Highlighting Interventions and User Differences: Informing Adaptive Information Visualization Support

Giuseppe Carenini, Cristina Conati, Enamul Hoque, Ben Steichen, Dereck Toker, James Enns*[1]
Department of Computer Science, *Department of Psychology
University of British Columbia, Vancouver, Canada
{carenini, conati, enamul, steichen, dtoker}@cs.ubc.ca, jenns@psych.ubc.ca

ABSTRACT

There is increasing evidence that the effectiveness of information visualization techniques can be impacted by the particular needs and abilities of each user. This suggests that it is important to investigate information visualization systems that can dynamically adapt to each user. In this paper, we address the question of *how to adapt*. In particular, we present a study to evaluate a variety of visual prompts, called 'interventions', that can be performed on a visualization to help users process it. Our results show that some of the tested interventions perform better than a condition in which no intervention is provided, both in terms of task performance as well as subjective user ratings. We also discuss findings on how intervention effectiveness is influenced by individual differences and task complexity.

Author Keywords

User characteristics, Adaptive Information Visualization.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Recent advances in visualization research have shown that individual user needs, abilities and preferences can have a significant impact on user performance and satisfaction during visualization usage (e.g., [6][14][19][33]). It is therefore important to investigate the potential of user-adaptive visualizations, i.e., visualization techniques and systems that support the provision of visual information personalized to each user's needs and differences.

The benefits of user-adaptive interaction have been shown in a variety of human-computer interaction tasks and applications such as operation of menu based interfaces, web search, desktop assistance, and human learning [22]. There are three key decisions that need to be made when designing a user-adaptive system: (1) what to adapt to, i.e., understanding which user features should be considered for adaptation, including stable, long-term user traits (e.g., cognitive abilities, personality, etc.), as well as transitory, short-term states (e.g., current task, cognitive load, attention); (2) when to adapt, i.e., understanding when it is appropriate and/or necessary to provide adaptive support to the user; (3) how to adapt, i.e., understanding how adaptation should be provided.

In this paper, we focus primarily on this latter question in the context of designing user-adaptive visualizations. In information visualization, the only research we are aware of targeting the question of how to adapt is on *recommending alternative visualizations* based on specific user, data, or task features (e.g., [16][18]). By contrast, in this paper we focus on adaptive interventions aimed at *improving the effectiveness of the visualization a user is currently working with*. In particular, we evaluate a set of four alternative *highlighting* interventions aimed at supporting analytical interaction by directing the user's attention to a specific subset of data within a visualization, while still retaining the context of the data as a whole [12].

Highlighting can be extremely useful in any scenario in which an agent (a system or a human) needs to communicate to a user *several* points about a possibly large and complex dataset. For instance, in a dataset of car sales, two key points could be that *"more cars were sold this year in China than in India"* and that *"Europe sales have been decreasing in the last 3 years"*. In these scenarios, the ability to highlight subsets of the data would naturally support a more effective communication. While the whole dataset can be compactly conveyed with an appropriate visualization, the information relevant to each point can be synchronously highlighted as the key points are sequentially expressed in language (written or spoken). For instance, in our example, *sales for China and India* would be highlighted first, followed by *sales for European countries in the last 3 years.*

The ability to generate highlighting interventions would be especially useful in computer-human communication, for instance, when a system has automatically analyzed and derived insights from a complex dataset (e.g., [5][30]), and needs to communicate this to a user. This functionality may also be beneficial to support a user in inspecting human-generated presentations combining visualizations with

CHI 2013, April 27–May 2, 2013, Paris, France.

[1] Author order is alphabetical except for last author.

textual material, which are quite common in documents ranging from newspaper articles to scientific papers. If a system could track what part of the text the reader is currently reading, and infer the corresponding point made (as it is being investigated in [10]), such a system could use one (or more) of the interventions evaluated in this paper to highlight the relevant visualization elements [7].

The interventions that we evaluated in our study were inspired by the analytical interaction techniques presented in Few [12] and by a taxonomy of post-hoc strategies for visual prompts presented by Mittal [26]. While both [12] and [26] provide valuable descriptions and taxonomies of different techniques, to the best of our knowledge there is no formal evaluation of which interventions may be most useful, both in general and under particular task/user contexts. The user study presented in this paper aims to answer the following research questions on the effectiveness of the four interventions that we target:

1. Can highlighting interventions improve information visualization processing?
2. Is there an intervention that is the most effective?
3. Are questions 1 & 2 above affected by individual user characteristics, by task complexity, and by when the interventions are delivered?

Generally speaking, if we find an intervention that is the most effective, it should be used whenever a system needs to draw the user's attention to a subset of the data. However, if intervention effectiveness is found to depend on the task and/or the user, the results of our study could inform adaptive highlighting for visualization support.

RELATED WORK

Three key decisions are involved in supporting user-adaptive interaction: *what to adapt to*, *when to adapt* and *how to adapt*. Deciding *what to adapt to* involves identifying which individual user features influence interaction performance enough to justify adaptation. In visualization, there are already results on the impact of a number of user characteristics on user performance and satisfaction. For example, user performance across different visualizations and task types has been linked to the cognitive measures of perceptual speed and spatial visualization ([6][33][37]), as well as to the personality trait of locus of control ([19][40]). Also, the cognitive abilities of visual/verbal working memory, as well as visualization expertise, have been shown to impact user satisfaction [33].

Addressing the decision of *when to adapt* involves formalizing adaptation strategies that identify those situations in which the benefits of providing adaptive interventions outweigh their cost (e.g., disrupting the interaction). *When to adapt* has been extensively investigated in fields such as Intelligent Tutoring [39] or Availability Management Systems [22]. In visualization, to our knowledge, [16] is the only work that actively monitors real-time user behavior in order to infer the need for intervention, although [7] describes a user study designed to capture instances of user confusion during visualization.

Addressing the question of *how to adapt*, which is the focus of this paper, has been studied outside information visualization to support, for example, display notifications [2], or hints provision [27]. In information visualization, researchers have so far focused on adaptivity that relates to suggesting alternative visualizations based on specific user or task features [16][18]. By contrast, in this paper we focus on interventions that relate to the current visualization.

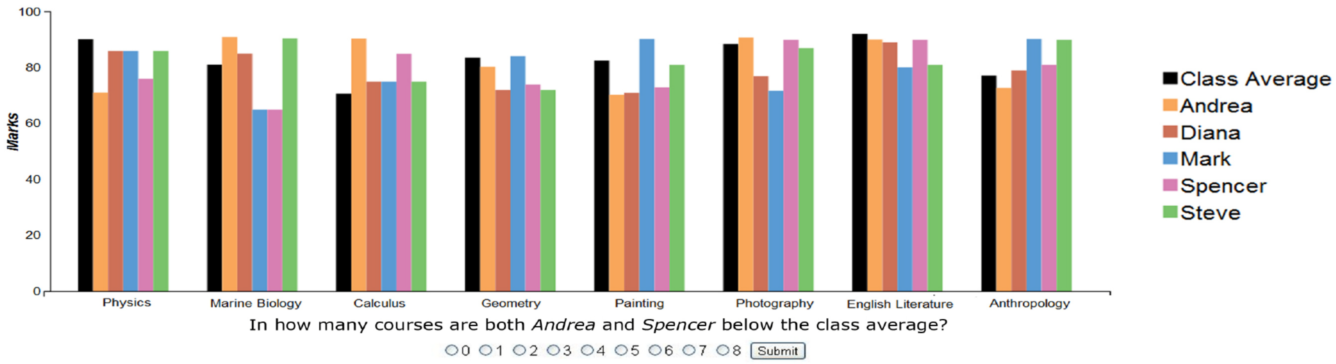
Highlighting interventions are the most relevant techniques to our goal of devising dynamic interventions, because by definition they can be added to an existing visualization as needed to emphasize a specific aspect. Our sources of inspiration for highlighting interventions were Few [12] and Mittal [26]. Mittal [26] was especially useful, as it presents a taxonomy of post-hoc strategies for visual prompts, which is based on a detailed analysis of previous InfoVis literature [3][25], and on the analysis of several thousand charts in newspapers, magazines and business/governmental reports.

USER STUDY

We conducted a user study to investigate the effectiveness of four different highlighting interventions that can be used to emphasize specific aspects of a visualization. We also look at how this effectiveness may be impacted by task complexity, user differences, and delivery time. To keep the number of conditions manageable, we only studied one visualization: bar graphs (see Figure 1 for an example). We focused on bar graphs for three reasons. First, bar graphs are one of the most ubiquitous and effective information visualization techniques. Second, there is already research showing that performance with and preferences for this basic form of visualization is influenced by individual differences such as perceptual speed, visual working memory, and verbal working memory [6][33][37]. Thus, it can be beneficial to investigate how to provide visual interventions for different users who may be working sub-optimally with bar graphs. Finally, as we argue at the end of the paper, results on bar graphs are likely to generalize to other information visualizations.

Experimental Tasks

Tasks were performed via dedicated software. Each task consisted of presenting the participant with a bar graph along with a textual question relating to the displayed data. Participants would select their answer from a horizontal list of radio buttons and click 'Submit' to advance to the next task (see Figure 1). The study questions related to comparing individuals against a group average (data points in the bar graph) on a set of dimensions (data series in the bar graph). Since all tasks require comparisons to the average, the corresponding bar was given a fixed color (black) and position (leftmost) across all tasks (see Fig. 1). In contrast, the color of the other bars was varied from task to task, and selected at random from a set of four color schemes optimized using ColorBrewer [20].



In how many courses are both *Andrea* and *Spencer* below the class average?

Figure 1. Example bar graph visualization as used in the experimental tasks.

For variety, the task questions were drawn from four different data sets: *i)* student grades by course; *ii)* movie revenue by city; *iii)* pet food nutritional values by vitamin and mineral content; and *iv)* company growth rates by business department. All tasks involved the same number of data points (six, including the average) and series dimensions (eight).

Task complexity was varied by making subjects perform two different types of task, chosen from a set of primitive data analysis tasks that Amar et al. [1] identified as "largely capturing people's activities while employing information visualization". The first task type was Retrieve Value (RV), described by Amar et al. [1] as "Given a set of specific cases, find attributes of those cases". This is one of the simplest task types in the Amar hierarchy, and thus it was selected to exemplify tasks of lower complexity. In our study, RV tasks required to retrieve a specific individual in the target domain and compare it against the group average (e.g., "Is Club Universe's revenue in Paris below the average movie revenue in that city?"). The second task type we chose was Compute Derived Value (CDV), defined in [1] as "Given a set of data cases, compute an aggregate numeric representation of those data cases". In our study, CDV tasks required users to first perform a set of comparisons, and then compute an aggregate of the comparison outcomes (e.g., "In how many departments is BioRestore above the average growth and Microfirm is below it? "). CDV tasks in our study are more complex than RV tasks because they require users to *i)* perform significantly more comparisons, *ii)* remember the comparison outcomes, and *iii)* compute an aggregate from the remembered comparisons.

Highlighting Interventions

Selected Interventions

Figure 2 shows the four highlighting interventions that were evaluated in the study, which are designed to guide a user's focus to a specific subset of data within the bar graph while still retaining the context of the data as a whole [12].

In our study, these interventions would highlight those bars that were relevant to answer the current question. For instance, the question "In how many cities are both Shark Swamp and Speed Freak above the average movie revenue?", the interventions would highlight the bars for Shark Swamp's revenue, Speed Freak's revenue, and the average movie revenue in each city. *Bolding* (Figure 2, top left) draws a thickened border around the relevant bars[2]. *De-Emphasis* (Figure 2, top right) fades all non-relevant bars. *Average Reference Lines* (Figure 2, bottom left) draws a horizontal line going from the top of the left-most bar (representing the average) to the last relevant bar, to facilitate comparison. *Connected Arrows* (Figure 2, bottom right) involves a series of connected arrows pointing downwards to the relevant bars.

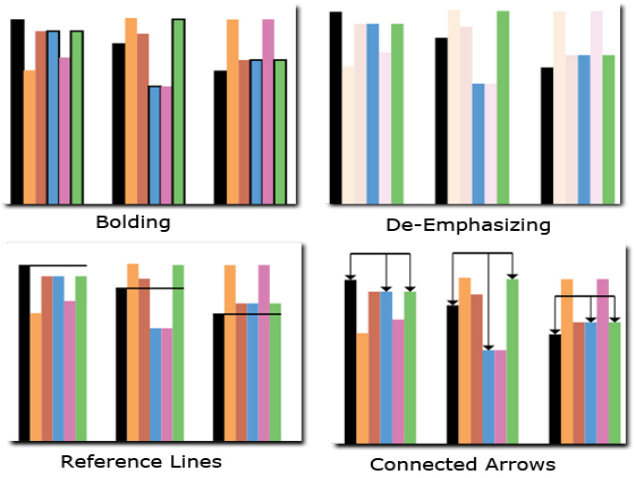


Figure 2. Interventions used in the study.

We focused on only four highlighting interventions to keep the number of conditions and trials within the limit of the available resources for this study. We selected the four interventions in Figure 2 because they were the most

[2] Notice that *Bolding* highlights the average only by thickening the bar, because of its black color. This is arguably not a serious confound, because the average is always relevant in the study tasks, and it was already made to stand out through its constant and distinctive color as well as a consistent leftmost position.

suitable to support our target tasks, as compared, for instance, to highlighting by *Color Change* (left out because color is already used to encode information in our visualizations), or to *Annotate Values*, i.e., adding their specific values on top of selected bars (left out because it can interfere with perceived bar height, and negatively affect our tasks).

Participants performed each of the two task types described earlier with each of the four highlighting interventions in Figure 2, as well as with *no intervention*, as a baseline.

Intervention timing

If highlighting interventions were to be used to provide real-time adaptive support, they would be superimposed on a visualization while the user is looking at it. This could possibly be disruptive, even if the adaptive system had a reliable mechanism to decide *when* the interventions should appear based on user needs. In this study, we wanted to evaluate the relative effectiveness of the selected interventions without this confound, as well as gain initial insights on whether this relative effectiveness changes when the interventions are provided dynamically. Thus, we added an experimental factor that varied *when* the interventions would be shown, consisting of two conditions, *Time zero* (T0) and *Time x* (TX).

In the **T0 condition**, the interventions are included in the bar graph from the beginning of the task, to evaluate them without the possible confound of the disruption that can be caused by a dynamic superimposition.

In contrast, the **TX condition** aims to gauge interventions' effectiveness when they are added dynamically. At the time of the study, however, we had no criterion implemented to decide when an intervention should appear based on a user's needs. Thus, we adopted a procedure designed to minimize the potential intrusiveness of an unjustified superimposition of visual prompts. Essentially, the idea is to add the visual prompts to the target bar graph as soon as the user has had a chance to look at both the bar graph and the related task question. This constraint is enforced in the TX condition by the following steps, which leverage the real-time gaze information provided by a Tobii T120 Eye-tracker installed on the experimental machine:

1. The bar graph appears, without the task question (and without intervention). It stays visible until a user has had a total of 5 eye fixations on the graph or more than 5 seconds have passed.
2. The graph disappears and the question text appears. The question stays visible until the user has had at least 6 fixations on it (2 fixations each in the first third, in the middle, and in the last third of the text), or more than 5 seconds have passed.
3. The graph reappears. At this point, the graph and question text are both visible.
4. After 500ms, the selected intervention is added. This slight delay aims to ensure that users recognize that the intervention is an added component to the graph they had seen so far.

Participants saw each intervention on each task type with both the T0 and the TX delivery strategy, thus generating 20 experimental conditions: 2 task types (RV vs. CDV), times 2 delivery times (T0 vs. TX), times 5 interventions (including no intervention). It should be noted that participants are expected to be slower in the TX delivery condition because of the delay before both graph and text are visible on the screen (which is necessary to complete the task). What we aim to understand with these two conditions is whether delivery time affects the *relative effectiveness* of the interventions.

User Characteristics Explored in the Study

The user characteristics investigated in this study include three cognitive abilities (perceptual speed, verbal and visual working memory), two measures of user visualization expertise with using bar graphs, as well as one personality trait (locus of control).

Perceptual speed (a measure of speed when performing simple perceptual tasks), visual working memory (a measure of storage and manipulation capacity of visual and spatial information), and verbal working memory (a measure of storage and manipulation capacity of verbal information) were selected because they were repeatedly shown to influence visualization performance or user satisfaction in studies involving bar graphs ([6][33][37]).

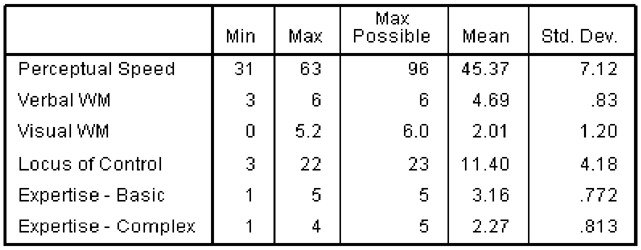
Besides cognitive abilities, the study in [33] also looked at the impact of visualization expertise, but results were inconclusive, possibly because they measured expertise via self-report questions asked *after* the experimental tasks. In this study, we aim to provide a more reliable investigation of the impact of user visualization expertise, not only on bar graph processing but also on the effectiveness of our visual interventions. We use two separate measures for expertise, captured in a pre-questionnaire: one that gauges user familiarity with simple bar graphs (*expertise-simple*) and one with complex ones (*expertise-complex*), elicited as described in the next section.

Locus of control (a measure of the degree to which individuals perceive outcomes as either a result from their own behavior, or from forces that are external to themselves) has been shown to impact user performance with visualizations other than bar graphs [19][40], e.g., list-like visualizations and visualizations with a strong containment metaphor. With this study we wanted to ascertain whether locus of control may also have an impact while interacting with simpler visualizations such as bar graphs and on the effectiveness of our visual interventions.

Study Procedure

62 subjects ranging in age from 18 to 42 participated in the experiment. We selected the number of participants by performing a power analysis [11] a priori on the parameters

of our experimental design, defined to detect a small effect size of at least $\eta_p^2 = .01$ with 0.8 power.



	Min	Max	Max Possible	Mean	Std. Dev.
Perceptual Speed	31	63	96	45.37	7.12
Verbal WM	3	6	6	4.69	.83
Visual WM	0	5.2	6.0	2.01	1.20
Locus of Control	3	22	23	11.40	4.18
Expertise - Basic	1	5	5	3.16	.772
Expertise - Complex	1	4	5	2.27	.813

Table 1. Descriptive statistics of user characteristics.

Participants were mostly recruited via dedicated systems at our university. This resulted in a variety of student participants from diverse backgrounds (e.g., Psychology, Forestry, Computer Science, Finance, Fine Art, German, Commerce). We also recruited 7 non-student participants such as a non-profit community connector, 3D artist, and air combat systems officer. Table 1 presents summary statistics on the user characteristics data collected from the study. A correlation analysis over our 6 user characteristics shows no significant correlations, except for a strong positive correlation ($r = 0.47$, $p < .01$) between *expertise-simple* and *expertise-complex*, and a weak negative correlation ($r = -0.27$, $p <.01$) between perceptual speed and locus of control. Because the expertise measures are highly correlated, we retain only *expertise-complex* as our measure of expertise for further analysis, given its higher variance.

The experiment was a within-subjects study, fitting in a single session lasting at most 90 minutes. There were 20 experimental conditions: 2 task types (RV vs. CDV), times 2 delivery times (T0 vs. TX), times 5 interventions (including no intervention). Participants were instructed to complete the tasks as quickly and accurately as possible. To account for within-subject variance, each participant repeated each condition 4 times, which is a well-established procedure in perceptual psychology experiments measuring performance in terms of time and accuracy [28][36]. Thus, there were a total of 80 trials per participant. To avoid participants getting bored, each of the four domains described earlier were randomly assigned to each task.

Participants began by filling out a pre-study questionnaire asking for demographic information as well as self-reported expertise with simple and complex bar graphs. *Expertise-simple* was elicited with the question *'How often do you look at simple Bar Graphs',* followed by a basic bar graph with 8 bars (values for one series over 8 dimensions); *Expertise-complex* was elicited with the question *'How often do you look at complex Bar Graphs'*, followed by a graph with 48 bars (values for 6 series over 8 dimensions), as used for the experimental tasks. Both questions had five answer options: *i)* Never, *ii)* Rarely (several times a year), *iii)* Occasionally (several times a month), *iv)* Frequently (several times a week), *v)* Very frequently (several times a day). Participants then completed standard computer-based tests for Locus of Control [29], Verbal Working Memory [35], Visual Working Memory [15], and a paper-based test for Perceptual Speed [9]. Next, participants underwent a training phase to expose them to bar graphs, the study tasks, and the highlighting interventions. Then participants underwent a calibration phase for the eye-tracker, before starting the study trials. Participants then performed 40 of the 80 study trials, followed by a 5 minute break. After the break, the eye-tracker was re-calibrated and the participant performed the remaining 40 trials. The 80 trials were fully randomized in terms of experimental conditions (i.e., task complexity, intervention delivery time, interventions). The experimental software was fully automated and ran in a web-browser, with the visualizations and interventions programmed using the D3 visualization framework [8].

Lastly, participants took a post-questionnaire asking for their evaluations of each intervention's usefulness, as well as their relative preferences. The questionnaire included:

- 10 rating statements in the form of "I found the X intervention useful for performing Y tasks", for each intervention and task type (i.e., simple vs. complex). The statements were rated on a Likert scale from 1 to 5.
- 2 ranking statements in the form of "Please rank your preference of interventions for simple/complex tasks (order from 1 to 5) (1: most preferred, 5:least preferred)".

ANALYSIS OF TASK PERFORMANCE

We look at both task *completion time* and task *accuracy* as performance measures. *Completion time* was normally distributed (M=19.5s, SD=10.2), whereas task accuracy indicated a ceiling effect with 91.4% correct answers, possibly due to the tasks being generally easy to solve, or due to participants focusing on generating the correct answer, while sacrificing their time on task. The ceiling effect on accuracy arguably makes a separate analysis of this performance measure not very informative. We nevertheless did not want to discard accuracy altogether, because trials that were answered incorrectly should be penalized accordingly**.** We opted to use a combined score for task performance, known as Inverse Efficiency Score [34]. Given that participants repeated each experimental condition 4 times, *task performance* is calculated by averaging completion time for the trial repetitions that were performed correctly, and then dividing this score by the percentage of correct repetitions[3]. Task performance values thus calculated can be essentially interpreted as completion times penalized for incorrect trials (the lower the percentage of correct trials, the higher the adjusted average time on these trials). Thus, performance is reported in seconds and a higher score represents a lower performance.

[3] When there are no correct repetitions, leading to a divide-by-0 problem, the participant for that trial must be discarded from the analysis. In our study, only one participant's data was removed from the analysis for this reason.

We use a General Linear Model (GLM) repeated measures to analyze our performance data. We first run a 2 (task complexity) by 2 (delivery time) by 5 (intervention type) General Linear Model (GLM) repeated measures to investigate the effects of our experimental factors alone. Next, we analyze the effects of each of our five co-variates separately (perceptual speed, visual WM, verbal WM, locus of control, and expertise-complex**)**, by running a GLM with the experimental factors and only that co-variate. Due to the high number of covariates in our study, this approach ensures that we do not overfit our models by including all co-variates at once. Each co-variate was discretized into three levels via a three-way split. *Low* represents the bottom quartile of the values distribution (i.e. lower 25%), *average* represents the values within the interquartile range (i.e., middle 50%), and *high* represents the upper quartile (top 25%). In the next sections, effect sizes (partial eta-squared) are reported as small for .01, medium for .09, and large as .25 [13]. All reported pairwise comparisons are corrected with the Bonferroni adjustment.

Results on Task Performance

Main Effects

We found main effects of task type, delivery time, intervention type, Perceptual Speed, and Verbal WM, as shown in Table 2.

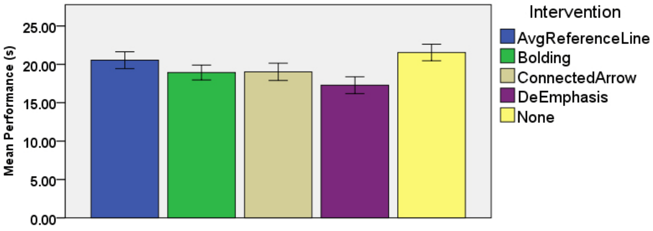


Figure 3. Performance score for each intervention. All bar graphs are shown with 95% confidence intervals.

Main Effect	F-Ratio	Effect Size	Sig. Value
Task Type	F(1,59) = 543.47	$\eta_p^2 = .902$	p< .001
Delivery Time	F(1,59) = 1277.94	$\eta_p^2 = .956$	p< .001
Intervention	F(4,236) = 44.44	$\eta_p^2 = .430$	p< .001
Percep. Speed	F(1,58) = 10.02	$\eta_p^2 = .147$	p < .01
VerbalWM	F(3,56) = 5.42	$\eta_p^2 = .225$	p < .01

Table 2. Significant main effects on task performance.

The main effect of task type confirms the difference in complexity between the two task types in the study, with Compute Derived Value having longer task performance values (M=25.2s, SD=7.5) than the simpler Retrieve Value tasks (M=13.8s, SD= 5.3). The main effect of delivery time is to be expected because of the delay in answering the question generated by the TX condition, as discussed earlier. The average performance for TX was 22.7s (SD=7.8), as opposed to 16.3s (SD=8.2) for T0.

There was also a main effect of intervention type. As shown in Figure 3, performance was best for *De-Emphasis*, and worst for *None*, (i.e., no intervention provided). Pairwise comparisons show that interventions are significantly different from one another except for *Bolding* and *Connected Arrows*, and for *None* and *Avg. Ref. Lines*. This result indicates that all interventions, except for *Avg. Ref. Lines*, were helping users solve the selected tasks more efficiently than when they received no intervention. These results will be further qualified by interactions with task type and delivery time described in the next section.

The main effect and related pairwise comparisons for Perceptual Speed indicate that performance was similar for users with low Perceptual Speed (M=20.6s, SD=8.6) and average perceptual speed (M=20.3s, SD=9.3)**,** whereas users with high perceptual speed were significantly better at completing tasks (M=17s, SD=6.7), hence confirming previous work [30]. The results for Verbal WM show similar directionality, except that the performance of users with low Verbal WM (M=24.1s, SD=12.0) was significantly worse than the scores of users in both the average group (M=19.4s, SD=8.4) and the high group (M=18.0s, SD=7.2). While [33] previously uncovered a link between Verbal WM and user preferences for different visualizations, and [32] has showed that low Verbal WM increases a user's gaze fixations on textual elements [30], our current result on Verbal WM is, to the best of our knowledge, the first to directly link this cognitive ability to task performance with information visualizations. The results for Perceptual Speed and Verbal Working memory will be further qualified by interactions with task type.

Interaction Effects

Table 3 shows a summary of the interaction effects.

Interaction Effect	F-Ratio	Effect Size	Sig. Value
Intervention*Task Type	F(4,236) = 8.65	$\eta_p^2 = .128$	p< .001
Perceptual Speed*TaskType	F(1,58) = 8.64	$\eta_p^2 = .130$	p < .01
VerbalWM*TaskType	F(3,56) = 5.79	$\eta_p^2 = .237$	p < .01
VisualWM*TaskType	F(1,58) = 3.81	$\eta_p^2 = .062$	p < .05
DeliveryTime* Intervention	F(4,236) = 7.56	$\eta_p^2 = .114$	p < .01

Table 3. Significant interaction effects for task performance.

Intervention*TaskType: Figure 4 shows that, for both task types, *None* is the intervention with the worst performance and *De-Emphasis* the one with the best. Pairwise comparisons, however, show that for simple tasks (RV), all interventions are significantly different from one another and are better than *None*, whereas for complex tasks (CDV), *Avg. Ref. Lines* is no longer significantly better than *None*. A possible explanation is that *Avg. Ref. Lines* helps the comparisons with the average bar, but it does not highlight the elements to be compared as well as the other interventions, except in the case when they are contiguous to the average bar and to each other. This may become a greater disadvantage with the more complex comparisons involved in our CDV tasks.

Additionally, there are no-longer significant differences between *Bolding* and *De-Emphasis*, nor among *Bolding, Connected Arrows*, and *Avg. Ref. Lines,* indicating that for more complex tasks, the relative performance between the interventions is less pronounced. For instance, feedback we gathered from participants indicates that *De-Emphasis* can make it hard to see bar groupings. Even though *Bolding* and *De-Emphasis* can be considered conceptually similar (emphasizing relevant bars vs. de-emphasizing non-relevant bars), it is possible that for complex tasks, the fading of 'irrelevant bars' removes some contextual cues for sample grouping, which would help solve the tasks (e.g., when many bars in the middle of a group are faded out, the outermost bars of a group may look disconnected).

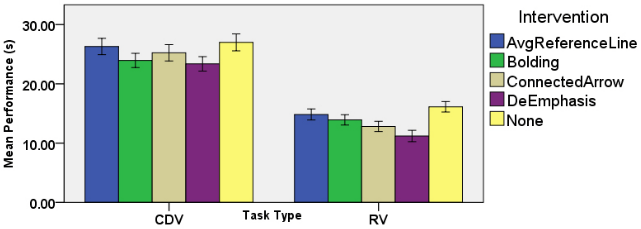


Figure 4. Interaction between Interventions and Task Type.

Perceptual Speed*TaskType, VerbalWM*TaskType, VisualWM*TaskType: There was no significant difference in performance with RV tasks for users with different values of Perceptual Speed, VerbalWM, and VisualWM**.** For CDV tasks, in contrast, users with higher values of these cognitive measures perform better. Figure 5 shows the interactions for Perceptual Speed and Verbal Working Memory.

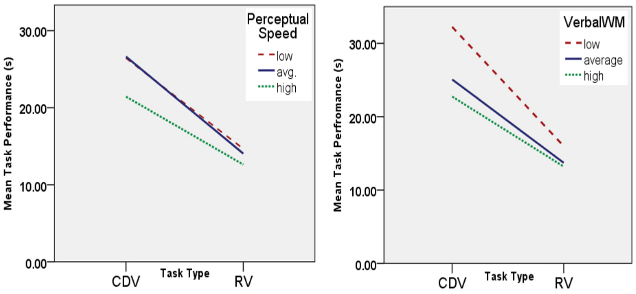


Figure 5. Interaction between: Perceptual Speed and Task Type (left); VerbalWM and Task Type (right).

The trends for Visual Working Memory are very similar to those in Figure 5. These results are likely due to the fact that CDV tasks requires processing and remembering more of the visual elements (bars, interventions, etc.) and more of the verbal information on the graph (i.e., legend items, labels, etc.). Thus, for CDV tasks, higher values of the cognitive measures may be having a stronger impact compared to RV tasks. The result for perceptual speed aligns with results in previous work [33], where it was also found that users with lower perceptual speed require more time to complete a complex task relative to their high perceptual speed counterparts. For visual and verbal working memory, this study is the first to connect these two cognitive traits to task performance (as opposed to user preferences) with a visualization, possibly because previous studies relied on tasks that were not complex enough to detect these effects.

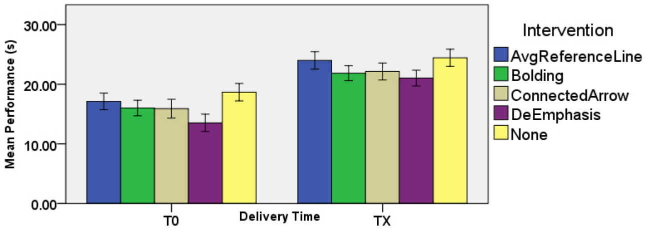


Figure 6. Interaction between Delivery Time and Interventions on task performance.

Delivery Time*Intervention: This interaction effect is shown in Figure 6 and indicates that for T0, *None* and *De-Emphasis* are, respectively, significantly worse and better than all other interventions (with no other significant differences). For TX, the difference between interventions are much smaller, with *Avg. Ref. Lines* no longer being significantly better than *None*, and *Bolding* and *Connected Arrows* no longer being worse than *De-Emphasis*. This result is important, because it suggests that when interventions are delivered dynamically, they may lose some of their value due to possible intrusiveness, and thus it is crucial to evaluate them in the right context of usage. On the other hand, even in the potentially intrusive TX condition, some interventions are still better than none, indicating that it is possible to provide dynamic adaptive interventions that can help improve effectiveness.

ANALYSIS OF SUBJECTIVE MEASURES

As we did for performance measures, we first ran a 2 (task type) x 5 (intervention) General Linear Model repeated measures on the usefulness *ratings* in order to investigate the effects of our experimental factors alone, followed by additional analyses on each of our five co-variates with the experimental factors. These ratings were corrected using the Aligned Rank Transformation (ART)-Tool [38] to make them suitable for parametric analysis. Results from this analysis are shown in Table 4. A similar set of analyses on the preference *rankings* yielded no significant results.

There was a significant main effect of intervention on usefulness ratings, shown in Figure 7. Pairwise comparisons reveal that all of the intervention ratings were significantly different from one-another (except for *Connected Arrows* and *Bolding*), and that all interventions were better than *None.*

Results	F-Ratio	Effect Size	Sig. Value
Intervention	F(4,236) = 100.23	η_p^2= .629	p< .001
VisualWM* Intervention	F(4,224) = 2.34	η_p^2= .040	p< .05

Table 4. Significant effects for intervention usefulness ratings.

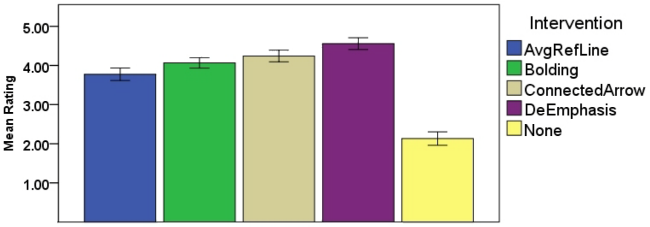


Figure 7. Usefulness ratings for each Intervention.

This main effect and the trends of the relative ratings between interventions correspond exactly to those for intervention on task performance found in the previous section (see Figure 3), showing a strong connection between objective and subjective effectiveness of the tested interventions. It is also worth noting that users found all the interventions more useful than no intervention, regardless of task type. This was not the case for task performance. There was, however, an interaction between intervention type and visual WM, as shown in Figure 8.

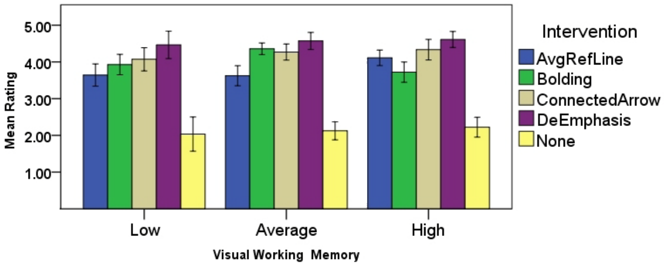


Figure 8. Usefulness ratings by VisualWM levels.

This is in line with previous work linking Visual WM to user subjective (preference) scores [32]. Pairwise comparisons show that users with either low or average Visual WM rated the usefulness of *Avg. Ref Lines* significantly lower than users with high Visual WM. A possible explanation for this result is that the added reference lines may have been 'visual distractors' for lower Visual WM users, given that the lines do not run only through the relevant bars, but also through any other bars between the average and the last relevant bar. We also find that users with average Visual WM rate *Bolding* significantly higher than users with either low or high Visual WM. While this finding further confirms the influence of Visual WM on subjective ratings, we currently do not have an intuition as to the directionality of the result.

DISCUSSION AND CONCLUSIONS

The goal of our study was to investigate the relative effectiveness of four visual prompts designed to support users in visualization processing by highlighting visualization elements relevant to performing target tasks. As we discussed in the introduction, this functionality can be extremely useful for scenarios in which users need to make a variety of inferences on a visualized dataset, and may benefit from having the most relevant subsets of graph elements emphasized in turn.

Although in our study, to keep the number of conditions manageable, we only considered one type of information visualization, i.e., bar graphs, there are at least three different arguments that support the potential generality of our results to other visualizations (but of course, generalizations should be eventually corroborated by empirical evidence). First, bar graphs are one of the most popular visualizations because they rely on length and 2-D position, the only two pre-attentive attributes that can be perceived quantitatively with a high degree of precision [12]. Thus, results on bar graphs can arguably generalize to other popular visualizations that rely on the same pre-attentive attributes, like line-graphs and scatter-plots. Second, since bar graphs are so effective and popular they have been used as building blocks of more complex visualizations. For instance, [17] recently presented LineUp, an interactive visualization supporting the very common and critical task of ranking items based on multiple heterogeneous. As another example, ValueCharts [4] is a visualization that has been applied to elicit user preferences in decision making in different domains, as well as a component of a sophisticated interface to query event sequences. We argue that our results may well generalize to these more complex and extremely useful visualizations based on bar graphs. Third, most of the interventions considered in the paper can be applied to other visualizations besides bar graphs. Thus, our results may well generalize also to these visualizations. For example, [24] demonstrated several example applications of reference line, bolding, and de-emphasis in pie charts and line charts in addition to bar graphs. Average reference lines have been used to visually compare individual marks to a predetermined value in various charts [26][24]. Bolding and de-emphasis form a perceptual group based on the Gestalt principle of similarity [24] and thus have been applied in various visualizations to relate items (e.g., TreeMap, Scatter Plot Matrix, Arc diagram [21]).

We now discuss the user study results with respect to our original research questions. In the study we wanted to ascertain (1) if highlighting interventions can improve visualization processing; (2) if there is a highlighting intervention that is the most effective, and (3) if questions 1 & 2 are affected by user characteristics, task complexity, and intervention timing. We investigated these questions in the context of performing visual tasks with bar graphs.

As for question 1, our results show that all the highlighting interventions we tested, except for *Avg. Ref Lines*, can improve visualization processing compared to receiving no interventions, both in terms of task performance and a user's perceived usefulness. Thus, these interventions should be further investigated as means of providing users with dynamic support during visualization tasks.

As for question 2, results show that no single highlighting intervention is the most effective in general. *De-Emphasis* always performed at the top, in terms of both performance and rated usefulness, but it was absolute best only with the simpler RV tasks, and when it was present from the beginning of the task (delivery condition T0).

Hence, we did find significant effects of task complexity and delivery time on intervention effectiveness (question 3). When considering task performance, there was no longer a significant difference between *De-Emphasis* and *Bolding* during complex tasks, or among *De-Emphasis*, *Bolding* and *Connected Arrows* when the interventions were delivered dynamically. For the long-term goal of providing adaptive highlighting interventions, this latter result suggests that future studies should focus on further investigating the effectiveness of *De-Emphasis*, *Bolding* and *Connected Arrows* in dynamic delivery conditions, and in particular in conjunction with delivery criteria based on actual user needs (e.g., at the onset of *confusion* as suggested in [7]). It is already a very encouraging result however, that delivering the interventions dynamically did not neutralize their effectiveness compared to no intervention, suggesting that their benefits outweigh their potential intrusiveness.

Still in relation to question 3, we also found an impact of user characteristics, in terms of an effect of Visual WM on ratings for perceived intervention usefulness. This result is in line with previous findings that Visual WM affects subjective ratings for visualizations [33]. Our results suggest that, if information on a user's Visual WM is available, higher perceived usefulness may be achieved by using *Bolding* as a highlighting intervention for users with average Visual WM.

Interesting effects of individual differences were also found when analyzing the interaction with task complexity for task performance. In particular, for each of the three cognitive abilities tested in the study, we found no significant difference in performance among participants with different levels of these abilities for simple tasks (RV). In contrast, for complex tasks (CDV) participants with high measures performed significantly better, indicating that complexity can significantly impact user performance depending on cognitive abilities. Similar results were found in previous work for Perceptual Speed [33], but this is the first study that extends them to Visual and Verbal WM, likely because of the increased complexity of our tasks.

The implication for user-adaptive visualizations is that participants with low-medium cognitive measures would benefit the most from help such as adaptive interventions. The fact that there were no interaction effects between cognitive abilities and the different highlighting interventions targeted in the study suggests that perhaps other types of interventions should be explored to help users with low-medium cognitive measures. For instance, previous work linking gaze patterns to performance when processing bar graphs [33], suggests that users with low Perceptual Speed may benefit from help in processing a graph's legend, whereas users with low Verbal WM may benefit from interventions that facilitate processing of the verbal elements of a graph.

Also to note, we did not find any significant results for the personality trait Locus of Control. A likely explanation is that most findings for this user characteristic were found when comparing list-like visualizations and visualizations with a strong containment metaphor [19], which were not the target of our interventions. We also did not find any significant results for the visualization expertise measures we collected from users. This could be due to the fact that some users were possibly biased when self-reporting their expertise, or that previous expertise was not a relevant factor with regard to a user's performance/preference with the visualization tasks administered in our study.

Our next step involves an analysis of user eye gaze behavior in order to verify and better qualify our findings, and to suggest further interventions for adaptive help. We also plan to run similar experiments on more complex visualizations and on a broader set of interventions.

REFERENCES

1. Amar, R.A., Eagan J., & Stasko, J.T. Low-Level Components of Analytic Activity in Information Visualization. 16th IEEE Info. Vis. Conf., 15-21, 2005.
2. Bartram, L., Ware, C., & Calvert, T. Moticons: detection, distraction and task. Int. J. Hum.-Comput. Stud. 58(5), 515-545, 2003.
3. Bertin, J. Semiology of Graphics. University of Wisconsin Press, 1983.
4. Carenini, G., Loyd, J. Valuecharts: Analyzing Linear Models Expressing Preferences and Evaluations. In Proc. of the working conf. on Advanced visual interfaces, 150–157, 2004.
5. Carenini G. & Rizoli L. A Multimedia Interface for Facilitating Comparisons of Opinions. Proc. 13th Int. Conf. on Intelligent User Interfaces, IUI 2009, 325-334.
6. Conati, C. & Maclaren, H. Exploring the Role of Individual Differences in Information Visualization. In Proc. Conf. on Advanced Vis. Interf., 199-206, 2008.
7. Conati, C., Hoque, E., Toker, D., & Steichen, B. When to Adapt: Detecting User's Confusion During Visualization Processing. Proc. 1st Int. Workshop on User-Adaptive Inf. Vis. (WUAV 2013), in conj. with UMAP 2013.

8. D3 javascript library. http:// http://d3js.org/.

9. Ekstrom, R., French, J., Harman, H. & Dermen, D, Manual from Kit of Factor-References Cognitive Tests. Educational Testing Service, Princeton, NJ, 1976.

10. Elzer, S., Carberry, S., & Zukerman, I. The automated understanding of simple bar charts. Artificial Intelligence Journal 175(2), 526-555, 2011.

11. Erdfelder, E., Faul, F., & Buchner, A. GPOWER: A general power analysis program. Behavior Research Methods, Instruments, & Computers, 28, 1-11, 1996.

12. Few, S. Now You See It: Simple Visualization Techniques for Quantitative Analysis, First Edition. Analytics Press, 2009.

13. Field, A., & Hole, G., How to Design and Report Experiments. Sage Publications, London, 2003.

14. Flatla, D.R., Gutwin, C. SSMRecolor: Improving Recoloring Tools with Situation-Specific Models of Color Differentiation. Proc. of Human factors in computing systems, CHI 2012, 2297-2306.

15. Fukuda, K., & Vogel, E.K. Human variation in overriding attentional capture. J. of Neurosc., 8726-8733, 2009.

16. Gotz D., & Wen, Z.. Behavior Driven Visualization Recommendation. Proc. Conf. on Intelligent User Interfaces, IUI 2009, 315-324.

17. Gratzl, S., Lex, A., Gehlenborg, N., Pfister, H., & Streit, M. LineUp: Visual Analysis of Multi-Attribute Rankings. Visualiz. & Comp. Graph., 2277-2286, 2013.

18. Grawemeyer, B. Evaluation of ERST – an external representation selection tutor. Proc. Conf. on Diagrammatic Represent. and Inference, 154-167, 2006.

19. Green, T. M. & Fisher, B. Towards the personal equation of interaction: The impact of personality factors on visual analytics interface interaction. In IEEE Visual Analytics Sc. and Technology, 203-210, 2010.

20. Harrower, M. & Brewer C. Colorbrewer.org: an online tool for selecting colour schemes for maps. The Cartographic Journal 40.1, 27-37, 2003.

21. Heer, J., Bostock, M., Ogievetsky, V. A Tour through the Visualization Zoo. Communications of the ACM, 53(6), 59-67, 2010.

22. Jameson, A. "Adaptive Interfaces and Agents" in Human-Computer Interface Handbook, eds J.A. Jacko and A. Sears, 305-330, 2003.

23. Kaptein, M.C., Nass, C., & Markopoulos, P. Powerful and Consistent Analysis of Likert-Type Rating Scales. Proc. Conf. on Human factors in Comp. Sys., CHI 2010, 2391-2394.

24. Kong, N., Agrawala, M. Graphical Overlays: Using Layered Elements to Aid Chart Reading. Visualization and Computer Graphics, 2631-2638, 2012.

25. Kosslyn, S.M. Elements of Graph Design. W.H. Freeman and Company, 1994.

26. Mittal, V.O. Visual Prompts and Graphical Design: A Framework for Exploring the Design Space of 2-D Charts and Graphs. Proc. AAAI/IAAI, 57-63, 1997.

27. Muir, M. & Conati, C.: An Analysis of Attention to Student – Adaptive Hints in an Educational Game. Proc. Intelligent Tutoring Systems, 112–122, 2012.

28. Palmer, E.M., Horowitz, T.S., Torralba A., Wolfe, J. What Are the Shapes of Response Time Distributions in Visual Search? J. of Exp. Psy.: Human Percep. and Perf., 37(1), 57-71, 2011.

29. Rotter, Julian B. Generalized Expectancies for Internal Versus External Control of Reinforcement. Psych. Monographs: General and Applied 80.1, 1-28, 1966.

30. Skytree Adviser. http://www.skytree.net/products-services/adviser-beta/faq/.

31. Steichen, B., Carenini, G., Conati, C. User-Adaptive Information Visualization - Using eye gaze data to infer visualization tasks and user cognitive abilities. Proc. Conf. on Intelligent User Interfaces, IUI 2013, 317-328.

32. Toker, D., Conati, B., Steichen, Carenini, G. Individual User Characteristics and Information Visualization: Connecting the Dots through Eye Tracking. Proc. Conf. on Human Factors in Comp. Sys., CHI 2013, 295-304.

33. Toker, D., Conati, C., Carenini, G., & Haraty, M. Towards Adaptive Information Visualization: On the Influence of User Characteristics. Proc. Conf. on User Model., Adapt., and Personaliz., UMAP 2012, 274-285.

34. Townsend, J. T., & Ashby, F. G. (1983). Stochastic modelling of elementary psychological processes. London: Cambridge University Press.

35. Turner, M. L., & Engle, R.W. Is working memory capacity task dependent? Journal of Memory and Language, 28(2), 127-154, 1989.

36. Van Zandt, T. How to fit a response time distribution. Psychonomic Bulletin & Review, 7(3), 424-465, 2002.

37. Velez, M.C., Silver, D., & Tremaine, M. Understanding visualization through spatial ability differences. Proc. of Visualization, 511-518, 2005.

38. Wobbrock, J., Findlater, L., Gergle, D., & Higgins, J. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only ANOVA Procedures. Proc. Conf. on Human Factors in Comp. Sys., CHI 2011, 143-146.

39. Wolf, B.P. Building Intelligent Interactive Tutors: Student-Centered Strategies for Revolutionizing E-Learning. Morgan Kaufmann Publishers Inc., 2008.

40. Ziemkiewicz, C., Crouser, R.J., Yauilla, A.R., Su, S.L., Ribarsky, W., & Chang, R. How Locus of Control Influences Compatibility with Visualization Style. Proc. of IEEE VAST 2011, 81-90.