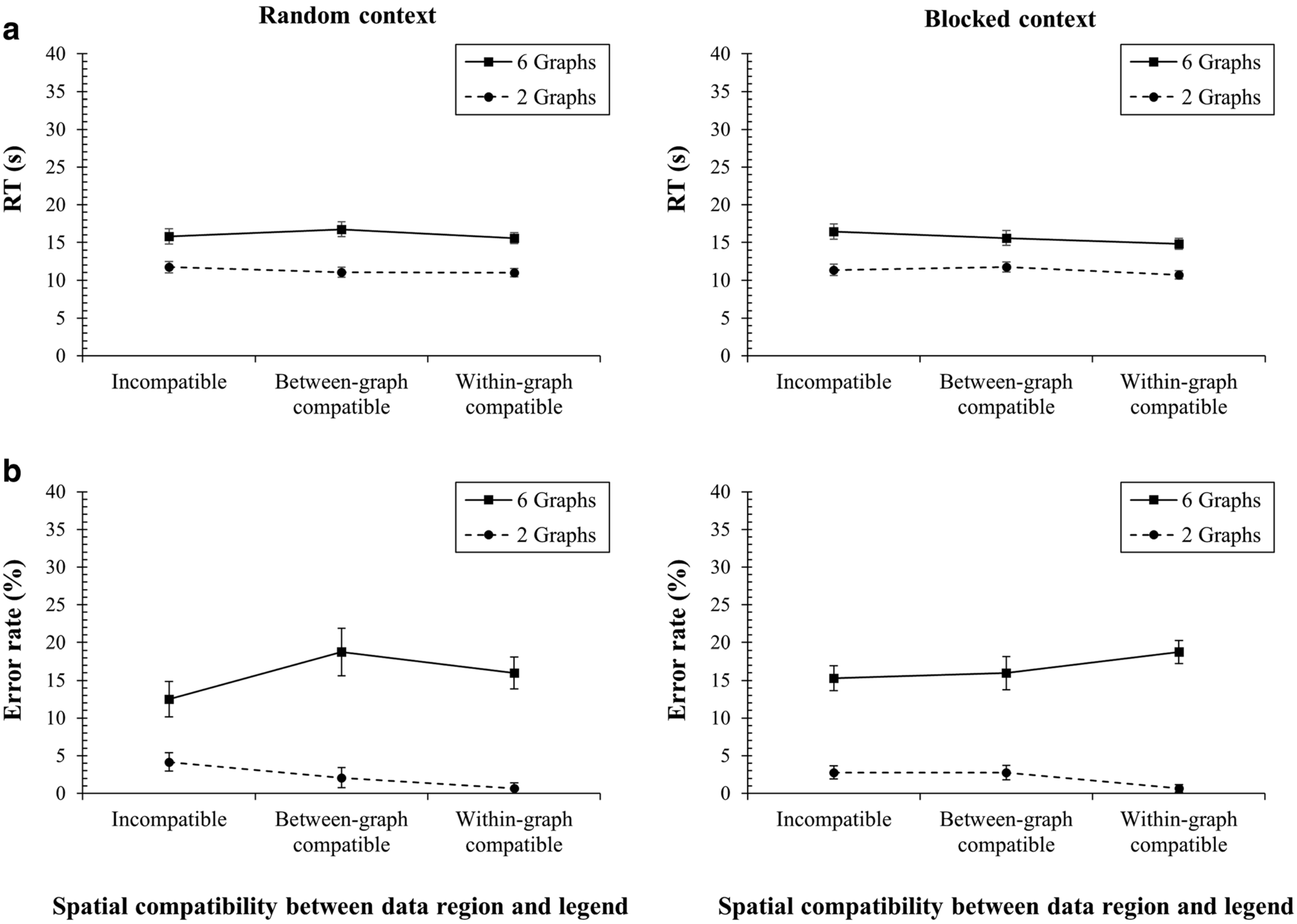

Fig. 4 Means and standard errors (*SE*s) for (**a**) response times (RTs, in seconds) and (**b**) error rates (in percentages) in Experiment 1, as a function of spatial compatibility (incompatible, between-graph compatible, within-graph compatible) between the data region and/or the legends, as well as presentation context (random vs. blocked context) for the judgment of multiple panels depicting two or six graphs

Method

Participants We recruited 26 new participants. Due to low accuracy (see the Results and discussion section for details), the data of only 24 participants (21 female, three men; age range: 18–28 years, *M* = 22.67, *SE* = 2.76) were finally analyzed. All participants reported normal or corrected-to-normal vision and had basic prior experience with statistics (similar to the sample in Exp. 1).

Stimuli, apparatus, task, procedure, and design The apparatus, task, procedure, and design were the same as in Experiment 1. Only the line graph stimuli differed. On the basis of the data material of Experiment 1, we created two additional legend entries per each graph in Experiment 2, thereby increasing visual complexity.

As a consequence of the additional legend entries, the legend's height increased to 2.9°, as compared to 1.5° in Experiment 1, but the other graph dimensions remained constant. In addition to the circular data point markers from Experiment 1, we used black and white triangles as data markers (Fig. 1b). We maintained the global tendency of the data pattern used in Experiment 1, but to avoid any ambiguity in answering the questions in Experiment 2, single data points had to be altered within the graphs taken from Experiment 1. We also slightly altered the questions, to incorporate the additional legend entries as correct answers and to avoid any ambiguity regarding the answers, but the types of questions remained unchanged.

Unlike in Experiment 1, there were several possible options for implementing the incompatibility and between-graph compatibility manipulations in Experiment 2, thus requiring some additional specifications. For the incompatible condition, we randomly selected two versions for each multiple panel (two-graph and six-graph multiple panels for each topic) out of all possible versions with incompatible legend arrangements. Thus, neither between-legend compatibility nor any individual data–legend compatibility was present in this condition.

Due to the presence of four legend entries in each graph, it was (unlike in Exp. 1) not necessary to design panels in the between-graph-compatible condition that also contained many individual data–legend-compatible graphs. However, we considered it more realistic (to ensure generalizability to real-world data contexts) to include one data–legend-compatible graph in each between-graph-compatible panel. Thus, similar to Experiment 1, between-graph compatibility was defined in terms of the absence of consistent (not of any) within-graph compatibility. Again, the assignment of the correct answer to a position within the multiple panels was equally distributed across trials.

Results and discussion

Outliers were defined in the same way as in Experiment 1. Altogether, 26 trials across all participants were excluded due to exceedingly long RTs, and two participants were excluded because of low accuracy (the cutoff value amounted to an error rate of 26.39%). In both the initial and global performance analyses, RTs and error rates were higher for multiple panels depicting six graphs than for multiple panels depicting two graphs, $F(1, 9) = 228.63$, $p < .001$, $\eta_p^2 = .96$, and $F(1, 9) = 9.59$, $p = .013$, $\eta_p^2 = .52$, respectively, for RTs and error rates in the initial performance analysis, and $F(1, 22) = 252.40$, $p < .001$, $\eta_p^2 = .92$, and $F(1, 22) = 19.41$, $p < .001$, $\eta_p^2 = .47$, respectively, for RTs and errors in the global performance analysis.

Initial performance analysis We found a significant main effect of compatibility on RTs, $F(2, 9) = 10.55$, $p = .004$, $\eta_p^2 = .70$ (see Fig. 5). On average, RTs were longest for the incompatibility condition ($M = 27.63$ s, $SE = 1.51$), followed by the between-graph compatibility condition ($M = 23.26$ s, $SE = 1.51$) and the within-graph compatibility condition ($M = 17.86$ s, $SE = 1.51$). The interaction of compatibility and number of graphs was also significant, $F(2, 9) = 4.99$, $p = .035$, $\eta_p^2 = .53$, revealing a larger influence of compatibility on RTs in the six-graph than in the two-graph condition. In the two-graph condition, pairwise follow-up comparisons showed that RTs were significantly longer in the incompatible condition than in the within-graph-compatible condition, $p = .004$, whereas marginally significant differences were found between the incompatible and the between-graph-compatible, $p = .089$, as well as between the between-graph- and within-graph-compatible conditions, $p = .084$. In the six-graph condition, RTs were also significantly longer in the incompatible condition than in the within-graph-compatible condition, $p = .002$, and marginally significant differences were obtained between the incompatible and the between-graph-compatible conditions, $p = .090$. Crucially, RTs were also significantly different between the between-graph- and the within-graph-compatible conditions, $p = .029$, yielding faster RTs for the within-graph-compatible condition.

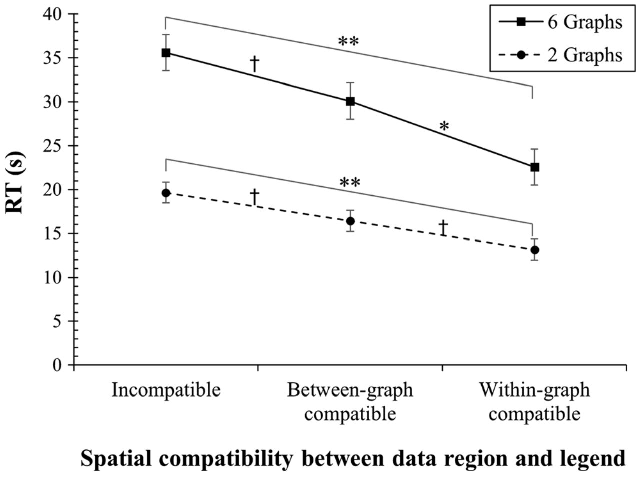

Fig. 5 Mean RTs (in seconds) and *SE*s in the initial performance analysis of Experiment 2, as a function of spatial compatibility (incompatible, between-graph compatible, within-graph compatible) between the data region and/or the legends, for the judgment of multiple panels depicting two or six graphs. Asterisks and daggers indicate the significance levels of two-tailed paired *t* tests. ***p* < .01, **p* < .05, † *p* < .10

Regarding error rates, we observed neither a significant main effect of compatibility, $F(2, 9) = 1.38$, $p = .301$, $\eta_p^2 = .23$, nor a significant interaction of compatibility and number of graphs, $F < 1$.

The results of the initial performance analysis in Experiment 2 indicated a strong general advantage of compatible designs for graph comprehension, which was particularly pronounced for within-graph-compatible designs in multiple panels of greater visual complexity (i.e., in the six-graph condition). Note that these effects were significant despite the relatively small sample size ($N = 12$) underlying this particular analysis.

Global performance analysis Regarding RTs, we observed a marginally significant main effect of compatibility, $F(2, 44) = 2.95$, $p = .063$, $\eta_p^2 = .12$ (see Fig. 6a). Pairwise post-hoc *t* tests revealed that incompatible designs ($M = 19.54$ s, $SE = 1.20$) elicited larger RTs than did within-graph-compatible designs ($M = 17.54$ s, $SE = 0.69$), $p = .020$. There were no significant differences between incompatible designs and between-graph-compatible designs ($M = 18.26$ s, $SE = 0.96$), as well as between the two compatible designs, both *p*s > .10. We found no significant RT effect of context, $F < 1$, nor did any of the interactions reach significance, all *F*s < 1.

The mean overall error rate amounted to 6.98% ($SE = 0.97$). We detected no significant effect of compatibility on error rates, $F(2, 44) = 1.11$, $p = .339$, $\eta_p^2 = .05$ (see Fig. 6b). Also, the main effect of context on error rates was not significant, $F < 1$, nor were the interactions of compatibility and context and of compatibility and number of graphs, both *F*s < 1. The interaction of context and number of graphs was only marginally significant, $F(1, 22) = 3.56$, $p = .072$, $\eta_p^2 = .14$.

Note that between-graph compatibility elicited worse initial performance than did within-graph compatibility, despite the fact that the former condition also contained one data–legend-

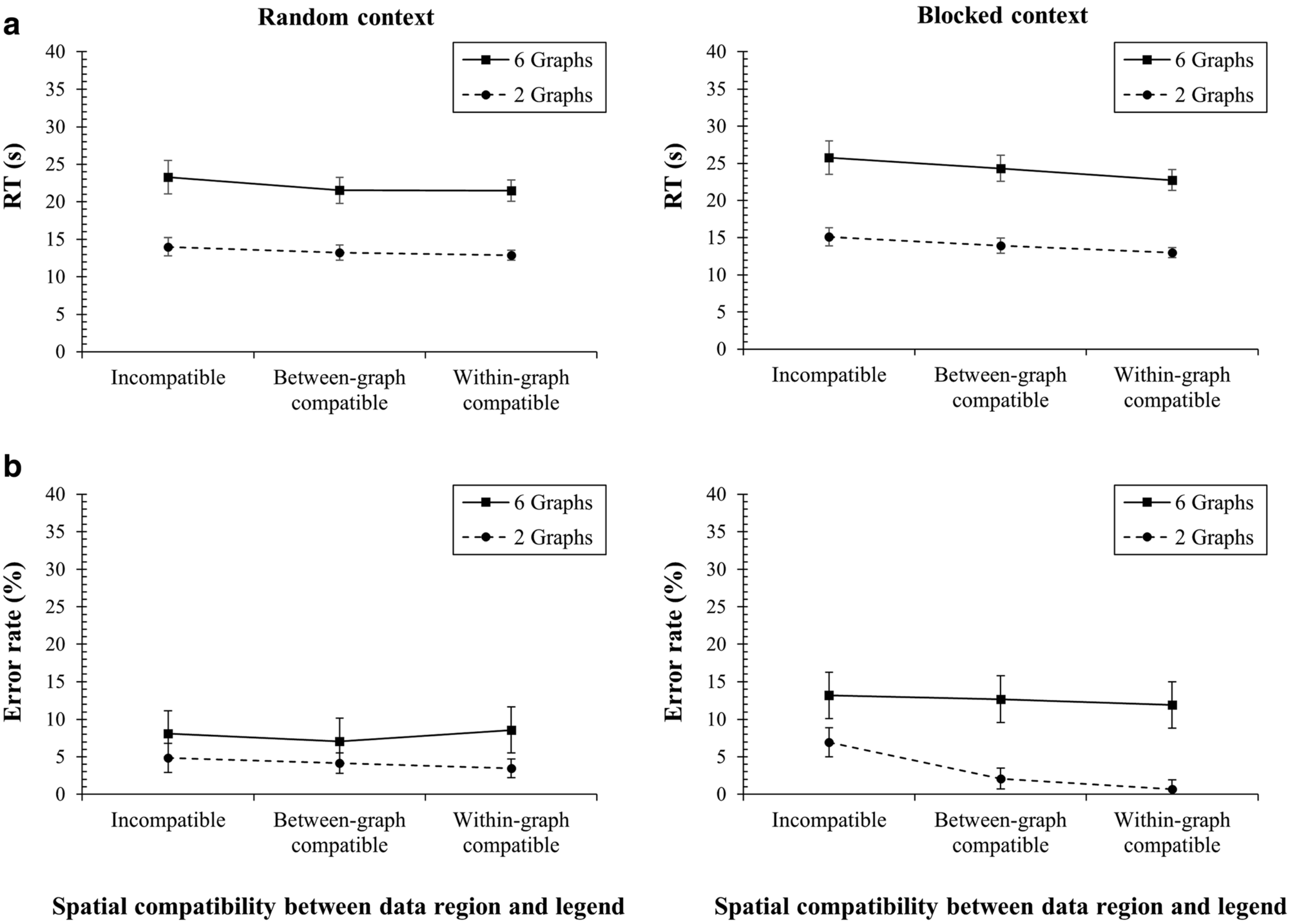

Fig. 6 Means and *SE*s for (**a**) RTs (in seconds) and (**b**) error rates (in percentages) in the global performance analysis of Experiment 2, as a function of spatial compatibility (incompatible, between-graph compatible, within-graph compatible) between the data region and/or the legends, as well as presentation context (random vs. blocked context) for the judgment of multiple panels depicting two or six graphs

compatible graph per panel (see the Method section). Therefore, the relative disadvantage of between-graph compatibility might even be slightly more pronounced in situations that lacked any within-graph compatibility in a panel.

In sum, the global performance analysis of Experiment 2 tended to confirm the superiority of data–legend-compatible graph design over incompatible design. However, as compared to the initial performance analysis, the compatibility effect was only marginally significant and was considerably smaller. Given that the difference in visual graph complexity between experiments was additionally associated with differences in other design features (e.g., visual clutter, data:ink ratio), we did not consider a direct comparison of data across experiments.

General discussion

The aim of the present study was to investigate legend compatibility effects on graph comprehension in multiple panels on the basis of performance measures. We manipulated spatial compatibility between the data region and the legend, in terms of either global (legend–legend) or local (data–legend) compatibility. The influence of visual complexity was addressed in two ways: within experiments, by manipulating the number of graphs depicted within a panel, and between experiments, by manipulating the amount of data (and legend entries) within each graph of the panel. We also considered potential effects of presentation context by varying the order of the stimulus presentation, and we deliberately planned an analysis of initial performance before participants had been repeatedly confronted with changing types of legend design.

Overall, the results supported our hypothesis that spatial multipanel legend compatibility speeds up integration processes in graph comprehension (Exp. 2). Importantly, the effect requires that a certain level of visual complexity emerges, since we could not find substantial effects of compatibility in Experiment 1 with relatively simple graphs (each with only two lines/legend entries). However, in Experiment 2, which involved more complex graphs, corresponding effects on RTs were present in both two-graph and six-graph panels. The assumption that compatibility effects scale with visual complexity is evident not only in the between-experiment comparison of result patterns but also in Experiment 2, in which compatibility effects were larger for six-graph than for two-graph panels. The interaction of compatibility and visual graph complexity is in line with previous reports (Huestegge & Philipp, 2011; Ratwani et al., 2008; Shah et al., 2005), and our results further emphasize that compatibility is an important factor (beside other factors related to task type or format) that needs to be taken into account when reasoning about the issue of graph complexity (Shah et al., 2005).

The present results also confirmed our hypothesis that encountering an unnatural situation with many changing types of graph design throughout the experiment (in the random group as well as in the blocked group, when all blocks were considered) may trigger a special strategy (e.g., a more time-consuming strategy involving very detailed encoding of the whole legend information in order to deal with all types of compatible/incompatible graph designs), yielding diluted compatibility effects. The initial performance analysis, which is more similar to real-life encounters with multiple graph panels, was much more sensitive to compatibility effects (despite relying on fewer participants). We thus propose that the initial performance analysis was more representative of real-life performance than the full analysis of all available data, which will be important to consider in future studies of graph design and its impact on comprehension processes. As expected, both experiments showed that a greater amount of information to process (six vs. two graphs) increases the overall processing demands, an effect reflected not only in RTs but also in error rates.

Our results bear interesting implications for the design of multiple-panel graphs. First, we advise that graph designers generally consider compatibility issues when designing multiple panels. Regarding complex multiple panels, however, it is compulsory to consider spatial data–legend compatibility: Our present results generalize the previous findings of Huestegge and Philipp (2011), who also reported evidence for legend compatibility effects in single graphs that scaled up with graph complexity to multiple panels. More specifically, the initial performance analysis in Experiment 2 revealed an advantage of within-graph-compatible data–legend designs over between-graph-compatible designs in multiple panels. Since this advantage was especially present in the six-graph panels, it is likely that this effect is due to serial (as opposed to parallel or integrated) processing of the individual graphs (due to working memory limitations). Specifically, participants appear to follow a graph-by-graph decoding strategy that benefits from spatial data–legend compatibility within each single graph. Spatial data–legend compatibility could reduce the need for mental transformations and working memory load necessary to associate the data lines with their respective meanings coded in the legend (Huestegge & Philipp, 2011). This superiority of within-graph compatibility is in line with the recommendations of Kosslyn (2006) to design each graph of multiple panels in conformity with the recommendations for that specific type of graph, suggesting a lower priority for the concept of global compatibility (Andre & Wickens, 1992) in multiple-panel graph design.

Note that we did not find effects of compatibility on error rates, except for the marginally significant interaction in the initial performance analysis of Experiment 1. Although error rate is a well-established and sensitive measure used in the study of graph comprehension processes (e.g., Körner, Höfler, Tröbinger, & Gilchrist, 2014; Meyer, 2000), it is possible that our setting, which did not impose any time limit for responding, together with the instruction to respond accurately, contributed to the selective effects on RTs. However, one should keep in mind that many real-life situations (e.g., attending a research presentation) provide graphs for only a limited amount of time; thus, in these situations one would expect more comprehension errors for incompatible panels.

Note that under specific conditions (specifically, the two-graph panels in the blocked within-graph condition), it is (at least in principle) possible to adopt a strategy of finding the right answer without integrating information across graphs, simply by focusing on spatial legend position. Furthermore, it is also possible in principle to answer the questions without explicitly comparing trend patterns across the graphs in the two-graph conditions (thus minimizing the benefits of between-graph-compatible designs). However, given the overall complexity of the task in general, we think it is relatively unlikely that most participants would really adopt these particular (theoretically optimal) strategies throughout. More importantly, even if we assume that such specific strategies might have been present when processing the two-graph panels, these alternative explanations could not account for any effects in the larger panels. In general, we are confident that the specific potential processing shortcuts in the two-panel condition do not endanger our overall conclusions.

Limitations and future implications

Apart from a few recommendations without strong empirical support (e.g., Kosslyn, 2006; Wickens et al., 2013), there has been a general lack of knowledge in the scientific community regarding the design of legends in multiple panels. To the best of our knowledge, our study was the first one to systematically study the role of legend compatibility in multiple panels, providing empirical evidence for the use of multipanel compatibility (specifically, local within-graph compatibility), especially when the graphs are visually complex.

We here analyzed RTs and error rates as objective performance measures, which is probably the most common approach in graph-processing research. However, in recent years the number of eyetracking studies focusing on integration processes in graph comprehension has increased (Huang, 2013; Körner et al., 2014; Renshaw, Finlay, Tyfa, & Ward, 2004; Strobel, Saß, Lindner, & Köller, 2016), which has also advanced our understanding of spatial legend compatibility effects (Huestegge & Philipp, 2011). The obvious advantage of tracking eye movements is that temporal and spatial information about how graph-readers allocate their visual attention to distinct elements in the graph (= graph-readers' strategies) can be revealed (Körner et al., 2014), in order to more precisely pinpoint underlying processing mechanisms (e.g., Carpenter & Shah, 1998). Thus, for future studies and an extension of theories of graph comprehension to multiple panels, an eye-movement-based analysis of data–legend compatibility will certainly be a promising approach to further specifying the mechanisms of compatibility effects.

One limitation of the present study is that we used only uncrossed line graphs, since these provided unambiguous solutions for a data–legend-compatible graph design. Thus, we cannot directly draw conclusions regarding graphs with crossed lines, which will inevitably occur in real-life data. Nevertheless, and in line with the discussion of Huestegge and Philipp (2011), the order of the legend entries for crossed line graphs (with the legend on the right side) should likely be analogous to the respective order of the rightmost endpoints of the lines. These considerations are based on the Gestalt principle of proximity (Wertheimer, 1923; i.e., that proximity is the result of elements being placed close together, whereby the elements tend to be perceived as belonging together), as well as on insights from eyetracking studies (Huestegge & Philipp, 2011, who showed that participants started graph comprehension by encoding the legend, thus reading the graph from right to left). Note, however, that this prediction should be explicitly addressed in a future empirical study.

A final limitation of the present study is that it did not consider the case of having only a single legend across multiple panels (e.g., in one of the corners of the figure; see Kosslyn, 2006). Here we did not implement such a condition because it would not have allowed us to manipulate the two types of compatibility in a controlled fashion. However, such a single legend should, with variable real-life data, nearly always result in incompatible spatial relations with most individual graphs in a figure. Thus, this design option would lose the advantage of within-graph-compatible legends, namely saving the mental transformation processes and working memory load involved with associating the data with their respective meanings. Again, this prediction should be tested explicitly in future research.

Conclusion

The present study allows us to make clear recommendations about the design of legends in multiple-panel graphs: One should avoid spatial incompatibility and, if possible, ensure local, within-graph-compatible data–legend relations, since this design option yielded no performance disadvantages in simple graphs (Exp. 1), but substantial performance advantages in more complex graphs (Exp. 2). With this study, we have contributed to resolving controversy about the suitability of different types of multipanel legend compatibility in graph comprehension, even though further research will be necessary for a deeper understanding of the underlying mechanisms. Given the omnipresence of graphs in general, and of multiple panels in particular (in science and in daily life), it is highly recommended that researchers consider empirically backed-up design recommendations more seriously.

References

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Andre, A. D., & Wickens, C. D. (1992). Compatibility and consistency in display-control systems: Implications for aircraft decision aid design. *Human Factors*, *34*(6), 639–653. http://hfs.sagepub.com/content/34/6/639.full.pdf

Baddeley, A. D. (1983). Working memory. *Philosophical Transactions of the Royal Society B*, *302*, 311–324. https://doi.org/10.1098/rstb.1983.0057

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47–89). New York: Academic Press. https://doi.org/10.1016/S0079-7421(08)60452-1

Bertin, J. (1983). *Semiology of graphics* (W. J. Berg, Trans.). Madison, WI: University of Wisconsin Press.

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, *4*, 75–100. https://doi.org/10.1037/1076-898X.4.2.75

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, *79*, 531–554. https://doi.org/10.1080/01621459.1984.10478080

Coghill, A. M., & Garson, L. R. (2006). *The ACS style guide*. Washington, DC: American Chemical Society.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. https://doi.org/10.3758/BF03193146

Fitts, P. M., & Simon, C. W. (1952). Some relations between stimulus patterns and performance in a continuous dual-pursuit task. *Journal of Experimental Psychology*, *43*, 428–436. https://doi.org/10.1037/h0058736

Franzblau, L. E., & Chung, K. C. (2012). Graphs, tables, and figures in scientific publications: The good, the bad, and how not to be the latter. *Journal of Hand Surgery*, *37*, 591–596. https://doi.org/10.1016/j.jhsa.2011.12.041

Gillan, D. J., Wickens, C. D., Hollands, J. G., & Carswell, C. M. (1998). Guidelines for presenting quantitative data in HFES publications. *Human Factors*, *40*, 28–41. https://doi.org/10.1518/001872098779480640

Hollands, J. G., & Spence, I. (1998). Judging proportion with graphs: The summation model. *Applied Cognitive Psychology*, *12*, 173–190. https://doi.org/10.1002/(SICI)1099-0720(199804)12:2<173::AID-ACP499>3.0.CO;2-K

Hommel, B., & Prinz, W. (1997). *Theoretical issues in stimulus–response compatibility* (Vol. 119). Amsterdam: Elsevier.

Huang, W. (2013). Establishing aesthetics based on human graph reading behavior: Two eye tracking studies. *Personal and Ubiquitous Computing*, *17*, 93–105. https://doi.org/10.1007/s00779-011-0473-2

Huestegge, L., & Philipp, A. M. (2011). Effects of spatial compatibility on integration processes in graph comprehension. *Attention, Perception, & Psychophysics*, *73*, 1903–1915. https://doi.org/10.3758/s13414-011-0155-1

Körner, C., Höfler, M., Tröbinger, B., & Gilchrist, I. D. (2014). Eye movements indicate the temporal organisation of information processing in graph comprehension. *Applied Cognitive Psychology*, *28*, 360–373. https://doi.org/10.1002/acp.3006

Kosslyn, S. M. (1989). Understanding charts and graphs. *Applied Cognitive Psychology*, *3*, 185–225. https://doi.org/10.1002/acp.2350030302

Kosslyn, S. M. (1994). *Elements of graph design*. New York: Freeman.

Kosslyn, S. M. (2006). *Graph design for the eye and mind*. New York: Oxford University Press.

Kumar, N., & Benbasat, I. (2004). The effect of relationship encoding, task type, and complexity on information representation: An empirical evaluation of 2D and 3D line graphs. *MIS Quarterly*, *28*, 255–281.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, *11*, 65–100. https://doi.org/10.1111/j.1551-6708.1987.tb00863.x

Lohse, G. L. (1993). A cognitive model for understanding graphical perception. *Human–Computer Interaction*, *8*, 353–388. https://doi.org/10.1207/s15327051hci0804_3

Meyer, J. (2000). Performance with tables and graphs: Effects of training and a visual search model. *Ergonomics*, *43*, 1840–1865. https://doi.org/10.1080/00140130050174509

Peirce, J. W. (2007). PsychoPy—psychophysics software in python. *Journal of Neuroscience Methods*, *162*, 8–13.https://doi.org/10.1016/j.jneumeth.2006.11.017

Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Ed.), *Artificial intelligence and the future of testing* (pp. 73–126). Hillsdale, NJ: Lawrence Erlbaum Associates.

Proctor, R. W., & Vu, K.-P. L. (2006). *Stimulus–response compatibility principles: Data, theory, and application*. Boca Raton: Taylor & Francis.

Purchase, H. C. (2014). Twelve years of diagrams research. *Journal of Visual Languages and Computing*, *25*, 57–75. https://doi.org/10.1016/j.jvlc.2013.11.004

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

Renshaw, J., Finlay, J., Tyfa, D., & Ward, R. (2004). Understanding visual influence in graph design through temporal and spatial eye movement characteristics. *Interacting with Computers*, *16*, 557–578. https://doi.org/10.1016/j.intcom.2004.03.001

Schmid, C. F., & Schmid, S. E. (1979). *Handbook of graphic presentation* (2nd ed.). New York: Wiley.

Shah, P., & Carpenter, P. A. (1995). Conceptual limitations in comprehending line graphs. *Journal of Experimental Psychology. General*, *124*, 43–61. https://doi.org/10.1037/0096-3445.124.1.43

Shah, P., Freedman, E. G., & Vekiri, I. (2005). The comprehension of quantitative information in graphical displays. In P. Shah & A. Miyake (Eds.), *The Cambridge handbook of visuospatial thinking* (pp. 426–476). New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511610448.012

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review*, *14*, 47–69. https://doi.org/10.1023/A:1013180410169

Simkin, D., & Hastie, R. (1987). An information-processing analysis of graph perception. *Journal of the American Statistical Association*, *82*, 454–465. https://doi.org/10.2307/2289447

Strobel, B., Saß, S., Lindner, M.A., & Köller, O. (2016). Do graph readers prefer the graph type most suited to a given task? Insights from eye tracking. *Journal of Eye Movement Research*, *9*(4), 1–15. https://doi.org/10.16910/jemr.9.4.4

Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt. *Psychologische Forschung*, *4*(1), 301–350. https://doi.org/10.1007/BF00410640

Wickens, C. D., & Carswell, C. M. (1995). The proximity compatibility principle: Its psychological foundation and relevance to display design. *Human Factors*, *37*, 473–494. https://doi.org/10.1518/001872095779049408

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2013). *Engineering psychology and human performance* (4th ed.). Boston: Pearson.

Zacks, J., Levy, E., Tversky, B., & Schiano, D. (2002). Graphs in print. In M. Anderson, B. Meyer, & P. Olivier (Eds.), *Diagrammatic representation and reasoning* (pp. 187–206). London: Springer. https://doi.org/10.1007/978-1-4471-0109-3_11